Population Genomics

Om P. Rajora *Editor*

Population Genomics

Concepts, Approaches and Applications

Springer

Population Genomics

Editor-in-Chief
Om P. Rajora
Faculty of Forestry and Environmental Management
University of New Brunswick
Fredericton, NB, Canada

This pioneering *Population Genomics Series* deals with the concepts and approaches of population genomics and their applications in addressing fundamental and applied topics in a wide variety of organisms. Population genomics is a fast emerging discipline, which has created a paradigm shift in many fields of life and medical sciences, including population biology, ecology, evolution, conservation, agriculture, horticulture, forestry, fisheries, human health and medicine.

Population genomics has revolutionized various disciplines of biology including population, evolutionary, ecological and conservation genetics, plant and animal breeding, human health, genetic medicine, and pharmacology by allowing to address novel and long-standing intractable questions with unprecedented power and accuracy. It employs large-scale or genome-wide genetic information across individuals and populations and bioinformatics, and provides a comprehensive genome-wide perspective and new insights that were not possible before.

Population genomics has provided novel conceptual approaches, and is tremendously advancing our understanding the roles of evolutionary processes, such as mutation, genetic drift, gene flow, and natural selection, in shaping up genetic variation at individual loci and across the genome and populations, disentangling the locus-specific effects from the genome-wide effects, detecting and localizing the functional genomic elements, improving the assessment of population genetic parameters or processes such as adaptive evolution, effective population size, gene flow, admixture, inbreeding and outbreeding depression, demography and biogeography, and resolving evolutionary histories and phylogenetic relationships of extant and extinct species. Population genomics research is also providing key insights into the genomic basis of fitness, local adaptation, ecological and climate acclimation and adaptation, speciation, complex ecologically and economically important traits, and disease and insect resistance in plants, animals and/or humans. In fact, population genomics research has enabled the identification of genes and genetic variants associated with many disease conditions in humans, and it is facilitating genetic medicine and pharmacology. Furthermore, application of population genomics concepts and approaches can facilitate plant and animal breeding, forensics, delineation of conservation genetic units, understanding evolutionary and genetic impacts of resource management practices and climate and environmental change, and conservation and sustainable management of plant and animal genetic resources.

The volume editors in this Series have been carefully selected and topics written by leading scholars from around the world.

Om P. Rajora
Editor

Population Genomics

Concepts, Approaches and Applications

Springer

Editor
Om P. Rajora
Faculty of Forestry and Environmental Management
University of New Brunswick
Fredericton, NB, Canada

ISSN 2364-6764 ISSN 2364-6772 (electronic)
Population Genomics
ISBN 978-3-030-04587-6 ISBN 978-3-030-04589-0 (eBook)
https://doi.org/10.1007/978-3-030-04589-0

Library of Congress Control Number: 2018965916

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Affectionately dedicated to my wife Malti
and children Apoorva, Anu, and Maneesha.

Preface

Recent novel advances in sequencing technologies, bioinformatics tools, statistical methods and software, and models have created a paradigm shift in several disciplines of biology (especially population biology, ecology, evolution, and conservation), agriculture, forestry, fisheries, human health, and medicine. Population genomics is an outcome of these advances, which is a fascinating and fast-growing discipline. Population genomics has revolutionized various disciplines of biology including population, evolutionary, ecological and conservation genetics, plant and animal breeding, human health, genetic medicine, and pharmacology by allowing to address novel and long-standing intractable questions with unprecedented power and accuracy. It employs large-scale or genome-wide genetic information and bioinformatics to address various fundamental and applied aspects in biology and related disciplines, and provides a comprehensive genome-wide perspective and new insights that were not possible before.

Population genomics has provided novel conceptual approaches and is tremendously advancing our understanding the roles of evolutionary processes, such as mutation, genetic drift, gene flow, and natural selection, in shaping up genetic variation at individual loci and across the genome, individuals and populations, disentangling the locus-specific effects from the genome-wide effects, detecting and localizing the functional genomic elements, improving the assessment of population genetic parameters or processes such as adaptive evolution, adaptive population genetic differentiation, effective population size, gene flow, admixture, inbreeding and outbreeding depression, demography and biogeography, and resolving evolutionary histories and phylogenetic relationships of extant and extinct species. Population genomics research is also providing key insights into the genomic basis of fitness, local adaptation, ecological and climate acclimation and adaptation, speciation, colonization, complex ecologically and economically important traits, and disease and insect resistance in plants, animals and/or humans. In fact, population genomics research has enabled the identification of genes and genetic variants associated with many disease conditions in humans, and it is facilitating genetic medicine and pharmacology. Furthermore, application of population genomics

concepts and approaches can facilitate plant and animal breeding, forensics, delineation of conservation genetic units, understanding evolutionary and genetic impacts of resource management practices and climate change, and conservation and sustainable management of plant and animal genetic resources.

I have been working on various aspects of molecular, population, evolutionary and conservation genetics, and genomics for about four decades. Recognizing the power and potential of population genomics, I started organizing a pioneering annual workshop on Population and Conservation Genomics in 2007 as a part of the premier annual International Plant and Animal Genome Conference. This Workshop has provided a platform for the presentation and sharing of the latest advances in population and conservation genomics at the international stage. I may have been the first to identify Conservation Genomics as a research area in 2004 when I used this term in my Senior (Tier 1) Canada Research Chair title. Leading and emerging scholars have been presenting their research results at the Population and Conservation Genomics workshop, which has grown to more than one Workshop session and has given rise to several offshoot workshops. The pool of the Workshop speakers provided a good resource for recruiting authors for the current *Population Genomics* book. Indeed, the chapters are written by prominent pioneering, leading and emerging research scholars in various fields of population genomics.

This *Population Genomics* book discusses the concepts, approaches and applications of population genomics in addressing various fundamental and applied crucial aspects outlined above in a variety of organisms from microorganisms to humans. The book provides insights into a range of emerging topics including population epigenomics, landscape genomics, paleogenomics, ecological and evolutionary genomics, seascape genomics, biogeography, demography, speciation, admixture, colonization and invasion, genomic selection, and plant and animal domestication. This book fills a vacuum in the field and is expected to become a primary reference in Population Genomics world-wide.

The book is organized into four parts. The first part provides an overview of the population genomics concepts, approaches, applications, challenges and future perspectives. The second part includes three chapters discussing sequencing and genotyping technologies, and bioinformatics methods as applied to population genomics. The third part focuses on various concepts and approaches in population genomics, such as population epigenomics, landscape genomics, paleogenomics, genome-wide association studies, and genomic selection. The fourth, the last part, includes nine chapters addressing population, evolutionary and ecological genetics applications and inferences, such as evolutionary and ecological genomics, demography, biogeography, seascape genomics, speciation, admixture, invasion and colonization, and plant and animal domestication and breed development. With such quite comprehensive and diverse topics, the book is envisioned for a wide readership, including undergraduate and graduate students, research scholars, and professionals and experts in the field.

I would like to thank all contributors to this volume and peer reviewers.

Fredericton, NB, Canada Om P. Rajora

Contents

Contributors

Abdulkadir Abakir Wolfson Centre for Stem Cells, Tissue Engineering and Modelling (STEM), Division of Cancer and Stem Cells, School of Medicine, Centre for Biomolecular Sciences, University of Nottingham, Nottingham, UK

Sally N. Aitken Centre for Forest Conservation Genetics, Faculty of Forestry, University of British Columbia, Vancouver, BC, Canada

Kimberly R. Andrews Department of Fish and Wildlife Sciences, University of Idaho, Moscow, ID, USA

Genetics and Genomics Group, NOAA Pacific Marine Environmental Lab, University of Washington JISAO, Seattle, WA, USA

Niko Balkenhol Wildlife Sciences, University of Goettingen, Göttingen, Germany

David M. Cairns Department of Geography, Texas A&M University, College Station, TX, USA

Philippe Cubry Institut de Recherche pour le développement, Université de Montpellier, Montpellier, France

Katrina M. Dlugosch Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA

Rachael Y. Dudaniec Department of Biological Sciences, Macquarie University, Sydney, NSW, Australia

Maria Eleftheriou Wolfson Centre for Stem Cells, Tissue Engineering and Modelling (STEM), Division of Cancer and Stem Cells, School of Medicine, Centre for Biomolecular Sciences, University of Nottingham, Nottingham, UK

Freddy Fikse Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

Keith D. Gaddis Department of Geography, Texas A&M University, College Station, TX, USA

D. C. Haak Department of Plant Pathology, Physiology, and Weed Science, Virginia Tech, Blacksburg, VA, USA

E. M. Hallerman Department of Fish and Wildlife Conservation, Virginia Tech, Blacksburg, VA, USA

Brian K. Hand Flathead Lake Biological Station, Conservation Genomics Group, Division of Biological Sciences, University of Montana, Polson, MT, USA

Rasmus Heller Department of Biology, University of Copenhagen, Copenhagen N, Denmark

Paul A. Hohenlohe Department of Biological Sciences, Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID, USA

J. A. Holliday Department of Forest Resources and Environmental Conservation, Virginia Tech, Blacksburg, VA, USA

Kristopher J. L. Irizarry The Applied Genomics Center, College of Veterinary Medicine, Western University of Health Sciences, Pomona, CA, USA

Jean-Luc Jannink United States Department of Agriculture, Agricultural Research Service, R.W. Holley Center for Agriculture and Health, Ithaca, NY, USA

Jeremy S. Johnson Department of Geography, Texas A&M University, College Station, TX, USA

School of Forestry, Northern Arizona University, Flagstaff, AZ, USA

Dorena Genetic Resource Center, Cottage Grove, OR, USA

Elisabeth Jonas Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

Stéphane Joost Laboratory of Geographic Information Systems (LASIG), School of Civil and Environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Marty Kardos Flathead Lake Biological Station, Conservation Genomics Group, Division of Biological Sciences, University of Montana, Polson, MT, USA

Takeshi Kawakami Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK

Department of Evolutionary Biology, Evolutionary Biology Centre (EBC), Uppsala University, Uppsala, Sweden

Konstantin V. Krutovsky Department of Forest Genetics and Forest Tree Breeding, Georg-August University of Göettingen, Göttingen, Germany

Department of Ecosystem Science and Management, Texas A&M University, College Station, TX, USA

Laboratory of Population Genetics, N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia

Genome Research and Education Center, Siberian Federal University, Krasnoyarsk, Russia

Tianying Lan Department of Biological Sciences, University at Buffalo (SUNY), Buffalo, NY, USA

School of Biological Sciences, Nanyang Technological University, Singapore, Singapore

Martin Lascoux Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden

Lara C. Lewis Wolfson Centre for Stem Cells, Tissue Engineering and Modelling (STEM), Division of Cancer and Stem Cells, School of Medicine, Centre for Biomolecular Sciences, University of Nottingham, Nottingham, UK

Charlotte Lindqvist Department of Biological Sciences, University at Buffalo (SUNY), Buffalo, NY, USA

School of Biological Sciences, Nanyang Technological University, Singapore, Singapore

Roberto Lozano Plant Breeding and Genetics, School for Integrative Plant Science, Cornell University, Ithaca, NY, USA

Gordon Luikart Flathead Lake Biological Station, Conservation Genomics Group, Division of Biological Sciences, University of Montana, Polson, MT, USA

Ehren R. V. Moler Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA

Elena Flavia Mouresan Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

Nicola J. Nadeau Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK

Marjorie F. Oleksiak Marine Biology and Ecology, Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL, USA

Dunia Pino Del Carpio Agriculture Research Division, Agriculture Victoria, Melbourne, VIC, Australia

Om P. Rajora Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, NB, Canada

Lars Rönnegård Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

Statistics Unit, School of Technology and Business Studies, Dalarna University, Falun, Sweden

Alexey Ruzov Wolfson Centre for Stem Cells, Tissue Engineering and Modelling (STEM), Division of Cancer and Stem Cells, School of Medicine, Centre for Biomolecular Sciences, University of Nottingham, Nottingham, UK

Jordi Salmona Laboratoire Evolution and Diversité Biologique, UMR 5174, CNRS/Université Toulouse III Paul Sabatier, Toulouse, France

Université de Toulouse, UMR 5174 EDB, Toulouse, France

Jarkko Salojärvi School of Biological Sciences, Nanyang Technological University, Singapore, Singapore

Gernot Segelbacher Wildlife Ecology and Management, University of Freiburg, Freiburg, Germany

Kimberly A. Selkoe Bren School of Environmental Science & Management, National Center for Ecological Analysis and Synthesis, University of California Santa Barbara, Santa Barbara, CA, USA

Oliver Selmoni Laboratory of Geographic Information Systems (LASIG), School of Civil and Environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Aaron Shafer Forensic Science and Environmental and Life Sciences, Trent University, Peterborough, ON, Canada

Alexandros Stamatakis Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

Elton J. R. Vasconcelos The Applied Genomics Center, College of Veterinary Medicine, Western University of Health Sciences, Pomona, CA, USA

Yves Vigouroux Institut de Recherche pour le développement, Université de Montpellier, Montpellier, France

Sophie von der Heyden Evolutionary Genomics Group, Department of Botany and Zoology, University of Stellenbosch, Stellenbosch, South Africa

Ian J. Wang Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA, USA

Shana R. Welles Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA

Amy V. Whipple Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA

Pamela Wiener The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, UK

Samantha Wilkinson The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, UK

Marnin D. Wolfe Section on Plant Breeding and Genetics, School of Integrative Plant Sciences, Cornell University, Ithaca, NY, USA

Part I
Introduction

Population Genomics: Advancing Understanding of Nature

Gordon Luikart, Marty Kardos, Brian K. Hand, Om P. Rajora, Sally N. Aitken, and Paul A. Hohenlohe

Abstract Population genomics is advancing our understanding of evolution, ecology, conservation, agriculture, forestry, and human health by allowing new and long-standing questions to be addressed with unprecedented power and accuracy. These advances result from plummeting costs for DNA sequencing, which makes genotyping feasible for hundreds to millions of individuals and loci, and also allows for the study of variation in gene expression, epigenetic variation, and proteins. The increased power also results from the development of innovative software, statistical approaches, and models to extract information from massive next-generation sequencing datasets. Among the most exciting developments are conceptually novel approaches that are advancing understanding about inbreeding and outbreeding depression, adaptive gene flow, population demographic history, and the genomic basis of local adaptation and speciation. Remaining challenges in applying genomics to natural and managed populations include the limited understanding and availability of validated bioinformatics pipelines for genotyping and analyzing genomic data. We also lack knowledge of best practices and general guidelines for filtering and genotyping genomic data including restriction site-associated DNA sequences (RAD), targeted DNA capture, and pooled sequencing. Finally, we emphasize the need for continued rigorous teaching of population genetics theory,

G. Luikart (✉) · M. Kardos · B. K. Hand
Flathead Lake Biological Station, Conservation Genomics Group, Division of Biological Sciences, University of Montana, Polson, MT, USA
e-mail: gordon.luikart@mso.umt.edu

O. P. Rajora
Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, NB, Canada

S. N. Aitken
Centre for Forest Conservation Genetics, Faculty of Forestry, University of British Columbia, Vancouver, BC, Canada

P. A. Hohenlohe
Department of Biological Sciences, Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID, USA

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_60,

so that the next generation of population genomicists can ask well-informed questions and interpret next-generation sequence datasets.

Keywords Adaptation · Community genetics · Conservation genetics · Ecological genomics · Epigenetics · Evolutionary genomics · Landscape genomics · Population genetics · Selection detection

> Molecular markers have totally changed our view of nature (Schlötterer 2004).
>
> Population genomics is a new term for a field of study that is as old as the field of genetics itself, assuming that it means the study of the amount and causes of genome-wide variability in natural populations (Charlesworth 2010).
>
> Population genomic tools have revolutionized many aspects of biology, as detailed throughout the chapters of this volume (Hohenlohe et al. 2018).

1 Introduction

New and long-standing questions in ecology, evolution, conservation biology, and related fields can now be addressed with unprecedented power and accuracy using population genomics approaches. This power results largely from new sequencing and genotyping technologies that produce enormous amounts of data (Schlötterer 2004; Narum et al. 2013; van Dijk et al. 2018; Sedlazeck et al. 2018) but also from new statistical approaches and software (Paradis et al. 2017; Ceballos et al. 2018; Cooke and Nakagome 2018; Faria et al. 2018; Gruber et al. 2018; Hendricks et al. 2018; Knaus and Grünwald 2017; Zhang et al. 2018). These molecular and computational approaches are now within reach of many biologists in terms of costs, ease of data production, and availability of computational tools. This chapter provides an overview of the concepts and primary approaches employed to study genome-wide genetic variation in natural and managed species and populations. Some of these approaches are not yet widely used but are emerging in the literature on population genomics (Hendricks et al. 2018).

Population genomics has been broadly defined as the simultaneous study of numerous loci and genome regions to better understand the roles of evolutionary processes (such as mutation, genetic drift, gene flow, and natural selection) that influence variation across genomes and populations (Black et al. 2001; Luikart et al. 2003). This definition emphasizes understanding of locus-specific effects like selection against the background of genome-wide effects such as demography and genetic drift in order to improve assessments of adaptive evolution, the effective population size, gene flow, admixture, inbreeding and outbreeding depression, speciation, and the genomic basis of fitness (Fig. 1) (Allendorf et al. 2010; McMahon et al. 2014; Hunter et al. 2018).

Hohenlohe et al. (2010a) outlined a novel conceptual framework for population genomics that emphasizes the understanding of patterns of genetic variation and

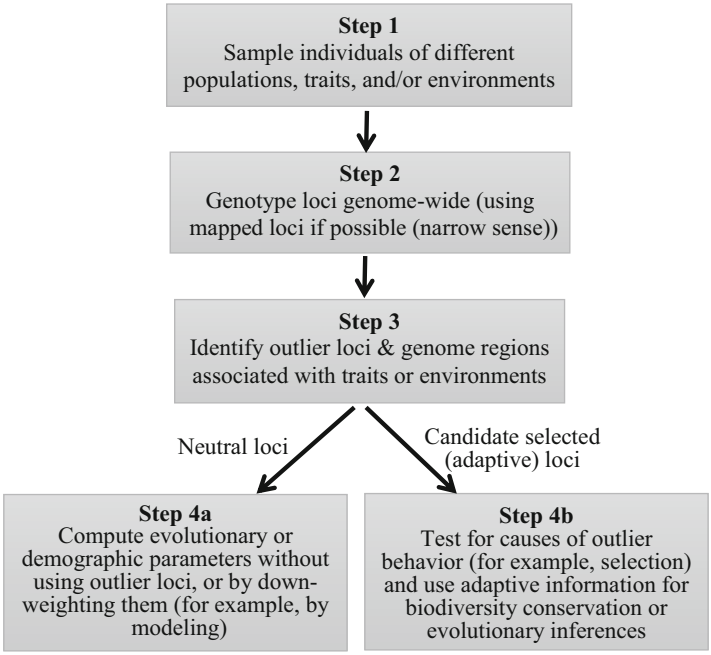

Fig. 1 Conceptual framework of main steps in a population genomics approach used to identify outlier loci under selection (or genotyping errors) and also to improve estimates of population history and demography using the selectively neutral loci. In Step 1, individuals can be sampled from different phenotypes or environments to help test for adaptive gene marker associations and to dissect the genomic basis of phenotypes, local adaptation, adaptation to captivity, artificial selection, or speciation. Step 2 requires a genetic linkage map or a physical map (Sects. 3.1 and 3.2) to localize genome regions under selection and to ensure high marker density (narrow sense approach). However many unmapped loci can be used in broad sense genomics (Figs. 3 and 4). Step 3 could employ conceptually novel approaches to identify "outlier loci" or chromosomal regions that behave unlike most other loci in the genome and therefore could be under selection or associated with phenotypic traits. Outlier loci under selection can bias estimates of neutral population genetic parameters (Step 4a) such as gene flow, effective population size, and structure. Figure modified from Luikart et al. (2003)

evolutionary processes in all genome regions by plotting population genetic statistics across each chromosome using many mapped loci (Fig. 2; Box 1). An example of a population genomics approach is measuring a population genetic summary statistic, e.g., genomic diversity, population differentiation, or gene expression, as a continuous variable along chromosomes to help identify loci under selection, chromosomal islands of adaptive divergence, or alleles associated with a phenotypic trait (see also Fig. 3 in Luikart et al. 2003; Hohenlohe et al. 2010b; Ellegren 2014; Kardos et al. 2015b).

Allendorf (2017) and Hohenlohe et al. (2018) defined population genomics as requiring a sufficient density of DNA markers to detect forces affecting any particular genomic region, e.g., genes under selection, regions of reduced recombination. Here, we provide a narrow sense *definition of population genomics as the use of*

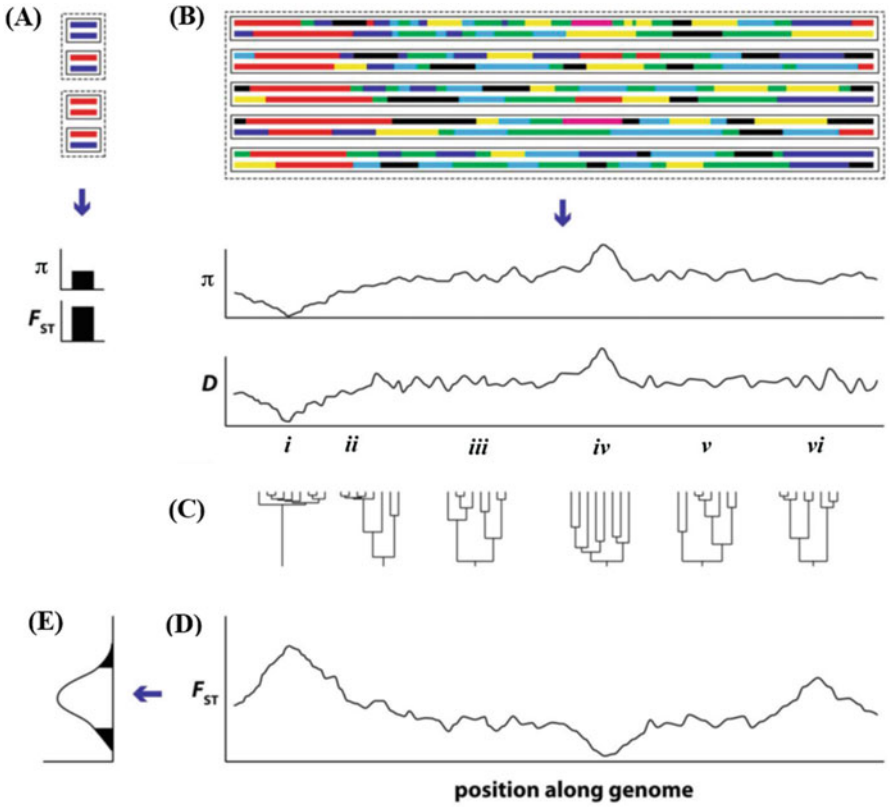

Fig. 2 A population genomics perspective and conceptual framework. (**A**) Traditional population genetics takes data on alleles (colored bars), grouped within individuals (solid boxes) and populations (dashed boxes), and calculates summary statistics to make inferences about evolution, such as nucleotide diversity (π) and population differentiation (F_{ST}). (**B**) Population genomics takes data on haplotypes within a population and calculates summary statistics as continuous variables along the length of the genome, such as π and the allele frequency spectrum (Tajima's D). The different types of evolutionary processes leave different signatures in these distributions: (i) hard selective sweep, (ii) region linked to hard sweep, (iii) neutral expectation, (iv) balancing selection, (v) neutral expectation, and (vi) soft sweep. (**C**) The coalescent structure of ancestral relationships among alleles within a population also reflects these processes along the genome. (**D**) Given these genomic processes within a population, statistics comparing genetic variation across populations, such as F_{ST}, can also indicate genomic patterns of selection. (**E**) Collapsing the genomic distribution of a statistic into a frequency distribution provides an estimate of the genome-wide average, allowing identification of statistically significant outliers (shaded regions). Reproduced with permission from Hohenlohe et al. (2010a)

conceptually novel approaches to address questions intractable by traditional genetic methods by using high-density genome-wide markers (e.g., DNA, RNA, epigenetic marks) to provide high power to detect genomic regions associated with traits or evolutionary processes such as fitness, phenotypes, and selection (Box 1). This definition combines the requirement for conceptual novelty aspect

from Garner et al. (2016) and Hohenlohe et al. (2010a), with the high-density marker requirement of Allendorf (2017); it also explicitly includes multiple omics approaches (transcriptomics, epigenomics, and proteomics).

Broad sense population genomics can be defined as the use of new genomics technology and numerous loci to address questions in population genetics (e.g., Shafer et al. 2015; Garner et al. 2016; Hohenlohe et al. 2018) (Box 1). We include broad sense approaches here because some are advancing understanding of genomics questions ranging from the discovery of genes underlying adaptive evolution to assessing population parameters and demography using thousands to millions of neutral markers that are often anonymous or not mapped.

Our main goals for this chapter are fourfold. First, we discuss the research topics and questions for which genomics tools are most valuable. We illustrate where genomics methods are improving our ability to address long-standing objectives and also to address previously intractable questions using conceptually novel approaches. Second, we give a brief introduction to new molecular techniques and computational approaches (including bioinformatics workflows and Bayesian methods) to help biologists understand this growing literature and to plan their projects. Third, we provide an overview of the emerging disciplines where population genomics concepts and approaches are being applied. Finally, we discuss future perspectives of applications of population genomics concepts and approaches and conclude the chapter. Throughout, we highlight the opportunities and challenges associated with population genomic analyses in studies of natural and managed populations.

Box 1 How is Narrow Sense Population *Genomics* Different from Broad Sense Genomics and Traditional Population *Genetics*?

Defining broad and narrow sense population genomics can be useful because there is often confusion among students and researchers as to what constitutes genomics and also because broad sense population genomics studies include traditional population genetic approaches and the use of more DNA markers (see Charlesworth 2010; Allendorf 2017). An example of a broad sense population genomics study would be using thousands or tens of thousands of anonymous SNPs (Fig. 3) to estimate the inbreeding coefficients of individuals using traditional parameters (e.g., individual heterozygosity; Hoffman et al. 2014; Kardos et al. 2016a), while a narrow sense study would be the mapping of runs of homozygosity (RoH) to infer recent and historical inbreeding (or population bottlenecks) (Bérénos et al. 2015; Howard et al. 2015; Palkopoulou et al. 2015; Pemberton et al. 2017; Kardos et al. 2017; Ceballos et al. 2018). The requirement for narrow sense genomics to include "conceptual novelty" and to address questions not tractable using traditional population genetics addresses the criticism of Charlesworth (2010) and of others saying that population genomics is nothing new.

(continued)

Box 1 (continued)

A narrow sense population genomics study precisely characterizes variation at *many specific (mapped) regions of the genome* (Allendorf 2017). The density of markers required (see below) varies and depends on phenomena that affect gametic disequilibrium along a chromosome such as mating system (e.g., selfing versus random mating), effective population size, population subdivision, gene flow or admixture, and recombination rates (Slatkin 2008).

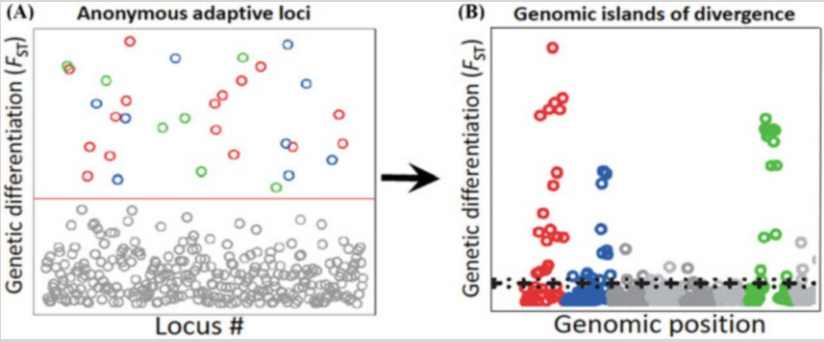

Fig. 3 Illustration of how (**A**) anonymous (unmapped) loci are often detected to be under directional selection (e.g., with high allele frequency differentiation, F_{ST}) among populations and how (**B**) a genetic linkage map or a physical map (genome assembly) helps to localize the genome regions under selection by positioning loci (SNPs) along a chromosome or entire genome. In panel **B**, each color represents a different chromosome (linkage group) including the different shades of gray. Knowing the genome position of SNPs allows for multiple, often linked, SNPs to be identified that result from the same selection process and signature (e.g., high F_{ST}), which increases our confidence that the SNPs or genome region are actually under selection and not false positives. Positional information also helps understand the number of loci or genome regions that are under selection. Further, if coding or annotated genes have also been mapped or physically located on a genome sequence, researchers can identify genes in the region of the selection signature, which represent candidate adaptive genes (e.g., Mckinney et al. 2016). Figure (**A**) represents a broad sense genomics approach, while (**B**) is narrow sense genomics. Figure modified from Garret McKinney (pers. comm., 2018)

2 When Is Population Genomics Most Valuable?

A wide array of fundamental and novel questions can now be reliably addressed thanks to developments in population genomics (Table 1). In this section, we describe several newly invigorated avenues of research in evolutionary biology and conservation genetics. The most exciting developments of population genomics involve using novel approaches to address previously unapproachable questions such as mapping adaptive variation genome wide and resolving the genomic basis of fitness and phenotypes (Hoban et al. 2016; Hendricks et al. 2018; Hunter et al.

Table 1 Questions or objectives that population genomics can help to address and examples of genomics approaches to address them

Question or objective	Genomics approach (example)/software	References
Identify candidate adaptive loci by detecting selective sweeps	Genome scan for low heterozygosity regions	Rubin et al. (2010), Axelsson et al. (2013), Kardos et al. (2015b)
	High differentiation (e.g., F_{ST}) outlier regions	Rochus et al. (2018)
	High gametic disequilibrium	Pérez O'Brien et al. (2014)
	Shifted site frequency spectrum (high-frequency-derived alleles)	Alachiotis and Pavlidis (2018), Tajima (1989), Fay and Wu (2000), DeGiorgio et al. (2016)
	Extended haplotype homozygosity	Sabeti et al. (2002, 2007), Voight et al. (2006)
	Scan for soft selective sweeps (sweeps of alleles that are already present on multiple haplotypes [standing genetic variation] or of positively selected alleles at multiple sites in the same region or gene)	Hermisson and Pennings (2017), Marques et al. (2018), Hodel et al. (2018)
	Scan for hard selective sweeps (sweeps of new (de novo) positively selected mutations)	Pennings and Hermisson (2006), Betts et al. (2018), Kreiner et al. (2018)
Identify candidate loci underlying local adaptation and speciation	See sweeps above (including F_{ST} outliers)	
	Heritable gene expression profile differences	Christie et al. (2016)
	Parallel evolution of gene expression	Yeaman et al. (2016)
	Parallel phenotypic or DNA sequence evolution	Prince et al. (2017)
Identify loci associated with environmental variation (landscape genomics)	Methods testing for gene-environment association can detect subtle signatures of adaptation that are not detectable using genome-wide selection scans	Joost et al. (2007), Coop et al. (2010), Hancock et al. (2011), Rellstab et al. (2015), Rajora et al. (2016), Harrisson et al. (2017), Rougeux et al. (2018), Schmidt et al. (2017)
Detect signatures of polygenic adaptation	Single vs multiple genes and the genomic basis of fitness	Berg and Coop (2014), Bourret et al. (2014), Brieuc et al. (2015), Laporte et al. (2016), Stölting et al. (2015), Sork (2016), Yeaman et al. (2016), Rajora et al. (2016), Harrisson et al. (2017)
Identify loci underlying species and landscape interactions (landscape community genomics)	Landscape community genomics	Hand et al. (2015b), De Kort et al. (2018), Kozakiewicz et al. (2018)

(continued)

Table 1 (continued)

Question or objective	Genomics approach (example)/software	References
Identify loci associated with traits within populations	Genome-wide association analysis	Smith and O'Brien (2005), Johnston et al. (2011), Johnston et al. (2013), Barson et al. (2015), Bérénos et al. (2015), Husby et al. (2015)
	Admixture mapping	Lamichhaney et al. (2015)
Quantify inbreeding and inbreeding depression and identify underlying loci	Individual heterozygosity; runs of homozygosity	Hoffman et al. (2014), Bérénos et al. (2016), Dobrynin et al. (2015), Hedrick and Garcia-Dorado (2016), Howard et al. (2015), Huisman et al. (2016), Kardos et al. (2017), (2018)
Quantify hybridization, outbreeding depression, adaptive introgression, and associated loci	Runs of hybridity	Guan (2014), Gompert (2016), Leitwein et al. (2018), Jones et al. (2018)
	Selection against introgression	Kovach et al. (2016)
Estimate effective population size (N_e) or Nb	Abundance and lengths of IBD segments (both within and between individuals) can be used to estimate N_e and historical changes in N_e	Kirin et al. (2010), Pemberton et al. (2012), Browning and Browning (2015), Kardos et al. (2017)
	Pairwise sequentially Markovian coalescent (estimate deep historical time series of N_e)	Li and Durbin (2011)
Detect population declines (reduction of N_e)	BOTTLENECK; ABC analyses in DIYABC	Cornuet and Luikart (1996), Hoban et al. (2013), Cammen et al. (2018)
Estimate contemporary gene flow rates (Nm)	BayesAss3-SNPs	Wilson and Rannala (2003), Waterhouse et al. (2018), Brauer et al. (2018)
Distinguishing continuous migration from strict isolation	Maximum-likelihood method based on the jSFS	Gutenkunst et al. (2009), Fraisse et al. (2018)
Identify adaptively differentiated populations such as ESUs (evolutionarily significant units)	Detecting a major gene (haplotype) for migration timing Chinook salmon and steelhead trout; multidisciplinary framework to delineate distinct populations of butterflies	Prince et al. (2017), Dupuis et al. (2018)

2018). Identifying loci underlying adaptive evolution is a long-standing goal in evolutionary biology, and doing so helps to understand the phenotypic traits, biochemical pathways, and nature of the selective forces that have resulted in the bewildering array of biodiversity.

A more common or widespread application of population genomics approaches is improving estimation of population genetic parameters and evolutionary relationships – including assessments of effective population size, population structure, phylogeography, and demography – which are largely broad sense genomics (Luikart et al. 2003). We first discuss these broader sense applications in Sect. 2.1. *We then discuss exciting and previously intractable applications including mapping of adaptive genomic variation* in Sects. 2.2 through 2.8.

2.1 Estimating Population Genetic Parameters with Genome-Wide Markers: Broad Sense Genomics Approaches

Genomics approaches can be used to address questions that have long been studied using traditional molecular markers such as allozymes or microsatellites (Box 1). In this section, we describe some of those population genetic questions and how genomics can be used to improve them. While traditional molecular markers provide information on a small fraction or subset of the genome, large-scale genomic data (thousands to hundreds of thousands of SNPs) provide a more complete picture of genetic parameters across the entire genome (e.g., Hohenlohe et al. 2010b; Brelsford et al. 2017).

Statistical inference can be used to estimate population genetic parameters, such as genetic diversity, effective population size, population differentiation, or phylogenetic relationships, and these population genetic metrics reflect processes that affect the genome as a whole. However, these metrics can vary tremendously across the genome, which suggests a narrow sense approach (e.g., mapped loci) is advisable. For example, genetic variation and population differentiation often vary tremendously across the genome due to variation in recombination rate, selection intensity (purifying and positive), and the mutation rate (Hohenlohe et al. 2010b).

The primary advantage of broad sense genomics is providing many more genetic markers, often by several orders of magnitude, than previous techniques, and often for similar cost and research effort. This results in the potential for much greater precision of estimates of population genetic parameters. Many more markers can also reduce bias of estimates of population genetic parameters by identifying loci under selection that often should not be used to estimate parameters requiring only neutral loci, such as gene flow, demographic history, and phylogenies. In some cases, recent *genomics techniques can also be more cost-effective than traditional techniques*, for instance, with the ability to simultaneously detect and genotype loci

using RADseq and RAD capture (see Sect. 4) in taxa for which microsatellite or other loci have not previously been developed (Andrews et al. 2016).

In population genomics studies, genome-wide estimates are often considered as the background against which outliers reflect adaptive or functionally important loci (Fig. 1; Luikart et al. 2003), and detection of these loci is central to narrow sense population genomics as described in the sections below (see also Hohenlohe et al. 2018 this volume). The genome-wide background, estimated by either traditional genetic or genomics techniques, is often interpreted to reflect selectively neutral processes. But it is important to remember that the effects of selection and genotype-phenotype relationships are pervasive across the genome due to processes, such as hitchhiking (Maynard Smith and Haigh 1974), background selection (Charlesworth et al. 1993), or isolation by adaptation (Nosil et al. 2008; Corbett-Detig et al. 2015). Whether techniques tend to avoid coding regions (e.g., microsatellites), focus on them (e.g., exon capture, RAD capture with targets in or near genes), or sample randomly across the genome (e.g., RADseq), it can be treacherous to interpret genome-wide patterns as solely reflecting "neutral" processes.

2.1.1 Genetic Variation and Effective Population Size

A central quantity in population genetics is the amount of genetic variation present in a population. This can be quantified in several ways, including expected heterozygosity (H_e) or nucleotide diversity (π), which can be estimated from genome-wide SNP data using many analysis programs, such as PLINK (Purcell et al. 2007). Genome-wide genetic variation is the result of multiple interacting processes, including mutation, genetic drift, selection, and population structure, that affect the genome as a whole.

The amount of genetic variation in a population is closely related to the effective population size (N_e), which is often a focus of population genomics studies, particularly those relevant to conservation (e.g., Hare et al. 2011; Cammen et al. 2018). While there are several ways to define N_e, a common definition derives from the amount of genetic drift in a local population relative to an idealized Wright-Fisher model (Charlesworth 2009; Allendorf et al. 2013). The most direct way to estimate the rate of genetic drift and N_e is with temporal genetic samples from a local population, which provide measurements of changes in allele frequencies over time (Wang 2005; Luikart et al. 2010). Often, however, multiple samples over time are not available from natural populations, so other estimation techniques are required.

Random genetic drift due to small population size also leads to nonrandom associations between alleles from different loci, known as gametic disequilibrium (GD). GD provides the basis for methods to estimate N_e from a single genetic sample collected at one time point, such as program LDNe in *NeEstimator*. LDNe requires independent loci such as those on different chromosomes (Do et al. 2014). With the large number of markers available from genomic data, it is likely that physically linked loci (those on the same chromosome) are included. Physically linked loci can downwardly bias estimates of N_e by increasing GD (Waples and Do 2010). If markers can be mapped to a reference genome assembly or linkage map, one locus

in physically linked pairs of loci can be removed (e.g., as done by Larson et al. (2017)) or a general correction for the number of chromosomes can be applied (Waples et al. 2016). An alternative class of methods uses coalescent-based inference of N_e; Nunziata and Weisrock (2018) found that GD methods require more individuals (e.g., $n > 30$), while coalescent methods require fewer individuals (e.g., $n = 15$) but more SNP markers (25,000–50,000). Estimates of N_e from different methods can vary, and knowledge of population demography or temporal data can improve estimates considerably (Gilbert and Whitlock 2015).

2.1.2 Population Structure and Phylogeography

Populations exist across space, and the spatial distribution of genetic variation is an important focus of population genetics. Quantifying population structure and levels of genetic differentiation among populations (e.g., estimating the parameter F_{ST}) has been tractable with traditional population genetic tools, but again genomic techniques provide greater statistical power and precision for estimating parameters (Hohenlohe et al. 2018 in this volume). Furthermore, the number of markers from genomic data can allow for estimates from fewer individual samples; for instance, Nazareno et al. (2017) report consistent estimates of F_{ST} when using as few as two individuals, genotyped at over 1,500 SNPs.

Many analytical tools are well-suited for assessing and visualizing population structure from large genomic SNP datasets, such as principal components analysis and Bayesian clustering methods, and applying multiple techniques to a single dataset can help reveal important patterns (Fig. 4). When applied to genome-wide data, these approaches illustrate the results of processes that affect the genome as a whole, such as population size and migration rates. In a landscape genetics framework, a combination of genomic and landscape data can identify landscape features associated with variation in dispersal patterns (see Johnson et al. 2018a, b in this volume for a review). Interpolating and mapping genetic similarity across landscapes can reveal areas of high versus low gene flow, e.g., using the estimated effective migration surface (EEMS) approach of Petkova et al. (2016). Recent genomics techniques also provide new power for understanding the relationship between landscape variables and functional genetic variation at specific loci, such as genes; Balkenhol et al. (2017) in this volume review this field of landscape genomics.

2.1.3 Demographic History

A goal of population genomics studies that was considerably less tractable with traditional genetic techniques is a detailed reconstruction of historical demographic patterns, including changes in effective population size and migration rates, using genetic data sampled only from the contemporary populations. A number of techniques have been developed for demographic reconstruction from genetic or genomic data, such as approximate Bayesian computation (ABC; Boitard et al. 2016;

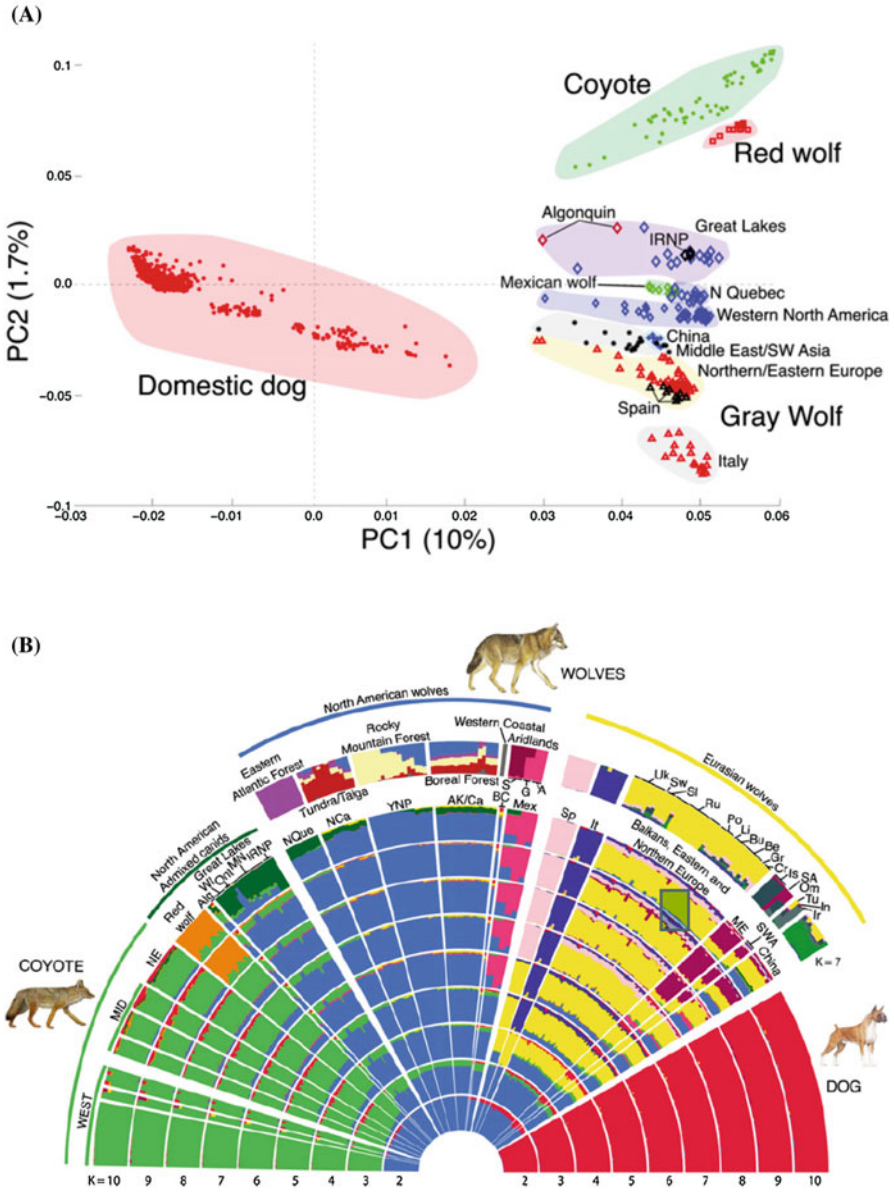

Fig. 4 Two methods for visualizing patterns of genetic differentiation among populations or closely related taxa: (**a**) principal components analysis and (**b**) Bayesian clustering analysis. Here these methods are applied to data from a 48,000 SNP genotyping array from wolves and their relatives. Reproduced with permission from VonHoldt et al. (2011)

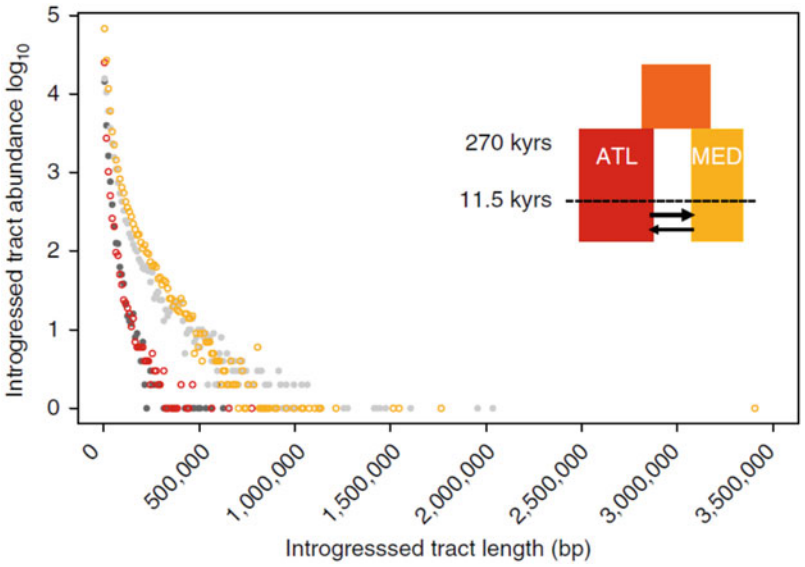

Fig. 5 Mapped genomic markers provide information on haplotype lengths, which are informative to assess historic admixture processes. Here the observed distributions of haplotype tract lengths in Atlantic and Mediterranean populations of European sea bass (*Dicentrarchus labrax*) (red and yellow dots) closely match simulated distributions (dark and light gray dots), allowing estimation of parameters in a model of historic isolation followed by secondary contact and gene flow. The haplotype information and modeling allows estimation of timing, directionality, and amount of gene flow. Reproduced with permission from Duranton et al. (2018)

Elleouet and Aitken 2018), sequential Markovian coalescent methods (Terhorst et al. 2017), and site frequency spectrum methods (Gutenkunst et al. 2009). See Salmona et al. (2017) in this volume as well as Beichman et al. (2017) for detailed reviews.

As an example, Duranton et al. (2018) estimated the parameters of a demographic model of two populations of European sea bass (*Dicentrarchus labrax*). Using genomic data mapped to a reference genome, the authors were able to characterize the distribution of lengths of haplotypes and fit model parameters to the observations (Fig. 5). Specifically, they identified tracts of migrant ancestry using the program ChromoPainter (Lawson et al. 2012) and estimated admixture parameters, and they used the method of Harris and Nielsen (2013) to infer demographic history from tracts of identity by state. These results reconstruct the historical details of population isolation and secondary gene flow between Atlantic and Mediterranean populations. This is a narrow sense genomics study because high-density mapped markers are used with a conceptually novel approach (haplotype tracts of immigrant ancestry).

2.1.4 Phylogenomics

Phylogenetic relationships among taxa can be estimated from a wide range of genetic data types, including genomic data. A complication is that many genetic markers spread across the genome may reflect different evolutionary histories because of recombination, particularly in recently diverged species and where incomplete lineage sorting and admixture play important roles (Edwards et al. 2016). Methods accounting for this, for instance, in estimating phylogeny from large SNP datasets, have been developed (Hohenlohe et al. 2018 this volume; McKain et al. 2018). Ideally, phylogenomic datasets are used not only to estimate a consensus tree among taxa but also to reveal patterns of hybridization and admixture (e.g., using analyses that allow for specific admixture events, such as TreeMix; Pickrell and Pritchard 2012).

2.2 Identifying Adaptive Genetic Variation Underlying Selective Sweeps

Population genomics makes it possible to identify "footprints" of natural selection in genome-wide patterns of genetic variation. The classical genomic signature of positive selection is the hard selective sweep, where fixation of a positively selected de novo mutation dramatically reduces genetic diversity at closely linked loci in a process referred to as genetic hitchhiking (Maynard Smith and Haigh 1974). The size of the region of reduced variation around the positively selected allele depends mainly on the strength of selection (and thus how quickly the sweep progressed) and the recombination rates on either side of the selected site (Jensen et al. 2016).

Hard selective sweeps are characterized by very low nucleotide diversity, and polymorphisms subsequently arising within a swept region display an excess of low-frequency-derived alleles compared to the genome-wide background. Thus, methods used to identify classical selective sweeps generally scan the genome for regions with low diversity (Maynard Smith and Haigh 1974), an excess of rare alleles (Tajima 1989), and a shifted site frequency spectrum (SFS) toward relatively high-frequency-derived alleles (DeGiorgio et al. 2016; Fay and Wu 2000; Huber et al. 2015; Kim and Stephan 2002).

While classical hard selective sweeps strongly reduce genetic variation around the selected site, soft selective sweeps arise from positive selection on standing genetic variation and leave a subtler genomic signature (Hermisson and Pennings 2005). In particular, soft sweeps usually do not strongly reduce genetic variation or result in a large shift in the site frequency spectrum around the selected site because the positively selected allele is present within multiple flanking haplotypes (Pennings and Hermisson 2006; Teshima et al. 2006). Soft sweeps appear to be a dominant mechanism of recent adaptation in humans (McCoy and Akey 2017; Schrider and Kern 2017). Methods based on extended haplotype homozygosity,

which look for derived alleles sitting on exceptionally long haplotypes, are thought to have substantially higher power to detect soft selective sweeps than diversity- or site frequency spectrum-based genome scans (Ferrer-Admetlla et al. 2014; Voight et al. 2006). Machine learning appears to also be a powerful method to detect soft sweeps (Schrider and Kern 2017).

Recent studies have detected putative selective sweeps in an array of organisms, ranging from domesticated livestock and humans to natural populations of non-model species. In some cases, these studies have helped to identify the phenotypes and underlying genetic and biochemical pathways involved with the response to positive selection. Recent studies using genome scans based on genome resequencing data have identified putative selective sweeps underlying adaptation to domestication in pigs (*Sus scrofa*; Rubin et al. 2012), dogs (*Canis lupus familiaris*; Axelsson et al. 2013), chickens (*Gallus gallus*; Rubin et al. 2010), and rabbits (*Oryctolagus cuniculus*; Carneiro et al. 2014).

Schweizer et al. (2016) identified putative selective sweeps in North American gray wolves (*Canis lupus*) related to coat color and environmental conditions by conducting genome scans via resequencing of exons and intergenic sequences. Kardos et al. (2015b) identified a putative selective sweep in wild bighorn sheep (*Ovis canadensis*) in the vicinity of the *RXFP2* gene associated with horn growth in domestic sheep (*RXFP2*). Their results suggested that horn morphology (or size) in bighorn sheep evolved at least in part via positive selection on a beneficial variant at *RXFP2*. See the chapter herein by Hohenlohe et al. (2018) for additional examples of selective sweeps and also Marques et al. (2018), Stetter et al. (2018), and Sugden et al. (2018).

2.3 Genetic Architecture Underlying Adaptive Differentiation

Positive selection acting differently among populations can result in exceptionally strong genetic differentiation in genomic regions containing loci subjected to selection (Lewontin and Krakauer 1973). For example, alleles conferring adaptation to high elevation in humans tend to be at high frequency in high-elevation populations but at low frequency in low-elevation populations in humans (e.g., Lorenzo et al. 2014; Hackinger et al. 2016). Genomic signatures of local adaptation can be detected by scanning a large number of densely mapped loci to detect genes or chromosome regions with exceptionally high genetic differentiation (e.g., F_{ST} outliers) among populations (Hohenlohe et al. 2010b; Paris et al. 2017). Small numbers (100s) of unmapped loci can be tested for adaptive signatures (broad sense genomics), particularly if candidate loci have been identified a priori (e.g., Holliday et al. 2010, 2012), but if adaptation is highly polygenic, some of the causal loci will likely be missed.

Many studies have analyzed large numbers of mapped SNPs to detect F_{ST} outlier chromosomal regions that represent candidate genomic regions for local adaptation (Hohenlohe et al. 2010b; Wang et al. 2016). Gene-environment association (GEA) analyses are also used to identify outlier loci associated with environmental

differences (Sect. 2.4; Figs. 3 and 5). Genomic regions displaying exceptionally high genetic differentiation between incipient species can also help to localize loci subjected to divergent selection during speciation (Burri et al. 2015; Ellegren et al. 2012; Harr 2006; Marques et al. 2016; Martin et al. 2013; Poelstra et al. 2014; Renaut et al. 2013; Turner et al. 2005; Wolf and Ellegren 2017).

Problems with F_{ST} outlier tests, and related tests for differentiation, include the use of the wrong null model resulting in false positives. For example, hierarchical population genetic structure can cause higher variance in F_{ST} (e.g., higher F_{ST}'s) than expected assuming a simpler model of population structure. The problem can be assessed and dealt with using simulations to simulate null distributions of F_{ST} (for 1,000s of neutral loci) for a hierarchical population structure (e.g., Lotterhos and Whitlock 2014). False negatives are another problem, which can also be caused by using the wrong or suboptimal spatial model. For example, to avoid many false negatives and increase power to detect selection, Foll et al. (2010) developed a hierarchical Bayesian to improve detection of genes involved in adaptation by humans to living at high altitude and hypoxia.

To avoid false negatives, researchers should use high SNP densities because variation in F_{ST} among SNPs is high even within a strongly selected gene. For example, SNP alleles from the lactose tolerance gene have been under strong positive selection in humans in Northern Europe (Beja-Pereira et al. 2003; Tishkoff et al. 2007). However, only 15 of 61 SNPs across the gene show significantly high F_{ST} (>0.45) between Europeans and other populations (Fig. 6). This suggests that many SNP genotyping strategies (e.g., SNP chips, restriction site-associated DNA sequencing, targeted sequencing) will often have too few SNPs per gene region to reliably detect molecular signatures of adaptive genetic differentiation and perhaps other selection signatures as well (Luikart et al. 2003).

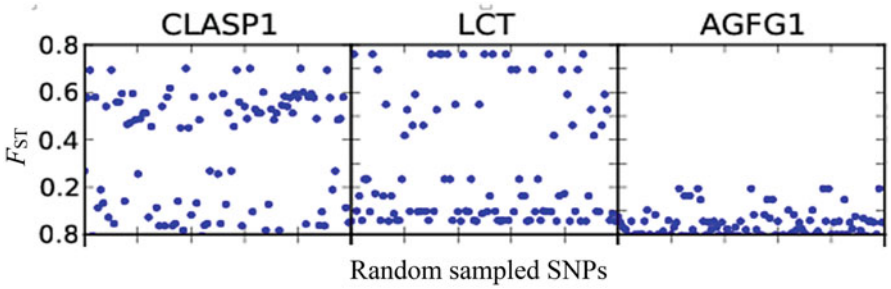

Fig. 6 F_{ST} for individual SNPs (dots) randomly sampled from across each of the two genes (CLASP1 and LCT, human chromosome #2) having the highest proportion of SNPs with F_{ST} above 0.45 between the Yorubans in Africa and Utahans representing North Western Europeans. AGFG1 is a typical gene without apparent selection signatures. CLASP1 and LCT are under strong directional selection. An F_{ST} value of 0.45 is approximately the upper 99.9 percentile of empirically observed SNP F_{ST} values across the genome and above which few neutral SNPs are expected. The *x*-axis represents a randomly chosen SNP (for instance, under random sampling with replacement). Unpublished manuscript by T. Antao and Luikart

2.4 Landscape Genomics

Landscape genomics is an emerging field or approach that strives to identify environmental factors that shape neutral and especially adaptive variation and the genes and their variants that underlie local adaptation (Rellstab et al. 2015; Balkenhol et al. 2017 this book). Environmental conditions vary across time and space, and local conditions can cause fitness differences among individuals that vary for phenotypic traits on which natural selection can act (Blanquart et al. 2013; Hoban et al. 2016). These differences in traits can be associated with underlying genotypic differences and with environmental conditions. Thus landscape genomics methods test for associations among environmental factors, geo-spatial location, or phenotypic traits and genomic variation. Landscape genomics studies focus on local adaptation to environmental conditions within and among different geographic locations (Rellstab et al. 2015; Hoban et al. 2016). The topic of landscape genomics is discussed in detail in the chapter by Balkenhol et al. (2017) in this book.

Genetic differentiation (e.g., F_{ST}) outlier tests alone do not identify the environmental factors or selective pressures driving local adaptation. However, *genotype-environment association (GEA) analyses can identify loci associated with specific environmental factors driving local adaptation*. Simulation-based studies have found that, in general, GEAs have more power than outlier-based approaches but higher rates (20–50%) of false positives (De Mita et al. 2013; Frichot et al. 2013; Forester et al. 2016). Examples of GEA-based programs are Bayenv2 (Gunther and Coop 2013) that adjusts for population structure using an independent set of markers that are assumed a priori to be neutral and the latent factors mixed model (LFMM, Frichot et al. 2013) approach that uses the covariance structure of all loci being tested to adjust for population history and demographics. There are a large number of tests and software packages available for detecting differentiation outliers and GEAs, and the number of publications using them has grown rapidly, especially for BayeScan, Bayesenv, and LFMM (Ahrens et al. 2018).

Lotterhos and Whitlock (2014) used simulations to show that reliable genetic differentiation test results vary depending on the number of individuals sampled. Their review suggests that F_{ST} outlier tests will detect a higher proportion of outliers as more individuals are sampled. This bias did not occur for GEA where the proportion of associations remained relatively constant as the total number of individuals increased. This finding implies that GEAs are more robust (see also Ahrens et al. 2018).

One recent use of multiple GEA approaches identified a congruent set of candidate genes (among approaches) that are potentially important in the local adaptation of Mediterranean striped red mullet (*Mullus surmuletus*) populations to their saline environment (Dalongeville et al. 2018). Brauer et al. (2018) used GEA analysis to test for adaptive divergence in the Murray river rainbowfish (*Melanotaenia fluviatilis*) genome associated with hydroclimate. Brauer et al. (2018) used 17,504 SNPs in a multivariate GEA framework accounting for structure of a river system to identify 146 candidate loci potentially underlying polygenic adaptive responses to

seasonal fluctuations in stream flow and periods of extreme temperature and precipitation.

Adjusting or accounting for neutral population structure is necessary to avoid a high rate of false positives with GEA analyses. However, such adjustments can result in false negatives if environmental factors driving local adaptation are correlated with population structure (e.g., from patterns of post-glacial recolonization). Yeaman et al. (2016) addressed this problem using a comparative genomics approach by identifying GEA candidate loci correlated with variation in low temperatures from exome capture and resequencing data based on raw GEA correlations in one conifer species (*Pinus contorta*). They then looked for significant GEA in those candidate loci in a second species complex (*Picea glauca*, *P. engelmannii*, and their hybrids) and vice versa. They also identified shared loci associated with phenotypic variation in cold hardiness. In this way, they identified 47 loci underlying local adaptation to cold in populations of both conifers. For additional examples involving gene expression and epigenetics, see below.

2.4.1 Spatial Signatures of Polygenic Adaptation

Adaptive traits are often polygenic and controlled by a large number of alleles from many loci each having small phenotypic effect (Bourret et al. 2014; Laporte et al. 2016; Stölting et al. 2015; Sork 2016; Yeaman et al. 2016; Boyle et al. 2017). However, methods for detecting adaptive genetic variation often only have the power to detect loci and alleles with large phenotypic effects (Wellenreuther and Hansson 2016). GEA methods can potentially detect weak signatures of adaptation but still might seldom detect alleles with small effect sizes (Coop et al. 2010; Joost et al. 2007).

Many of the early gene-environment association (GEA) methods tested only a single locus at a time, rather than looking at the combined effects of multiple loci simultaneously (Rellstab et al. 2015). More recent work has suggested that multivariate approaches (e.g., redundancy analysis (RDA), canonical correlation analysis (CCA), or using a population graph approach) might help reduce the number of false positives and maintain reasonable power to detect associations under even conditions of weak, multilocus selection (Rajora et al. 2016; Forester et al. 2018). However, multivariate approaches remain seldom used in population genomics literature (Rajora et al. 2016; Wellenreuther and Hansson 2016).

A recent study tested for polygenic signatures of local adaptation using multivariate approaches and 6605 RADseq SNPs in an Australian endemic fish, Murray cod (*Maccullochella peelii*) (Harrisson et al. 2017). The polygenic multivariate method (redundancy analysis, RDA) supported comparable roles of climate (temperature- and precipitation-related variables) and geography in shaping the distribution of multiple SNP genotypes across the range of Murray cod. Among the candidate SNPs identified by these multivariate and the univariate methods, the top 5% of SNPs contributing to significant RDA axes included 67% of the SNPs identified by univariate methods. The results highlight the value of using a combination of

different approaches, including polygenic methods, when looking for signatures of local adaptation in landscape genomics studies.

2.4.2 Landscape Community Genomics: Identifying Loci Underlying Both Species and Landscape Interactions

Genomic variation is influenced by complex interactions between abiotic (e.g., environmental) and biotic (e.g., community) effects. Researchers should consider the effects of both environmental and community factors on evolutionary dynamics simultaneously to avoid potentially incomplete, spurious, or erroneous conclusions about the mechanisms driving patterns of genomic variation among and within populations. Any study of genomic variation and adaptation in nature would ideally begin with a set of predicted abiotic and biotic drivers, including interactions between these two fundamental categories of effects (Hand et al. 2015b). Despite the value of studying concordant patterns of genetic variation in interacting species, there are relatively few empirical examples, in part because of the expense of conducting population genomics on multiple interacting species across heterogeneous landscapes or environmental gradients. Few examples exist but will become more common as it becomes feasible to conduct landscape genomics on multiple interacting species (e.g., see Beja-Pereira et al. 2003).

One recent example of landscape community genomics is a study of the parasitic Alcon blue butterfly (*Phengaris alcon*) and its two hosts: an ant species (*Myrmica scabrinodis*) and the marsh gentian (*Gentiana pneumonanthe*) (De Kort et al. 2018). The female butterfly lays its eggs onto gentian flower buds which develop into caterpillars at the expense of the gentian's ovules. This has led to coevolutionary shifts in flowering phenology to escape peak times of infestation by the Alcon butterflies (Valdés and Ehrlén 2017). When the caterpillars leave the plant, they are adopted by *Myrmica* ants as the caterpillar's chemical signature misleads the ants into accepting and rearing the caterpillar in preference to their own brood. This social parasitism of ants has also lead to coevolutionary changes in the surface chemistry of Myrmica and in the Alcon butterfly larvae (Nash et al. 2008). De Kort et al. (2018) focused on the impact of habitat fragmentation on the Alcon butterfly and subsequently the possible effect on its two obligatory host species (ants and gentians). Some of the among-population genetic variation in the host species could be explained by abiotic variables (e.g., altitude). Additional analyses showed a substantial amount of variation in Alcon butterfly genetic structure could be explained by host genetic structure. De Kort et al. (2018) then suggested that coevolutionary selection has been important in synchronizing genetic structure of this host-parasite system. Habitat fragmentation is impacting the Alcon butterfly (*Phengaris rebeli*) and will likely impact the genetic structure of its host species as well.

2.5 Genome-Wide Association Studies: Loci Associated with Traits Within Populations

A growing number of population genomics studies have identified loci contributing to phenotypic variation among individuals, including in traits that strongly affect fitness and local adaptation, via genome-wide association studies (GWAS). GWAS typically use a regression model (e.g., a linear mixed-effects [LME] model) to identify loci where genotypes are associated with a trait of interest (Gibson 2018). Population structure is accounted for by fitting a genomic-relatedness matrix (GRM) as a random effect; other potentially informative predictor variables can be included as needed in the random or fixed effects parts of the model. Additional discussions of GWAS and heritability estimation, with emphasis on functional genomics, is provided in the chapter by Pino Del Carpio et al. (2018) in this book (see also Santure and Garant 2018; Armstrong et al. 2018).

The number of studies finding loci associated with variation in fitness-related traits in natural populations is proliferating. Trait-associated loci are often identified in regions that show strong genetic differentiation between individuals with stark differences in morphology. For example, SNPs around the *RXFP2* gene included on a 50K SNP array were found to be associated with horn morphology in wild feral Soay sheep (*Ovis aries*) (Johnston et al. 2011, 2013). Horn morphology strongly affects fitness in Soay sheep (*Ovis aries*) and in natural populations of wild mountain sheep (e.g., bighorn sheep, *Ovis canadensis*; Hogg 1984). Thus identifying loci associated with horn size provides an interesting look into the genetic basis of fitness-related variation.

In another recent GWAS example, Brelsford et al. (2017) studied a natural hybrid zone between Audubon's and myrtle warblers (*Setophaga coronata auduboni* x *S. c. coronata*) to identify genomic regions associated with color pigmentation potentially associated with mating success and fitness. RADseq produced 154,683 to 393,755 SNPs, depending on the filtering criteria. For each of five plumage coloration traits studied (eye spot, throat color, eye line, wing bar, and auricular), the authors detected highly significant associations with multiple SNPs genome wide that clustered into chromosomal regions (Fig. 7). The high success in identifying loci associated with these traits likely resulted from the relatively high gametic disequilibrium along chromosomal stretches resulting from hybridization.

In another study, Husby et al. (2015) identified a locus that was associated with clutch size (a life history trait) in the collared flycatcher (*Ficedula albicollis*). Similarly, Bérénos et al. (2015) identified two SNPs in Soay sheep (*Ovis aries*) associated with leg length (a measure of body size), with each of the two SNPs explaining $>10\%$ of the additive genetic variance in the trait. One of the SNPs found to be associated with leg length by Bérénos et al. (2015) was also associated with female reproductive success, providing evidence for a link between genotype, phenotype, and fitness in Soay sheep. Lamichhaney et al. (2015) and Küpper et al. (2015) simultaneously identified a large (~4.5 Mb) inversion that controlled mating morphology in the ruff (*Philomachus pugnax*). Barson et al. (2015) identified a locus with sex-specific dominance and large effects on age at maturation in wild Atlantic salmon.

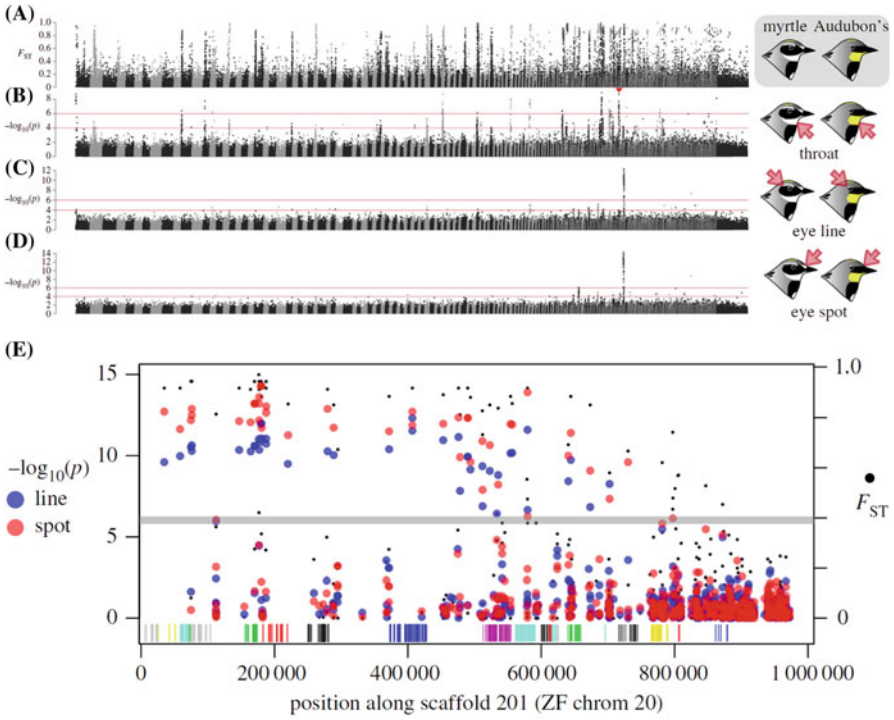

Fig. 7 Manhattan plots of genomic differentiation (**A**) and plumage associations (**B**, **C**, **D**). (**A**) F_{ST} between allopatric myrtle and Audubon's warblers at 393,755 SNPs across the genome with scaffolds ordered by size. Adjacent scaffolds across the genomes are distinguished by alternating gray or black coloration. Panels **B**, **C**, and **D** are phenotype-genotype associations for three of the five plumage characters studied. The tiny red triangle near the top right of panel (**B**) shows the cluster of loci that aligns to the zebra finch chromosome 15. This region includes the SCARF2 gene, which is a strong candidate gene for carotenoid pigment transport. Panel (**E**) shows patterns of divergence and genotype-phenotype associations for eye line (blue points) and eye spot (red points) for a region of chromosome 20. Associations between these two traits are highly correlated with each other as well as patterns of divergence (F_{ST}, small black dots). Coding regions (exons) for genes are shown by the vertical bars, with different adjacent genes colored differently with arbitrarily chosen colors. Modified from Brelsford et al. (2017)

GWAS methods are also being widely used in conjunction with common garden experiments containing natural or seminatural populations of plants, fish, and other taxa. For example, in black cottonwood (*Populus trichocarpa*), Mckown et al. (2014) conducted GWAS using 29,355 filtered SNPs using a unified mixed model accounting for population structure effects. They uncovered 410 significant SNPs (from 275 genes) across 19 chromosomes that explained 1–13% of trait variation in trait associations, mostly associations with phenology genes (240 genes) but also biomass (53 genes) and ecophysiology (25 genes).

In the future, association studies will continually find more loci, including loci of small effects associated with adaptive traits, thanks to improved power from sequencing strategies like pool-seq with a reference genome that allow high-density

genotyping of populations or lineages (Haussler et al. 2009; Schlötterer et al. 2014; Wessinger et al. 2018; Pruisscher et al. 2018). For example, Narum et al. (2018) used a new genome assembly (2.8 Gb) and pool-seq resequencing for Chinook salmon (*Oncorhynchus tshawytscha*) to conduct association mapping of important life history traits. The authors pooled individuals from populations of each of three phylogenetic lineages that exhibit different maturation and run-timing phenotypes. Their whole-genome resequencing of pooled (barcoded) individuals suggested that divergent selection was extensive at many loci genome wide within and among phylogenetic lineages. Association mapping with millions of SNPs revealed a genomic region of major effect associated with phenotypes for migration timing. This study illustrates how a genome assembly and high-density markers can help resolve the genetic basis of important phenotypes.

2.6 *Quantifying Inbreeding, Inbreeding Depression, and Historical Bottlenecks*

The availability of population genomic data is improving our understanding of inbreeding (mating between relatives) and inbreeding depression in the wild (Hedrick and Garcia-Dorado 2016; Kardos et al. 2016a). Inbreeding causes offspring to be homozygous and "identical by descent" (IBD) across large chromosomal segments where the two inherited DNA copies arise from a single DNA copy in a common ancestor of the parents (Kardos et al. 2016a; Speed and Balding 2015; Thompson 2013). The increased homozygosity arising from IBD causes inbreeding depression: reduced fitness of inbred individuals (Charlesworth and Willis 2009).

The pedigree inbreeding coefficient (F_P) is a traditional measure of individual inbreeding and predicts the fraction of the genome that is IBD, assuming that pedigree founders are unrelated and noninbred (Keller and Waller 2002; Malécot 1970; Wright 1922). However, F_P can be an imprecise measure of the realized fraction of the genome that is IBD (F) due pedigree errors, the stochastic nature of Mendelian segregation and recombination, and the presence of related and inbred pedigree founders (Fisher 1965; Franklin 1977; Stam 1980; Kardos et al. 2016a; Knief et al. 2017; Forstmeier et al. 2012; Goudet et al. 2018). The imprecision of F_P and the recent availability of genomic data have led to increased application of genomic estimates of individual inbreeding and inbreeding depression (Hoffman et al. 2014; Huisman et al. 2016; Bérénos et al. 2016).

Genomic measures of individual inbreeding have the advantage that they directly measure patterns of homozygosity across the genome, thus making pedigrees unnecessary to estimate individual inbreeding. Encouragingly, *only a few thousand unmapped SNP loci can provide more precise estimates of F (IBD) than a pedigree five to ten generations deep* (Kardos et al. 2015a, 2018). Even more powerful, the analysis of many tens of thousands of mapped loci allows the use of runs of homozygosity (ROH) residing within chromosomal segments that are IBD to assess inbreeding (IBD) with very high precision (Kardos et al. 2015a). Genomics studies of inbreeding are greatly advancing our understanding of the extent of inbreeding

depression in humans, domestic animals and plants, and natural populations of non-model organisms (Palkopoulou et al. 2015; Xue et al. 2015; Kardos et al. 2018).

ROH can be used to identify and map loci contributing to inbreeding depression by testing for associations between the presence of ROH and individual fitness-related traits (Keller et al. 2012; Kijas 2013; Lander and Botstein 1987; Pryce et al. 2014). Large-scale genomics studies of inbreeding depression (sample sizes >100,000 individuals) based on ROH and other genomic measures of inbreeding are now being done to precisely estimate inbreeding effects on a wide range of human traits (Wessinger et al. 2018; Johnson et al. 2018a, b). Thus, population genomics is beginning to contribute substantially to our understanding of the evolution of fitness-related phenotypes and the genetic basis of inbreeding depression in many species. This understanding has the potential to guide conservation and management of wild population and captive breeding programs, for example, to avoid inbreeding depression and invoke genetic rescue through restoring gene flow (Tallmon et al. 2004; Whiteley et al. 2015).

In another step to identify contributing loci, exons identified by ROH can also be used to bioinformatically identify likely deleterious alleles based on the likely effects of amino acid substitutions and whether such substitutions are common in homologous genes in other organisms using software such as PROVEAN (Choi and Chan 2015). The frequencies of these alleles can be compared among individuals and populations. For example, Conte et al. (2017) found over 13% of all SNP alleles in *Picea engelmannii*, *P. glauca*, and hybrid populations had amino acid substitutions predicted to be deleterious, but homozygous genotypes for deleterious alleles were less frequent in hybrid populations due to complementation.

Historical effective population size can be qualitatively inferred from the abundance and length distribution of runs of homozygosity (Fig. 8). For example, analyses of genome-wide runs of homozygosity (ROH) showed inbreeding arising from recent common ancestors of parents (due to small population size) in individuals of recently reintroduced populations of alpine ibex (*Capra ibex*). The detected ROH were associated with small population size during captive breeding and the founding of small wild populations approximately 20 generations ago. In spite of a rapid population growth in the wild, the ibex carried a genomic signature of their small recent historical population size (Fig. 8). The authors thus suggested that genomic monitoring for ROH could provide an improved indicator for early detection of inbreeding in wild and managed populations (Grossen et al. 2018).

Historical population bottlenecks can also be inferred and approximately dated using ROH and coalescent modeling (Ceballos et al. 2018). Palkopoulou et al. (2015) sequenced genomes from two wooly mammoths from distant populations in terms of both geography (northeastern Siberia versus Wrangel Island, Alaska) and time (~44,800 versus ~4,300 YBP). Intriguingly, both yielded very similar genomic signatures of a nearly identical population decline at the start of the Holocene. One mammoth individual sample was dated to have died just before the species' went extinct approximately 4,000 years ago. From coalescent modeling, a second genomic signature of a reduced population effective size (and inbreeding) was inferred just before the extinction at the start of the Holocene. The analyses suggested that the wooly mammoth was subject to reduced genetic variation prior to its extinction.

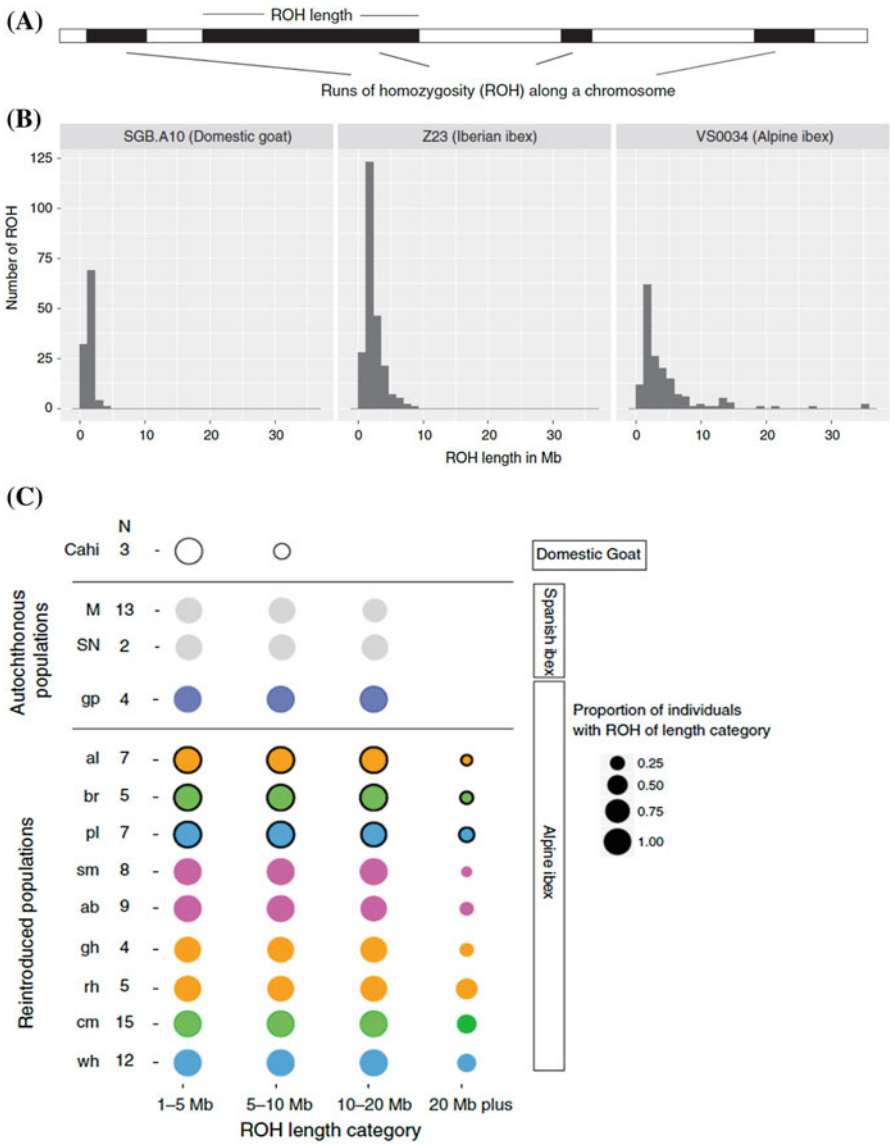

Fig. 8 (**A**) Schematic showing runs of homozygosity (ROH) along a chromosome. (**B**) Distribution of total genome-wide runs of homozygosity in one representative individual from each of three species including domestic goat (SGB A10), Iberian ibex (Z23), and Alpine ibex (VS0034). The distribution is right-shifted to have longer ROH, >10–20 Mb, in the reintroduced Alpine ibex. (**C**) Tract length distribution of ROH in wild and reintroduced populations. ROH for individuals from different populations show a range of different tract lengths. Only the reintroduced (captive bred, bottlenecked) individuals have 20 Mb tracts. The wild source population GP (Gran Paradiso) never suffered the captive breeding founder effects, but it did decline to ~100 individuals approximately 100 years ago. Black-outlined circles show the three primary reintroduced populations Albris (orange), Pleureur (light blue), and Brienzer Rothorn (green). Secondary reintroductions established from the primary reintroduced populations share the same color. Populations with mixed ancestry are shown in purple. N, sample size per population. Reproduced with permission from Grossen et al. (2018)

2.7 Delineating Adaptively Differentiated Populations

Population genomics can help identify locally adapted, differentiated populations that are difficult to delineate using selectively neutral markers, especially in high gene flow species, such as forest trees and marine organisms. Prince et al. (2017) used RADseq to assess the evolutionary basis of premature migration among individuals within local populations of Pacific salmonids. Chinook salmon and also steelhead trout exhibit two major migration strategies: premature migrators enter freshwater in the spring with high fat content and stay in freshwater for months until spawning, and mature migrators which enter freshwater sexually mature just prior to the spawn. Gene flow was relatively high between the two very different forms (premature vs normal migration) within a stream ($F_{ST} \sim 0.03$); F_{ST} between streams was far higher ($F_{ST} \sim 0.13$). The authors found the same single locus associated with premature migration in multiple populations in each of two different species, Chinook salmon and also steelhead trout.

Results from this study suggest conservation implications. While many traits involved in local adaptation are polygenic, in this case a single locus appears to control migration timing and has significant economic, ecological, and cultural importance (Fig. 9). In particular, extirpation of the premature migration allele and phenotype are unlikely to re-evolve once extirpated from a population in the absence of immigrants carrying the allele from elsewhere. Mutations producing a given important allele are rare evolutionarily, suggesting such alleles will not re-evolve quickly or easily if lost. Furthermore, spatial patterns of adaptive allelic variation can differ from patterns of overall population genetic differentiation. Taken together, these results suggest that conservation units based on genome-wide patterns of genetic differentiation will sometimes fail to protect evolutionarily significant genetic and phenotypic variation.

Adaptively differentiated populations can be identified and prioritized for conservation and breeding (Funk et al. 2012). Population genomics and landscape genomics approaches are often necessary to identify adaptively differentiated populations because common garden or reciprocal transplant experiments are not feasible for many species. Bonin et al. (2007) devised a population adaptive index (PAI), which uses both neutral and adaptive distinctiveness to assess the adaptive value of the population. They suggested that outlier tests could help identify adaptive loci and alleles to then use to identify and prioritize or rank populations for conservation values. In species to which they applied the index (PAI), the neutral and adaptive marker variation among populations were not correlated; Therefore the authors concluded that conservation strategies based on the neutral and adaptive indexes would not protect the same populations.

Other authors have suggested genomics approaches be used to identify adaptively differentiated populations (Funk et al. 2018; Razgour et al. 2018; Hoban 2018). Approaches include genotype-environment associations and gene expression analysis (e.g., Hansen 2010; Chen et al. 2018, see Sects. 2.4 and 4.4). *Including environmental variables improves power over differentiation-based methods, helps identify the environmental drivers of adaptation, and facilitates detection of*

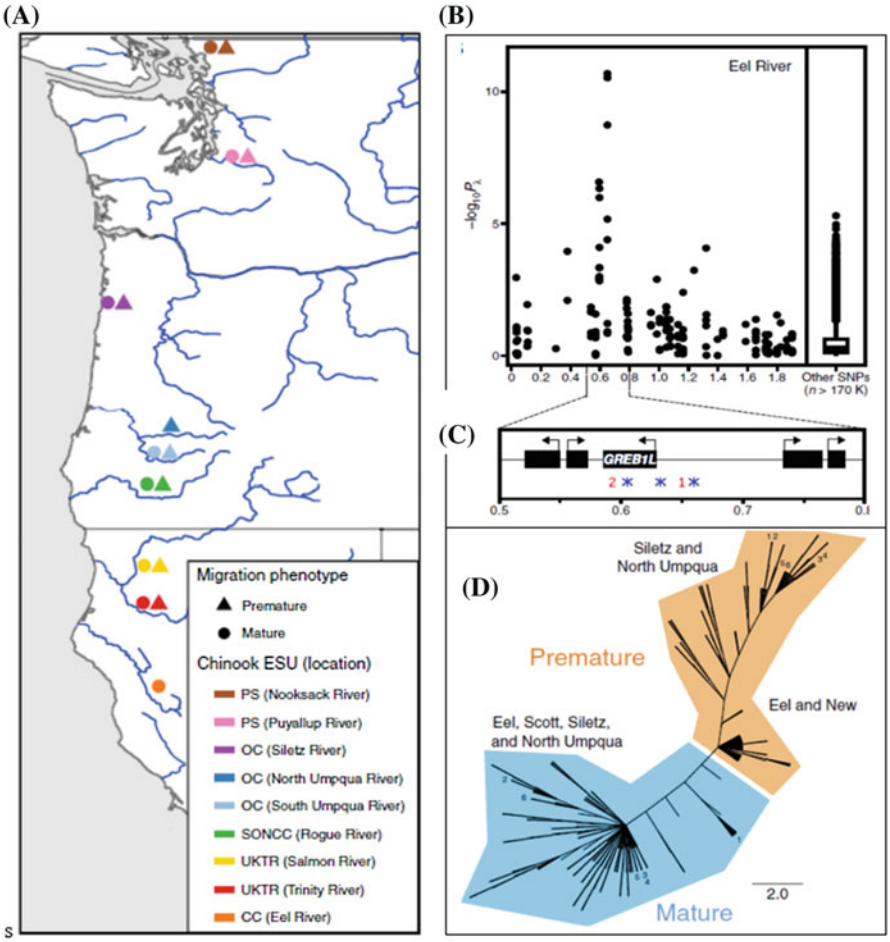

Fig. 9 Genomic basis of premature migration in steelhead. (**A**) Map of sampling locations of early versus mature (normal) migration types of steelhead trout sampled together in each of many drainages. (**B**) Association mapping of early vs normal migration of the Eel River steelhead trout with gene annotation, with the (**C**) gene annotation of a region with strong association; red numbers show genomic locations of the two RAD restriction sites with strongest associated SNPs, and blue asterisks indicate positions of amplicon sequencing, with the candidate gene GREB1L. (**D**) Phylogenetic tree depicting maximum parsimony of phased amplicon sequences from all individuals; branch lengths, with the exception of terminal tips, reflect nucleotide differences between haplotypes; numbers identify individuals with one haplotype in each migration category clade (i.e., heterozygotes for premature and normal migration haplotypes). Reproduced with permission from Prince et al. (2017)

contemporary (and historical) selection (Forester et al. 2018). Ideally, multiple independent data types would be combined to maximize power and reliability of delineating adaptively differentiated populations (geography, environment, behavior, ecology, physiology, transcriptomics, and genomics; Allendorf et al. 2013).

There is enormous risk of prioritizing populations for conservation based on population genomics (or outlier) approaches alone. It can be extremely difficult or

impossible to verify whether genes that behave as outliers are genuinely adaptive. The genomic signatures expected from local adaptation (e.g., F_{ST} outliers, GEA) can arise from genetic drift, particularly when small populations and low migration rates are involved. Further, genuine genomic signatures of selection may be due to selective forces in deep history that have since disappeared and thus are irrelevant to adaptation in current or future environments. Third, prioritizing certain populations based on certain particular alleles (even if they are genuinely relevant to adaptation) could actually reduce diversity across the rest of the genome that is necessary for future adaptation (Luikart et al. 2003; Allendorf 2017; Kardos and Shafer 2018).

2.8 *Speciation, Hybrid Zones, Admixture, and Adaptive Introgression*

Population genomics approaches have opened new avenues to study speciation, admixture events, and hybrid zones in all organisms. A detailed account of this topic is presented by Nadeau and Kawakami (2018) in this book. Here we introduce the topic and provide a few relevant examples.

The European bison (*Bison bonasus*), Europe's largest land mammal, was recently shown to be a hybrid of two previously recognized subspecies, by authors using low coverage genome sequence alignments of historical and modern individuals (Wecek et al. 2017). Admixture occurred between subspecies prior to extinction in the wild and also subsequently during recent captive breeding. Admixture with domestic cattle was also significant but was ancient rather than from recent hybridization with domestics. These discoveries would have been difficult or impossible without genome-wide mapped loci and both historical and modern samples.

Kovach et al. (2016) studied genome-wide patterns of admixture and natural selection across recently formed hybrid zones between native cutthroat trout and invasive rainbow trout (*Oncorhynchus clarki lewisi and O. mykiss*) by genotyping 9,380 species-diagnostic RADtag SNP loci. A significantly greater proportion of the genome appeared to be under selection favoring native cutthroat trout (rather than rainbow trout), in the local native environments. This negative selection against rainbow introgression was found on most chromosomes and was consistent among populations and environments, even in warmer environments where rainbow trout were predicted to have a selective advantage. These data are consistent with previous findings that admixed fish have reduced reproductive success (Muhlfeld et al. 2009). Future studies could use far more loci to precisely map tracts of hybridity and infer timing of introgression of the rainbow haplotype segments into the native cutthroat trout.

Among the most intriguing examples of natural selection favoring "adaptive introgression" of certain alleles following admixture is the introgression of advantageous alleles from Neanderthals (and Denisovans) into modern humans. Genes involved in sugar metabolism, muscle contraction, and oocyte meiosis have been influenced by adaptive introgression from Neanderthals. For example, *EPAS1* which influences hemoglobin concentration and response to hypoxia has introgressed from Denisovans into Tibetans, facilitating adaptation to life at high altitude through ancient

admixture (Huerta-Sánchez et al. 2014). Other benefits of archaic (Neanderthal) introgression in the past are associated with several neurological and dermatological traits (Kelso and Prüfer 2014; Racimo et al. 2015; Vattathil and Akey 2015).

Evidence for adaptive introgression in nonhuman populations is growing. For example, adaptive introgression was detected in the Tibetan mastiff (*Canis domesticus*). Alleles for adaptation to high elevation (hypoxia) were identified at several loci, including the EPAS1 and HBB, which were introgression from Tibetan gray wolves (*Canis lupus*) (Miao et al. 2017). This demonstrates that domestic animals could rapidly become locally adapted by secondary contact with their wild relatives.

Adaptive introgression was also associated with the evolution of seasonal variation in coat color in snowshoe hares (Jones et al. 2018). Snowshoe hare populations molt to white during winter in order to maintain camouflage in environments with consistent winter snow cover. However, snowshoe hares in areas that remain snow-free year round often retain their brown coat color during the winter, thus maintaining effective camouflage in the absence of winter snow. The brown winter coat in snowshoe hares appears to arise from an allele that has introgressed from black-tailed jackrabbits (Jones et al. 2018). Other studies have also shown interesting genome-wide patterns of adaptive introgression (Song et al. 2011; Rieseberg 2011; Pardo-Diaz et al. 2012; Norris et al. 2015; Ozerov et al. 2016; Saint-Pé et al. 2018).

New approaches to analyze mapped loci will advance understanding of hybridization and evolution in hybrid zones. For example, large numbers of mapped loci can be analyzed to infer "local ancestry" across genomes of individuals. This involves mapping the locations of haplotypes arising from different source populations across the genomes of hybrids (Guan 2014; Leitwein et al. 2018). Such *ancestry tracts can be used to estimate individual hybridity and population level admixture at both the genome wide and local scale across chromosomes*. Additionally, local ancestry information is highly useful for trait mapping in mixed-ancestry populations (Smith and O'Brien 2005). Because the introgressing haplotypes decay in length at a predictable rate with increasing generations since hybridization, analyses of ancestry tract lengths can be informative of the historical timing of admixture events. For example, Leitwein et al. (2018) used 75,684 mapped SNPs obtained from double-digested RAD to identify ancestry tracts and estimate individual admixture proportions along with the timing of admixture in brown trout (*Salmo trutta*).

3 Benefits of Mapped Loci in Population Genomics

Information on the location of loci in the genome is a defining characteristic of population genomics (narrow sense), as mentioned above (Allendorf 2017). Loci can be mapped in terms of physical and/or genetic (linkage) positions in the genome. Producing both physical and linkage maps is far more tractable with modern genomics methods in non-model organisms than a few years ago. As a result, population genomics research efforts can now feasibly include the construction of a physical or linkage map for most study systems. Below we describe the key

features of physical and genetic maps and the relative value of each for population genomic analyses.

Physical and genetic (linkage) mapping are two separate but complementary ways of describing the locations of loci in the genome. A physical map is a genome sequence. Long sequence reads from new (third)-generation sequencers enable high-quality genome assemblies, discovery of novel fitness-affecting structural variation, and the ability to sequence through previously "unsequenceable" repetitive DNA to allow mapping between distant loci along each chromosome. Reference genomes for non-model organisms often, however, are not assembled into chromosomal units, especially when genomes are large and contain a high fraction of highly repetitive content (i.e., retrotransposons) (Ellegren 2014; Epstein et al. 2016).

A linkage map describes the gene order based on the recombination frequency between loci along each chromosome. Linkage maps are constructed by genotyping pedigreed individuals and using linkage analysis, which quantifies how often adjacent loci co-segregate versus segregate independently due to recombination during meiosis. The distance between loci on a linkage map is described in terms of centimorgans (cM), where 1 cM is defined as a 1% recombination frequency between two adjacent loci inherited from a parent. Linkage maps can be developed in some cases where assembly of physical maps remains difficult (e.g., large conifer genomes, De La Torre et al. 2014).

Both physical and linkage maps facilitate population genomics research in at least five ways. First, having large numbers of mapped loci improves the power to identify and localize loci influencing phenotypic variation, fitness, and adaptation (e.g., Burri et al. 2016; Rastas et al. 2016). For example, the availability of densely mapped SNPs along a chromosome allows for localization of the chromosomal region(s) and genes underlying traits or adaptations (Figs. 2, 3, 7, and 10). This helps determine the genetic basis of adaptations or phenotypic variation, including determination of the number, kind, and effect size of genes underlying an adaptation or trait.

Second, physical and linkage maps also help identify independent loci, e.g., loci far apart on the same chromosome or on different chromosomes (although statistical tests for independence can identify independent loci without a map). Independent loci are required for some population genetic inferences, including analyses of effective population size (N_e), gene flow, or population relationships (Landry et al. 2002; Storz et al. 2002; Luikart et al. 2003). For example, Larson et al. (2014) estimated N_e for wild Chinook salmon using ~10,000 SNPs and the LDNe method (based on gametic disequilibrium) that assumes all loci are independent or not physically linked (Waples and Do 2010). Estimates using only pairs of SNPs from different chromosomes (<1,000 SNPs) consistently gave estimates of N_e that were higher than when using all pairs of SNPs; for example, an N_e estimate was 1,909 for unlinked SNPs versus only 808 for all SNPs (including linked SNPs), as expected because gametic disequilibrium is stronger for physically linked SNPs, which drives (biases) lower N_e estimates.

Third, the combination of a linkage map and a physical genome assembly allows understanding variation in the recombination rate across the genome. This is important because the recombination rate affects the extent of GD (gametic

disequilibrium) and genetic diversity across a chromosome. Lower recombination rates result in GD extending over longer physical distances across a chromosome. As described below, the extent of significant GD strongly influences the power to detect footprints of natural selection and the ability to map loci contributing to phenotypic variation.

The recombination rate influences genetic diversity and differentiation among populations or species via its interaction with natural selection. *Knowing how recombination rate varies across the genome is, therefore, crucial for interpreting genomic patterns of genetic diversity and differentiation*. For example, the recombination rate is known to interact with background selection to generate chromosomal islands of reduced diversity (Charlesworth et al. 1993) and increased differentiation (high $F_{ST,}$ Burri et al. 2015), which might be erroneously interpreted as resulting from positive selection.

Fourth, physical and linkage maps both help researchers determine if they have a sufficient density of loci in the genome to have high power to detect loci subjected to positive selection or genotype-phenotype associations. With a linkage map, researchers can compute how far in centimorgans (cM) significant GD spans across chromosomes or linkage groups. Similarly, with a physical map, researchers can compute how far in base pairs (or kb) GD spans across chromosomes. Knowing the extent of GD is important because detection of phenotype-genotype associations and signatures of selection required GD between genotyped loci and causal loci. In addition, detecting phenotype-genotype associations requires GD between genotyped marker loci and causal loci, and so a relatively high density of markers is needed (Box 2).

Box 2 Importance of Gametic Disequilibrium and Marker Density for Identifying Adaptive Loci

Researchers recently resequenced 81 whole genomes in flycatchers with extreme phenotypes and also genotyped 50K SNP in 415 individuals. Birds were phenotyped for forehead patch size, a sexually selected trait associated with reproductive success. No SNPs were significantly associated with patch size (Fig. 10A). One reason for the failure to detect loci (QTL, quantitative trait loci) using association mapping could be that gametic (linkage) disequilibrium extends only over short chromosome distances (Fig. 10B), which makes the chances of strong associations between a DNA marker and trait loci small even when genotyping many SNP markers (Lowry et al. 2017, but see McKinney et al. 2017a; Catchen et al. 2017).

These results suggest that reliably detecting large-effect trait loci in large natural populations will often require thousands of individuals and the genotyping of hundreds of thousands of loci across the genome. Encouragingly, far fewer individuals and loci will often be sufficient to achieve high power to detect large-effect loci in small populations that typically have widespread strong gametic disequilibrium. This study illustrates the

(continued)

Box 2 (continued)
importance of knowing if strong gametic disequilibrium extends over long chromosome distances (e.g., due to low recombination rates, small effective populations size and drift, or perhaps admixture).

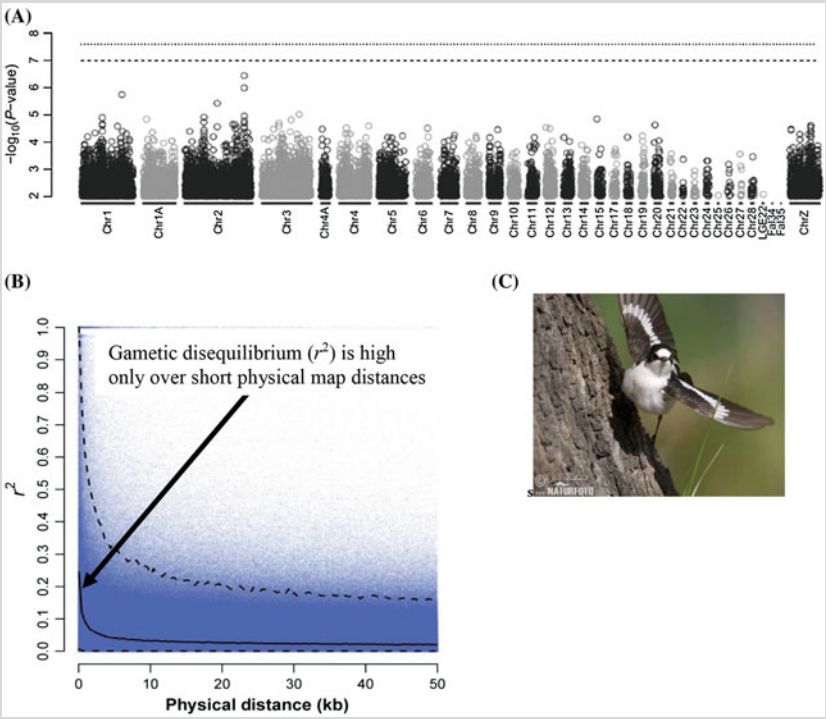

Fig. 10 (**A**) Manhattan plot of −log10 (*P*-value) from a genome-wide association (GWA) analyses of color-patch size based on whole-genome resequencing of 81 male flycatchers. Chromosome identity is shown on the *x*-axis, and the *P*-values (open circles) are arranged according to physical SNP positions on each chromosome. Horizontal dashed lines are permutation-based statistical significance thresholds, and the dotted lines are the Bonferroni statistical significance thresholds of statistical significance (no points above the dashed line). (**B**) The relationship between the strength of linkage disequilibrium (r^2 or nonrandom association between loci) and physical distance in 81 whole-genome resequenced collared flycatcher males. r^2 is shown for each pair of SNPs separated by 50 or fewer kb. The solid line represents a function fitted to the rolling mean of r^2 calculated in nonoverlapping windows of 100 bp. The arrow shows where the mean of r^2 drops below 0.20. The dashed lines represent loess functions fitted to the rolling 5% and 95% quantiles of r^2 in the same nonoverlapping 100 bp windows. (**C**) Collared flycatcher photo (note forehead patch). (**A**, **B**) Reproduced with permission from Kardos et al. (2016b). (**C**) Copyrighted license and permission to use photo from Jiri Bohdal

We caution that while maps allow quantification of the extent of GD along chromosomes, this quantification must be conducted for each study population of interest because the extent of GD varies among populations with in a species (Table 2) (Whiteley et al. 2011; Gray et al. 2009). GD will be relatively higher (genome wide) in populations with small N_e and/or recent admixture (Fig. 11). Quantifying GD along chromosomes also allows researchers to identify hotspots of recombination (low GD) and thus to know which genome regions will require higher densities of markers when screening for loci associated with adaptation or phenotypic variation.

Directional selection is expected to reduce genetic variation and to alter the site frequency spectrum at the selected site and at closely linked loci (Charlesworth et al. 1993). The expected physical distance over which selection affects genetic variation depends on the local recombination rate. We expect directional selection to affect genetic variation across larger regions when the local recombination rate is low. As described below, accounting for recombination rate variation across the genome is necessary in order to assess differentiation among populations (e.g., F_{ST}) measured across each chromosome. Information on recombination patterns (genome wide) improves interpretation of population genomic tests (GWAS, F_{ST} outliers, etc.) because recombination can influence outlier locus behavior. For example, the rate of recombination is expected to correlate positively with local nucleotide diversity and rates of adaptive evolution, which could influence tests for selective sweeps using heterozygosity or F_{ST} outlier loci (Cutter and Payseur 2013; Campos et al. 2014).

Fifth and finally, GD information from linkage or physical maps can improve theoretical models to advance population genetics beyond bean-bag genetics. Models parameterized with chromosomally explicit GD information can help to understand issues such as the importance of interactions of gene flow, recombination, and selection in adaptation and speciation. Some models stress the importance of recombination and distance among loci in the establishment and maintenance of adaptive alleles in a population (Bürger and Akerman 2011; Yeaman and Whitlock 2011; Feder et al. 2012).

3.1 What Can Physical Maps Provide that Linkage Maps Cannot?

Physical maps (reference genomes) generally provide higher power than linkage maps for detecting selective sweeps or genotype-phenotype associations because millions of SNPs can be mapped (positioned) via sequencing, whereas it is difficult to produce linkage maps with more than approximately 20–30K SNPs. Linkage mapping for tens of thousands of SNPs can require genotyping of many families, which is difficult or impossible in most species due to small family sizes, unavailability of families, or large expense of genotyping tens of thousands of loci

Table 2 Estimated chromosomal length in kilobase pairs (kb) with moderate gametic disequilibrium ($r^2 = 0.2$) in populations from diverse species

Species [population] (Genus species)	Mean kb with moderate GD[a]	Reference
Flycatchers (*Ficedula albicollis*)	<2	Kardos et al. (2016b)
Mosquito (*Anopheles arabiensis*)	1	Marsden et al. (2014)
Mosquito (*Anopheles gambiae*)	<0.5	Harris et al. (2010)
Honey bee (*Apis mellifera*)	0.5	Wallberg et al. (2014)
Bighorn sheep [bison range population] (*Ovis canadensis*)	>4,000	Miller et al. (2014)
Bighorn sheep [Ram Mountain population] (*O. canadensis*)	<2,000	Miller et al. (2014)
Zebra fish [lab strain] (*Danio rerio*)	>3,000	Whiteley et al. (2011)
Zebra fish [wild population] (*Danio rerio*)	<20	Whiteley et al. (2011)
Murry cod (*Maccullochella peelii*)	5	Harrisson et al. (2017)
Rainbow trout (*Oncorhynchus mykiss*)	2,000	Vallejo et al. (2018)
[b]Pig breeds [China] (*Sus scrofa*)	~10	Amaral et al. (2008)
[b]Pig breeds [Europe] (*Sus scrofa*)	~400	Amaral et al. (2008)
Pig breed [Korea] (*Sus scrofa*)	3,700	Shin et al. (2018)
Common bean [Mesoamerican] (*Phaseolus vulgaris*)	<100	Valdisser et al. (2017)
Common bean [Andean] (*Phaseolus vulgaris*)	~0.5	Valdisser et al. (2017)
Mung bean (cultivated) (*Vigna radiata*)	100	Noble et al. (2018)
Mung bean (wild) (*Vigna radiata*)	60	Noble et al. (2018)
[c]Antarctic fur seal (*Arctocephalus gazella*)	15	Humble et al. (2018)
Wolves [Alaska, Minnesota, or Canada] (*Canis lupus*)	<10	Gray et al. (2009)
Wolves [Isle Royal] (*Canis lupus*)	>5,000	Gray et al. (2009)
Wolves [Yellowstone National Park] (*Canis lupus*)	<10	Gray et al. (2009)
Wolves [Spain] (*Canis lupus*)	>1,000	Gray et al. (2009)
[d]Melon (*Cucumis melo*)	<100	Gur et al. (2017)

Gametic disequilibrium (GD) in different populations of the same species can differ by orders of magnitude, as seen here for zebra fish, pigs, beans, and wolves. The distance that moderate-to-strong GD extends along chromosomes can vary due to different effective sizes (drift), substructure, gene flow and admixture, mating system (inbreeding), recombination rates, and selection (e.g., sweeps)

[a]Publications here generally define a "moderate GD" to be $r^2 = 0.20$. Mean chromosomal distance in kilobases (kb) at which r^2 decayed to 0.2

[b]GD stretched 10 kb in physical distance and 0.5 cM in linkage map distance. In European pig breeds GD extended 400 kb physical distance and 2 cM in linkage map distance

[c]Moderate GD extended to 15 kb; strong GD ($r^2 = 0.5$) extended to 5 kb

[d]GD of $r^2 \geq 0.2$ extended from ~75 to 100 kb (and 1 to 6 cM) in different strains

in many large families. For example, a map from a single family of Chinook salmon had 5,400 SNP loci while increasing to four families allowed mapping of 13,800 loci (G. McKinney, unpublished data, 2018; see also Mckinney et al. 2016). There are diminishing returns from adding families for mapping because the number of

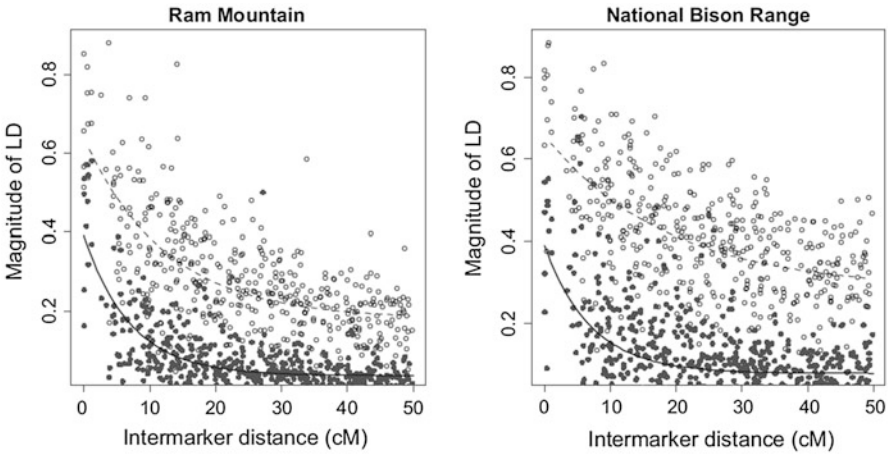

Fig. 11 Gametic disequilibrium is stronger and more variable between loci (dots) in small populations of bighorn sheep (National Bison Range, n ~50–75) compared to the moderately larger population (Ram Mountain; n ~100–200). Strong LD (magnitude >0.4, see upper dashed line) stretches over ~30 cM in the National Bison Range population but only to over ~10 cM in Ram Mountain population. Reproduced with permission from Miller et al. (2014)

additional loci that can be mapped declines as the number of families increases (unless perhaps genetically divergent families, with different variable loci are mapped).

Physical maps are also useful for improving both the process of discovery of SNP loci and of the subsequent genotyping of SNPs when using next-generation sequencing approaches such as RADseq (Sect. 4.1). For example, physical maps help identify paralogues and duplicated genes to avoid them or genotype them by allowing the alignment of sequencing reads to the physical map. If samples sizes are large, paralogs can be identified in RADseq data (e.g., see HDplot method of McKinney et al. 2017b).

Physical maps can improve genotyping by allowing the alignment of sequencing reads to the entire reference genome during the genotyping process, instead of using only a limited number of putative loci or de novo assembled loci (Hand et al. 2015a; Shafer et al. 2017). A caveat is that reference genomes are never 100% complete, and loci from missing sections of the genome will not be genotyped if doing only reference alignments for genotyping. If a genome is 90% complete, it is possible that 10% of your loci would not be mapped or genotyped when using the reference for genotyping.

Importantly, a physical map (assembly) can be used for genotyping next-generation sequencing reads from a closely related species to help improve genotyping (Cosart et al. 2011; Shafer et al. 2017). In this scenario, reads from one species are aligned to the genome for another for genotyping. This is a benefit of initiatives like Genome 10k that is providing a genome assembly for one species per genus or family of vertebrate, which provides related species a reference genome for mapping and genotyping (Haussler et al. 2009).

3.2 What Can Linkage Maps Provide that Physical Maps Cannot?

A high-density linkage map enables understanding of mechanisms (background/negative selection, positive selection, gene flow, and recombination) that cause heterogeneity along chromosomes in diversity within and differentiation between populations (Burri et al. 2015). A linkage map reveals recombination hot and cold spots which are known to interact with background selection to generate chromosomal islands of divergence (high F_{ST}). Thus, a linkage map can help prevent false positives for local adaptation and improve detection of islands of divergence that are truly indicative of local adaptation (not false positives) (Cruickshank and Hahn 2014). Regional estimates of the recombination rate also help interpret data on runs of homozygosity (RoH) to detect inbreeding and to infer demographic history because recombination hotspots influence the lengths of RoH (Thompson 2013) and the density of SNPs (Charlesworth et al. 1993) and thus the power to detect RoH in genome regions.

Chromosomal level assemblies are often not possible without a linkage map, especially for large genomes with many repetitive sequences (Amores et al. 2011). Assembled chromosomes in turn can be used for identification of chromosomal synteny and structural polymorphisms such as rearrangements (e.g., inversions) within or between species (Amores et al. 2011; O'Quin et al. 2013; Rondeau et al. 2014). Structural changes or polymorphisms can influence fitness and adaptation and thus are important to discover and map (Wellenreuther and Bernatchez 2018). Additionally, assembled chromosomes can improve genome scans for loci associated with adaptation and phenotypic variation, by allowing computation of chromosome-specific distributions of summary statistics (continuously along each chromosome), which can increase power and reliability of outlier tests.

3.3 Combining Linkage and Physical Maps: The Ideal Genomics Approach

Having both a reference genome assembly and linkage map is ideal because they complement each other, and the linkage map improves the accuracy and contiguity of the assembly. Perhaps the most important point is that *a linkage map must be combined with a physical map to estimate and map recombination rates across a genome*. If researchers must choose between map types when developing genome resources for their species, the *physical genome assembly will often be the map of choice because many more SNPs can be mapped* physically; It is difficult to build linkage maps including extremely large numbers of SNPs (e.g., because many mapping families are required), as mentioned above.

3.4 Apply Genomics Approaches Without Maps

Many of the methods mentioned above can be applied to sequences from known genes or loci with unmapped locations in the genome. For example, we can conduct tests for loci under selection by testing for different kinds of outlier behavior (F_{ST}, GD, allele frequency skew or heterozygosity excess, excessive locus-specific introgression; Luikart et al. 2003). We can also test for population adaptive differentiation (Bonin et al. 2007) and test for associations between genotypes and the environment or phenotypes (Fig. 1, Step 4a) (Fig. 3a).

4 Genotyping and Sequencing Technologies for Population Genomics

This quote by Schlötterer (2004) at the start of this chapter emphasizes the importance of molecular genetic methods and implies the importance of choosing an appropriate DNA marker or sequencing method for your research question (as did Sunnucks 2000; Benestan et al. 2016). The methods continue to evolve and improve our understanding of nature. SNPs and other markers from a variety of partial genome (and transcriptome) sequencing methods are the mainstay in population genomics studies. Here we provide a short introduction to key marker technologies likely to be most widely useful for non-model species. Low-cost genotyping, including RAD capture, DArT (diversity array technology), and related methods will continue to make population genomics increasingly feasible and widely used. Later in this book, Holliday et al. (2018, Chapter 2) provide more details and merits and demerits of different genotyping and sequencing technologies (see also Andrews et al. 2016; Jones and Good 2016; Holliday et al. 2018). For information on the promising approach of multiplex sequencing of many pooled individuals (pool-seq), see Box 3, Sect. 2.5, Schlötterer et al. (2014), and Narum et al. (2018).

4.1 Reduced Representation and Genotyping-by-Sequencing

Reduced representation sequencing is revolutionizing population genetics, molecular ecology, and conservation biology by making feasible and affordable use of massively parallel sequencing (MPS) on many individuals and loci genome wide (Narum et al. 2013). *We can now use MPS to discover and genotype thousands of SNP loci for less cost than genotyping of only ~20 microsatellites. This makes population genomics research feasible for nearly any species*. Understanding the strengths and limitations of the many reduced representation approaches is crucial to choose the best method for your research question (Andrews et al. 2016).

Approaches for reduced representation sequencing include general and targeted approaches (Jones and Good 2016). Anonymous approaches include unmapped restriction site-associated DNA sequencing (RADs) and transcriptome sequencing. Targeted approaches allow direct sequencing of loci of interest such as genes or informative RAD loci using capture arrays (below). Informative RAD loci are those in candidate adaptive genes and/or loci that are evenly spaced (mapped) across chromosomes to ensure genome wide coverage and high power for outlier tests, GEA, and association studies (e.g., GWAS) (e.g., Hohenlohe et al. 2010b; Kovach et al. 2016; Simons et al. 2018; Gibson 2018).

4.1.1 RADseq

The development of restriction site-associated DNA sequencing (referred to as RADseq and genotyping-by-sequencing, GBS) was considered among the most important scientific breakthroughs in the first decade of the twenty-first century because it allowed for simultaneous discovery and genotyping of many thousands of SNPs in a single experiment, in non-model species with no genomic resources (Science 2010). It involves the cutting of DNA through digestion with one or more restriction enzymes, labeling fragments from each individual with a unique barcode (short 6–12 bp reads), amplifying fragments using PCR, and high-throughput sequencing of pooled samples from multiple individuals (Andrews et al. 2016).

Another advantage of RADseq is its flexibility in the number of loci that can be genotyped – from hundreds to tens of thousands – by choosing among different restriction enzymes and >15 different RADseq-based techniques (Andrews et al. 2016). A main disadvantage is that there is typically highly uneven coverage of genotypic data among individuals and among loci, with many individuals missing data for many loci unless very stringent filtering is conducted with deep coverage sequencing.

This method has become extremely popular and has been applied to many taxa and questions in conservation, ecology, and evolution including quantifying inbreeding, genomic diversity, effective population size (N_e), and for discovery of adaptive genes and genome regions (reviewed in Andrews et al. 2016; see also Lowry et al. 2017; McKinney et al. 2017a; Catchen et al. 2017; Hohenlohe et al. 2010b; Nadeau et al. 2014; Benestan et al. 2016; Candy et al. 2015; Sovic et al. 2016; and also subsequent chapter by Holliday et al. (2018) in this volume).

4.1.2 Targeted Sequence Capture

Sequence capture allows targeted sequencing of any region of a genome for which DNA sequence information exists. Sequence capture is often called “exon capture” because it is often used to sequence coding regions of the genome, including candidate adaptive genes (Flanagan et al. 2018). It is more expensive than RAD but a cheaper and more efficient alternative to whole-genome sequencing and results in more uniform sequencing of individuals and loci (and therefore less missing data)

than restriction enzyme-based methods. It can be scaled to sequence hundreds to tens of thousands of genes (Hodges et al. 2007; Jones and Good 2016). Another advantage of sequence capture is in the genotyping of degraded DNA such as ancient, historical, and fecal DNA (Castellano et al. 2014; Bi et al. 2012; Bos et al. 2015).

Targeted capture enriches for DNA of interest and washes away nontarget DNA, as mentioned. This is important for genotyping fecal DNA because a majority (>90%) of DNA can be from bacteria (e.g., Perry et al. 2010). Recent examples of sequence capture include a wide range of question from phylogenetics to the detection of adaption signatures in humans, wolves, sharks, wild sheep, ungulates, birds, amphibians, trees, aquatic invertebrates, and host-parasites simultaneously (Cosart et al. 2011; Schweizer et al. 2016, Roffler et al. 2016; Gasc et al. 2016; Portik et al. 2016; McCartney-Melstad et al. 2016; Syring et al. 2016; Dowle et al. 2016: Campana et al. 2016; Manthey et al. 2016; Suren et al. 2016; Gauthier et al. 2017; see also Chapter 2 by Holliday et al. 2018).

4.1.3 RAD Capture

RAD capture ("Rapture") combines the primary advantages of RADseq with advantages of targeted sequence capture. For example, the relatively inexpensive and rapid DNA library preparation methods of RADseq (Ali et al. 2015) are combined with the high specificity in targeting hundreds or thousands of loci. Loci are of high value (in genes, evenly spaced genome wide) for addressing nearly any questions of interest, focusing sequencing effort on those loci (Andrews et al. 2016; Jones and Good 2016; Hoffberg et al. 2016; Peek et al. 2018; see also Chiou and Bergey 2018). Another advantage is that a single Rapture array (e.g., for trout) works for genotyping in multiple divergent species such as salmon and trout (M. Miller, pers. comm., 2018).

The Rapture method was first used to successfully study SNP variation in lake trout (M. Miller, unpublished, 2018) and rainbow trout (Ali et al. 2015). This study used a capture array targeting 500 loci that were distributed across 29 chromosomes (Ali et al. 2015). All 1,440 individuals genotyped for the 500 loci were sequenced in a single Illumina HiSeq lane.

4.1.4 DArT

Diversity array technology (DArT) is another sequencing-based approach (a modification of GBS) allowing affordable discovery and genotyping of thousands of SNPs in hundreds of individuals (Elbasyoni et al. 2018). DArT has been used mainly in agriculturally important species and plants (Valdisser et al. 2017). This technology is similar to RADseq. Commercial companies exist, as for RADseq, to facilitate the discovery and application of genome-wide markers for population genomics approaches.

4.2 Reference Genomes

A reference genome sequence (i.e., genome assembly) is the portion of the genome that has been sequenced and assembled, i.e., pieced together, from short sequence reads. A reference genome is important in population genomics because it improves mapping of NGS reads to facilitate both the initial discovery of loci and the eventual genotyping of loci from many individuals. For example, if the reads from a RADseq project can be mapped to a reference genome, it can improve the detection of SNPs and duplicated genes or chromosomal regions that will be difficult to genotype because reads from duplicated regions often will stack up (align) together as if from a putative single locus (Hand et al. 2015a; Shafer et al. 2017). Shafer et al. (2017) observed large differences between reference-based and de novo approaches; use of a reference genome yielded more SNPs and reduced estimates of F_{IS} and Ts/Tv.

Genome assembly is difficult in large genomes of plants where repetitive elements (e.g., retrotransposons) constitute >50% of the genome (Nystedt et al. 2013). In loblolly pine (*Pinus taeda*), 62% of the 22 Gb genome is made up of retrotransposons, and other conifers have similarly large repeat element content (De La Torre et al. 2014). Similarly, for genomes resulting from recent polyploidization events, as in many fish and plants, the assembly is difficult because, for example, in a tetraploid four similar copies exist for much of the genome. Most eukaryotic genomes contain complex repetitive sequences that are difficult to sequence and assemble as mentioned above (Ellegren 2014).

Assembly is becoming vastly easier thanks to new long-read technology as suggested by the following quote: "Long reads enable near reference-quality genome assemblies, discovery of novel disease-causing structural variation, and the ability to sequence through previously 'unsequenceable' repetitive DNA contents of clinical utility" (Ameur et al. 2018).

A reference genome sequence is not a standardized concept or item (Ellegren 2014). Even for well-characterized genomes, large parts are often not yet included in the genomic contigs (small assembled chromosomal regions) or the scaffolds (sets of contigs linked into larger regions) that have been ordered and linked into chromosomes. For example, the first published rainbow trout genome had only ~50% of sequences assembled and ordered into chromosomes; in fact one entire chromosome (#25) was unassembled such that no sequences were known from that chromosome (Berthelot et al. 2014). Similarly, chromosome 16 in the collared flycatcher genome is unassembled (Kawakami et al. 2014). In the rainbow trout and flycatcher examples, much of the one unassembled chromosome was likely sequenced and exists among the many contigs that have not been incorporated (assembled) into chromosomes. The quality and completeness of reference genomes vary widely among species.

Importantly, even partially assembled genomes are useful for many research questions. Partial genomes facilitate discovery of non-duplicated (versus duplicated) SNP loci for marker discovery. Partial (draft) genomes also increase quality of genotyping (e.g., with RADSeq or DNA capture data). Finally, draft genomes help design probes for exon sequence capture (e.g., when exons are identified from RNAseq data), and are useful for estimating the rate or distance of decay of gametic disequilibrium (Hand et al. 2015a; Shafer et al. 2017). Even *a draft assembly (N50*

>50 kb) usually contains well-assembled coding gene regions because coding genes have few repetitive elements and low heterozygosity, making a draft assembly relatively feasible and highly useful. In Tasmanian devils (*Sarcophilus harrisii*), researchers used a partially assembled draft genome (containing thousands of scaffolds not anchored on chromosomes) to successfully identify genomic regions and candidate genes underlying cancer risk, along with concordant signatures of selection including increased GD (gametic disequilibrium) and changes in allele frequencies (Epstein et al. 2016).

Another advantage of having at least a draft reference genome is that it allows estimation of the rate of decay of gametic disequilibrium, which is crucial for knowing the number of markers needed to adequately cover the genome to address particularly interesting or challenging research questions (narrow sense genomics). Having even only a hundred long scaffolds (>100 kb) with multiple DNA markers provides information on whether long stretches of GD exist genome wide, which is crucial for assessing the number of markers needed to achieve high density (Hendricks et al. 2018).

4.3 Whole-Genome Sequencing (WGS) and Resequencing

A main reason for sequencing (i.e., resequencing) entire genomes from many individuals is to maximize power to discover and localize DNA loci underlying fitness, adaptation, and phenotypic variation important for population persistence and growth (e.g., Kardos et al. 2016b). Increased power results from detecting most SNPs in the species and from being able to compute summary statistics (H, F_{ST}, GD) for those SNPs and other polymorphisms (e.g., indels) in a sliding window across genomic regions (e.g., Box 3).

Having only one individual's genome sequence (e.g., from one male) will not allow understanding of genome structural diversity or variation. This could bias subsequent comparisons of diversity among individuals (e.g., males and females), populations, and species, for example, when using GBS or RAD seq methods and mapping reads to the one genome reference sequence.

Box 3 Whole-Genome Sequencing Identifies Selective Sweeps and Candidate Genes

Researchers used whole-genome sequencing of wild Rocky Mountain bighorn sheep (*Ovis canadensis*) to identify 3.2 million SNPs and genomic regions with signatures of historical directional selection, i.e., selective sweeps (Kardos et al. 2015b). Sweeps were detected as chromosomal regions with low heterozygosity. Heterozygosity-based sweep analysis revealed evidence for strong historical selection at a gene (*RXFP2*) that affects horn size in domestic sheep, cattle, and goats (Johnston et al. 2011, 2013). The massive horns carried by bighorn sheep rams appear to have evolved in part via strong selection at the *RXFP2* gene (Fig. 12).

(continued)

Box 3 (continued)

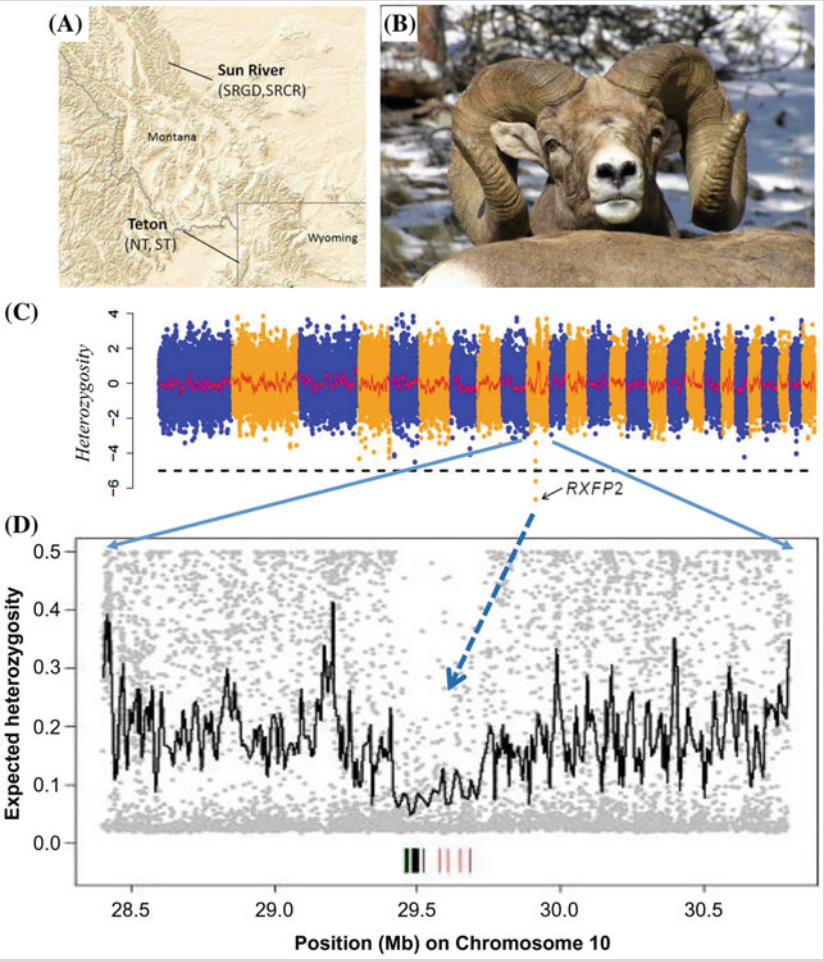

Fig. 12 Sequencing-based (pool-seq) genome-wide scan for selective sweeps that reduced heterozygosity in Montana and Wyoming populations (**A**) of bighorn sheep (**B**). Sliding window estimates of heterozygosity (**C**) across the bighorn sheep genome from an analysis of three populations pooled from Montana and Wyoming. Chromosomes (linkage groups) are arranged from 1 to 26 (left to right with alternative color (blue then orange) shading). The horizontal jagged red line represents the rolling mean across 100 adjacent sliding windows. The horizontal dashed line is 5 standard deviations below the mean heterozygosity. (**D**) Sweep on chromosome 10 spanning the *RXFP2* gene (vertical black lines at 29.5 Mb near the *x*-axis are exons). Expected heterozygosity is plotted for individual SNPs (gray dots) located across 2 Mb on chromosome 10. The location of exons (vertical lines) of *EEF1A1*, *RXFP2*, and an uncharacterized predicted gene ("*UNC*") is shown below the plot. Gene and exon positions were obtained from the Ensemble gene models generated during annotation of OARv3.1. The continuous horizontal jagged line shows mean expected heterozygosity calculated for nonoverlapping windows of 20 SNPs. The lowest genetic variation in the region occurred in a window centered at position 29,473,544 between exons 3 and 4 of *RXFP2 (dashed line arrow).* Reproduced with permission from Kardos et al. (2015b)

(continued)

Box 3 (continued)

The authors also identified evidence for selection at genes affecting early body growth and cellular response to hypoxia which is consistent with adaptation to life at high altitude. These results provide examples of strong genomic signatures of selection identified at genes with known function in wild populations of a non-model species.

A comparison of SNP diversity between the X chromosome and the autosomes also indicated that bighorn males had a dramatically reduced long-term effective population size compared to females. This likely reflects a long history of intense sexual selection mediated by male-male competition for mates, which reduces the effective population.

The approach of heterozygosity-based sweep analysis had been previously used successfully in domestic animals where breed formation and subsequent strong artificial selection have generated selective sweeps for genes that influence a spectrum of phenotypic traits (Rubin et al. 2010, 2012; Axelsson et al. 2013). In wildlife, genome sequencing of gray wolves from the high altitude plateaus of western Asia recently detected selective sweeps surrounding genes involved with adaptation to hypoxia (Zhang et al. 2014). Together, these studies provide encouragement that genome sequencing in carefully selected wild populations will continue to yield valuable insights into the genetics of adaptation (Kardos et al. 2015b).

The results illustrate the value of quality reference genome assemblies from agricultural or model species for studies of the genomic basis of adaptation in closely related wild taxa (domestic sheep in this case). This study also illustrates the use of genome sequencing of pooled DNA from many individuals (per population). This saves money and can be an efficient way to estimate allele frequencies at nearly all SNPs in the genome. However, drawbacks include imprecision in estimates of allele frequencies arising from uneven contribution individuals to sequencing (pool-seq without barcoded individuals). For more information, see discussions by Schlötterer et al. 2014; Kardos et al. 2015b; Narum et al. 2018).

Certain questions can only be reliably addressed by using whole-genome sequencing. For example, structural polymorphisms such as gene duplications (copy number variants) cannot be reliably detected with GBS (e.g., RADseq) or sequence capture but can be detected by whole-genome assemblies and ideally with a linkage map (Wellenreuther and Bernatchez 2018). Additionally, adequately covering the genome for applications, such as GWAS, will sometimes require whole-genome sequencing for populations in which gametic disequilibrium is low and decays rapidly along chromosomes, e.g., in populations with very large N_e or high recombination rates (Kardos et al. 2016b; Miles et al. 2017; Table 2).

Nonetheless, *most questions in population genetics, molecular ecology, and conservation genetics can be addressed sufficiently without whole-genome sequencing* and by using a population genomics approach (Allendorf et al. 2010). These include estimating individual inbreeding, detecting hybridization, quantifying population structure, and inferring gene flow. Whole-genome or exome resequencing is most useful for questions, such as determining the genomic basis (architecture) of local adaption or fitness when only a limited amount of gametic disequilibrium exists along chromosomes and thus millions of SNPs are required, for example.

4.4 Population Transcriptomics, Gene Expression, and Adaptation

Transcriptomics is the study of all RNA transcripts (transcriptome) that are produced by the genome. Population transcriptomics is the use of transcriptome-wide data to study variation in gene expression within and among populations to understand mechanisms underlying evolutionary change, for example, in response to environmental change. Such mechanisms can include plasticity in gene expression if it underlies adaptive evolutionary responses to new environments (Ghalambor et al. 2015) or if the amount or nature of plasticity itself evolves in response to selection. Here, we discuss the two main tools of population transcriptomics, microarray analysis and RNA sequencing (RNAseq), with examples of applications to natural populations.

cDNA microarrays and oligonucleotide microarrays can measure expression of thousands of genes simultaneously by quantifying levels of mRNA present in different tissues or individuals. Thousands or tens of thousands of different short DNA fragments are spotted onto a glass slide or other template, and cDNA from the individuals being studied, labeled with fluorescent dyes or other markers, is hybridized with that array. The intensity of fluorescence provides a quantification of the relative expression levels of targeted genes. Results are often validated with more precise estimates of RNA abundance (expression) using quantitative PCR for a subset of genes.

Gene expression profiles can be viewed as phenotypes because they are the product of both genetic and environmental variation (Hansen 2010). To assess genetic differences underlying gene expression, individuals can be reared in a common environment. Information on gene expression differences among populations can be used to complement data on neutral or adaptive genetic markers and adaptive traits for circumscribing conservation units. For example, Vandersteen Tymchuk et al. (2010) quantified gene expression for populations of Atlantic salmon in and around the Bay of Fundy, Newfoundland, using a 16,000 gene cDNA microarray. They found consistent year-to-year population differences in the expression of 389 genes when fish were reared in common environments. Population

differentiation for gene expression was stronger, and patterns were somewhat different than those observed for seven microsatellite loci.

RNAseq (also called whole transcriptome shotgun sequencing) is replacing hybridization-based microarray technologies for many applications thanks to lowering costs of next-generation sequencing (Ozsolak and Milos 2011; Wang et al. 2009; Oomen and Hutchings 2017). RNAseq can more comprehensively assess the entire repertoire of RNA molecules expressed from genomes over a wider range of expression levels than can microarrays. We note that RNAseq can also be used for SNP discovery, for SNP genotyping, or for probe design for exon capture. For example, Bi et al. (2012) use RNAseq to discover SNPs within coding genes. They then used the gene sequences to design DNA sequence capture baits to test for SNPs associated with adaptive differentiation in chipmunks.

RNAseq and RADseq were used by Chen et al. (2018) to test for genetic variation in thermal adaptation in redband trout populations (*Oncorhynchus mykiss gairdneri*) from warm versus cool environments. In a common garden, fish from a desert climate had significantly higher thermal tolerance and aerobic scope (>3°C) for higher cardiac performance (e.g., without arrhythmia) than fish from the cooler montane climate. In addition, the desert fish had the highest maximum heart rate during warming, indicating improved capacity to deliver oxygen to internal tissues. Following heat stress, distinct sets of cardiac genes were induced, which helped explain the differences in cardiorespiratory function. Candidate RADseq SNP markers and nearby genes underlying these physiological adaptations were identified, including genes involved in metabolic activity and stress response (such as heat shock genes hsp40, ldh-b, and camkk2). This kind of study is rare in that it identified both transcriptomic and genomic mechanisms of evolutionary adaptation that allow populations to persist in the difficult environmental conditions of desert streams.

5 Bioinformatics for Filtering, Genotyping, and Data Analyses

Bioinformatics skills and understanding are crucial to analyze the increasingly massive DNA sequence datasets. Bioinformatics involves intensive computations to analyze DNA, RNA, and protein sequence datasets. The field of bioinformatics underwent explosive growth starting in the mid-1990s, driven largely by the Human Genome Project and rapid advances in DNA sequencing technology. Thus, the need for bioinformatics training and approaches has increased greatly in the last decade as the data produced by massive parallel sequencing approaches has grown exponentially. However, while the costs of genome sequencing are plummeting, time and money spent on bioinformatic data filtering and analysis (and production of bioinformatics platforms) have increased more slowly over time (Sboner et al. 2011). Given the many advantages and increasing ease of generating massively parallel sequencing (MPS) data, it has become crucial for population geneticists to be trained

in computer programming and scripting to take full advantage of the growing catalog of bioinformatics tools (Andrews and Luikart 2014).

There are four major bioinformatics steps, often referred to as a bioinformatics pipeline, that occur in most population genomics studies including (1) sequence read filtering, (2) assignment of reads to loci (e.g., alignment to a reference genome or de novo loci assembly), (3) genotype calling, and (4) final filtering for problematic loci that do not meet biological expectations (e.g., Hardy-Weinberg proportions, high numbers of SNPs per locus (or 100 bp) usually resulting from alignment error, high observed heterozygosity, or more than two observed alleles) (Benestan et al. 2016). A major challenge in bioinformatic analysis is the creation of standardized pipelines (e.g., see the Broad Institute webpage for best practices – software.broadinstitute.org) that would improve consistency and comparison of results among species (and studies within species) but also even within the same species. Worrisome is the fact that different pipelines often result in very different results (for a given dataset) such that the number SNPs discovered and basic summary statistics and conclusions can change between pipelines (Shafer et al. 2017).

Analysis of up to entire genomes (millions of SNPs) presents challenges in filtering out loci that could lead to erroneous results and conclusions. There are no concrete rules for what criteria should be used for filtering loci from genomic datasets. The current state of filtering in population genomics has led to some colorful terms for filtering such as labeling filtering as the "F-word" or that filtering of genomic data is the "wild west" of population genomics (Benestan et al. 2016). Indeed, the potential effects of locus filtering approaches on downstream analyses and research conclusions have only recently started to be investigated (e.g., Lowry et al. 2017; Rodríguez-Ezpeleta et al. 2016; Shafer et al. 2017). However, it has also been suggested and shown empirically that filtering is helped greatly by the existence of a reference genome (Ellegren 2014; Hand et al. 2015a, b; Shafer et al. 2017).

Despite recent attempts to build conceptual and practical frameworks for MPS data analysis, a standardized pipeline remains elusive, and perhaps infeasible, given the nature of data variability present in most genomic datasets (Benestan et al. 2016). There has also been a move toward web-based platform analysis and filtering tools such as Galaxy which has gained users and popularity in recent years (Giardine et al. 2005; Afgan et al. 2016). Galaxy offers a more user friendly graphical interface for easy visualization and reproducibility of results through the tracking (logging) of all bioinformatic analysis steps and user-created and shared workflows. Workflows are flowchart-style representations of bioinformatics pipelines with drag and drop functionality that allows for easy customization, reproduction, and even publication of bioinformatics pipelines (Catchen et al. 2013; Eaton 2014). Galaxy also offers tools across a range of datatypes including RAD and RNAseq, WGS, and exon capture (Blankenberg et al. 2010; Pogorelcnik et al. 2018; Tranchant-Dubreuil et al. 2018). See the chapter "Computational Tools for Population Genomics" by Salojärvi (2018) in this book for more information.

6 Emerging Population Genomics Approaches

Here, we discuss emerging approaches that will become more widely used as costs decrease and technologies improve. These include population metagenomics, transcriptomics, epigenomics, proteomics, and paleogenomics.

6.1 Metagenomics

Metagenomics is the sequencing and analysis of DNA from all species in an environmental or gut sample (Srivathsan et al. 2016; Stat et al. 2017; Laforest-Lapointe et al. 2017; Waite et al. 2018). Metagenomics has usually been defined more narrowly as the study of DNA from microbial communities in environmental samples, perhaps because the initial studies were in microbes (Venter et al. 2004; Garcia et al. 2018). Metagenomics can be used to describe the diversity and relative abundance of taxonomic groups present within a single sample, experiment, or local population (DeLong 2009). These techniques have been applied widely to microbes in environmental samples, including water, soil, fecal, or gut samples, and subjected to high-throughput sequencing. Further, analysis of the functional groups of genes and their relative abundance, without requiring knowledge of which organism each sequence fragment came from, can provide a functional metabolic profile of the microbial community (Dinsdale et al. 2008).

From a population genomics perspective, metagenomics can allow the application of population genomics approaches (e.g., Fig. 1 or Fig. 2) on each of multiple microbial species, simultaneously. Further, if the microbial species are sampled from across a heterogeneous environment (or gradient), it facilitates the application of a landscape community genomics approach to improve understanding of eco-evolution interactions (Sect. 2.4; Hand et al. 2015b). Another application of metagenomic data is to describe a microbial community as an essential part of an individual host's phenotype, influencing the health and fitness of the host. The application of metagenomics in ecology, evolution, and conservation is in its early stages, but a few specific areas show promise for the future. A chapter in this book series volume describes how population genomics approaches can be applied to metagenomic data to delineate microbial populations in the environment and to study evolutionary processes within them (Denef 2018).

Metagenomic surveillance systems are increasingly being used to improve monitoring and determine mechanisms driving the spread of infectious diseases. Portable genomic sequencers provide rapid near real-time diagnostics that can resolve important epidemiological and genomic characteristics of an outbreak or epidemic's dynamics. As pathogens replicate and spread, mutations accumulate in their genomes. The whole-genome sequencing of spatially referenced samples allows researchers to track and reconstruct geo-spatial pathways of spread. Genomic epidemiology surveillance and rapid response programs can now take a more anticipatory approach to outbreak prevention and control.

Genomics-informed DNA detection assays have been developed to track a wide range of important fungal plant pathogens, including introduced, invasive species causing widespread diseases and mortality in natural populations and crop species (Feau et al. 2018). Monitoring and understanding which strains are emerging and associated with different environments and species (including humans) would also help to model, predict, and manage outbreaks and spread of pathogens. Whole-genome data from individual pathogen species in each of many host individuals can be used in population genomics approaches (or landscape community genomics approaches) to better understand the genomic basis of adaptation to hosts and local environments and to predict the effects of environmental change on a pathogen population and microbial community (Hand et al. 2015b).

Another application of metagenomics is to monitor or predict physiological condition, health, or fitness of individual organisms. For instance, Vega Thurber et al. (2009) have found shifts in the endosymbiont community of corals in response to stressors, such as reduced pH, increased nutrients, and increased temperature. Such shifts in the endosymbiont community could serve as indicators or predictors of reef health, and they could also suggest mechanisms by which coral condition affects other taxa in the reef ecosystem (Roitman et al. 2018; Leite et al. 2018).

Finally, a large-scale study used metagenomic techniques on fecal samples to catalog 3.3 million microbial genomes in the human gut fauna (Qin et al. 2010). The study found significant differences in the microbial metagenome between healthy individuals and those with two types of inflammatory bowel disease (Qin et al. 2010). In the future metagenomic techniques will be applied to noninvasively-collected fecal samples from wildlife species to assess their health status, such as starvation or disease infection, and to understand mechanisms underlying host and microbe interactions, population genomics, and coevolution (e.g., Beja-Pereira et al. 2009; Chiou and Bergey 2018; Waite et al. 2018).

6.2 Metatranscriptomics

While metagenomics focuses on detecting the presence of microbial species, metatranscriptomics investigates their gene expression profiles to address questions such as which genes are expressed in different environments or conditions. Thus, metatranscriptomics investigates the function and activity of the entire set of transcripts (RNAseq) from environmental, fecal, gut, or other samples. It is often used to identify sequences of genes expressed within natural microbial communities to advance understanding of microbial ecology and drivers of gene expression variation.

Assessing all the microbial community transcripts from a particular time and location, including bacteria, archaea, or small eukaryotes in the ocean, soil, or an organism's gut, can help understand the complex microbial processes simultaneously occurring in natural or disturbed environments. This allows "eavesdropping

on microbial ecology," a promising new approach for researchers in ecosystem ecology, animal health, and functional biodiversity monitoring (Moran 2009).

From a population genomics perspective, metatranscriptomics – like metagenomics – can facilitate landscape community genomics approaches to improve understanding of eco-evolutionary processes (Hand et al. 2015b). Transcriptomic and metatranscriptomic data can detect gene expression shifts in both host and microbes simultaneously (e.g., lung tissue and lung parasites, gut tissue and gut parasites, blood and malaria, etc.) and thus can help understand, model, and predict host-parasite interactions (e.g., Matthews et al. 2018; Lee et al. 2018; Campbell et al. 2018).

Metatranscriptomics and metagenomics together can provide entire transcriptome and genome repertoires of microorganisms through sequencing total DNA/RNA from samples; this provides taxonomic and also functional information with high resolution. These two approaches together with new bioinformatics tools can help us better understand mechanisms of adaption, coevolution, and processes like rumen fermentation, digestion, and community adaption to environmental change. A challenge for "meta" approaches is that only a small percentage of the many ecologically important genes has been annotated or identified. Sequence datasets often contain only the abundant genes from a limited number of natural microbial communities (Moran 2009).

6.3 Population Epigenomics

> While epigenetic inheritance is well documented the adaptive significance, if any, of such a complementary inheritance system remains enigmatic (Lind and Spagopoulou 2018).

Among the most intriguing and perhaps controversial areas of population genomics research involves understanding the role of transgenerational epigenetic inheritance in adaptive evolution. *Can a strong environment change produce transgenerational epigenetic adaptation?* Epigenetics has been defined as the study of heritable changes in a trait or phenotype caused by mechanisms other than DNA mutation. We focus here on transgenerational epigenetic inheritance, which is defined as changes in gene expression and resulting phenotypic variation that are transmitted between generations through germline, but do not involve changes in the underlying DNA sequence (Horsthemke 2018).

If environmentally caused shifts in gene expression are adaptive and transmitted to subsequent generations, it could represent a Lamarckian-type mechanism facilitating adaptation to environmental challenges, such as climate warming (e.g., Christie et al. 2016; Lind and Spagopoulou 2018; Horsthemke 2018). This idea could perhaps provide hope to conservation biologists that rapid adaption to climate warming is more likely than previously thought based on adaptation through natural selection. This idea is perhaps intriguing but still farfetched given the lack of evidence. The explosive growth in research on this topic results in part from the

question of whether "epigenetic mechanisms might provide a basis for the inheritance of acquired traits" (Horsthemke 2018).

Charlesworth et al. (2017) state that "allele frequency change caused by natural selection is the only credible process underlying the evolution of adaptive organismal traits." Similarly, Horsthemke (2018) states that the evidence for transgenerational epigenetic inheritance, "is not (yet) conclusive," in mammals, even though "it has been observed in plants, nematodes and fruit flies." While there is strong evidence for environmentally induced transgenerational inheritance of epigenetic gene expression changes that influence fitness traits, there is not yet evidence that such epigenetic changes persist in the longer term (many generations) or that they influence population genetic or evolutionary processes.

Questions outlined by Charlesworth et al. (2017) can help guide future research to investigate the potential role of transgenerational epigenetic inheritance in evolutionary adaptation. These questions include the following: How many generations do inherited epigenetic marks persist, and do they spread within and among populations? Also, are transgenerational epigenetics changes an important source of adaptive change, relative to DNA sequence change (Charlesworth et al. 2017)? These are population epigenetics questions, which now can be addressed using densely distributed epigenetic marks genome wide, thereby representing "narrow sense" population epigenomics.

Here we discuss recent evidence for environmentally induced multigenerational epigenetic inheritance. We also discuss the role or importance of this inheritance in population genomics research and understanding.

Evidence is growing rapidly for multigenerational transmission of environmentally induced epigenetic changes that influence fitness traits. Environmental factors observed to cause transgenerational epigenetic inheritance of phenotypic variation include heat shock or other thermal stresses, drought, salt stress, low-calorie diet, high-fat diet, smoking, and exposure to toxins, such as hydrocarbons from plastics, atrazine, tributyltinthe, pesticide DDT (dichlorodiphenyltrichloroethane), and the agricultural fungicide vinclozolin. Many of these stressors have caused transgenerational epigenetic inheritance in humans, fish, birds, plants, and insects.

Genome-wide environmentally induced transgenerational epigenetic inheritance of disease was documented in a recent study in rats. Ben Maamar et al. (2018) exposed one generation of gestating female rats to DDT or alternatively vinclozolin. The offspring (F_1 generation) were bred to generate the F_2 generation that was then bred to generate the F_3 generation (keeping separate the populations exposed – in the F_0 generation – to vinclozolin, DDT, or control treatments). The F_3 generation males' sperm revealed persistent environmentally induced histone modification genome wide (Fig. 13), which influences gene expression to cause disease. The fact that two different environmental toxins, each promoted transgenerational epigenetic (histone) changes, suggest that histone sites have a role in epigenetic transgenerational inheritance.

A particularly interesting study of epigenetic changes suggested that a single generation in an extreme environment (captivity, in a hatchery) can translate into heritable differences in expression at hundreds of genes. Christie et al. (2016)

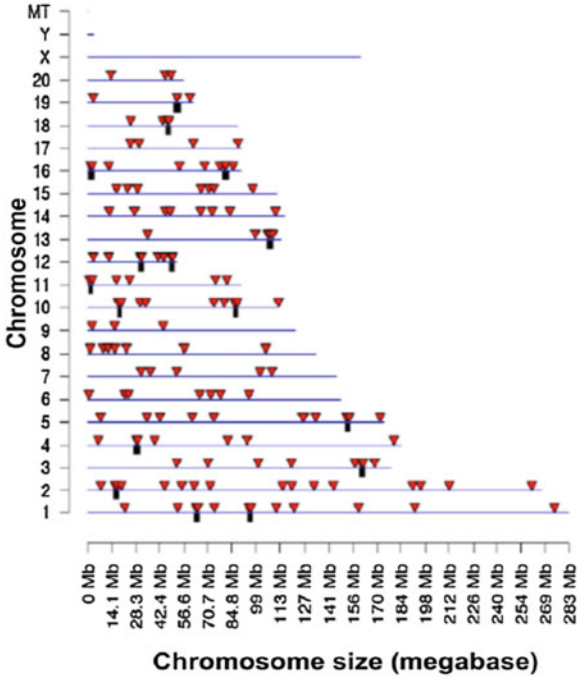

Fig. 13 Sperm histone site differences (site retention) caused by DDT (dichlorodiphenyltrichloroethane) and transmitted over multiple generations. Red arrowheads are individual chromosome locations of histone differences in sperm. DDT-induced histone differences cause transgenerational epigenetic inheritance of disease. Purified cauda epididymal sperm were collected from the transgenerational F3 generation male rats for histone analysis. Reproduced with permission from Ben Maamar et al. (2018)

measured differential gene expression in the offspring of wild and first-generation hatchery steelhead trout (*Oncorhynchus mykiss*) and found 723 differentially expressed genes in the two groups of offspring reared in the same common environment. Functional analyses of the 723 genes revealed that most genes involved responses in immunity, wound healing, and metabolism. The large proportion of immunity and healing genes being differentially methylated suggest that the high density, rapid growth (and diet change), and aggression among fish in captivity lead to disease and wounds. Finally, wild-born fish that had only one hatchery parent had much lower reproductive success in the wild (compared to fish with two wild parents), suggesting that adaptation to captivity leads to transmission of maladaptive gene expression to wild-born offspring. These findings suggest that rapid environmental adaptation is possible and might be transmitted to offspring through "heritable" (transmitted) epigenetic changes.

It is becoming clear that multiple ancestral environmental influences, such as toxins, stress, or unusual nutrition, can sometimes induce germline epigenome changes called epimutations that are transmitted to descendants. These epimutations

often occur in the germline and thus are transmitted (Gapp and Bohacek 2018). The germline epigenetic changes are often imprinted, and avoid epigenetic reprogramming (resetting/removal), and thus transgenerational inheritance occurs. Sperm RNAs are a mechanism for transfer of acquired complex phenotypes from father to offspring (Gapp et al. 2014). Stressful experiences were shown to cause metabolic and behavioral changes in mice that can be transmitted through RNAs in sperm to the offspring (Gapp and Bohacek 2018). Long-term studies are needed in natural populations to understand if inherited epigenetic marks persist across enough generations to significantly affect evolutionary processes, such as individual fitness, local adaption, gene flow, and population persistence.

6.3.1 Epigenetic Variation and Mechanisms

Here we discuss epigenetic variation that is potentially important evolutionarily but for which limited transgenerational inheritance information exists. Epigenomic variation is widespread in wild populations of plants (Schmitz et al. 2013a, b; Niederhuth et al. 2016) and animals (review in Hu and Barrett 2017). Epigenetic mechanisms causing gene expression shifts include DNA methylation, histone modifications, as well as variation in small RNAs. DNA methylation is the most frequently studied and best-understood epigenetic process to date. With the development of massive parallel sequencing techniques to examine genome-wide epigenetic marks, such as bisulfite DNA sequencing, epigenomics has progressed from investigating individual epigenomes to studying epigenomic variation across populations and species (e.g., Gavery and Roberts 2017).

The sources of epigenetic/epigenomic variation include genetic factors, environmental factors, or stochastic epimutations (reviews in Taudt et al. 2016; Yi 2017; Richards et al. 2017; Martin and Fry 2018). Recent studies have identified both the *cis* and *trans* regulatory genetic mechanisms conditioning population epigenomic variation at individual epigenetic marks to integrated chromatin state maps in a wide variety of species (review in Taudt et al. 2016). A number of methylation quantitative trait loci (*me*QTL) and histone quantitative trait loci (*h*QTL) have been identified in humans, plants, and animals (Taudt et al. 2016). Most of the work has been done on understanding the association of genetic (SNP, *me*QTL) and epigenetic variants for DNA methylation (DMR, differentially methylated region; DMP, differentially methylated polymorphism; SMV, single methylation variant; SMP, single methylation polymorphism). Nearly all of the detected *me*QTL in human mapped in *cis* association (review in Taudt et al. 2016).

Schmitz et al. (2013a), in the first plant population epigenomics study, examined the genome-wide DMRs in natural accessions of *Arabidopsis* worldwide and integrated these data with the whole-genome DNA sequences of the same accessions. They reported that 35% of the DMRs could be associated with *me*QTL, and 26% of the associations could be mapped to methylation changes in *cis*. In maize (*Zea mays*) about 50% of DMRs were associated in *cis*, with SNPs found within or near the DMR (Eichten et al. 2013). Similarly, *cis* *me*QTL-DMR associations were

widespread in soybean (*Glycine max*) (Schmitz et al. 2013b). Heritable variation in methylation can be genetically based (and not sensitive to the environment), or environmentally induced, or a combination of both. Additionally, random epimutations can cause epigenetic variation as well.

6.3.2 Associations Between Epigenomic Variation and Phenotypic, Ecological, and Disease Traits

There is growing evidence that epigenetic mechanisms and epigenomic variation contribute significantly to phenotypes, abiotic and biotic stress responses, disease conditions, adaptation to habitat, and range distributions in a variety of organisms (review in Richards et al. 2017). This has significance in the context of acclimation and adaptation to climate change. Epigenomic differences are often correlated with ecological and environmental factors (see Richards et al. 2017). For example, DNA methylation patterns were found to be associated with a climate gradient in *Quercus lobata* (Gugger et al. 2016).

Recent population epigenomics studies have concentrated on associations between epigenomic variation and phenotypic, ecological, disease, and other traits in humans, plants, and animals through epigenome-wide association studies (EWAS) and epigenome environment association analysis (epiEAA), and a number of significant associations have been identified. In particular, substantial EWAS work has been done in the past few years to identify the association of DNA methylation with common human disease conditions.

DNA methylation has been found to be significantly associated with kidney function (Chu et al. 2017), type 2 diabetes (Meeks et al. 2017), panic disorder (Shimada-Sugimoto et al. 2017), cardiovascular diseases (Nakatochi et al. 2017), cancer (Xu et al. 2013), chronic obstructive pulmonary disease and lung function (Lee et al. 2017), and other conditions. Population epigenomics has a role to play in pharmacogenomics and personal medicine (see Kabekkodu et al. 2017). In plants epigenetic variation has been associated with various phenotypic, phenological, and disease and adaptive traits, such as salt tolerance (Foust et al. 2016), disease susceptibility (Sollars and Buggs 2018), and flowering time (Aller et al. 2018).

Population epigenomics, as such, is an emerging approach in population genomics. The detailed discussion of various aspects of population epigenomics is presented in the chapter by Moler et al. (2018) later in this book. This includes the molecular basis of epigenetic mechanism, sources and evolution of population epigenomic variation, intra- and interspecific epigenomic variation, molecular and bioinformatics methods in population epigenomics, and association of epigenomic variation with phenotypic, ecological, and disease traits and pharmacogenomics. See also recent reviews (e.g., Gapp and Bohacek 2018) and the special edition set of papers on the evolutionary consequences of epigenetic inheritance (Lind and Spagopoulou 2018).

6.4 Population Proteomics

Population proteomics is the study of structural and functional variation (qualitative and quantitative) in proteins within and among populations to better understand their role in individual fitness, phenotypic variation, local adaption, and population performance (see also Biron et al. 2006; Nedelkov et al. 2006; Nedelkov 2008). Enzyme protein polymorphisms (isoenzymes, isozymes, allozymes) provided the first molecular markers for population genetic studies. Protein electrophoresis studies were widely conducted for several decades before DNA markers became available (Charlesworth et al. 2016).

Although population proteomics gained attention around 2005 (e.g., Biron et al. 2006; Nedelkov et al. 2006; Nedelkov 2008), especially for biomarker discovery for human disease conditions, it has not kept pace with population genomics owing to the rapid advances in high-throughput DNA and RNA sequencing technologies. However, the development of 2D gel electrophoresis, mass spectrophotometry methodologies (such as MALDI TOF), and shotgun proteomics methods has made high-throughput protein analysis possible. This has accelerated population proteomics studies across different species (e.g., Ma et al. 2015; Armengaud 2016; Di et al. 2016; Hidalgo-Galiana et al. 2016; Colinet et al. 2017; Gamboa et al. 2017; Suhre et al. 2017).

Since proteins influence important phenotypes and are the products of genes and epigenetic or posttranslational mechanisms, population proteomics has the potential to provide key insights into functional and metapopulation ecology, adaptation, and acclimation processes under various climate and environment conditions (e.g., Biron et al. 2006; Karr 2008; Di et al. 2016; Colinet et al. 2017; Gamboa et al. 2017; Trapp et al. 2018). Population proteomics approaches also help identify genetic loci underlying risk of disease and for clinical biomarkers for many human disease conditions (Nedelkov et al. 2006; Suhre et al. 2017).

Most population proteomics studies to date have been focused on humans, especially for discovering and validating biomarkers for clinical disease conditions. High levels of protein diversity have been reported in humans. For example, a total of 76 structural forms variants were observed for the 25 plasma proteins (an average of 3 variants per protein) in a cohort of 96 individuals (Nedelkov et al. 2005). Proteomics-based genome-wide association studies have identified many associations between protein levels and gene variants (protein QTLs, pQTLs) in different population cohorts (summary provided in the supplementary table in Suhre et al. 2017 and updated on http://www.metabolomix.com/a-table-of-all-published-gwas-with-proteomics/). For example, Suhre et al. (2017) reported 539 pQTLs in German, Asian, and Arab cohorts, and associations overlapped with 57 genetic risk loci for 42 unique diseases.

Proteomics approaches have also been useful in nonhuman systems. For example, clear ecotype-specific protein variation was found among eight *Arabidopsis* ecotypes that were related to their physiological status (Chevalier et al. 2004). Rees et al. (2011) reported significant within and among population variation in

proteins in three species of the teleost fish *Fundulus*; The authors suggested that the patterns of protein expression have evolved by natural selection.

Gamboa et al. (2017) investigated protein expression in five stream stonefly species (*Plecoptera*) from four geographic regions along a latitudinal gradient in Japan with varying climatic conditions. They found high spatial variation in protein expression among four geographic regions that were positively correlated with water temperature. However, low interspecific variation was observed in proteins within geographical regions, suggesting regulation of protein expression varied with environment and relates to local adaptation.

In Drosophila, Colinet et al. (2017) studied the regulatory mechanisms involved in the acquisition of thermal tolerance. They note that reversible phosphorylation is a common posttranslational modification that can rapidly alter proteins functions. They conducted a large-scale comparative study of phosphorylation networks in control versus cold-acclimated adult Drosophila and found that acclimation evoked a strong phosphoproteomic signal characterized by large sets of unique and differential phosphoproteins. In diving beetles (*Agabus ramblae* and *A. brunneus*), Hidalgo-Galiana et al. (2016) found protein expression parallels thermal tolerance and ecological conditions in the diversification of these two *Agabus* species.

These studies suggest that research on proteomic variation among natural populations along environmental gradients can provide insights into mechanisms underlying eco-evolutionary processes such as local adaptation, diversification, range shifts, and speciation. Future studies including genome-wide proteome data combined with population and landscape genomics approaches on multiple species (e.g., landscape community proteogenomics) will be especially helpful for understanding and predicting adaptive evolution, population performance, coevolution, and adaptive divergence.

6.5 Paleogenomics

Paleogenomics is the study of genomes of ancient organisms from fossil remains or specimen excavated from caves, permafrost, ice cores, or archeological or paleontological sites or stored in museum and herbarium collections (Heintzman et al. 2015; Lan and Lindqvist 2018). Paleogenetics and paleogenomics are recent fields of research relying on the extraction and analysis of preserved ancient DNA (aDNA). Early paleogenetics research was based on sequencing of mitochondrial DNA (mtDNA) fragments because of high copy numbers of the mitochondrial genomes in a cell. This research has provided quite useful information on phylogenetic relationships and timing of divergence among organisms and biographical patterns (Lan and Lindqvist 2018).

Paleogenomic studies are providing insights into complex evolutionary histories of ancient and extinct organisms, including humans (*Homo sapiens*) (Rasmussen

et al. 2010; Meyer et al. 2012; Prüfer et al. 2014), phylogenetic and evolutionary relationships of extinct organisms with living species and populations (e.g., Prüfer et al. 2014; Heintzman et al. 2015; Lan and Lindqvist 2018), inferences of demographic patterns and ancient admixtures in human and other organisms (Meyer et al. 2012; Prüfer et al. 2014; Shapiro and Hofreiter 2014; Lan and Lindqvist 2018), reconstruction of ancient adaptive phenotypes and inferences of extinction causes, such as in wooly mammoth (*Mammuthus primigenius*) (Palkopoulou et al. 2015; Rogers and Slatkin 2017), and causal agents and evolutionary history of ancient pandemics, such as Black Death (bubonic plague), small pox, tuberculosis and leprosy (reviewed in Lan and Lindqvist 2018), ancient pathogens through human history (Marciniak and Poinar 2018), and structural variants in ancient genomes (Resendez et al. 2018).

Paleogenomic investigations have provided key insights into the origin and history or domestication of crop plants (reviewed in Lan and Lindqvist 2018) and animals, such as dogs (*Canis lupus familiaris*) (Frantz et al. 2016; Thalmann and Perri 2018), cats (*Felis catus*) (Geigl and Grange 2018), and horses (*Equus caballus*) (Orlando et al. 2013; Orlando 2018), origins and genetic legacy of Neolithic farmers and human settlement in Europe (Skoglund et al. 2012), reconstruction of ancient plant communities (Parducci et al. 2018), and epigenomics of ancient species (Hanghøj et al. 2018). Most of the above paleogenomics aspects are discussed later in this book in the chapter "Paleogenomics: Genome-scale Analysis of Ancient DNA and Population and Evolutionary Genomic Inferences" by Lan and Lindqvist (2018).

One of the most studied topics in paleogenomics is the evolution of human species and its phylogenetic and evolutionary relationships with its closest evolutionary relatives. The first ancient human genome was sequenced by Rasmussen et al. (2010) from permafrost-preserved hair of a ~4-kyr-old Paleo-Eskimo. Then paleogenomes from archaic hominins, Neanderthal and Denisovan, were sequenced and published (Meyer et al. 2012; Prüfer et al. 2014). These paleogenomics studies suggested that that Neanderthal and Denisovan populations shared a common origin, that their common ancestor diverged from the ancestors of modern humans, and that admixture had taken place between archaic hominins and the ancestors of modern humans most likely after the dispersal of modern non-African humans out of Africa (Meyer et al. 2012; Prüfer et al. 2014). The analysis also indicated that this gene flow was from Neanderthal into the common ancestor of modern Eurasians.

Another example of paleogenomics applications is the inferences of the causes of extinction of the iconic ancient animal wooly mammoth, which was an abundant megafaunal species of the Northern Hemisphere. As mentioned above (Sect. 2.6), paleogenomics studies provided evidence that genetic stochasticity due to small population size could have contributed to the extinction of this species (Palkopoulou et al. 2015; Rogers and Slatkin 2017).

7 Does the Field of Population Genomics Promise More Than It Can Deliver?

Population genomics holds a great deal of promise for increasing our understanding of the genetic basis of phenotypic variation and adaptation in natural populations. However, *population genomics is not a panacea for addressing the outstanding fundamental questions* in many areas of biology. Genomes are tremendously complex, and traits related to fitness are often highly polygenic. Researchers need to better recognize the limitations of some methods and the opportunities for misleading or misinterpreted results (e.g., false positives and false negatives for selection tests). For example, the hallmark genomic signatures of positive selection (e.g., highly reduced genetic variation, shifted site frequency spectrum, or alleles associated with environmental variation) can arise from forces other than positive selection.

False signatures of positive selection can occur where purifying (background) selection has reduced genetic variation, particularly in genomic regions with low recombination (Charlesworth et al. 1993; Wolf and Ellegren 2017). Regions with low genetic variation can be caused by a locally low mutation rate, or where large haplotypes have drifted to high frequency or fixation in populations with small N_e (Nielsen et al. 2005; Kardos et al. 2015b). Regions with very high F_{ST} relative to the genome-wide background can occur between insipient species as a result of selection within lineages (e.g., background selection or recent selective sweeps), rather than via divergent selection during speciation (Burri et al. 2015; Charlesworth et al. 1993; Cruickshank and Hahn 2014; Payseur and Rieseberg 2016; Wolf and Ellegren 2017). Thus, genomic signatures of positive selection, including selective sweep signals and F_{ST} outlying regions must be interpreted cautiously.

Population genomics studies can have low power to detect loci related to adaptation or variation in phenotypes among individuals, especially for highly polygenic traits. The relatively low density of SNPs generated, in certain species, when using some technologies (e.g., some RADseq or sequence capture) means that selective sweeps, F_{ST} outliers, associations between markers and environmental variables, and QTLs may be missed because of low or no gametic disequilibrium between the genotyped SNPs and causal loci (Kardos et al. 2016a; (Catchen et al. 2017; McKinney et al. 2017a). Associations and outliers can also be missed by genotyping only a limited number of SNPs from an adaptive gene or a selected genome region (Fig. 6).

Additionally, the relatively low sample sizes that are frequent in studies of non-model organisms in the wild means that power to detect loci with relatively large effects may often be low, even when whole-genome sequencing is used in natural populations (Kardos et al. 2015a; Lotterhos and Whitlock 2014; Hunter et al. 2018; Flanagan et al. 2018). Finally, to help increase the understanding of the genetic basis of ecological and evolutionary traits and processes, we recommend applying multiple population genomics and related approaches (at different functional levels from DNA to RNA and proteins), as in Vasemagi and Primmer (2005).

8 Future Perspectives and Needs

Among the most exciting advances from "neutral" marker studies will be our improved understanding of inbreeding depression and genetic rescue in natural and managed populations. This will result from the fact that only 5000–10000 SNP loci are required to vastly improve precision of estimation of individual inbreeding compared to traditional marker-based and pedigree approaches (Kardos et al. 2016a). There will soon be many publications that use genomic data to estimate inbreeding depression (and genetic rescue) in many populations, which could change our view of the importance of inbreeding in conservation and evolution. Interestingly, most publications in the vast inbreeding literature had low power and precision to estimate inbreeding and inbreeding depression effects.

Even more exciting will be the use of novel, more informative statistical estimators such as ROH (runs of homozygosity), which measures inbreeding and effective population size change (Palkopoulou et al. 2015; Kardos et al. 2018; Grossen et al. 2018). The bioinformatic prediction of deleterious alleles from sequence data will also increase our ability to understand the genomic architecture of inbreeding depression and to predict and compare populations for genetic load.

An interesting advance will be the improved understanding of the importance of transgenerational epigenetic inheritance in adaptive traits (Charlesworth et al. 2017). Advances are likely to give the explosion of research and publications, following the controversy and calls to test the relevance of epigenetic "inheritance" in evolutionary processes and given lower costs for next-generation (bisulfite) sequencing (Christie et al. 2016; Le Luyer et al. 2017; Nilsson et al. 2018; Horsthemke 2018). Can environmentally induced transgenerational epigenetic inheritance contribute substantially to adaption to changing environments?

Another general advancement in power and precision will result from calling of microhaplotypes from short-read data. Most publications that use next-generation short-read data (e.g., RADseq) have not called haplotypes but rather scored only one SNP (or two independent SNPs) per locus, even though multiple SNPs exist per locus, e.g., RAD loci (Hendricks et al. 2018). Haplotype calling will yield more alleles (haplotypes), additional genealogical or phylogenetic information, and thus more power for many applications in population genetics (Sunnucks 2000). Longer single-end and paired-end reads and new software for haplotype calling will also improve power (Baetscher et al. 2018).

Understanding of the importance of structural polymorphisms in fitness and adaptation will increase soon (Wellenreuther and Bernatchez 2018). Genotyping and detection of inversions and copy number variants are becoming more feasible thanks to longer-read sequencing, reference genomes, linkage maps, and improved software for discovering and genotyping structural polymorphisms (e.g., Farek et al. 2018). This will help population genomics move beyond SNPs. This is an important advancement because structural variations are often involved with fitness-related phenotypic variation (e.g., Küpper et al. 2015) and are thought to play a key role in

sex chromosome evolution, local adaptation, and speciation (Kirkpatrick 2010; Wellenreuther and Bernatchez 2018).

Many studies will estimate gametic disequilibrium along chromosomes (or contigs) using *draft genome assemblies, thereby allowing more informative "narrow sense" population genomics studies with mapped high-density markers*. Even a few hundred contigs of 50–500 kb and 1,000s of marker loci will provide quantification of genome-wide GD (gametic disequilibrium) required for some narrow sense genomics approaches. Depending on the genome size and complexity, an investment of $10k to $20k can achieve a useful draft reference genome with an N50 of >50 kb for many species (Catchen et al. 2017; McKinney et al. 2017a; Hendricks et al. 2018).

There is a need to train researchers and students in data analysis including the initial filtering, genotyping, and data interpretation steps which requires an understanding of population genetics theory (Andrews and Luikart 2014; Allendorf 2017; Shafer et al. 2015; Hendricks et al. 2018). The trend toward learning the latest molecular techniques (RAD approaches, DNA capture, pool-seq, etc.) at the expense of a solid grounding in population genetics theory is worrisome (Allendorf 2017). *Training in theoretical and conceptual aspects of population genetics enables researcher to ask good questions* and to adequately test and interpret the massive and growing datasets against appropriate null models (Benestan et al. 2016; Allendorf 2017).

There is an urgent need for understanding the effects of data analysis choices on downstream biological inferences (Farek et al. 2018), because these choices can dramatically influence downstream statistical results and inferences (Shafer et al. 2017; Hendricks et al. 2018). We need to validate pipelines and downstream genomic statistical estimators, ensuring they are unbiased, by analyzing raw simulated and empirical data from populations with known genotypes and evolutionary parameters (N_e, Nm, S) in order to verify that we can recover or estimate the true (known) genotypes and parameters. Related to this, the field needs to develop a set of best practices for identifying possible genotyping errors, quantifying error rates, and quantifying effects of data analysis choices on downstream results and conclusions. The most rigorous approach for ensuring data quality can vary substantially from dataset to dataset and will change through time as the structure and quality or data change; thus we need the next generation of population genomicists to be well trained in bioinformatics and programming (Andrews and Luikart 2014).

Finally, new computational approaches and modeling made easy by ABC (approximate Bayesian computation) will vastly improve data analysis and inference from population genomic data (Cabrera and Palsbøll 2017; Elleouet and Aitken 2018). However, extensive model performance evaluations are required to ensure computational approaches are applied reliably and competently to natural populations (e.g., Lotterhos and Whitlock 2014; Forester et al. 2018; see Appendix in Allendorf et al. 2013).

9 Conclusions

Population genomics is transforming many sub-disciplines in biology and vastly improving our understanding of nature (Schlötterer 2004; Hohenlohe et al. 2018). The greatest advances in our fundamental understanding of populations and the translation of that knowledge to decisions around managing and conserving populations will result from applications of conceptually novel "narrow sense" genomics studies. This revolution will continue to accelerate for many years as more studies combine population genomics, transcriptomics, transgenerational epigenomics, and proteomics approaches simultaneously to multiple species co-distributed across environments (Chen et al. 2018; De Kort et al. 2018). This increase in strategic applications of narrow sense and multiple omics approaches combined with phenotypic and environmental data (e.g., from sensor networks and remote sensing) will ensure *we will soon be answering long-standing questions along with novel questions yet to be imagined by humanity*. It is an exciting time to be a population genomicist!

Acknowledgments We thank G. McKinney for helpful comments and information on linkage mapping and Fred Allendorf for discussions and ideas regarding population genomics concepts and definitions. GL, MK, and BKH were supported in part by funding from US National Science Foundation grants DEB-1258203 and DoB-1639014. Montana Fish Wildlife and Parks provided supported GL and MK through contract #199101903. GL and BKH were also supported in part by funding from NASA grant number NNX14AB84G. OPR received support from a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN 2017-04589. PAH received support from National Science Foundation grants DEB-1316549 and DEB-1655809.

References

Afgan E, Baker D, van den Beek M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 2016;44:W3–W10.

Ahrens CW, Rymer PD, Stow A, et al. The search for loci under selection: trends, biases and progress. Mol Ecol. 2018;27:1342–56.

Alachiotis N, Pavlidis P. RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. Commun Biol. 2018;1:79.

Ali OA, O'Rourke SM, Amish SJ, et al. RAD capture (rapture): flexible and efficient sequence-based genotyping. BioRxiv. 2015;52:4–7.

Allendorf FW. Genetics and the conservation of natural populations: allozymes to genomes. Mol Ecol. 2017;26:420–30.

Allendorf FW, Hohenlohe PA, Luikart G. Genomics and the future of conservation genetics. Nat Rev Genet. 2010;11:697–709.

Allendorf FW, et al. Conservation and the genetics of populations. Hoboken: Wiley; 2013.

Aller EST, Jagd LM, Kliebenstein DJ, Burow M. Comparison of the relative potential for epigenetic and genetic variation to contribute to trait stability. G3. 2018;8:1733–46.

Amaral AJ, Megens H-J, Crooijmans RPMA, Heuven HCM, Groenen MAM. Linkage disequilibrium decay and haplotype block structure in the pig. Genetics. 2008;179:569–79.

Ameur A, Kloosterman WP, Hestand MS. Single-molecule sequencing: towards clinical applications. Trends Biotechnol. 2018. In press.
Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH. Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. Genetics. 2011;188:799–808.
Andrews KR, Luikart G. Recent novel approaches for population genomics data analysis. Mol Ecol. 2014;23:1661–7.
Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016;17:81–92.
Armengaud J. Next-generation proteomics faces new challenges in environmental biotechnology. Curr Opin Biotechnol. 2016;38:174–82.
Armstrong C, Richardson DS, Hipperson H, et al. Genomic associations with bill length and disease reveal drift and selection across island bird populations. Evol Lett. 2018;2(1):22–36.
Axelsson E, Ratnakumar A, Arendt ML, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 2013;495:360–4.
Baetscher DS, Clemento AJ, Ng TC, Anderson EC, Garza JC. Microhaplotypes provide increased power from short-read DNA sequences for relationship inference. Mol Ecol Resour. 2018;18:296–305.
Balkenhol N, Dudaniec RY, Krutovsky KV, Johnson JS, Cairns DM, Segelbacher G, et al. Landscape genomics: understanding relationships between environmental heterogeneity and genomic characteristics of populations. In: Om PR, editor. Population genomics: concepts, approaches and applications. Cham: Springer International Publishing AG; 2017. https://doi.org/10.1111/eva.12672.
Barson NJ, Aykanat T, Hindar K, et al. Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. Nature. 2015;528:405–8.
Beichman AC, Phung TN, Lohmueller KE. Comparison of single genome and allele frequency data reveals discordant demographic histories. G3. 2017;7:3605–20.
Beja-Pereira A, Luikart G, England PR, et al. Gene-culture coevolution between cattle milk protein genes and human lactase genes. Nat Genet. 2003;35:311–3.
Beja-Pereira A, et al. Advancing ecological understandings through technological transformations in noninvasive genetics. Mol Ecol Resour. 2009;9:1279–301.
Ben Maamar M, Sadler-Riggleman I, Beck D, Skinner MK. Epigenetic transgenerational inheritance of altered sperm histone retention sites. Sci Rep. 2018;8:5308.
Benestan LM, Ferchaud AL, Hohenlohe PA, et al. Conservation genomics of natural and managed populations: building a conceptual and practical framework. Mol Ecol. 2016;25:2967–77.
Bérénos C, Ellis PA, Pilkington JG, et al. Heterogeneity of genetic architecture of body size traits in a free-living population. Mol Ecol. 2015;24:1810–30.
Bérénos C, Ellis PA, Pilkington JG, Pemberton JM. Genomic analysis reveals depression due to both individual and maternal inbreeding in a free-living mammal population. Mol Ecol. 2016;25:3152–68.
Berg JJ, Coop G. A population genetic signal of polygenic adaptation. PLoS Genet. 2014;10:e1004412.
Berthelot C, Brunet F, Chalopin D, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun. 2014;5:3657.
Betts A, Gray C, Zelek M, MacLean RC, King KC. High parasite diversity accelerates host adaptation and diversification. Science. 2018;360:907–11.
Bi K, Vanderpool D, Singhal S, et al. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics. 2012;13:403.
Biron D, et al. Population proteomics: an emerging discipline to study metapopulation ecology. Proteomics. 2006;6:1712–5.
Black WC, Baer CF, Antolin MF, DuTeau NM. Population genomics : genome-wide sampling of insect populations. Annu Rev Entomol. 2001;46:441–69.

Blankenberg D, Von KG, Coraor N, et al. Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. 2010;89:1–21.
Blanquart F, Kaltz O, Nuismer SL, Gandon S. A practical guide to measuring local adaptation. Ecol Lett. 2013;16:1195–205.
Boitard S, Rodríguez W, Jay F, Mona S, Austerlitz F. Inferring population size history from large samples of genome-wide molecular data – an approximate Bayesian computation approach. PLoS Genet. 2016;12:e1005877.
Bonin A, Nicole F, Pompanon F, Miaud C, Taberlet P. Population adaptive index: a new method to help measure intraspecific genetic diversity and prioritize populations for conservation. Conserv Biol. 2007;21:697–708.
Bos K, et al. Parallel detection of ancient pathogens via array-based DNA capture. Philos Trans R Soc Lond B Biol Sci. 2015;370:20130375.
Bourret V, Dionne M, Bernatchez L. Detecting genotypic changes associated with selective mortality at sea in Atlantic salmon: polygenic multilocus analysis surpasses genome scan. Mol Ecol. 2014;23:4444–57.
Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169:1177–86.
Brauer CJ, Unmack PJ, Smith S, Bernatchez L, Beheregaray LB. On the roles of landscape heterogeneity and environmental variation in determining population genomic structure in a dendritic system. Mol Ecol. 2018;27:3484–97.
Brelsford A, Toews DPL, Irwin DE. Admixture mapping in a hybrid zone reveals loci associated with avian feather coloration. Proc Roy Soc B Biol Sci. 2017;284:20171106.
Brieuc MSO, Ono K, Drinan DP, Naish KA. Integration of random forest with population-based outlier analyses provides insight on the genomic basis and evolution of run timing in Chinook salmon (*Oncorhynchus tshawytscha*). Mol Ecol. 2015;24:2729–46.
Browning SR, Browning BL. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. Am J Hum Genet. 2015;97:404–18.
Bürger R, Akerman A. The effects of linkage and gene flow on local adaptation: a two-locus continent-island model. Theor Popul Biol. 2011;80:272–88.
Burri R, Nater A, Kawakami T, et al. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome Res. 2015;25:1656–65.
Burri R, Antoniazza S, Gaigher A, et al. The genetic basis of color-related local adaptation in a ring-like colonization around the Mediterranean. Evolution. 2016;70:140–53.
Cabrera AA, Palsbøll PJ. Inferring past demographic changes from contemporary genetic data: a simulation-based evaluation of the ABC methods implemented in diyabc. Mol Ecol Resour. 2017;17:e94–e110.
Cammen KM, Schultz TF, Don Bowen W, et al. Genomic signatures of population bottleneck and recovery in Northwest Atlantic pinnipeds. Ecol Evol. 2018;8:6599–614.
Campana MG, Hawkins MTR, Henson LH, et al. Simultaneous identification of host, ectoparasite and pathogen DNA via in-solution capture. Mol Ecol Resour. 2016;16:1224–39.
Campbell LJ, Hammond SA, Price SJ, et al. A novel approach to wildlife transcriptomics provides evidence of disease-mediated differential expression and changes to the microbiome of amphibian populations. Mol Ecol. 2018;27:1413–27.
Campos JL, Halligan DL, Haddrill PR, Charlesworth B. The relation between recombination rate and patterns of molecular evolution and variation in *drosophila melanogaster*. Mol Biol Evol. 2014;31:1010–28.
Candy JR, Campbell NR, Grinnell MH, et al. Population differentiation determined from putative neutral and divergent adaptive genetic markers in Eulachon (*Thaleichthys pacificus*, Osmeridae), an anadromous Pacific smelt. Mol Ecol Resour. 2015;15:1421–34.
Carneiro M, Albert FW, Afonso S, et al. The genomic architecture of population divergence between subspecies of the European rabbit. PLoS Genet. 2014;10:e1003519.

Castellano S, Parra G, Sanchez-Quinto FA, et al. Patterns of coding variation in the complete exomes of three Neanderthals. Proc Natl Acad Sci. 2014;111:6666–71.
Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. Mol Ecol. 2013;22:3124–40.
Catchen JM, Hohenlohe PA, Bernatchez L, et al. Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. Mol Ecol Resour. 2017;17:362–5.
Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: windows into population history and trait architecture. Nat Rev Genet. 2018;19:220–34.
Charlesworth B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009;10:195–205.
Charlesworth B. Molecular population genomics: a short history. Genet Res. 2010;92:397–411.
Charlesworth D, Willis JH. The genetics of inbreeding depression. Nat Rev Genet. 2009;10:783–96.
Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134:1289–303.
Charlesworth B, Charlesworth D, Coyne JA, Langley CH. Hubby and Lewontin on protein variation in natural populations: when molecular genetics came to the rescue of population genetics. Genetics. 2016;203:1497–503.
Charlesworth D, Barton NH, Charlesworth B. The sources of adaptive variation. Proc Roy Soc B Biol Sci. 2017;284:20162864.
Chen Z, Farrell AP, Matala A, Hoffman N, Narum SR. Physiological and genomic signatures of evolutionary thermal adaptation in redband trout from extreme climates. Evol Appl. 2018. https://doi.org/10.1111/eva.12672.
Chevalier F, Martin O, Rofidal V, et al. Proteomic investigation of natural variation between Arabidopsis ecotypes. Proteomics. 2004;4:1372–81.
Chiou KL, Bergey CM. Methylation-based enrichment facilitates low-cost, noninvasive genomic scale sequencing of populations from feces. Sci Rep. 2018;8:1975.
Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31:2745–7.
Christie MR, Marine ML, Fox SE, French RA, Blouin MS. A single generation of domestication heritably alters the expression of hundreds of genes. Nat Commun. 2016;7:10676.
Chu AY, Tin A, Schlosser P, et al. Epigenome-wide association studies identify DNA methylation associated with kidney function. Nat Commun. 2017;8:1286.
Colinet H, Pineau C, Com E. Large scale phosphoprotein profiling to explore Drosophila cold acclimation regulatory mechanisms. Sci Rep. 2017;7:1713.
Conte GL, Hodgins KA, Yeaman S, et al. Bioinformatically predicted deleterious mutations reveal complementation in the interior spruce hybrid complex. BMC Genomics. 2017;18:970.
Cooke NP, Nakagome S. Fine-tuning of approximate Bayesian computation for human population genomics. Curr Opin Genet Dev. 2018;53:60–9.
Coop G, Witonsky D, Di Rienzo A, Pritchard JK. Using environmental correlations to identify loci underlying local adaptation. Genetics. 2010;185:1411–23.
Corbett-Detig RB, Hartl DL, Sackton TB. Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 2015;13:e1002112.
Cornuet JM, Luikart G. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics. 1996;144:2001–14.
Cosart T, Beja-Pereira A, Chen S, et al. Exome-wide DNA capture and next generation sequencing in domestic and wild species. BMC Genomics. 2011;12:347–55.
Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 2014;23:3133–57.
Cutter AD, Payseur BA. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet. 2013;14:262–74.

Dalongeville A, Benestan L, Mouillot D, Lobreaux S, Manel S. Combining six genome scan methods to detect candidate genes to salinity in the Mediterranean striped red mullet (Mullus surmuletus). BMC Genomics. 2018;19:217.
De Kort H, Baguette M, Prunier JG, et al. Genetic costructure in a meta-community under threat of habitat fragmentation. Mol Ecol. 2018;27:2193–203.
De La Torre AR, Birol I, Bousquet J, et al. Insights into conifer giga-genomes. Plant Physiol. 2014;166:1724–32.
De Mita S, Thuillet AC, Gay L, et al. Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. Mol Ecol. 2013;22:1383–99.
Degiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. SweepFinder2: increased sensitivity, robustness and flexibility. Bioinformatics. 2016;32:1895–7.
DeLong EF. The microbial ocean from genomes to biomes. Nature. 2009;459:200–6.
Denef VJ. Peering into the genetic makeup of natural microbial populations using metagenomics. In: Polz MF, Om PR, editors. Population genomics: microorganisms. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2018_14.
Di G, Miao X, Ke C, et al. Protein changes in abalone foot muscle from three geographical populations of *Haliotis diversicolor* based on proteomic approach. Ecol Evol. 2016;6:3645–57.
Dinsdale EA, Edwards RA, Hall D, et al. Functional metagenomic profiling of nine biomes. Nature. 2008;452:629–32.
Do C, Waples RS, Peel D, et al. NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Mol Ecol Resour. 2014;14:209–14.
Dobrynin P, Liu S, Tamazian G, et al. Genomic legacy of the African cheetah, *Acinonyx jubatus*. Genome Biol. 2015;16:277.
Dowle EJ, Pochon X, C Banks J, Shearer K, Wood SA. Targeted gene enrichment and high-throughput sequencing for environmental biomonitoring: a case study using freshwater macroinvertebrates. Mol Ecol Resour. 2016;16:1240–54.
Dupuis JR, Oliver JC, Brunet BMT, et al. Genomic data indicate ubiquitous evolutionary distinctiveness among populations of California metalmark butterflies. Conserv Genet. 2018. In press.
Duranton M, Allal F, Fraïsse C, et al. The origin and remolding of genomic islands of differentiation in the European sea bass. Nat Commun. 2018;9:2518.
Eaton DAR. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics. 2014;30:1844–9.
Edwards SV, Potter S, Schmitt CJ, Bragg JG, Moritz C. Reticulation, divergence, and the phylogeography–phylogenetics continuum. Proc Natl Acad Sci. 2016;113:8025–32.
Eichten SR, Briskine R, Song J, et al. Epigenetic and genetic influences on DNA methylation variation in maize populations. Plant Cell. 2013;25:2783–97.
Elbasyoni IS, Lorenz AJ, Guttieri M, et al. A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. Plant Sci. 2018;270:123–30.
Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 2014;29:51–63.
Ellegren H, Smeds L, Burri R, et al. The genomic landscape of species divergence in *Ficedula* flycatchers. Nature. 2012;491:756–60.
Elleouet JS, Aitken SN. Exploring approximate Bayesian computation for inferring recent demographic history with genomic markers in nonmodel species. Mol Ecol Resour. 2018;18:525–40.
Epstein B, et al. Rapid evolutionary response to a transmissible cancer in Tasmanian devils. Microbiome. 2016;6(1):168.
Farek J, Hughes D, Mansfield A, et al. xAtlas: scalable small variant calling across heterogeneous next-generation sequencing experiments. BioRxiv. 2018:295071.
Faria NR, Kraemer MUG, Hill S, et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. BioRxiv. 2018:299842.

Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000;155:1405–13.
Feau N, Beauseigle S, Bergeron M-J, et al. Genome-enhanced detection and identification (GEDI) of plant pathogens. PeerJ. 2018;6:e4392.
Feder JL, Egan SP, Nosil P. The genomics of speciation-with-gene-flow. Trends Genet. 2012;28:342–50.
Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol. 2014;31:1275–91.
Fisher R. The theory of inbreeding. 2nd ed. Edinburgh: Oliver & Boyd; 1965.
Flanagan SP, Forester BR, Latch EK, Aitken SN, Hoban S. Guidelines for planning genomic assessment and monitoring of locally adaptive variation to inform species conservation. Evol Appl. 2018;11:1035–52.
Foll M, Fischer MC, Heckel G, Excoffier L. Estimating population structure from AFLP amplification intensity. Mol Ecol. 2010;19:4638–47.
Forester BR, Jones MR, Joost S, Landguth EL, Lasky JR. Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. Mol Ecol. 2016;25:104–20.
Forester BR, Lasky JR, Wagner HH, Urban DL. Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. Mol Ecol. 2018;27:2215–33.
Forstmeier W, Schielzeth H, Mueller JC, Ellegren H, Kempenaers B. Heterozygosity-fitness correlations in zebra finches: microsatellite markers can be better than their reputation. Mol Ecol. 2012;21:3237–49.
Foust CM, Preite V, Schrey AW, et al. Genetic and epigenetic differences associated with environmental gradients in replicate populations of two salt marsh perennials. Mol Ecol. 2016;25:1639–52.
Fraïsse C, Roux C, Gagnaire P-A, et al. The divergence history of European blue mussel species reconstructed from approximate Bayesian computation: the effects of sequencing techniques and sampling strategies. PeerJ. 2018;6:e5198.
Franklin IR. The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. Theor Popul Biol. 1977;11:60–80.
Frantz LAF, Mullin VE, Pionnier-Capitan M, et al. Genomic and archaeological evidence suggests a dual origin of domestic dogs. Science. 2016;352:1228–31.
Frichot E, Schoville SD, Bouchard G, François O. Testing for associations between loci and environmental gradients using latent factor mixed models. Mol Biol Evol. 2013;30:1687–99.
Funk WC, McKay JK, Hohenlohe PA, Allendorf FW. Harnessing genomics for delineating conservation units. Trends Ecol Evol. 2012;27:489–96.
Funk WC, Forester BR, Converse SJ, Darst C, Morey S. Improving conservation policy with genomics: a guide to integrating adaptive potential into U.S. Endangered Species Act decisions for conservation practitioners and geneticists. Conserv Genet. 2018. In press.
Gamboa M, Tsuchiya MC, Matsumoto S, Iwata H, Watanabe K. Differences in protein expression among five species of stream stonefly (Plecoptera) along a latitudinal gradient in Japan. Arch Insect Biochem Physiol. 2017;96:e21422.
Gapp K, Bohacek J. Epigenetic germline inheritance in mammals: looking to the past to understand the future. Genes Brain Behav. 2018;17:e12407.
Gapp K, Jawaid A, Sarkies P, et al. Implication of sperm RNAs in transgenerational inheritance of the effects of early trauma in mice. Nat Neurosci. 2014;17:667–9.
Garcia SL, Stevens SLR, Crary B, Martinez-Garcia M, Stepanauskas R, et al. Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations. ISME J. 2018;12:742–55. https://doi.org/10.1038/s41396-017-0001-0.
Garner BA, Hand BK, Amish SJ, et al. Genomics in conservation: case studies and bridging the gap between data and application. Trends Ecol Evol. 2016;31:81–2.
Gasc C, Peyretaillade E, Peyret P. Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. Nucleic Acids Res. 2016;44:4504–18.
Gauthier J, Mouden C, Suchan T, et al. DiscoSnp-RAD: de novo detection of small variants for population genomics. BioRxiv. 2017:216747.
Gavery MR, Roberts SB. Epigenetic considerations in aquaculture. PeerJ. 2017;5:e4147.

Geigl E-M, Grange T. Of cats and men: ancient dNA reveals how the cat conquered the ancient world. In: Lindqvist C, Om PR, editors. Paleogenomics. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2018_26.
Ghalambor CK, Hoke KL, Ruell EW, et al. Non-adaptive plasticity potentiates rapid adaptive evolution of gene expression in nature. Nature. 2015;525:372–5.
Giardine B, Riemer C, Hardison RC, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005;15:1451–5.
Gibson G. Population genetics and GWAS: a primer. PLoS Biol. 2018;16:e2005485.
Gilbert KJ, Whitlock MC. Evaluating methods for estimating local effective population size with and without migration. Evolution. 2015;69:2154–66.
Gompert Z. A continuous correlated beta process model for genetic ancestry in admixed populations. PLoS One. 2016;11:e0151047.
Goudet J, Kay T, Weir BS. How to estimate kinship. Mol Ecol. 2018. In press.
Gray MM, Granka JM, Bustamante CD, et al. Linkage disequilibrium and demographic history of wild and domestic canids. Genetics. 2009;181:1493–505.
Grossen C, Biebach I, Angelone-Alasaad S, Keller LF, Croll D. Population genomics analyses of European ibex species show lower diversity and higher inbreeding in reintroduced populations. Evol Appl. 2018;11:123–39.
Gruber B, Unmack PJ, Berry OF, Georges A. dartr: an r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. Mol Ecol Resour. 2018;18:691–9.
Guan Y. Detecting structure of haplotypes and local ancestry. Genetics. 2014;196:625–42.
Gugger PF, Fitz-Gibbon S, Pellegrini M, Sork VL. Species-wide patterns of DNA methylation variation in *Quercus lobata* and their association with climate gradients. Mol Ecol. 2016;25:1665–80.
Gunther T, Coop G. Robust identification of local adaptation from allele frequencies. Genetics. 2013;195:205–20.
Gur A, Tzuri G, Meir A, et al. Genome-wide linkage-disequilibrium mapping to the candidate gene level in melon (*Cucumis melo*). Sci Rep. 2017;7:9770.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5:e1000695.
Hackinger S, Kraaijenbrink T, Xue Y, et al. Wide distribution and altitude correlation of an archaic high-altitude-adaptive EPAS1 haplotype in the Himalayas. Hum Genet. 2016;135:393–402.
Hancock AM, et al. Adaptation to climate across the Arabidopsis thaliana genome. Science. 2011;334:83–6.
Hand BK, Hether TD, Kovach RP, et al. Genomics and introgression: discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. Curr Zool. 2015a;61:146–54.
Hand BK, Lowe WH, Kovach RP, Muhlfeld CC, Luikart G. Landscape community genomics: understanding eco-evolutionary processes in complex environments. Trends Ecol Evol. 2015b;30:161–8.
Hanghøj K, Orlando L, Hanghøj K, Orlando ÁL. Ancient epigenomics. In: Lindqvist C, Om PR, editors. Paleogenomics. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2018_18.
Hansen MM. Expression of interest: transcriptomics and the designation of conservation units. Mol Ecol. 2010;19:1757–9.
Hare MP, Nunney L, Schwartz MK, et al. Understanding and estimating effective population size for practical application in marine species management. Conserv Biol. 2011;25:438–49.
Harr B. Genomic islands of differentiation between house mouse subspecies. Genome Res. 2006;16:730–7.
Harris K, Nielsen R. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 2013;9:e1003521.
Harris C, Rousset F, Morlais I, Fontenille D, Cohuet A. Low linkage disequilibrium in wild *Anopheles gambiae* s.l. populations. BMC Genet. 2010;11:81.

Harrisson KA, Amish SJ, Pavlova A, et al. Signatures of polygenic adaptation associated with climate across the range of a threatened fish species with high genetic connectivity. Mol Ecol. 2017;26:6253–69.
Haussler D, O'Brien SJ, Ryder OA, et al. Genome 10K: a proposal to obtain whole-genome sequence for 10000 vertebrate species. J Hered. 2009;100:659–74.
Hedrick PW, Garcia-Dorado A. Understanding inbreeding depression, purging, and genetic rescue. Trends Ecol Evol. 2016;31:940–52.
Heintzman PD, Soares AER, Chang D, Shapiro B. Paleogenomics. Rev Cell Biol Mol Med. 2015;1:243–67.
Hendricks S, Anderson EC, Antao T, et al. Recent advances in conservation and population genomics data analysis. Evol Appl. 2018;11:1197–211.
Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 2005;169:2335–52.
Hermisson J, Pennings PS. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods Ecol Evol. 2017;8:700–16.
Hidalgo-Galiana A, Monge M, Biron DG, et al. Protein expression parallels thermal tolerance and ecologic changes in the diversification of a diving beetle species complex. Heredity. 2016;116:114–23.
Hoban S. Integrative conservation genetics: prioritizing populations using climate predictions, adaptive potential and habitat connectivity. Mol Ecol Resour. 2018;18:14–7.
Hoban SM, Gaggiotti OE, Bertorelle G. The number of markers and samples needed for detecting bottlenecks under realistic scenarios, with and without recovery: a simulation-based study. Mol Ecol. 2013;22:3444–50.
Hoban S, Kelley JL, Lotterhos KE, et al. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. Am Nat. 2016;188:379–97.
Hodel RG, Chandler LM, Fahrenkrog AM, et al. Linking genome signatures of selection and adaptation in non-model plants: exploring potential and limitations in the angiosperm *Amborella*. Curr Opin Plant Biol. 2018;42:81–9.
Hodges E, Xuan Z, Balija V, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet. 2007;39:1522–7.
Hoffberg SL, Kieran TJ, Catchen JM, et al. RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. Mol Ecol Resour. 2016;16:1264–78.
Hoffman JI, Simpson F, David P, et al. High-throughput sequencing reveals inbreeding depression in a natural population. Proc Natl Acad Sci. 2014;111:3775–80.
Hogg JT. Mating in bighorn sheep: multiple creative male strategies. Science. 1984;225:526–9.
Hohenlohe PA, Bassham S, Etter PD, et al. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 2010a;6:e1000862.
Hohenlohe PA, Phillips PC, Cresko WA. Using population genomics to detect selection in natural populations: key concepts and methodological considerations. Int J Plant Sci. 2010b;171:1059–71.
Hohenlohe PA, Hand BK, Andrews KR, Luikart G. Population genomics provides key insights in ecology and evolution. In: Om PR, editor. Population genomics: concepts, approaches and applications. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2018_20.
Holliday JA, Ritland K, Aitken SN. Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (*Picea sitchensis*). New Phytol. 2010;188:501–14.
Holliday JA, Wang T, Aitken S. Predicting adaptive phenotypes from multilocus genotypes in Sitka spruce (*Picea sitchensis*) using random forest. G3. 2012;2:1085–93.
Holliday JA, Hallerman EM, Haak DC. Genotyping and sequencing technologies in population genetics and genomics. In: Om PR, editor. Population genomics: concepts, approaches and

applications. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2017_5.
Horsthemke B. A critical view on transgenerational epigenetic inheritance in humans. Nat Commun. 2018;9:2973.
Howard JT, Haile-Mariam M, Pryce JE, Maltecca C. Investigation of regions impacting inbreeding depression and their association with the additive genetic effect for United States and Australia Jersey dairy cattle. BMC Genomics. 2015;16:813.
Hu J, Barrett RDH. Epigenetics in natural animal populations. J Evol Biol. 2017;30:1612–32.
Huber B, Whibley A, Poul YL, et al. Conservatism and novelty in the genetic architecture of adaptation in *Heliconius* butterflies. Heredity. 2015;114:515–24.
Huerta-Sánchez E, Jin X, Asan, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014;512:194–7.
Huisman J, Kruuk LEB, Ellis PA, Clutton-Brock T, Pemberton JM. Inbreeding depression across the lifespan in a wild mammal population. Proc Natl Acad Sci U S A. 2016;113:3585–90.
Humble E, Dasmahapatra KK, Martinez-Barrio A, et al. RAD sequencing and a hybrid antarctic fur seal genome assembly reveal rapidly decaying linkage disequilibrium, global population structure and evidence for inbreeding. G3. 2018;8:2709–22.
Hunter ME, Hoban SM, Bruford MW, Segelbacher G, Bernatchez L. Next-generation conservation genetics and biodiversity monitoring. Evol Appl. 2018;11:1029–34.
Husby A, Kawakami T, Rönnegård L, et al. Genome-wide association mapping in a wild avian population identifies a link between genetic and phenotypic variation in a life-history trait. Proc Biol Sci. 2015;282:20150156.
Jensen JD, Foll M, Bernatchez L. The past, present and future of genomic scans for selection. Mol Ecol. 2016;25:1–4.
Johnson EC, Evans LM, Keller MC. Relationships between estimated autozygosity and complex traits in the UK Biobank. PLoS Genet. 2018a;14:e1007556.
Johnson JS, Krutovsky KV, Rajora OP, Gaddis KD, Cairns DM. Advancing biogeography through population genomics. In: Om PR, editor. Population genomics: concepts, approaches and applications. Cham: Springer International Publishing AG; 2018b. https://doi.org/10.1007/13836_2018_39.
Johnston SE, McEwan JC, Pickering NK, et al. Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. Mol Ecol. 2011;20:2555–66.
Johnston SE, Gratten J, Berenos C, et al. Life history trade-offs at a single locus maintain sexually selected genetic variation. Nature. 2013;502:93–5.
Jones MR, Good JM. Targeted capture in evolutionary and ecological genomics. Mol Ecol. 2016;25:185–202.
Jones MR, Scott Mills L, Alves PC, et al. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. Science. 2018;360:1355–8.
Joost S, Bonin A, Bruford MW, et al. A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. Mol Ecol. 2007;16:3955–69.
Kabekkodu SP, Chakrabarty S, Ghosh S, Brand A, Satyamoorthy K. Epigenomics, pharmacoepigenomics, and personalized medicine in cervical cancer. Public Health Genomics. 2017;20:100–15.
Kardos M, Shafer ABA. The peril of gene-targeted conservation. Trends Ecol Evol. 2018. https://doi.org/10.1016/j.tree.2018.08.011.
Kardos M, Luikart G, Allendorf FW. Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. Heredity. 2015a;115:63–72.
Kardos M, Luikart G, Bunch R, et al. Whole-genome resequencing uncovers molecular signatures of natural and sexual selection in wild bighorn sheep. Mol Ecol. 2015b;24:5616–32.
Kardos M, Husby A, Mcfarlane SE, Qvarnstrom A, Ellegren H. Whole-genome resequencing of extreme phenotypes in collared flycatchers highlights the difficulty of detecting quantitative trait loci in natural populations. Mol Ecol Resour. 2016a;16:727–41.

Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW. Genomics advances the study of inbreeding depression in the wild. Evol Appl. 2016b;9:1205–18.
Kardos M, Qvarnström A, Ellegren H. Inferring individual inbreeding and demographic history from segments of identity by descent in *Ficedula* flycatcher genome sequences. Genetics. 2017;205:1319–34.
Kardos M, Åkesson M, Fountain T, et al. Genomic consequences of intensive inbreeding in an isolated wolf population article. Nat Ecol Evol. 2018;2:124–31.
Karr TL. Application of proteomics to ecology and population biology. Heredity. 2008;100:200–6.
Kawakami T, Smeds L, Backström N, et al. A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. Mol Ecol. 2014;23:4035–58.
Keller LF, Waller DM. Inbreeding effects in wild populations. Trends Ecol Evol. 2002;17:230–41.
Keller MC, Simonson MA, Ripke S, et al. Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. PLoS Genet. 2012;8:e1002656.
Kelso J, Prüfer K. Ancient humans and the origin of modern humans. Curr Opin Genet Dev. 2014;29:133–8.
Kijas JW. Detecting regions of homozygosity to map the cause of recessively inherited disease. Methods Mol Biol. 2013;1019:331–45.
Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. 2002;160:765–77.
Kirin M, McQuillan R, Franklin CS, et al. Genomic runs of homozygosity record population history and consanguinity. PLoS One. 2010;5:e13996.
Kirkpatrick M. How and why chromosome inversions evolve. PLoS Biol. 2010;8:e1000501.
Knaus BJ, Grünwald NJ. vcfr: a package to manipulate and visualize variant call format data in R. Mol Ecol Resour. 2017;17:44–53.
Knief U, Kempenaers B, Forstmeier W. Meiotic recombination shapes precision of pedigree- and marker-based estimates of inbreeding. Heredity. 2017;118:239–48.
Kovach RP, Hand BK, Hohenlohe PA, et al. Vive la résistance: genome-wide selection against introduced alleles in invasive hybrid zones. Proc Roy Soc B Biol Sci. 2016;283:20161380.
Kozakiewicz CP, Burridge CP, Funk WC, et al. Pathogens in space: advancing understanding of pathogen dynamics and disease ecology through landscape genetics. Evol Appl. 2018. In press.
Kreiner JM, Stinchcombe JR, Wright SI. Population genomics of herbicide resistance: adaptation via evolutionary rescue. Annu Rev Plant Biol. 2018;69:611–35.
Küpper C, Stocks M, Risse JE, et al. A supergene determines highly divergent male reproductive morphs in the ruff. Nat Genet. 2015;48:79–83.
Laforest-Lapointe I, Paquette A, Messier C, Kembel SW. Leaf bacterial diversity mediates plant diversity and ecosystem function relationships. Nature. 2017;546(7656):145. https://doi.org/10.1038/nature22399.
Lamichhaney S, Fan G, Widemo F, et al. Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). Nat Genet. 2015;48:84–8.
Lan T, Lindqvist C. Paleogenomics: genome-scale analysis of ancient DNA and population and evolutionary genomic inferences. In: Om PR, editor. Population genomics: concepts, approaches and applications. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2017_7.
Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science. 1987;236:1567–70.
Landry PA, Koskinen MT, Primmed CR. Deriving evolutionary relationships among populations using microsatellites and (δμ)2: all loci are equal, but some are more equal than others. Genetics. 2002;161:1339–47.
Laporte M, Pavey SA, Rougeux C, et al. RAD sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in North Atlantic Eels. Mol Ecol. 2016;25:219–37.

Larson WA, Seeb LW, Everett MV, et al. Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). Evol Appl. 2014;7:355–69.
Larson WA, Limborg MT, McKinney GJ, et al. Genomic islands of divergence linked to ecotypic variation in sockeye salmon. Mol Ecol. 2017;26:554–70.
Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8:e1002453.
Le Luyer J, Laporte M, Beacham TD, et al. Parallel epigenetic modifications induced by hatchery rearing in a Pacific salmon. Proc Natl Acad Sci. 2017;114:12964–9.
Lee MK, Hong Y, Kim S-Y, Kim WJ, London SJ. Epigenome-wide association study of chronic obstructive pulmonary disease and lung function in Koreans. Epigenomics. 2017;9:971–84.
Lee HJ, Georgiadou A, Otto TD, et al. Transcriptomic studies of malaria: a paradigm for investigation of systemic host-pathogen interactions. Microbiol Mol Biol Rev. 2018;82:e00071–17.
Leite DCA, Salles JF, Calderon EN, et al. Coral bacterial-core abundance and network complexity as proxies for anthropogenic pollution. Front Microbiol. 2018;9:833.
Leitwein M, Gagnaire P-A, Desmarais E, Berrebi P, Guinand B. Genomic consequences of a recent three-way admixture in supplemented wild brown trout populations revealed by local ancestry tracts. Mol Ecol. 2018;27:3466–83.
Lewontin RC, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics. 1973;74:175–95.
Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–6.
Lind MI, Spagopoulou F. Evolutionary consequences of epigenetic inheritance. Heredity. 2018;121:205–9.
Lorenzo FR, Huff C, Myllymäki M, et al. A genetic mechanism for Tibetan high-altitude adaptation. Nat Genet. 2014;46:951–6.
Lotterhos KE, Whitlock MC. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. Mol Ecol. 2014;23(9):2178–92.
Lowry DB, Hoban S, Kelley JL, et al. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. Mol Ecol Resour. 2017;17:142–52.
Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet. 2003;4:981–94.
Luikart G, Ryman N, Tallmon DA, Schwartz MK, Allendorf FW. Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. Conserv Genet. 2010;11:355–73.
Ma L, Sun X, Kong X, et al. Physiological, biochemical and proteomics analysis reveals the adaptation strategies of the alpine plant *Potentilla saundersiana* at altitude gradient of the Northwestern Tibetan Plateau. J Proteomics. 2015;112:63–82.
Malécot G. The mathematics of heredity. San Francisco: W.H. Freeman; 1970.
Manthey JD, Campillo LC, Burns KJ, Moyle RG. Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: a test in cardinalid tanagers (Aves, Genus: Piranga). Syst Biol. 2016;65:640–50.
Marciniak S, Poinar H. Ancient pathogens through human history: a paleogenomic perspective. In: Lindqvist C, Rajora OP, editors. Paleogenomics. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2018.
Marques DA, Lucek K, Meier JI, et al. Genomics of rapid incipient speciation in sympatric threespine stickleback. PLoS Genet. 2016;12:e1005887.
Marques DA, Jones FC, Di Palma F, Kingsley DM, Reimchen TE. Experimental evidence for rapid genomic adaptation to a new niche in an adaptive radiation. Nat Ecol Evol. 2018;2:1128–38.
Marsden CD, Lee Y, Kreppel K, et al. Diversity, differentiation, and linkage disequilibrium: prospects for association mapping in the malaria vector *Anopheles arabiensis*. G3. 2014;4:121–31.

Martin EM, Fry RC. Environmental influences on the epigenome: exposure-associated DNA methylation in human populations. Annu Rev Public Health. 2018;39:309–33.
Martin SH, Dasmahapatra KK, Nadeau NJ, et al. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 2013;23:1817–28.
Matthews B, Best RJ, Feulner PGD, Narwani A, Limberger R. Evolution as an ecosystem process: insights from genomics. Genome. 2018;61:298–309.
Maynard Smith J, Haigh J. The hitch-hiking effect of a favorable gene. Genet Res. 1974;23:23–35.
McCartney-Melstad E, Mount GG, Shaffer HB. Exon capture optimization in amphibians with large genomes. Mol Ecol Resour. 2016;16:1084–94.
McCoy RC, Akey JM. Selection plays the hand it was dealt: evidence that human adaptation commonly targets standing genetic variation. Genome Biol. 2017;18:139.
McKain MR, Johnson MG, Uribe-Convers S, Eaton D, Yang Y. Practical considerations for plant phylogenomics. Appl Plant Sci. 2018;6:e1038.
Mckinney GJ, Seeb LW, Larson WA, et al. An integrated linkage map reveals candidate genes underlying adaptive variation in Chinook salmon (*Oncorhynchus tshawytscha*). Mol Ecol Resour. 2016;16:769–83.
McKinney GJ, Larson WA, Seeb LW, Seeb JE. RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry et al. (2016). Mol Ecol Resour. 2017a;17:356–61.
McKinney GJ, Waples RK, Seeb LW, Seeb JE. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. Mol Ecol Resour. 2017b;17:656–69.
Mckown AD, Klápště J, Guy RD, et al. Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. New Phytol. 2014;203:535–53.
McMahon BJ, Teeling EC, Höglund J. How and why should we implement genomics into conservation? Evol Appl. 2014;7:999–1007.
Meeks KAC, Henneman P, Venema A, et al. An epigenome-wide association study in whole blood of measures of adiposity among Ghanaians: the RODAM study. Clin Epigenetics. 2017;9:103.
Meyer M, Kircher M, Gansauge MT, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338:222–6.
Miao B, Wang Z, Li Y. Genomic analysis reveals hypoxia adaptation in the tibetan mastiff by introgression of the gray wolf from the Tibetan plateau. Mol Biol Evol. 2017;34:734–43.
Miles A, Harding NJ, Bottà G, et al. Genetic diversity of the African malaria vector *anopheles gambiae*. Nature. 2017;552:96–100.
Miller JM, Malenfant RM, David P, et al. Estimating genome-wide heterozygosity: effects of demographic history and marker type. Heredity. 2014;112:240–7.
Moler ERV, Abakir A, Eleftheriou M, Johnson JS, Krutovsky KV, Lewis LC, Ruzov A, Whipple AV, Rajora OP. Population epigenomics. In: Om PR, editor. Population genomics: concepts, approaches and applications. Cham: Springer International Publishing AG; 2018.
Moran MA. Metatranscriptomics: eavesdropping on complex microbial communities. Microbe Mag. 2009;4:329–35.
Muhlfeld CC, Kalinowski ST, McMahon TE, et al. Hybridization rapidly reduces fitness of a native trout in the wild. Biol Lett. 2009;5:328–31.
Nadeau NJ, Kawakami T. Population genomics of speciation and admixture. In: Om PR, editor. Population genomics: concepts, approaches and applications. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2018_24.
Nadeau NJ, Ruiz M, Salazar P, et al. Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. Genome Res. 2014;24:1316–33.
Nakatochi M, Ichihara S, Yamamoto K, et al. Epigenome-wide association of myocardial infarction with DNA methylation sites at loci related to cardiovascular disease. Clin Epigenetics. 2017;9:54.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA. Genotyping-by-sequencing in ecological and conservation genomics. Mol Ecol. 2013;22:2841–7.
Narum SR, Di Genova A, Micheletti SJ, Maass A. Genomic variation underlying complex life-history traits revealed by genome sequencing in Chinook salmon. Proc Roy Soc B Biol Sci. 2018;285:20180935.
Nash DR, Als TD, Maile R, Jones GR, Boomsma JJ. A mosaic of chemical coevolution in a large blue butterfly. Science. 2008;319:88–90.
Nazareno AG, Bemmels JB, Dick CW, Lohmann LG. Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. Mol Ecol Resour. 2017;17:1136–47.
Nedelkov D. Population proteomics: investigation of protein diversity in human populations. Proteomics. 2008;8:779–86.
Nedelkov D, Kiernan UA, Niederkofler EE, Tubbs KA, Nelson RW. Investigating diversity in human plasma proteins. Proc Natl Acad Sci. 2005;102:10852–7.
Nedelkov D, U A K, Niederkofler EE, Tubbs KA, Nelson RW. Population proteomics: the concept, attributes, and potential for cancer biomarker research. Mol Cell Proteomics. 2006;5:1811–8.
Niederhuth CE, Bewick AJ, Ji L, et al. Widespread natural variation of DNA methylation within angiosperms. Genome Biol. 2016;17:174.
Nielsen R, Williamson S, Kim Y, et al. Genomic scans for selective sweeps using SNP data. Genome Res. 2005;15:1566–75.
Nilsson E, et al. Environmental toxicant induced epigenetic transgenerational inheritance of ovarian pathology and granulosa cell epigenome and transcriptome alterations: ancestral origins of polycystic ovarian syndrome and primary ovarian insufiency. Epigenetics. 2018;13:875–95.
Noble TJ, Tao Y, Mace ES, et al. Characterization of linkage disequilibrium and population structure in a mungbean diversity panel. Front Plant Sci. 2018;8:2102.
Norris LC, Main BJ, Lee Y, et al. Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. Proc Natl Acad Sci. 2015;112:815–20.
Nosil P, Egan SP, Funk DJ. Heterogeneous genomic differentiation between walking-stick ecotypes: "isolation by adaptation" and multiple roles for divergent selection. Evolution. 2008;62:316–36.
Nunziata SO, Weisrock DW. Estimation of contemporary effective population size and population declines using RAD sequence data. Heredity. 2018;120:196–207.
Nystedt B, Street NR, Wetterbom A, et al. The Norway spruce genome sequence and conifer genome evolution. Nature. 2013;497:579–84.
O'Quin KE, Yoshizawa M, Doshi P, Jeffery WR. Quantitative genetic analysis of retinal degeneration in the blind cavefish *Astyanax mexicanus*. PLoS One. 2013;8:e57281.
Oomen RA, Hutchings JA. Transcriptomic responses to environmental change in fishes: insights from RNA sequencing. FACETS. 2017;2:610–41.
Orlando L. An ancient DNA perspective on horse evolution. In: Lindqvist C, Om PR, editors. Paleogenomics. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2018_23.
Orlando L, Ginolhac A, Zhang G, et al. Recalibrating *equus* evolution using the genome sequence of an early Middle Pleistocene horse. Nature. 2013;499:74–8.
Ozerov MY, Gross R, Bruneaux M, et al. Genomewide introgressive hybridization patterns in wild Atlantic salmon influenced by inadvertent gene flow from hatchery releases. Mol Ecol. 2016;25:1275–93.
Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 2011;12:87–98.
Palkopoulou E, Mallick S, Skoglund P, et al. Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. Curr Biol. 2015;25:1395–400.
Paradis E, Gosselin T, Goudet J, Jombart T, Schliep K. Linking genomics and population genetics with R. Mol Ecol Resour. 2017;17:54–66.

Pardo-Diaz C, Salazar C, Baxter SW, et al. Adaptive introgression across species boundaries in *Heliconius* butterflies. PLoS Genet. 2012;8:e1002752.
Parducci L, Nota K, Wood J. Reconstructing past vegetation communities using ancient DNA from lake sediments. In: Lindqvist C, Om PR, editors. Paleogenomics. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2018_38.
Paris JR, Stevens JR, Catchen JM. Lost in parameter space: a road map for stacks. Meth Ecol Evol. 2017;8:1360–73.
Payseur BA, Rieseberg LH. A genomic perspective on hybridization and speciation. Mol Ecol. 2016;25:2337–60.
Peek RA, O'Rourke SM, Miller MR. Flow regulation associated with decreased genetic health of a river-breeding frog species. BioRxiv. 2018;316604.
Pemberton TJ, Absher D, Feldman MW, et al. Genomic patterns of homozygosity in worldwide human populations. Am J Hum Genet. 2012;91:275–92.
Pemberton JM, Ellis PE, Pilkington JG, Bérénos C. Inbreeding depression by environment interactions in a free-living mammal population. Heredity. 2017;118:64–77.
Pennings PS, Hermisson J. Soft sweeps II – molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol. 2006;23:1076–84.
Pérez O'Brien AM, Utsunomiya YT, Mészáros G, et al. Assessing signatures of selection through variation in linkage disequilibrium between taurine and indicine cattle. Genet Sel Evol. 2014;46:19.
Perry GH, Marioni JC, Melsted P, Gilad Y. Genomic-scale capture and sequencing of endogenous DNA from feces. Mol Ecol. 2010;19:5332–44.
Petkova D, Novembre J, Stephens M. Visualizing spatial population structure with estimated effective migration surfaces. Nat Genet. 2016;48:94–100.
Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8:e1002967.
Pino Del Carpio D, Lozano R, Wolfe MD, Jannink J-L. Genome-wide associationstudies and heritability estimation in the functional genomics era. In: Rajora OP, editor. Population genomics: concepts, approaches and applications. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2018_12.
Poelstra JW, Vijay N, Bossu CM, et al. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science. 2014;344:1410–4.
Pogorelcnik R, Vaury C, Pouchin P, Jensen S, Brasset E. SRNAPipe: a Galaxy-based pipeline for bioinformatic in-depth exploration of small RNAseq data. Mob DNA. 2018;9:25.
Portik DM, Smith LL, Bi K. An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). Mol Ecol Resour. 2016;16:1069–83.
Prince DJ, O'Rourke SM, Thompson TQ, et al. The evolutionary basis of premature migration in Pacific salmon highlights the utility of genomics for informing conservation. Sci Adv. 2017;3:e1603198.
Prüfer K, Racimo F, Patterson N, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505:43–9.
Pruisscher P, Nylin S, Gotthard K, Wheat CW. Genetic variation underlying local adaptation of diapause induction along a cline in a butterfly. Mol Ecol. 2018. In press.
Pryce JE, Haile-Mariam M, Goddard ME, Hayes BJ. Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. Genet Sel Evol. 2014;46:71.
Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464:59–65.
Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for archaic adaptive introgression in humans. Nat Rev Genet. 2015;16:359–71.

Rajora OP, Eckert AJ, Zinck JWR. Single-locus versus multilocus patterns of local adaptation to climate in eastern white pine (*Pinus strobus, Pinaceae*). PLoS One. 2016;11:e0158691.
Rasmussen M, Li Y, Lindgreen S, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010;463:757–62.
Rastas P, Calboli FCF, Guo B, Shikano T, Merilä J. Construction of ultradense linkage maps with Lep-MAP2: stickleback F2 recombinant crosses as an example. Genome Biol Evol. 2016;8:78–93.
Razgour O, Taggart JB, Manel S, et al. An integrated framework to identify wildlife populations under threat from climate change. Mol Ecol Resour. 2018;18:18–31.
Rees BB, Andacht T, Skripnikova E, Crawford DL. Population proteomics: quantitative variation within and among populations in cardiac protein expression. Mol Biol Evol. 2011;28:1271–9.
Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R. A practical guide to environmental association analysis in landscape genomics. Mol Ecol. 2015;24:4348–70.
Renaut S, Grassa CJ, Yeaman S, et al. Genomic islands of divergence are not affected by geography of speciation in sunflowers. Nat Commun. 2013;4:1827.
Resendez SD, Bradley JR, Xu D, Gokcumen O. Structural variants in ancient genomes. In: Lindqvist C, Om PR, editors. Paleogenomics. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2018_34.
Richards CL, Alonso C, Becker C, et al. Ecological plant epigenetics: evidence from model and non-model species, and the way forward. Ecol Lett. 2017;20:1576–90.
Rieseberg L. Adaptive introgression: the seeds of resistance. Curr Biol. 2011;21:R581–3.
Rochus CM, Tortereau F, Plisson-Petit F, et al. Revealing the selection history of adaptive loci using genome-wide scans for selection: an example from domestic sheep. BMC Genomics. 2018;19:71.
Rodríguez-Ezpeleta N, Bradbury IR, Mendibil I, et al. Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP markers: effects of sequence clustering parameters and hierarchical SNP selection. Mol Ecol Resour. 2016;16:991–1001.
Roffler GH, Amish SJ, Smith S, et al. SNP discovery in candidate adaptive genes using exon capture in a free-ranging alpine ungulate. Mol Ecol Resour. 2016;16:1147–64.
Rogers RL, Slatkin M. Excess of genomic defects in a woolly mammoth on Wrangel island. PLoS Genet. 2017;13:e1006601.
Roitman S, Joseph Pollock F, Medina M. Coral microbiomes as bioindicators of reef health. In: Population genomics. Cham: Springer; 2018. p. 1–19.
Rondeau EB, Minkley DR, Leong JS, et al. The genome and linkage map of the northern pike (Esox lucius): conserved synteny revealed between the salmonid sister group and the Neoteleostei. PLoS One. 2014;e102089:9.
Rougeux C, Gagnaire P-A, Praebel K, Seehausen O, Bernatchez L. Convergent transcriptomic landscapes under polygenic selection accompany inter- continental parallel evolution within a Nearctic Coregonus (Salmonidae) sister-species complex. BioRxiv. 2018. https://doi.org/10.1101/311464.
Rubin C-J, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 2010;464:587.
Rubin C-J, Megens H-J, Barrio AM, et al. Strong signatures of selection in the domestic pig genome. Proc Natl Acad Sci. 2012;109:19529–36.
Sabeti PC, Reich DE, Higgins JM, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419:832–7.
Sabeti PC, Varilly P, Fry B, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449:913–8.
Saint-Pé K, Blanchet S, Tissot L, et al. Genetic admixture between captive-bred and wild individuals affects patterns of dispersal in a brown trout (*Salmo trutta*) population. Conserv Genet. 2018;5:1269–79.

Salmona J, Heller R, Lascoux M, Shafer A. Inferring demographic history using genomic data. In: Rajora OP, editor. Population genomics: concepts, approaches and applications. Cham: Springer International Publishing AG; 2017. https://doi.org/10.1007/13836_2017_1.
Salojärvi J. Computational tools for population genomics. In: Om PR, editor. Population genomics: concepts, approaches and applications. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2018_57.
Santure AW, Garant D. Wild GWAS-association mapping in natural populations. Mol Ecol Resour. 2018;18:729–38.
Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! Genome Biol. 2011;12:125.
Schlötterer C. The evolution of molecular markers – just a matter of fashion? Nat Rev Genet. 2004;5:63–9.
Schlötterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. Nat Rev Genet. 2014;15:749–63.
Schmidt TL, Filipovi I, Hoffmann AA, Rašić G. Fine-scale landscape genomics of *Aedes aegypti* reveals loss of Wolbachia transinfection, dispersal barrier and potential for occasional long distance movement. BioRxiv. 2017. https://doi.org/10.1101/103598.
Schmitz RJ, He Y, Valdés-López O, et al. Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. Genome Res. 2013a;23:1663–74.
Schmitz RJ, Schultz MD, Urich MA, et al. Patterns of population epigenomic diversity. Nature. 2013b;495:193–8.
Schrider DR, Kern AD. Soft sweeps are the dominant mode of adaptation in the human genome. Mol Biol Evol. 2017;34:1863–77.
Schweizer RM, VonHoldt BM, Harrigan R, et al. Genetic subdivision and candidate genes under selection in North American grey wolves. Mol Ecol. 2016;25:380–402.
Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018;19:329–46.
Shafer ABA, Wolf JBW, Alves PC, et al. Genomics and the challenging translation into conservation practice. Trends Ecol Evol. 2015;30:78–87.
Shafer ABA, Peart CR, Tusso S, et al. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. Methods Ecol Evol. 2017;8:907–17. https://doi.org/10.1111/2041-210X.12700.
Shapiro B, Hofreiter M. A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. Science. 2014;343
Shimada-Sugimoto M, Otowa T, Miyagawa T, et al. Epigenome-wide association study of DNA methylation in panic disorder. Clin Epigenetics. 2017;9:6.
Shin D, Kim S-H, Park J, Lee H-K, Song K-D. Extent of linkage disequilibrium and effective population size of the Landrace population in Korea. Asian Australas J Anim Sci. 2018;31:1078–87.
Simons YB, Bullaughey K, Hudson RR, Sella G. A population genetic interpretation of GWAS findings for human quantitative traits. PLoS Biol. 2018;16:e2002985.
Skoglund P, Malmström H, Raghavan M, et al. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science. 2012;336:466–9.
Slatkin M. Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008;9:477–85.
Smith MW, O'Brien SJ. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat Rev Genet. 2005;6:623–32.
Sollars ESA, Buggs RJA. Genome-wide epigenetic variation among ash trees differing in susceptibility to a fungal disease. BMC Genomics. 2018;19:502.
Song Y, Endepols S, Klemann N, et al. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. Curr Biol. 2011;21:1296–301.
Sork VL. Gene flow and natural selection shape spatial patterns of genes in tree populations: implications for evolutionary processes and applications. Evol Appl. 2016;9:291–310.

Sovic MG, Carstens BC, Gibbs HL. Genetic diversity in migratory bats: results from RADseq data for three tree bat species at an Ohio windfarm. PeerJ. 2016;4:e1647.
Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet. 2015;16:33–44.
Srivathsan A, Ang A, Vogler AP, Meier R. Fecal metagenomics for the simultaneous assessment of diet, parasites, and population genetics of an understudied primate. Front Zool. 2016;13:17.
Stam P. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genet Res. 1980;35:131–55.
Stat M, Huggett MJ, Bernasconi R, et al. Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. Sci Rep. 2017;7:12240.
Stetter MG, Thornton K, Ross-Ibarra J. Genetic architecture and selective sweeps after polygenic adaptation to distant trait optima. BioRxiv. 2018:313247.
Stölting KN, Paris M, Meier C, et al. Genome-wide patterns of differentiation and spatially varying selection between postglacial recolonization lineages of *Populus alba* (*Salicaceae*), a widespread forest tree. New Phytol. 2015;207:723–34.
Storz JF, Beaumont MA, Alberts SC. Genetic evidence for long-term population decline in a savannah-dwelling primate: inferences from a hierarchical Bayesian model. Mol Biol Evol. 2002;19:1981–90.
Sugden LA, Atkinson EG, Fischer AP, et al. Localization of adaptive variants in human genomes using averaged one-dependence estimation. Nat Commun. 2018;9:703.
Suhre K, Arnold M, Bhagwat AM, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun. 2017;8:14357.
Sunnucks P. Efficient genetic markers for population biology. Trends Ecol Evol. 2000;15:199–203.
Suren H, Hodgins KA, Yeaman S, et al. Exome capture from the spruce and pine giga-genomes. Mol Ecol Resour. 2016;16:1136–46.
Syring JV, Tennessen JA, Jennings TN, et al. Targeted capture sequencing in whitebark pine reveals range-wide demographic and adaptive patterns despite challenges of a large, repetitive genome. Front Plant Sci. 2016;7:484.
Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123:585–95.
Tallmon DA, Luikart G, Waples RS. The alluring simplicity and complex reality of genetic rescue. Trends Ecol Evol. 2004;19:489–96.
Taudt A, Colomé-Tatché M, Johannes F. Genetic sources of population epigenomic variation. Nat Rev Genet. 2016;17:319–32.
Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. 2017;49:303–9.
Teshima KM, Coop G, Przeworski M. How reliable are empirical genomic scans for selective sweeps? Genome Res. 2006;16:702–12.
Thalmann O, Perri AR. Paleogenomic inferences of dog domestication. In: Lindqvist C, Om PR, editors. Paleogenomics. Cham: Springer International Publishing AG; 2018. https://doi.org/10.1007/13836_2018_27.
Thompson EA. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics. 2013;194:301–26.
Thurber RV, Willner-Hall D, Rodriguez-Mueller B, et al. Metagenomic analysis of stressed coral holobionts. Environ Microbiol. 2009;11:2148–63.
Tishkoff SA, Reed FA, Ranciaro A, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007;39:31–40.
Tranchant-Dubreuil C, Ravel S, Monat C, et al. TOGGLe, a flexible framework for easily building complex workflows and performing robust large-scale NGS analyses. BioRxiv. 2018. https://doi.org/10.1101/245480.
Trapp J, Gouveia D, Almunia C, et al. Digging deeper into the pyriproxyfen-response of the amphipod *gammarus fossarum* with a next-generation ultra-high-field orbitrap analyser: new perspectives for environmental toxicoproteomics. Front Environ Sci. 2018;6:54.

Turner TL, Hahn MW, Nuzhdin SV. Genomic islands of speciation in *Anopheles gambiae*. PLoS Biol. 2005;3:1572–8.
Valdés A, Ehrlén J. Caterpillar seed predators mediate shifts in selection on flowering phenology in their host plant. Ecology. 2017;98:228–38.
Valdisser PAMR, Pereira WJ, Almeida Filho JE, et al. In-depth genome characterization of a Brazilian common bean core collection using DArTseq high-density SNP genotyping. BMC Genomics. 2017;18:423.
Vallejo RL, Silva RMO, Evenhuis JP, et al. Accurate genomic predictions for BCWD resistance in rainbow trout are achieved using low-density SNP panels: evidence that long-range LD is a major contributing factor. J Anim Breed Genet. 2018;135:263–74.
van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. Trends Genet. 2018;34:666–81.
Vandersteen Tymchuk W, O'Reilly P, Bittman J, MacDonald D, Schulte P. Conservation genomics of Atlantic salmon: variation in gene expression between and within regions of the Bay of Fundy. Mol Ecol. 2010;19:1842–59.
Vasemagi A, Primmer CR. Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (Salmo salar L.). Mol Biol Evol. 2005;22:1067–76.
Vattathil S, Akey JM. Small amounts of archaic admixture provide big insights into human history. Cell. 2015;163:281–4.
Venter J, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004;304:66–74.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4:e154.
VonHoldt BM, Pollinger JP, Earl DA, et al. A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. Genome Res. 2011;21:1294–305.
Waite DW, Dsouza M, Sekiguchi Y, Hugenholtz P, Taylor MW. Network-guided genomic and metagenomic analysis of the faecal microbiota of the critically endangered kakapo. Sci Rep. 2018;8:8228.
Wallberg A, Han F, Wellhagen G, et al. A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. Nat Genet. 2014;46:1081–8.
Wang J. Estimation of effective population sizes from data on genetic markers. Phil Trans Roy Soc B Biol Sci. 2005;360:1395–409.
Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
Wang J, Street NR, Scofield DG, Ingvarsson PK. Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American aspens. Mol Biol Evol. 2016;33:1754–67.
Waples RS, Do C. Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. Evol Appl. 2010;3:244–62.
Waples RK, Larson WA, Waples RS. Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. Heredity. 2016;117:233–40.
Waterhouse MD, Erb LP, Beever EA, Russello MA. Adaptive population divergence and directional gene flow across steep elevational gradients in a climate-sensitive mammal. Mol Ecol. 2018;27:2512–28.
Wecek K, Hartmann S, Paijmans JLA, et al. Complex admixture preceded and followed the extinction of wisent in the wild. Mol Biol Evol. 2017;34:598–612.
Wellenreuther M, Bernatchez L. Eco-evolutionary genomics of chromosomal inversions. Trends Ecol Evol. 2018;33:427–40.

Wellenreuther M, Hansson B. Detecting polygenic evolution: problems, pitfalls, and promises. Trends Genet. 2016;32:155–64.
Wessinger CA, Kelly JK, Jiang P, Rausher MD, Hileman LC. SNP-skimming: a fast approach to map loci generating quantitative variation in natural populations. Mol Ecol Resour. 2018. https://doi.org/10.1111/1755-0998.12930.
Whiteley AR, Bhat A, Martins EP, et al. Population genomics of wild and laboratory zebrafish (*Danio rerio*). Mol Ecol. 2011;20:4259–76.
Whiteley AR, Fitzpatrick SW, Funk WC, Tallmon DA. Genetic rescue to the rescue. Trends Ecol Evol. 2015;30:42–9.
Wilson G, Rannala B. Bayesian inference of recent migration rates using multilocus genotyoes. Genetics. 2003;163:1177–91.
Wolf JBW, Ellegren H. Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet. 2017;18:87–100.
Wright S. Coefficients of inbreeding and relationship. Am Nat. 1922;56:330–8.
Xu Z, Bolick SCE, Deroo LA, et al. Epigenome-wide association study of breast cancer using prospectively collected sister study samples. J Natl Cancer Inst. 2013;105:694–700.
Xue Y, Prado-Martinez J, Sudmant PH, et al. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. Science. 2015;348:242–5.
Yeaman S, Whitlock MC. The genetic architecture of adaptation under migration-selection balance. Evolution. 2011;65:1897–911.
Yeaman S, Hodgins KA, Lotterhos KE, et al. Convergent local adaptation to climate in distantly related conifers. Science. 2016;353:1431–3.
Yi SV. Insights into epigenome evolution from animal and plant methylomes. Genome Biol Evol. 2017;9:3189–201.
Zhang W, Fan Z, Han E, et al. Hypoxia adaptations in the grey wolf (*Canis lupus chanco*) from Qinghai-Tibet plateau. PLoS Genet. 2014;10:e1004466.
Zhang W, Zhang H, Yang H, et al. Computational resources associating diseases with genotypes, phenotypes and exposures. Brief Bioinform. 2018:bby071.

Part II
Methods

Genotyping and Sequencing Technologies in Population Genetics and Genomics

J.A. Holliday, E.M. Hallerman, and D.C. Haak

Abstract Genotypes are the central data to any population genetic and genomic study, and genotyping methods have steadily evolved since the first direct glimpses of genetic variation were enabled through enzyme protein electrophoresis. Following the development of the polymerase chain reaction, allozymes were supplanted by methods that directly measured allelic variation in nuclear and organellar DNA, most notably through the use of restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs), and microsatellites. At the turn of the millennium, genome-scale polymorphism detection and scoring still was hampered by the low-throughput nature of Sanger sequencing. This limitation changed with the advent of genotyping microarrays that at first yielded hundreds of data points per sample – a revolution at the time – and that subsequently improved to the point where hundreds of thousands of genetic variants could be scored simultaneously. These methods suffered a major flaw, however, in that their cost put them out of reach for studies of most ecologically important but economically unimportant species. The democratization of population genomics arrived with the advent of high-throughput, short-read sequencers and subsequent development of DNA library techniques to subsample the genome in a large number of individuals. Today, such methods – genotyping-by-sequencing, restriction site-associated DNA sequencing, RNA sequencing, and sequence capture – have become mainstays of the population geneticist's toolkit. Refinements to existing library and sequencing methods continue to emerge at a rapid pace, and novel sequencing platforms may soon put the gold standard of long-read, genome-

J.A. Holliday (✉)
Department of Forest Resources and Environmental Conservation, Virginia Tech, Blacksburg, VA, USA
e-mail: jah1@vt.edu

E.M. Hallerman
Department of Fish and Wildlife Conservation, Virginia Tech, Blacksburg, VA, USA

D.C. Haak
Department of Plant Pathology, Physiology, and Weed Science, Virginia Tech, Blacksburg, VA, USA

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2017_5, © Springer International Publishing AG 2018

wide coverage within a broader reach. In this chapter, we comprehensively review genotyping methods used in population genetics, beginning with allozymes and progressing through AFLPs, microsatellites, and SNP arrays. We subsequently turn to a detailed discussion of methods that leverage next-generation technologies to enable truly genome-scale genotyping. Finally, we discuss recent developments and emerging technologies that constitute the "third wave" of sequencing and genotyping methods. Throughout, our aim is to provide methodological details that will be of use to population geneticists.

Keywords Ecological genomics · Genotyping by sequencing · Illumina · Population genomics · Sequence capture

1 Introduction

The central goal of population genetics is to document and understand the significance of intraspecific genetic variation. Four key questions underlie this objective: How much variation exists in a population? What is the origin of the variation? How is the variation maintained? What is the ecological and evolutionary significance of this variation? At the emergence of the field in the early decades of the twentieth century, methods for observing genetic variation were limited. In addition to classical Mendelian traits, geneticists could observe variation of chromosome number, chromosome morphology, and quantitative traits. Geneticists focused on laboratory model species and on those species for which fully inbred lines could be developed; the methods available were not well suited to screening of outbred, wild populations. These limitations, so prominent early in the development of genetics, were addressed by the rise of molecular genetics in the second half of the twentieth century. Indeed, the progress of molecular genetics has been marked by the advent of critical laboratory techniques. In this chapter, we review the emergence and refinement of genotyping technologies, from early protein markers through modern high-throughput sequencing approaches, and discuss the potential and limitations of each.

2 Early Molecular Genetic Markers: Allozymes

Molecular population genetics emerged as a field with the development of methods for observing variation of enzyme proteins. Starch gel electrophoresis and histochemical staining techniques were developed primarily for detecting variant forms of blood proteins, especially enzymes, as a means to study their biochemical function (Lewontin and Hubby 1966). These bidirectional catalytic enzymes convert one substrate to another without themselves being affected and control much of cell metabolism. Examples include enzymes of the glycolytic pathway in which

glucose is broken down and of the Krebs cycle in which energy is generated in the mitochondria. A brief description of the background and methodology will support an understanding of the strengths and limitations of allozyme genetic markers. *Enzymes* mediate a specific biochemical reaction; e.g., lactate dehydrogenase (LDH) removes a hydrogen atom from a lactate molecule. Some enzymes are encoded by multiple genes; *isozymes* are different forms of an enzyme encoded by different loci, which often are differentially expressed among the tissues of an organism. For example, in most fishes, LDH is encoded by three isozyme loci: *LDH-A* is expressed in almost all tissues, *LDH-B* predominately in the liver, and *LDH-C* in the eye. *Allozymes* are allelic forms of isoenzymes, encoded by different alleles at a particular locus, e.g., *LDH-A1* and *LDH-A2* would be different alleles expressed at the *LDH-A* locus. It is this variation that is sought and interpreted in screenings of allozymes. To conduct a screening of a population, small samples of tissues are dissected out, placed in a buffer, and homogenized. Filter paper wicks with a bit of homogenate from each individual in the collection are placed along a slot in a starch gel, to which an electric current is applied, leading molecules in the homogenate to migrate. Different molecular forms of an enzyme encoded by different alleles and genes migrate different distances through the gel depending upon molecular weight and net electrical charge. The gel is then sliced into slabs and the activity of a particular enzyme visualized with a histochemical stain. The stain includes the substrate and cofactors for the enzyme and a suite of chemicals that take a simple product of the reaction and change color, resulting in a banding pattern that can be interpreted to yield presumptive alleles and genotypes. Thorough technical reviews of visualization and interpretation of allozyme markers are provided by Buth (1990), Morizot and Schmidt (1990), Maxam and Gilbert (1977), Murphy et al. (1996), and May (1998). The genotype data are then subjected to statistical analyses to determine various population genetic parameters.

Allozymes have the favorable property of being the products of codominant gene expression – homozygotes and heterozygotes can be distinguished – and can be developed relatively easily for species of interest. Freed of the need for outwardly observable phenotypic traits in plants and animals, screening of genetic variation in any population of interest became a viable technical possibility. It led to the discovery of unexpectedly high levels of genetic variation in a wide range of natural populations, which in turn revolutionized geneticists' view of the world particularly regarding the adaptive significance of genetic variation (Kimura 1983). Although allozyme applications revolutionized molecular population genetics, allozyme methods also posed limitations. Sampling of multiple tissues – e.g., liver, muscle, and eye – at least for animals is generally lethal to the sampled individual, which limits application of allozymes for studies of imperiled species. Only the variation of enzymes for which we have histochemical assays can be screened, which represents but a tiny portion of the genome. Most studies involved screening of about 30 loci for diploid organisms, with some exceptions of up to 54 allozyme loci (Buchert et al. 1997), and not all loci were found to be polymorphic. Furthermore, the most common allele at many loci often showed a frequency greater than 0.9, limiting the power of statistical analyses. The evolutionary

relations of allelic variants cannot be precisely inferred, limiting study of phylogenetics within a lineage. Finally, because of the redundancy of the genetic code, not all DNA sequence-level substitutions lead to protein-level variation, and only a subset of amino acid changes lead to detectable differences in net electric charge or molecular weight. For these reasons, with the evolution of the molecular genetic techniques, DNA-based markers were developed and have been the preferred approach to study patterns of genetic variation within and among populations.

3 DNA Markers

3.1 Restriction Fragment Length Polymorphisms

The discovery of restriction endonucleases revolutionized molecular biology (Avise 2004). Type II restriction enzymes (Kessler 1987) cleave double-stranded DNA at particular base-pair sequences, typically four to six base pairs in length (Roberts 1984). DNA sequence polymorphisms among individuals may result in differences in the presence or absence of restriction sites and hence in the sizes of the respective restriction fragments. These differences are termed restriction fragment length polymorphisms or RFLPs. The methods for visualization of RFLPs differ for organellar and nuclear DNA. Early RFLP studies were conducted using organellar DNA and not genomic DNA because of the relative simplicity owing to the small size of organellar genomes. The animal mitochondrial DNA molecule is relatively small (~15–20 kb) and circular, properties that contribute to its isolation and made analysis by restriction site variability easy (Lansman et al. 1981; Hoelzel 1992; Dowling et al. 1996; Avise 2004). Following digestion with one or more restriction enzymes, fragments may be visualized by using gel electrophoresis (either ethidium bromide staining or end-labeling with radioactive nucleotides followed by autoradiography). The fragment sizes are observed, and inferences are made of haplotypes (haploid genotypes), i.e., combinations of the presence or absence of restriction sites. Early surveys of animal populations revealed the haploid character, maternal inheritance, and rapid evolution of animal mitochondrial DNA. Analysis of mitochondrial DNA variation is useful for tracking matrilineages and inference of the origins of species or populations, patterns of population dispersal, and occurrence of past population bottlenecks. In humans, for example, Cann et al. (1987) analyzed mitochondrial DNAs from 147 people drawn from five geographic populations. All of the mitochondrial DNA variants were inferred to stem from one woman, popularly referred to as the "mitochondrial Eve," who lived approximately 200,000 years ago, most likely in Africa. Lansman et al. (1983) analyzed mtDNA sequence variation in 135 deer mice *Peromyscus maniculatus* collected across their range in North America and distinguished five major genetic assemblages within the species, as well as extensive diversity within each of those assemblages. Phylogenies derived from mtDNA restriction fragment

analysis were not generally concordant with those derived from morphological characters.

Mitochondrial DNA of plants exhibits surprising contrasts with that of animals (Avise 2004). Plant mtDNA is highly variable in size, ranging from about 200 kb to 2,500 kb among species (Ward et al. 1981; Palmer 1985; Pring and Lonsdale 1985). Within an individual, mtDNA sequences typically exist as a heterogeneous collection of circular molecules that arise from extensive recombination (Palmer and Herbon 1986; Backert et al. 1997; Lonsdale et al. 1988). Inheritance is often, but not always, maternal (Birky 1978; Neale et al. 1989). Plant mtDNA gene order evolves rapidly but about a hundredfold more slowly in nucleotide sequence compared with nuclear DNA (Birky 1988; Palmer and Herbon 1988; Palmer 1992; Palmer et al. 2000). These properties, and the technical difficulties of laboratory assays, have limited the utility of plant mtDNA for molecular systematics (Knoop 2004) and population biology (Avise 2004). Nevertheless, RFLP markers have been used, for example, to demonstrate changes in the mtDNA molecule associated with restoration of fertility in cytoplasmic male-sterile maize (Schardl et al. 1985) and common bean (Johns et al. 1992) and determine the phylogenetic relationships and maternal parentage of natural interspecific hybrids in *Populus* (Barrett et al. 1993). Screenings of the geographical distribution of mtDNA haplotypes led to insights into the natural history of plants. For example, modern populations of Scots pine (*Pinus sylvestris*) are derived from dispersal from three different refugia following deglaciation (Sinclair et al. 1999). Olson and McCauley (2002) observed 13 mtDNA haplotypes among 250 individuals in 18 populations of bladder campion, *Silene vulgaris*, a flowering plant, within a 20-km region in western Virginia, and found that the populations were highly differentiated. Sex was determined by an interaction between cytoplasmic male sterility factors and autosomal male fertility restorers, with indications of population genetic structuring for the male fertility restorer genes.

Chloroplast DNA (cpDNA) exhibits its own unique molecular biology (Palmer 1985). It is transmitted maternally in some species (Birky 1978; Gillham 1978), biparentally in some (Metzlaff et al. 1981; Harris and Ingram 1991), and paternally in yet others (Chat et al. 1999), including in most gymnosperms (Wagner et al. 1987; Neale and Sederoff 1989). The circular molecule varies greatly in size, from 120 to 217 kb among photosynthetic land plants (Zurawski and Clegg 1987). The rate of molecular evolution is slow in terms of both gene order and nucleotide sequence (Palmer and Thompson 1981; Curtis and Clegg 1984), which makes cpDNA suitable for phylogenetic studies (Palmer and Zamir 1982; Clegg et al. 1986; Sytsma and Gottlieb 1986; Zurawski and Clegg 1987). CpDNA variation has been characterized in wild (e.g., barley (*Hordeum vulgare*), Clegg et al. 1984) and cultured (e.g., barley, Clegg et al. 1984; maize (*Zea mays*), Doebley et al. 1987) populations and has demonstrated interspecific hybridization in wild (e.g., pine, Wagner et al. 1987) and cultured (e.g., cotton, Wendel 1989) species. Among studies of phylogeographic variation, a consortium of laboratories (Petit et al. 2002) screened four PCR-amplified cpDNA fragments among 12,214 individuals from 2,613 European oak populations representing eight species. Six cpDNA

lineages were identified, with distinct geographic distributions along a longitudinal gradient reflecting patterns of colonization of the European landscape following deglaciation, a pattern corroborated and dated with fossil pollen evidence (Petit et al. 2002). RFLP analyses of mitochondrial and chloroplast DNA were common until around 1990, when direct sequencing of PCR amplicons become possible, which effectively replaced the whole-molecule, RFLP approach.

The complexity of nuclear DNA (e.g., three billion base pairs in human) is much greater than for mitochondrial DNA (16.6 kb), and Southern (1975) blotting using specific probes is needed to investigate RFLP variation of genomic DNA (gDNA). Probe hybridization patterns are interpreted to infer which bands represent restriction site alleles at a given locus. RFLPs were initially developed as markers for human diseases and disorders (e.g., β-thalassemia, Little et al. 1980; sickle-cell anemia, Phillips et al. 1980; Huntington's disease, Gusella et al. 1983) and subsequently extended to many nonhuman genomes, including livestock (e.g., cattle; prolactin, Camper et al. 1984; growth hormone, Beckmann et al. 1986) and crop plants (maize, Rivin et al. 1983; barley, Saghai-Maroof et al. 1984). Southern blot hybridization of the repeated sequence to *Eco*RI-restricted human DNA yielded numerous hybridization fragments which showed Mendelian inheritance and hypervariability. This multi-locus DNA fingerprinting approach found applications in forensics (Gill et al. 1985), breeding, population genetic, and other contexts. The advantage of the RFLP approach is that investigators can seek polymorphism at any genomic site for which there is a hybridization probe, and RFLP markers display codominant patterns. As early as 1980, Botstein et al. (1980) described a basis for using RFLP variation at random, single-copy loci to construct a genetic linkage map of the human genome. The approach was applied, for example, to map the genomes of several crop plants (including maize and tomato (*Solanum lycopersicum*), Helentjaris et al. 1986; Ritter et al. 1990), which are well suited for producing the requisite mapping populations. However, the disadvantage of the RFLP approach is that Southern blot hybridization is laborious and not well suited to the cost-effective, high-throughput genotyping required for many applications (Kashi et al. 1990).

3.2 PCR-Based Fingerprinting and Genotyping

Invention of the polymerase chain reaction (PCR) (Saiki et al. 1985, 1988; Mullis and Faloona 1987) revolutionized molecular biology, and subsequently organismal and population biology (Avise 2004), largely by stimulating new approaches to genetic marker screenings. A number of PCR-based genotyping methods fall under the general category of DNA fingerprinting.

A number of fingerprinting methods have been developed based on the amplification of random genomic DNA (gDNA) fragments using PCR primers of arbitrary sequence. The patterns generated depend on the sequence of the PCR primers and the nature of the template DNA. PCR is performed at low annealing

temperatures to allow the primers to anneal to multiple loci on the sample DNA. These PCR-based fingerprinting methods have the major disadvantage that they are very sensitive to reaction conditions, template DNA quality, PCR temperature profiles, and detection system, which limits their repeatability among laboratories and ultimately their range of utility. Williams et al. (1990) described one such procedure in which gDNA was PCR-amplified using single primers of arbitrary nucleotide sequence. The DNA segments that amplify are inherited in a Mendelian fashion from one or both parents. The polymorphisms so visualized are termed RAPD (random amplified polymorphic DNA) markers and have been used to discover variation in many species, including humans, corn (*Zea mays*), soybean (*Glycine max*), and *Neurospora* (Williams et al. 1990). While RAPD markers are an inexpensive, readily adapted method for assessing genetic variation in a yet-uncharacterized genome, they also pose several disadvantages. The sensitivity of the assay to reaction conditions leads to issues of repeatability of results among laboratories working with the same organism. Further, RAPD fragment patterns are expressed and interpreted as dominant genetic markers, which limits our ability to test for departures from Hardy-Weinberg equilibrium or to apply many classical population genetic tests. For these reasons, the RAPD approach is no longer widely used in population genetic studies.

The AFLP (amplified fragment length polymorphism) technique (Vos et al. 1995), based on the selective PCR amplification of restriction fragments from a total digest of gDNA, addresses some of the repeatability issues of RAPDs. The technique involves three steps: (1) restriction of the DNA and ligation of oligonucleotide adapters onto the restriction fragments, (2) selective amplification of subsets of the restriction fragments, and (3) electrophoretic analysis of the amplified fragments in a large polyacrylamide gel. Selective amplification is achieved by using primers that extend into the restriction fragments, amplifying only that subset of fragments in which the primer extensions match the nucleotides flanking the restriction sites. The key advantage of this method is that sets of restriction fragments may be visualized by PCR without previous knowledge of nucleotide sequence within the genome of interest. The method allows the specific co-amplification of high numbers of restriction fragments. The number of fragments that can be analyzed depends on the resolution of the detection system; typically, 50–100 restriction fragments are amplified and detected on denaturing polyacrylamide gels. The AFLP method has been applied primarily in studies of plants and microbes, with a strong bias toward economically important cultivated species and their pests. For example, AFLPs were widely used to construct single-tree genetic linkage maps in conifers by assaying haploid megagametophytes (Travis et al. 1998; Remington et al. 1999). Spooner et al. (2005) presented phylogenetic analyses of 261 wild and 98 landrace potatoes and three outgroup relatives, genotyped with 438 robust amplified fragment length polymorphisms. The AFLP data supported a monophyletic origin of the landrace potato cultivars from the northern component of the *Solanum brevicaule* complex in Peru, rather than from multiple independent origins from various northern and southern members. Cervera et al. (2005) applied AFLPs for examining genus-wide intraspecific and interspecific phylogenetic and

genetic relationships in *Populus*. Beismann et al. (1997) applied AFLP analysis to 26 individuals of white willow *Salix alba*, crack willow *S. fragilis*, and several individuals that were difficult to identify morphologically. Analysis of the AFLP data revealed distinct clusters corresponding to the nominal species and to interspecific hybrids. Kang et al. (2010) used AFLPs and other markers to construct high-density genetic maps in black spruce (*Picea mariana*) using a three-generation outbred pedigree and a black spruce x red spruce (*P. rubens*) hybrid using a BC_1 mapping population (Kang et al. 2011). AFLPs were widely used to construct single-tree genetic linkage maps in conifers by assaying haploid megagametophytes (e.g., Travis et al. 1998; Remington et al. 1999). Bensch and Åkesson (2005) identified a number of research areas where the AFLP method would be a valuable tool in the study of wild species of animals, including studies of population genetic structure and phylogenetic reconstructions, finding markers for genes governing adaptation, and the distribution of DNA methylation. However, with multiple technical steps, the procedure is prone to failure. Like RAPDs, AFLP bands are interpreted as dominant genetic markers. In addition, AFLPs reflect anonymous restriction sites, which are of less interest than markers within or linked to genes. Repeatability among laboratories is reliable only to the degree that electrophoretic conditions are standardized. AFLP markers were considered genome-wide markers before the development of genome-scale SNPs and genotyping-by-sequencing techniques.

3.3 Microsatellites

Due to the limitations of fingerprinting, alternative methods were sought that were both highly repeatable and enabled direct scoring of heterozygotes. The discovery of microsatellite sequences provided such an alternative. Microsatellites are tracts with tandem repeats of simple motifs of one to four nucleotides, first noted in the myoglobin gene, the zeta-globin pseudogene, the insulin gene, and the *X*-gene region of hepatitis B virus (Nakamura et al. 1987). Such tandem repeat tracts were subsequently found in all genomes and have been widely used as genetic markers. Different communities of geneticists have termed them STRs (short tandem repeats), SSRs (simple sequence repeats), or microsatellites, the term we use in this chapter. Microsatellite loci are PCR-amplified by using primers that anneal to unique genomic sequences flanking the tandem repeat tract. The amplification products may be visualized by standard electrophoresis or by labeling forward PCR primers with fluorescent dyes and using a sequencing instrument to score amplification products (Fig. 1). The latter allows precise estimation of the molecular weight of each DNA fragment, and the method is well suited to high-throughput genotyping. Microsatellite markers provide strong advantages for many applications. There is normally high variation at each locus, often ten or more alleles, providing great power for studies of population genetic variation, structure, and differentiation. Because most individuals are heterozygous at such loci,

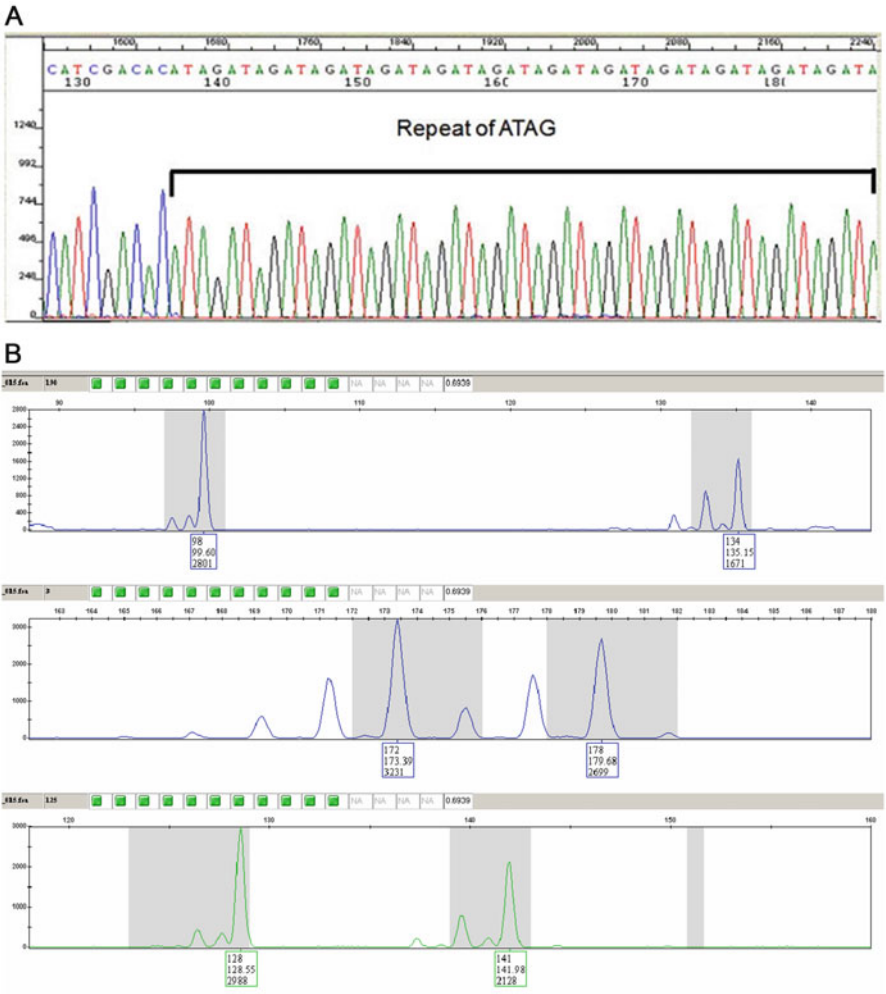

Fig. 1 Methodology for observing microsatellite DNA variation. (**a**) Scattered about the genome are regions bearing tandem repeats of simple nucleotide motifs, e.g., ATAG (NCTC 2014). (**b**) A particular microsatellite-bearing locus is PCR-amplified using primers specific to regions flanking the locus, and the amplification products are size-scored using a DNA sequencer (PMGF 2017)

microsatellite markers can provide linkage information in almost all families. There are many microsatellite loci potentially available as markers. The loci exhibit codominant expression of alleles, allowing use of a full range of data analysis approaches. Screening of microsatellite loci is economical once the primers for a locus have been developed. Several loci may be screened in one amplification mix or visualized in one lane in a DNA sequencer, an approach termed multiplexing (Chamberlain et al. 1988). Primer pairs used for amplifying microsatellite loci often are useful for screening the genome of related species of plants (Dayanandan et al.

1997; Peakall et al. 1998) and animals (Wilson et al. 1997). Microsatellite screening protocols are transferable among laboratories, especially if a few samples of known genotype are shared so that allele calls are standardized. Finally, variation at microsatellite loci is generally selectively neutral, hence appropriate for assessing the effects of population genetic processes such as migration and random genetic drift. It should be noted, however, that selection on microsatellite variation has been detected (Chhatre and Rajora 2014; Edelist et al. 2006), and care must therefore be taken to ensure that a given marker meets this assumption of neutrality. In humans, the number of SSR repeats has been found associated with disease conditions, including Crohn's disease (Hugot et al. 2001) and Behcet disease (Mizuki et al. 1997).

With such methodological strengths, microsatellite markers became the marker of choice for population genetics through the latter years of the twentieth century. They have been widely applied in studies of population structure in wild (Streiff et al. 1998) and cultivated (Morgante and Olivieri 1993; Eujayl et al. 2002; Ghislain et al. 2004) plants, as well as wild (Paetkau et al. 1995, 1998; Estoup et al. 1998; DeWoody and Avise 2000; King et al. 2001) and domesticated (Parker et al. 2004) animals. Microsatellite markers have been used to infer the origins (Vila et al. 2001) and to map the genomes (Bishop et al. 1994; Barendse et al. 1994) of domesticated species. Microsatellite-based inference of parentage and relatedness opened up studies of fitness and dispersal in wild populations (Blouin et al. 1996; Lawson Handley and Perrin 2007) and of seed dispersal by frugivores (Godoy and Jordano 2001). They have been used for noninvasive tracking of secretive (Kohn et al. 1999) or dangerous (Taberlet et al. 1997) animals by genotyping of sloughed host cells in feces. Microsatellites are also suited for genotyping of archived samples, such as fish scales (Nielsen et al. 1997), for use in forensic cases (Craft et al. 2007), and have been applied to determine the genetic impacts of forest harvest and management practices (Fageria and Rajora 2013; Rajora et al. 2000). The key disadvantage of microsatellite markers is the need to invest in the identification of microsatellite loci and development of useful primer pairs, although modern genomic sequencing technologies make their identification much easier (see next section). Microsatellite loci can have null alleles (Callen et al. 1993), i.e., alleles that do not amplify because a primer does not anneal to the sequence flanking the targeted microsatellite region; because this allele will not amplify during PCR, the individual will in error be regarded as a homozygote for the amplifying allele. While analytical protocols exist for identifying loci with null alleles (Van Oosterhout et al. 2004), the loss of data from such loci can limit the power of microsatellite-based studies. Expressed sequence tag and whole-genome and RNA sequencing have resulted in large numbers of candidate microsatellite loci, which can address this issue.

3.4 DNA Sequencing

The ability to determine the sequence of DNA opened up the entire genome for analysis. Originally, the target sequence had to be cloned but with the advent of PCR that was no longer necessary. Two methods have been available since the mid-1970s for sequencing target DNA. In the Maxam and Gilbert (1977, 1980) approach, the DNA was radioactively end-labeled and divided into four aliquots, which are treated with different chemical reagents that cleaved the DNA strand at base-specific positions. The fragments for all reactions are separated electrophoretically in a large polyacrylamide gel, visualized by autoradiography, and the DNA sequence is read directly from the ladder-like bands in the autoradiograph. Then the Sanger et al. (1977) method quickly became more widely used, which involves denaturing the DNA and dividing the mixture into four aliquots, each with a single dideoxynucleotide lacking the 3′ OH group needed for strand elongation. Strand elongation upon a particular template DNA molecule goes forward until a dideoxynucleotide becomes incorporated into the growing strand and then is arrested, resulting in different DNA molecules in the reaction mixture reaching different lengths before their elongation is terminated. The mixtures of fragments for the four different nucleotides are subjected to electrophoresis through an acrylamide gels, visualized through autoradiography, and the DNA sequence is read directly. Development of a fluorescent labeling technique enabling all four dideoxynucleotides to be identified in a single lane (Prober et al. 1987) led to the development of automated DNA sequencers. The system is based on the Sanger dideoxy chain termination method except that each dideoxynucleotide has a different fluorescein dye. The DNA fragments are resolved by polyacrylamide gel electrophoresis in one filament. Fluorescence is elicited by a laser and detected by a fluorescence detection system matched to the emission characteristics of the dye set. The output shows a sequence of fluorescence peaks with different colors for each nucleotide. Automation of DNA sequencing brought down its cost, opening the technique to cost-effective application to a wide range of issues and organisms and to a huge increase in DNA sequence information available. As of June 2017, approximately 200 million DNA sequences – over 231 billion nucleotides – have been archived in GenBank (http://www.ncbi.nlm.nih.gov/genbank/statistics/).

Early DNA sequencing efforts targeted organellar genomes due to their small sizes. Direct sequencing made screenings of mitochondrial DNA much more powerful than RFLP-based screenings, as much more information became available. PCR primers annealing to conserved sequences (Kocher et al. 1989; Meyer et al. 1990; Normark et al. 1991; Meyer 1993) enabled ready sequencing of selected mitochondrial regions with contrasting mutation rates. For example, after isolating gDNA from single plucked human hair, Vigilant et al. (1989) used mtDNA sequence variation to construct a genealogical tree relating Khoisan-speaking southern Africans to 68 other humans. Results were consistent with an African origin of human mtDNA and suggested that during hunter-gatherer times, female lineages moved their home bases very little. Certain regions of animal mtDNA

evolve at a rate suitable for phylogenetic inference. The mitochondrial sequences of cattle (Loftus et al. 1994) fell into two distinct geographic lineages – European and African breeds in one lineage and all Indian breeds are in the other – that did not correspond with the taurine-zebu dichotomy. The two major mtDNA clades diverged 200,000 to 1 million years ago, suggesting two separate domestication events, presumably of different subspecies of the aurochs, *Bos primigenius*. Lake Victoria and its satellite lakes harbor roughly 200 endemic forms of haplochromine cichlid fishes. After sequencing mitochondrial DNA from 14 representative Victorian species and 23 additional African species, Meyer et al. (1990) suggested a monophyletic origin for the haplochromines within the past million years. Mitochondrial DNA sequence data are well suited for application of a molecular phylogenetic approach to inference of natural history events and identification of conservation units. Screening sequence variation in the mitochondrial control region for 151 individuals representing 24 populations of European brown trout *Salmo trutta*, Bernatchez et al. (1992) observed monomorphism across all Atlantic basin populations and high inter-drainage diversity in more southerly populations, likely reflecting dispersal from different glacial refugia. In animals, mtDNA provides the basis for DNA barcoding (Hebert et al. 2013; Kress and Erickson 2012), in which the investigator sequences the cytochrome oxidase I subunit 3 gene and compares it against reference sequences in a taxonomic database (BOL 2016). Among many applications, Moran et al. (2015) used DNA barcoding to identify prey items in the stomach of invasive catfishes in eastern Virginia. While traditional morphological identification led to species-level identification of 65% of fish prey items, addition of DNA barcoding resulted in identification to species of 88% of fish prey items overall, including anadromous striped bass, herrings, and shads that are the focus of fishery restoration programs in these rivers.

The availability of consensus primers for amplifying genes and introns (Duminil et al. 2002) has eased screenings of plant mitochondrial DNA. For example, two polymorphic mitochondrial tandem repeats in the second intron of the *nad1* gene of Norway spruce (*Picea abies*) showed pronounced population genetic differentiation (Sperisen et al. 2001), with lineage A in north-northeastern and lineage B in Central and Southern Europe. Building on this work, Tollefsrud et al. (2008) used fossil pollen data and assessed variation in *nad1* among 4,876 trees in 369 populations. Observing 28 mitochondrial variants, patterns of population subdivision superimposed on interpolated fossil pollen distributions indicated that survival in separate refugia and postglacial colonization led to significant structuring of genetic variation in the southern range of the species. Shallow genetic structure consistent with the fossil pollen data suggested that the vast northern range was colonized from a single refugium. In the Alps, the diversity decreased over short distances, probably as a result of population bottlenecks caused by the presence of competing tree species. Increased genetic diversity north of the Carpathians probably resulted from admixture of expanding populations from two separate refugia.

Screening for variation in cpDNA has been facilitated by the development of universal PCR primers by Taberlet et al. (1991), Demesure et al. (1995), Dumolin-Lapegue et al. (1997), and Hamilton (1999). Often, selected chloroplast sequences

are amplified and characterized for restriction fragment length polymorphisms. Using such tools, Palmé et al. (2003) inferred the geographic patterns of postglacial recolonization of silver birch (*Betula pendula*), and Palmé et al. (2004) showed hybridization among the birches *Betula pendula*, *B. pubescens*, and *B. nana*. Similarly, Heuertz et al. (2004) showed the routes of postglacial recolonization of common ash *Fraxinus excelsior* in Europe. Cavers et al. (2003) explained the observed population structure in Spanish cedar *Cedrela odorata* in Central America as the result of repeated colonizations from South American source populations, first by a dry-adapted type and later by moist-adapted types. DNA barcoding of plants, which is based on comparing the sequence of the *trnL* gene of the chloroplast (Taberlet et al. 1991) between an unknown sample and a reference database, has proven useful for many species identification applications. Among them, Quéméré et al. (2013) used the approach to show that the golden-crowned sifaka, *Propithecus tattersalli*, an endangered lemur in Madagascar, exhibits remarkable dietary diversity, consuming at least 130 plant species belonging to 80 genera and 49 families, suggesting a high flexibility of foraging strategies.

4 SNP Genotyping Arrays

Direct PCR-based DNA sequencing opened the path for new approaches to genomic characterization, most notably for the discovery of single nucleotide polymorphisms (SNPs). SNPs are the most abundant and widespread type of polymorphism in both coding and noncoding regions, and they evolve in a manner well described by simple mutation models (Vignal et al. 2002). Prior to the availability of high-throughput sequencing methods (see below), SNPs emerged as the marker of choice for population genomic studies during the first decade of the twenty-first century, superseding microsatellites. Many factors caused this transition. While microsatellites provide high resolution for inference of neutral processes (migration, drift, inbreeding), their anonymous nature means that while some are surely under selection, the a priori expectation is that microsatellites behave neutrally. A multitude of enzymatic and detection methods were developed early in the SNP era (Kim and Misra 2007; Kwok 2001), but most were rather labor intensive and expensive per data point.

While initial efforts to study SNP variation were hampered by being relatively low-throughput and expensive, genotyping arrays changed this. The original method developed for humans and termed "variant detector arrays" (VDAs) involved fixing oligomeric probes ("oligos") to a glass surface (the "chip") (Wang et al. 1998). Similar to gene expression microarrays, VDA oligos were complementary to a target sequence but differed at a single site, which first was used to detect the presence of a SNP via changes in hybridization patterns when biotin-labeled samples were hybridized to the chip. However, it was the application of this method to known SNPs that would lead to a revolution in the field of population genomics. While the first report involved typing of only ~500 SNPs,

two companies – Affymetrix and Illumina – soon developed assays that approached the genomic scale for the first time. Current high-throughput SNP array platforms include standard panels for model species (human, *Arabidopsis*, various crop species) and can assay up to several million SNPs. Of more interest to population genomicists are custom solutions, which use the same chemistry as the standard chips but are designed and fabricated using SNP locations and flanking sequences for the particular species of interest. While the up-front cost of developing such an array is typically very high, when large numbers of samples are expected to be genotyped, the per-sample cost can be quite competitive with sequencing-based genotyping (see below). An additional advantage is that all assay steps are completed by a core facility as part of the overall cost of the genotyping effort. While next-generation library preparation can be completed at the sequencing core, this adds greatly to the cost of the project.

Early custom array-based genotyping solutions typically contained a few hundred to a few thousand SNPs. One widely used platform was Illumina GoldenGate (Fan et al. 2003; Shen et al. 2005), which used three oligos per SNP for allele discrimination: one locus-specific oligo (LSO) and two allele-specific oligos (ASOs). Each ASO carried one of the possible SNP alleles at their 3′ end. Following solution hybridization of the sample gDNA with these oligos, PCR amplification of the target loci was carried out, whereby amplification for a given sample/SNP proceeded using the ASO corresponding to the allele at that locus. The LSO contained an address sequence that enabled hybridization to the array and subsequent imaging and genotype calling. The GoldenGate assay was used extensively in population genomics to understand patterns of neutral and especially adaptive genetic diversity across species ranges (Eckert et al. 2009; Holliday et al. 2010; Loridon et al. 2013; Pavy et al. 2008). At the time, this platform had the advantage of being relatively inexpensive per data point, scalable, and much less laborious than previous methods. Recent developments in SNP array technology and application have focused on characterizing larger numbers of loci.

The two most widely used genome-scale methods for custom SNP genotyping are Illumina Infinium iSelect BeadChip (Illumina, Inc., San Diego, CA, USA) and Affymetrix Axiom (Affymetrix, Inc., Santa Clara, CA, USA). The former allows for up to 700,000 SNPs and relies on hybridization of fragmented DNA to a bead array, where each bead contains identical oligos that terminate one base prior to the expected SNP site. Allelic discrimination is achieved by single-base extension of the probes using fluorescent nucleotides with subsequent imaging and genotype calling. The Axiom platform allows up to 650,000 SNP targets and works on a similar principle. gDNA is fragmented and hybridized to oligomers on the array that end one base upstream of the SNP site. Instead of single-base extension, labeled probes complementary to the region including and downstream of the expected SNP are ligated to the array oligomer and hybridized to the cognate sample DNA. Two probe types correspond to the different expected alleles, each having a different fluorescent moiety, and allele discrimination is, therefore, achieved by respective hybridization of the probe or probes that are complementary to the sample DNA sequence. Numerous examples of the use of these two methods

have been reported for non-model species, with a bias toward the Illumina platform (Faivre-Rampant et al. 2016; Johnston et al. 2014; Lepoittevin et al. 2015; Malenfant et al. 2015; Pavy et al. 2013, 2016; Plomion et al. 2016; Yanez et al. 2016). Due in part to their economic importance, forest trees and commercially relevant fish species are overrepresented among these studies. Conversion rates vary widely and depend on the quality of the data used for SNP discovery and the stringency of filtering prior to array design. For example, a study in *Populus nigra* reported a >90% success rate for a 12k Infinium array (Faivre-Rampant et al. 2016), whereas a study in polar bears (*Ursus maritimus*) achieved a conversion rate of ~60% for a 9k array. As most candidate SNPs currently arise from high-throughput sequencing of discovery panels, understanding the parameters in those sequencing data that affect conversion is crucial. Goncalves da Silva et al. (2015) developed an Infinium array for orange roughy (*Hoplostethus atlanticus*) and found that standard SNP filtering metrics (e.g., depth of coverage) fail to address systematic sequencing errors. Rather, their data show that it is more important to filter for strand bias, where one allele is overrepresented among sequencing reads, polymorphism type (A/C and T/G polymorphisms had especially poor conversion rates), and the interaction between these two parameters.

5 High-Throughput Sequencing Methods

Genotyping arrays revolutionized our ability to score large numbers of variants in a cost-effective manner. These methods are still in use where data on a large, fixed panel of SNPs is desirable. For example, in populations with high levels of linkage disequilibrium (LD), such as agricultural breeding populations, generating dense, genome-wide data may be a waste of effort. On the other hand, for natural populations with low LD, and for which we lack the infrastructure and funding necessary to develop such arrays, the emergence of high-throughput, sequencing-based genotyping methods has enabled relatively inexpensive genome-scale studies. High-throughput sequencing began at the turn of the twenty-first century with the development of pyrosequencing (i.e., 454) technology (Ronaghi et al. 1996, 1998), followed by the sequencing by oligonucleotide ligation and detection (SOLiD) (McKernan et al. 2009) and Solexa (sequencing by synthesis; now Illumina) systems (Bentley et al. 2008). For a variety of reasons – cost, throughput, error rate, and run time – the Illumina platforms have captured much of the market at present (though see the section entitled "Emerging Sequencing and Genotyping Platforms"). We will, therefore, focus this section on sequence-based genotyping methodologies that make use of Illumina instruments/platforms.

5.1 General Features of Library Preparation

While there is a great diversity of DNA and RNA library preparation approaches, with the specific choice depending on the project goals, all share some common attributes. The first step, in most library preparation protocols, is fragmentation to achieve a desired insert size range, which may be achieved by physical or enzymatic means. The most common method for physical fragmentation is acoustic shearing with an ultrasonicator (e.g., instruments manufactured by Covaris, Inc.). While this method yields relatively consistent fragment pools and is not dependent on template sequence, it can be costly ($5–10 per sample) and require optimization. Recently, enzymatic fragmentation has gained traction as an alternative to shearing. This is most apparent in genotyping-by-sequencing (GBS) protocols in which the genomic DNA is digested with one or more restriction enzymes, giving rise to a heterogeneous pool of fragments that is subsequently size-selected (see below for more details). General fragmentation also may be achieved with restriction enzymes (e.g., NEB Fragmentase) or transposase-based systems (e.g., Illumina Nextera), with the latter introducing less bias (Picelli et al. 2014). Following fragmentation, library preparation involves ligation of oligonucleotide adapters to either end of a pool of DNA fragments. Sequencing adapters serve several functions: (1) they contain sequences complementary to oligos affixed to the Illumina flow cell, which enables their immobilization on the flow cell for sequencing; (2) they contain primers both for their amplification prior to sequencing (cluster generation) and for the sequencing reaction itself (Bentley et al. 2008); and (3) they frequently contain individual- or population-specific barcodes that allow for multiplexing within a single flow-cell lane (Fig. 2). More details on adapter design for specific applications are presented throughout the remainder of this section.

Following adapter ligation, the library is usually amplified by PCR, although PCR-free protocols do exist (Kozarewa et al. 2009). The number of PCR cycles at this stage should be limited, as each additional cycle introduces fragments that are exact duplicates of one another, which can bias SNP calling if one allele is preferentially amplified. For this reason, PCR duplicates are usually filtered out

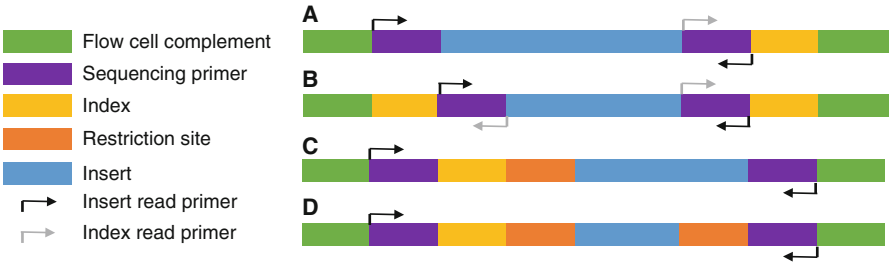

Fig. 2 Adapter configurations for (**a**) generic paired-end sequencing (e.g., WGS, RNA-Seq, sequence capture) with a single separate index reads, (**b**) paired-end sequencing with dual barcodes, (**c**) single-enzyme RAD-Seq, and (**d**) dual-enzyme genotyping-by-sequencing

computationally, but this filtering means wasted sequencing effort (only one copy of each duplicate read is usually retained). Moreover, PCR duplicates can only be identified when at least one end of gDNA fragments was generated randomly by shearing or by a non-specific enzyme (in which case, we expect that no two gDNA fragments will be identical, and when they exist in the data, they must have arisen due to the PCR step). When dual restriction enzymes are used to fragment the genome (known as two-enzyme GBS or double-digest RAD-Seq; see below), we expect multiple exact copies of each fragment to arise from the multiple copies of each chromosome in the gDNA extraction, and it is not possible to separate these natural duplicates from PCR duplicates.

At each step of library preparation, it is useful to check the fragment size range on a digital electrophoresis appliance (e.g., Agilent Bioanalyzer) and is essential prior to sequencing. Standard gel electrophoresis may be used as a "quick and dirty" means to assess the success of the ligation and PCR steps, but the higher resolution of a Bioanalyzer is recommended to estimate the library size range, to determine whether (and how much) adapter dimers may be present, and to assess the presence of high-molecular-weight fragments that can bias assessment of the molar concentration of the library. The latter will not interfere with sequencing, but their presence in the library may lead to underclustering on the flow cell and hence a reduced data yield. While the Bioanalyzer also can integrate the library concentration, fluorescence-based methods are generally preferred by sequencing centers (spectrophotometers are considered inaccurate for this purpose). If fragmentation and adapter ligation were optimal, as evidenced by a tight Bioanalyzer trace centered on the desired fragment size, the library should be ready for sequencing following an appropriate cleanup step, usually with paramagnetic beads. The rationale behind this approach is that beads are coated with carboxyl molecules, which bind DNA in the presence of polyethylene glycol (PEG). When placed on a magnetic stand, the supernatant-containing contaminants can be removed, leaving only the desired DNA fragments, which are then washed from the beads.

Often the Bioanalyzer reveals a suboptimal fragment size distribution – either too broad or containing adapter dimers. Shorter fragments, especially adapter dimers, will preferentially bind to the flow cell (most likely because of their increased mobility during flow-cell loading relative to larger fragments) and hence may be overrepresented in the sequence data. A library containing 10% adapter dimers will yield a disproportionate amount of useless data from these sequences. Libraries with a faint adapter dimer band yielded upward of 90% of reads from these sequences (J. Holliday, unpublished data). More generally, the Bioanalyzer trace also provides precise information on the size distribution of insert-containing fragments. If skewed toward smaller fragments, a large number of reads while yielding useable data will also contain adapter sequence, which means wasted sequencing resources. If the Bioanalyzer trace reveals one or more of the undesirable properties described above, size selection must be undertaken. It cannot be emphasized enough that any detectable adapter dimer band is unacceptable and will not only cause many reads to be discarded but may result in very little useable data. While the intuitive (and time-consuming) way to remove large or

small fragments is through gel extraction, on its own, this approach is unlikely to remove all of the problematic fragments, unless a polyacrylamide gel is used. One of the preferred methods is to size using the same magnetic beads used for general library purification, which can be achieved by adjusting the ratio of beads to DNA. Higher bead concentrations will capture both large and small fragments, whereas lower concentrations tend to favor larger fragments (due to the preferential electrostatic interaction between beads and larger DNA fragments that have larger total negative charges per molecule). The obvious advantage of this approach is that it can be used (at least in theory) to deplete the library of any adapter dimers revealed by the Bioanalyzer. However, in practice, it is not always effective. When other approaches have failed, size selection may be achieved by running the sample on an automated electrophoresis instrument (e.g., Pippin Prep, Sage Science, Inc.), which has the ability to target a narrow range of fragment sizes with much greater specificity than conventional gel extraction.

5.2 Library Strategies: Length, Sequencing Mode, and Multiplexing

The original Solexa sequencing strategy involved generating 35-bp reads from only one end of each fragment, with a single biological sample in a single lane. While many applications still rely on one or more of these parameters, it is more common to see some combination of longer reads, paired-end sequencing, and multiplexing in a single lane. We will consider each of these options in turn and their use for different applications. Current single-end read lengths for Illumina instruments vary between 36 and 300 bp depending on the instrument, with run times and cost scaling proportional to length. While it may appear obvious that longer is better, longer read length is more costly and may not be necessary. The 36-bp read length is clearly all that is needed for small RNA (21–24 bp in length) expression studies. For transcriptome studies in which counting transcripts (differential expression) is the objective, single-end 36-bp read lengths remain a cost-effective option. However, single-end reads do not provide information on alternative splicing and may not be sufficient to uniquely map transcripts arising from tandem or whole-genome duplication events. Paired-end sequencing (Korbel et al. 2007), in which sequences are sequentially read from each end of the insert, is the method of choice for gDNA and also frequently used for transcriptomics (Fullwood et al. 2009). Sequencing both ends of fragments that are typically 300–500 bp in length enables more precise read mapping (where a reference genome or transcriptome exists) and de novo assembly (where a reference is unavailable). Both the forward (read 1 or R1) and reverse (read 2 or R2) reads from each cluster on the flow cell are synchronized in the resulting data file, and this information can therefore be used to constrain their mapping/assembly. For example, if library insert size ranges from 300 to 500 bp, the software used for mapping/assembly will only allow a given read

pair to be placed within this approximate distance from one another. This strategy allows for efficient assembly of contiguous regions (e.g., genes) or even whole genomes (although for eukaryotes, shotgun sequencing is much more involved and beyond the scope of this chapter). Paired-end sequencing also enables mapping of splice junctions for transcriptome studies and can resolve structural genomic variation in some cases.

While some experiments, particularly whole-genome sequencing/re-sequencing, call for including only a single sample in each flow-cell lane, the high output currently available means that most population genomic experiments entail multiplexing. In a multiplexed run, each sample is prepared separately with one or two unique 4–8-bp barcodes (also called indexes) prior to pooling. Information from these barcode sequences enables bioinformatic demultiplexing. Barcode location depends on the type of experiment, with several options available. Library kits sold by Illumina and other manufacturers place the barcode upstream of the sequencing primer for R1, and these sequences are read in a separate index read. In recent iterations, there may be dual barcodes, one on each adapter, which enables more precise demultiplexing and fewer reads lost due to barcode sequencing errors. In some cases, particularly for custom GBS adapters, barcodes may be placed downstream of the sequencing primer, in which case the first 4–8 bp of R1 provide the sample information for demultiplexing. This approach has the disadvantage that a small amount of data on R1 is lost to sequencing of the barcode. Besides barcode position, an important consideration is the base composition of the barcodes. The Illumina imaging software tends to get confused by stretches with high GC content, in situations where many clusters are signaling the same base, and by homopolymer runs (the same base repeated multiple times). Commercial barcodes are designed with these considerations in mind, but if ever there is cause to design custom barcodes, these constraints must be accounted for – it is not as simple as generating a random list of 4-bp oligonucleotides. One consideration specific to GBS and inline barcodes is that the first sequence read after the barcode is the restriction site that generated the fragment. If the barcodes are all the same length, then every cluster will be reading the restriction site in the same sequencing cycle. To avoid this, it is advisable to use barcodes of heterogeneous length so that different clusters reach the restriction sites at different points in the run.

The number of samples that can be multiplexed in a single lane depends on the expected throughput of the sequencer. To determine an appropriate "plexity," the amount of data provided by the sequencing instrument and chemistry to be used is divided by the product of the desired coverage depth (e.g., 15 or 30X) and the cumulative length of the DNA fragments in the sample. For example, on an Illumina HiSeq 2,500 with version-four chemistry in high-output mode, approximately 200 million reads are expected per lane. For an RNA-Seq experiment, for example, in a species with a transcriptome comprised of 50,000 unique transcripts, with an average length 1,000 bp, we have 50,000,000 bp of unique sequence data to be gathered per sample. For a single-end 125 bp run, we expect approximately 25 billion bases of data (200,000,000 reads $\times$ 125 bp). We can therefore theoretically expect ~500X coverage of the transcriptome (25 billion divided by 50 million) if we

place a single sample in this lane. If we would like 30X coverage of our transcriptome per sample (a reasonable target), we could place 16 samples in a single lane (500/30). Of course, highly heterogeneous coverage depth, as a result of natural variation in the abundance of each transcript, is expected for transcriptome studies. It should be noted that for expression studies, we would not multiply the output of the sequencer by two for paired-end sequencing, because while R2 is useful for read mapping and splice variant detection, it does not give additional information about transcript abundance, since it is read from the same transcript as R1. For gDNA, we would multiply the output by two for paired-end sequencing, since the goal is to generate sufficient coverage for a given cohort of gDNA fragments (rather than to assess the abundance of those fragments). When estimating plexity levels, we usually discount the data output somewhat to allow for possible underclustering or a greater number of low-quality reads than expected. In the example above, we might conservatively multiplex 12 samples in a lane. This calculation works well for RNA-Seq because we generally know, or can make an educated guess, as to the cumulative length of the transcriptome. Whole-genome resequencing is similarly straightforward when we know the total size of the genome (which would generally also be the case). For GBS, which we discuss in detail below, the number of fragments expected can be somewhat nebulous. With a reference genome, we can scan for the relevant restriction sites and count the number of fragments that are expected to fall within our library size range. This would give an upper limit on, but would probably overestimate (perhaps substantially), the complexity of the sample due to variation in how efficiently the enzyme cuts in different areas of the genome or if a methylation-sensitive enzyme is used. Hence, it is often useful to do a test run to empirically estimate a reasonable level of multiplexing.

5.3 Genome Complexity Reduction

Many strategies have been developed in recent years to generate reduced representation libraries. These genome complexity reduction (GCR) methods fall into two general categories, those based on digestion of genomic DNA with restriction enzymes and those based on capture of desired gDNA fragments with synthetic baits. Each method has advantages and disadvantages, which are summarized in Table 1. In general, restriction enzyme-based methods – restriction site-associated DNA sequencing (RAD-Seq) and genotyping-by-sequencing (GBS) – are relatively inexpensive but do not allow for target selection and often result in much missing data, while sequence capture is more costly but allows for target selection and usually yields more complete datasets with less missing data. A summary of the laboratory workflows associated with each of these techniques is illustrated in Fig. 3.

Table 1 Summary of library preparation methods relevant to population genomics

Method	Cost	Advantages	Disadvantages
Whole genome re-sequencing	$$$ $	Full genome coverage; possibility to reconstruct full haplotype space	Costly, especially for large genomes; much greater computational resources needed
Sequence capture	$$$	Reliable coverage of target regions; typically recovers adjacent regions	High cost relative to enzyme-based methods
RAD-Seq	$$	Relatively inexpensive, though shearing increases per-sample cost relative to GBS	Missing data; inability to target particular areas of the genome
Genotyping by sequencing	$	Low cost	Missing data; inability to target particular areas of the genome
RNA-Seq	$$	Relatively inexpensive, though more costly than GBS if commercial kits used	Missing data due to differential abundance of transcripts; allele-specific expression may bias SNP calling
Genome skimming	$$	Inexpensive option for whole-genome coverage	Individual genotype calling may be inaccurate in wild species with low LD
Pool-Seq	$	Inexpensive option for whole-genome coverage; accurate allele frequency estimation for populations	Individual genotypes not resolved

5.3.1 Restriction Enzyme-Based Methods

Numerous approaches to GCR using restriction enzymes have been reported. The original method, RAD-Seq (Baird et al. 2008), involves single-enzyme digestion of gDNA followed by shearing and purification of a particular fragment size range from the digested DNA. The subset of the genome thus sequenced includes fragments of a particular size (usually 300–500 bp) that are flanked on one end by the enzyme cut site, with the other end the result of the random shearing process. The library is then further enriched using custom adapters that complement the restriction site sequence. The complexity of the fragment pool depends on the frequency with which the enzyme cuts, which is determined by the length of recognition site and other properties of the enzyme (methylation sensitivity, star activity). Enzymes with a 4-bp recognition sequence will cut much more frequently than 5-bp cutters. The pool may be enriched for euchromatic (i.e., genic) sequences through the use of a methylation-sensitive enzyme that does not cut highly methylated gDNA (i.e., the heterochromatin). This strategy is related to the AFLP method described in the previous section, with the obvious difference being that the fragments are sequenced rather than scored by size on a gel. Indeed, the similarity between the patented AFLP technology and GBS has led the United States patent office to grant sole rights to the GBS procedure to Keygene Inc. At

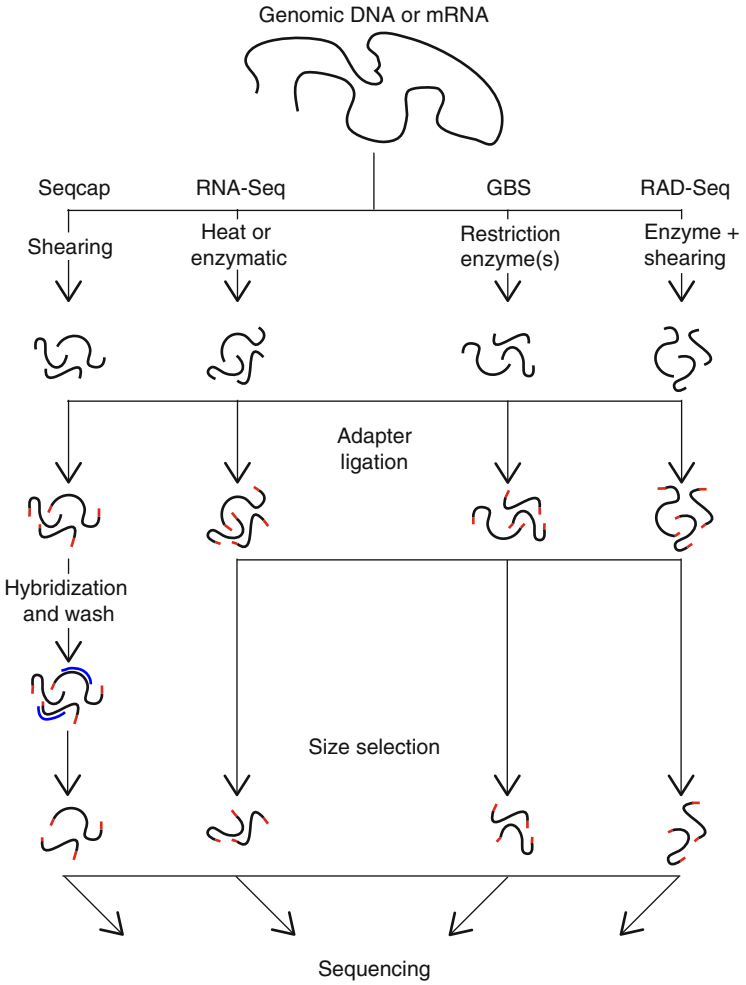

Fig. 3 Illustration of laboratory procedures involved in each genome complexity reduction technique discussed. Each method begins with some form of fragmentation in order to achieve the appropriate size range of DNA fragments for sequencing (usually 300–500 bp). In the case of GBS and RAD-Seq, fragmentation with one or two selected restriction enzymes also provides the means of reducing genome complexity when combined with size selection. Following fragmentation, all protocols involve adapter ligation, which may include multiplex indexes. In the case of sequence capture, there is one additional step – hybridization to synthetic baits and subsequent wash steps to remove unbound (nontarget) fragments. Finally, size selection is usually required to remove adapter dimers. In the case of GBS/RAD-Seq, size selection is also integral to the method, as it removes the many fragments of undesirable sizes, leaving only those of in the desired range that were flanked by one or more enzyme cut sites. Following size selection, visualization of the finished libraries is essential to ensure the correct size range has been achieved and there is no evidence of adapter dimers

present, the use of the GBS method requires a license from Keygene, though it is unclear if this applies to academic laboratories. GBS yields two types of polymorphisms. The first are presence/absence variants, in which the restriction site itself is polymorphic. These variants are dominant in the same way as AFLP fragment patterns are dominant. In addition, the sequenced gDNA fragments yield all of the usual types of sequence-based, codominant markers (SNPs, indels, etc.), which in many cases are the only variants used in downstream analyses.

Numerous elaborations of the RAD-Seq principle have been developed. The most widely used among these are GBS (Elshire et al. 2011) and double-digest RAD-Seq (ddRAD) (Peterson et al. 2012). GBS involves no shearing and only a single enzyme, which leads to lower-diversity libraries (i.e., only fragments of a certain size flanked by the enzyme site are sequenced) that may be desirable in some applications. An advantage of ddRAD over conventional RAD-Seq is that no shearing or end repair is required, which reduces library development costs. As noted above, multiplexing levels for restriction enzyme-based methods generally need to be determined empirically. Many investigators have had success placing up to 96 samples in a single flow-cell lane. The advantage of starting here is that useful data will be obtained even in the event that coverage depth is insufficient at this level of plexity. If such a test run indicates that fewer samples must be included per lane (e.g., 48), the original library may be sequenced a second time to provide an equivalent amount of data as two 48-sample runs, and the rest of the samples then can be processed as 48 plexes.

GBS and RAD-Seq have become a mainstay of population genomic studies across a wide variety of taxa, largely due to their flexibility and cost-effectiveness (Gagnaire et al. 2013; Pascoal et al. 2014; Rheindt et al. 2014; Sobel and Streisfeld 2015). However, these methods have some limitations. The random nature of the gDNA fragmentation process means that while a large number of variants may be genotyped (typically 5,000–100,000), many fragments will arise from intergenic regions not under selection. For species with reference genomes, or where there is a reference genome for a closely related species, the GBS fragments can be positioned, and their relationship to genic regions that may be under selection can be ascertained. However, GBS data frequently are generated for species that lack a reference genome, and de novo assembly of such data, while possible, means that the markers remain anonymous. In these situations, GBS may be better suited to questions about neutral than selective processes. An additional issue with enzyme-based methods is missing data, which may arise due to polymorphic restriction sites that yield a fragment in some samples but not others, as well as due to insufficient depth of coverage (Fig. 4). The latter can be overcome with more sequencing, but the former cannot. It is not uncommon to see 90% or more of SNPs called from a GBS dataset lost even with relatively liberal filters for missing data, quality, and depth of coverage filters, and few SNPs typically remain where a complete dataset is required. One way to overcome this limitation is by imputing missing data (Browning and Browning 2016; Scheet and Stephens 2006), which may or may not be successful depending on the haplotype structure of the population. Care must be taken when imputing data, because a complete dataset will be output from the relevant software, but the accuracy may be low and can only be ascertained

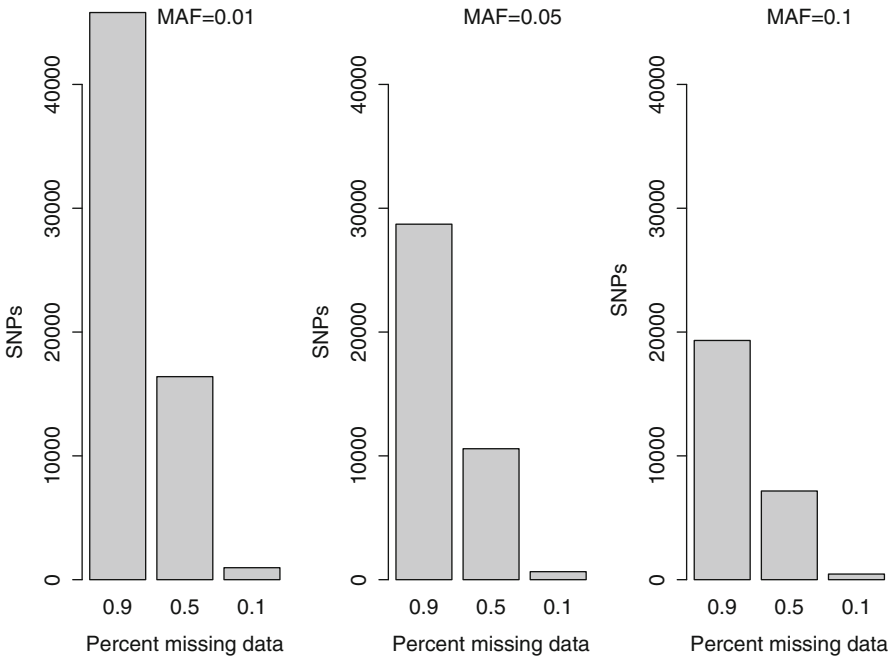

Fig. 4 Illustration of the missing data problem for genotyping-by-sequencing/RAD-Seq. In this experiment, libraries for 96 samples from an American chestnut (*Castanea dentata*) breeding program were prepared using two restriction enzymes, a common cutter (MspI) and a rare cutter (PstI). Shown are the number of SNPs retained at various minor allele frequency (MAF) and missing data cutoffs. For missing data, three thresholds were tested, corresponding to a maximum of 90, 50, and 10% missingness across samples for a given SNP (J. Holliday, unpublished data). Very few SNPs remained when the most stringent cutoff of <10% was used. The biological samples for this experiment were a multigenerational pedigree that arose from an initial hybridization event between American chestnut and Chinese chestnut (*Castanea mollissima*), which was aimed at introgressing alleles for resistance to *Cryphonectria parasitica* from Chinese chestnut. The progeny of the initial cross was backcrossed over three generations to American chestnut and subsequently intercrossed for two generations. The resulting high linkage disequilibrium in this pedigree enabled a relatively high degree of accuracy in the imputation of missing genotypes. This was assessed by randomly masking known genotypes and subsequently using Beagle software (Browning and Browning 2016) to impute these masked genotypes. In this case, accuracy of imputation was considered acceptable (~90%) for sites with <50% missing data. This illustrates that while missing data is a significant problem for GBS and related genotyping approaches, the genetic characteristics of the population can in some cases mitigate this issue

empirically, for example, by masking a subset of known genotypes, imputing them, and comparing the genotypes called by the software with those known from the sequencing run. A better way to assess the accuracy of imputation is to use a reference panel of known SNPs ascertained by other means (e.g., a SNP array) and to similarly compare the known and imputed genotypes (Li et al. 2009). While this is sometimes possible, such reference panels are not usually available for non-

model species of interest to population genomicists, at least not for the same SNPs assayed in the GBS experiment. In spite of these limitations, enzyme-based library preparation and sequencing methods are an attractive approach where funding is limited and can yield a reasonably complete dataset comprising thousands of SNP loci at a fraction of the cost of array-based genotyping.

5.3.2 Sequence Capture

Sequence capture is an alternative method for GCR that involves identifying regions of interest, synthesizing complementary oligonucleotide baits (usually 60–120 bp), and using these baits to retrieve the genomic intervals of interest through hybridization. While GBS can provide a reasonably complete dataset of a few hundred to a few thousand SNPs at a relatively low cost, sequence capture offers a number of advantages that make it the method of choice for GCR, in our view. The original sequence capture method involved immobilizing the baits on a glass slide, similar to early gene expression microarrays, hybridizing fragmented, adapter-ligated gDNA to those fixed baits, and subsequently eluting the captured fragments prior to sequencing (Hodges et al. 2007; Okou et al. 2007). The microarray method of capture was soon superseded by solution capture (Gnirke et al. 2009), which is more flexible in that array synthesis is not required. As all available platforms have moved toward solution capture, we hereafter focus on this technology.

The first step in a hybrid capture study is bait design. The nature and number of baits are determined by the research question and may include a few genes or genomic intervals (Nadeau et al. 2012), a broad selection of candidate genes (Hebert et al. 2013), or the entire gene space (known as the "exome") (Evans et al. 2014a; Suren et al. 2016; Zhou and Holliday 2012). The usual process is to identify regions of interest, which may be genomic intervals (where a reference genome is available) or a cohort of sequence files, either from shotgun genome or RNA sequencing. The primary companies that synthesize capture baits are Agilent (Santa Clara, CA, USA) and NimbleGen (Madison, WI, USA), and each uses proprietary software to determine exact bait positions relative to the sequence to be captured. These software packages provide a score based on the likelihood that capture will be successful. Because the capture hybridization is highly multiplexed, there is a trade-off between matching the melting point of the different baits as closely as possible and exclusion of desired targets on this basis. Hybridization conditions must be fairly relaxed to allow for this variation, but not so lenient as to encourage non-specific hybridization. Bait length factors into this specificity, and the two companies have different strategies in this regard. Agilent designs baits of a fixed length, 120 bp, whereas NimbleGen allows some variation in length to enable design for a greater number of target regions. The problem with the latter strategy is that short baits are more likely to hybridize with off-target regions of the genome (Kiialainen et al. 2011; Suren et al. 2016). Off-target hybridization is not a problem as long as the resulting data can be accurately mapped back to the reference

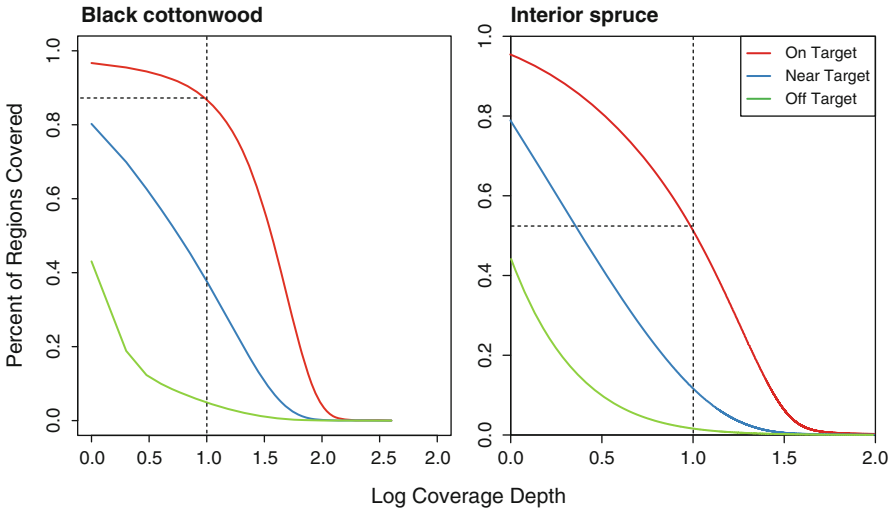

Fig. 5 Comparison of exome capture in black cottonwood (*Populus trichocarpa*) (Zhou and Holliday 2012) and interior spruce (*Picea engelmannii x glauca*) (Suren et al. 2016). Black cottonwood has a high quality and well-annotated reference genome (Tuskan et al. 2006), which was used to design baits around most exons and regions immediately upstream of genes. By contrast, at the time of bait design, no reference genome existed for interior spruce (though a draft genome has since been published (Birol et al. 2013)). Shown are cumulative distributions of sequencing depth for on-target data (arising from designed bait regions), near-target data (that mapped within a few hundred base pairs of a bait), and off-target data (not proximal to any designed bait). Dashed lines illustrate the percent of target regions covered at 10X sequencing depth and show much better recovery and sequencing of target regions for cottonwood than for spruce. Almost 90% of targeted bases are covered at 10X or greater depth in cottonwood, whereas for spruce only ~50% of targeted bases are covered at 10X or greater depth. Each of these experiments comprised nearly the full exome for the respective species, with 16 samples pooled in each lane of an Illumina HiSeq instrument. The black cottonwood example is adapted from data reported in Zhou and Holliday (2012), and the interior spruce example is adapted from Fig. 1 of Suren et al. 2016.

sequence, but it does introduce additional targets for sequencing that will affect mean coverage if not accounted for in the multiplexing strategy. Another issue that arises in bait design particular to species without quality reference genomes is the presence of unknown intron-exon boundaries. Neves et al. (2013) designed baits for capture in loblolly pine (*Pinus taeda*) prior to the completion of the reference genome and found that as baits became more centered on subsequently identified intron-exon boundaries, coverage decreased dramatically. While ignorance of these splice junctions does not make sequence capture impossible, it does result in wasted resources in terms of bait synthesis. The effects of designing an exome capture experiment for species with and without a high quality reference genome are illustrated in Fig. 5.

The issues of off-target hybridization and bait overlap with intron-exon boundaries illustrate another important point about sequence capture more generally:

mismatches are to some extent tolerated. This feature has several implications. First, paralogous genes (i.e., genes related by descent from a common ancestral DNA sequence) are likely to be simultaneously captured with baits designed for just one of the paralogs. Where tandem or whole-genome duplication has led to significant levels of paralogy in the study species, this fact needs to be considered in the multiplexing strategy. Off-target capture varies depending on the platform and species but should be assumed to comprise between 10% (for exome capture in species with well-curated reference genomes) and 50% or more (for species lacking a reference genome or with significant paralogy issues) of the completed library. The second way that mismatch tolerance manifests itself is in what we might call "near-target" capture. Near-target capture involves hybridization of a bait to a gDNA fragment, which arose through random shearing, comprised of both the target sequence and flanking sequence. Studies routinely find near-target capture up to ~200 bp from the nearest designed bait and in some cases more than that (Suren et al. 2016; Zhou and Holliday 2012). In the case of exome capture, for example, where baits are designed only for coding regions, near-target data may include introns, untranslated regions (UTRs), and regions up- or downstream of UTRs. How can a 120-bp bait capture gDNA >200 bp from the nearest target sequence? The random process of shearing yields diversity in the length of gDNA fragments as well as diversity in how much target sequence each contains. Shearing to mean fragment size of 300 bp will yield some fragments >500 bp. When one of those 500-bp fragments contains the 120-bp target sequence, it is possible that it will be captured in spite of its long, unpaired "tails." Although truly off-target capture may provide useful data, it is generally considered undesirable due to the sequencing effort that is used for regions that were not designed to be part of the study. However, near-target capture can prove an advantage. If the goal is to capture ecologically relevant variation, which may reside in regulatory regions of introns, UTRs, or promoters, capturing those regions without specifically targeting them reduces the overall cost of the study (i.e., as the number of baits required increases, so does the cost). We specifically took advantage of this outcome in revisions to an exome capture bait design that first was used in a test for a few dozen samples and subsequently for several hundred. Based on the results from the first cohort, we strategically left gaps between baits of between 100 and 200 bp under the assumption that those regions ultimately would be captured in spite of not having been specifically targeted by baits, which worked well (Holliday et al. 2016; Zhou and Holliday 2012).

The final consideration in the context of mismatch tolerance lies at the interspecific level. Baits designed for one species are not solely useful for that species but also for related species. This fact has been leveraged extensively in the realm of phylogenomics, where baits for relatively conserved genes have been used for capture in multiple species, even across relatively deep evolutionary divergences. While such studies require a careful selection of a limited number of slowly evolving regions of the nuclear or, more often, organellar genomes, capture of more diverse targets can be successfully achieved in closely related species. For example, a study in the spruce and pine genera yielded reasonably complete

datasets in spite of several million years of intrageneric divergence (Suren et al. 2016). Alternatively, where the species of interest does not have a reference genome, but a congener does, baits may be designed for the congener and used in capture of gDNA for the focal species. Congeneric exome capture works reasonably well for divergence times up to a few million years at least, and there is often very little difference in the efficiency of capture between the species for which the baits were designed and congeners that are closely related.

With an appropriate bait design and synthesis completed, sequence capture library preparation involves many of the same steps as other methods, including fragmentation of gDNA, ligation of adapters, and amplification. Of course, the primary difference is the hybridization itself, which usually involves incubation in a thermocycler at an appropriate annealing temperature (usually ~65°C) for several hours. One key consideration in this context is whether to pool different samples before or after the capture hybridization step. The original solution-based hybridization protocols called for pooling after capture, which obviously increases the sample-handling burden. This approach also gave the bait provider greater control over pricing per sample. Today, pre-capture pooling is common. In the case of Agilent, pricing is still on a per-sample basis but assumes pooling of 12–16 samples in a single hybridization. Other providers leave the degree of multiplexing up to individual labs, although they may provide guidance. We have had good results pooling up to 16 samples in each exome capture (Zhou et al. 2014), which also happens to be a reasonable number to multiplex in a single flow-cell lane for exome re-sequencing. However, much greater levels of multiplexing are possible. For example, a study in humans pooled 96 samples in a single capture when a small number of gDNA targets were used (Neiman et al. 2012). The success of such high levels of multiplexing depends on careful quantification of the individual samples prior to pooling such that the bait:target ratio remains constant across samples. Pooling such a large number of samples into a full exome capture would seem to be advantageous in terms of cost even though the resulting library would need to be sequenced several times (since a single lane would not yield sufficient coverage). However, it is likely that competition/interference between baits and targets in such a complex hybridization enforces limits on the number of samples that can be pooled prior to capture. Moreover, the physical number of baits in a given aliquot of the bait library may be limiting, although to our knowledge it is not possible to calculate this stoichiometry (i.e., the information is proprietary). We are not aware of any systematic studies that have investigated the limits of multiplexing for exome capture across the different platforms, though it is likely to exceed the 16-sample "rule of thumb" noted above. Such a study would have tremendous practical value.

5.3.3 RNA Sequencing as "Natural" Genome Complexity Reduction

Given the trade-offs for GBS and sequence capture noted above, one wonders that RNA-Seq is not more widely used in population genomics as a way to gather

sequence data from the areas of the genome of most interest, the gene space. While RNA-Seq has some of its own issues from the perspective of exome-wide variant discovery, it is an underutilized tool in this context (De Wit et al. 2012). The obvious advantage of sequencing cDNA is that it provides information on coding regions of the genome without the need for synthesis of expensive baits and the associated hybridization step. Baits are expensive, and while commercial RNA-Seq library kits are also costly, the component reagents lend themselves to design of custom kits that lower the price significantly. The main difference between sequencing gDNA compared with RNA is the necessity to deplete the latter of highly expressed ribosomal RNA. Fortunately, mRNAs have a natural molecular tag, the poly-A tail, which has been used for decades as a means to separate mRNAs from ribosomes. Poly-A isolation tends to be the most expensive step in custom RNA-Seq protocols. The current approach typically employed involves oligo-dT-bound magnetic beads, which are available from several providers (e.g., Dynabeads from Thermo Fisher and Oligo $d(T)_{25}$ Magnetic Beads from New England Biolabs). Because mRNA is labile, fragmentation can be achieved simply by the application of heat (85–95°C) for a few minutes, which can be folded into the reverse transcription reaction to produce cDNA (Hou et al. 2015). With custom-synthesized adapters, coupled with the above methods, it is quite possible to put together an RNA-Seq protocol that costs ~$20 per sample (perhaps even less). By contrast, the list price of the NEXTflex RNA-Seq Kit (Bioo Scientific; one of the best values currently) is more double than that. Although RNA-Seq for variant genotyping has the advantage over GBS in that it targets the gene space and an advantage over sequence capture in that it is less expensive, there are limitations. The most important of these is the heterogeneity of the transcriptome. Natural abundance varies among transcripts over several orders of magnitude, and the composition of the transcriptome varies across tissues and over time. For these reasons, it is advisable to capture the widest variety of tissues and conditions possible where the goal is exome-wide coverage. In spite of this, some transcripts will comprise a much greater proportion of the library than others, which presents two issues. First, highly abundant transcripts will be disproportionately represented on the flow cell, which means wasted sequencing effort. On the other hand, some transcripts will be present at such low levels that they may be sequenced at a depth insufficient for variant calling. Nevertheless, each method for GCR carries trade-offs, and RNA-Seq is probably underutilized in population genomics to gain genotype data on coding regions of transcripts that are reasonably abundant and whose abundance is reasonably consistent across genetic backgrounds.

5.4 Whole-Genome Sequencing and Re-sequencing

The GCR methods described above represent the vast majority of population genomic studies reported to date and reflect the relatively high cost of simply sequencing the entire genome of a focal population. While whole-genome

sequencing (WGS) has been used in a few cases for population genomics (Evans et al. 2014b; Jones et al. 2012; Soria-Carrasco et al. 2014), it is unlikely to become common, at least at high coverage depth, using sequencing platforms currently available. While the cost per base pair of sequence data has dropped precipitously since the emergence of the Solexa technology, recent iterations of sequencing chemistries and instruments have flattened this curve somewhat. As such, we do not see the price of whole-genome sequencing becoming comparable with that of genome complexity-reduced libraries, although this calculus depends on the size of the genome under study. Many species of ecological interest have genome sizes that exceed 1 Gb, and for such species WGS is currently vastly more expensive than GBS and somewhat more expensive than sequence capture. On the other hand, if one is fortunate enough to be interested in a species with a genome size on the order of a few hundred Mb, WGS is a viable option (still much more expensive than GBS and RNA-Seq, comparable to exome capture when done according to manufacturer specifications, although see above). For species that fall in the >1 Gb category, or where true genome-wide data is desired but funds are limited, other options exist. These generally fall into two categories: genome skimming and pooled sequencing. Genome skimming involves individually barcoding samples and sequencing them to low average depth (e.g., 1X), which yields a very incomplete dataset (an average depth of 1X leaves many bases unsequenced in each sample) (Straub et al. 2012). However, such missing data can be overcome using advanced probabilistic bioinformatics approaches (Buerkle and Gompert 2013), particularly in populations with high linkage disequilibrium. The other option, pooled sequencing, involves forgoing individual barcoding. Instead, the goal is to sequence entire populations to a reasonable depth (Schlotterer et al. 2014). In this case, individual genotypes are not resolved, but accurate population allele frequencies may be obtained. As population allele frequencies are the basis for many downstream analyses, this approach is of use where the population is the ecological unit of interest. This approach has been applied to a number of systems (Christe et al. 2016; Fabian et al. 2012; Fischer et al. 2013; Kofler et al. 2012). For example, a study in white poplar (*Populus alba*) and European aspen (*Populus tremula*) scored approximately eight million SNPs from pools of 24 samples and used these data to infer demographic history (Christe et al. 2016). Importantly, results from the demographic models were similar to those inferred from RAD-Seq data, which showed that the Pool-Seq method is robust. Theory and software tools are emerging to handle Pool-Seq data (Boitard et al. 2012; Kofler et al. 2011, 2016a, b), and we expect this method see extensive use in the coming years.

6 Emerging Sequencing and Genotyping Platforms

The introduction of commercially available high-throughput sequencing platforms in 2005 led to a boom in the amount of sequence data and drastically reduced costs. Concordant with the precipitous decline in per-base sequencing costs is an

exponential increase in the ability to conduct population genomic studies in model and non-model systems. For example, the Illumina platform provides an unparalleled ability to generate informative markers from natural populations with little to no preexisting genetic/genomic information, at very low cost. Along with these advances, however, come new challenges: the shorter read lengths (35–700 bp) limit inference by restricting analyses to smaller variants and also introduce biases in the sequence data (Benjamini and Speed 2012). Additionally, de novo genome assemblies are often more fragmented and gapped than those generated using older approaches (Lee et al. 2016). Thus, the biggest innovations in sequencing technology are platforms/techniques that cost-effectively generate reads in the range of 10–100 kb, enabling the discovery of novel variants, improving the accuracy of sequence capture, and expanding the ability to uncover variation in non-model systems.

6.1 Illumina

Illumina offers a wide range of sequencing by synthesis (SBS) applications with the throughput and turnaround time tuned to applications. Sequencing across these systems proceeds through clonal amplification of DNA fragments containing adaptor sequences and reversible dye termination (for a detailed review, see (Bentley et al. 2008). The available suite of sequencers, MiniSeq, MiSeq, NextSeq, HiSeq, and NovaSeq (released in 2017), can generate from 7.5 to 6,000 Gb of data in as few as 4 h to 4 days, respectively. The benchtop sequencers MiniSeq and MiSeq range in data output (7.5–15 Gb) and read length (150 and 300 bp); these systems are best suited to small genomes (microbe, virus) and targeted sequencing (including microbiome 16S sequencing on the MiSeq). The other benchtop system, NextSeq, produces up to 120 Gb in 30 h and is particularly well suited to exome sequencing and whole-transcriptome sequencing (from simple experimental designs). Perhaps the most well-known system, HiSeq, is also the most versatile offering a rapid run mode on the 2,500 where ca 50 Gb can be generated from a two-lane flow cell in about 7 h, providing a great cost/turnaround time balance for genotyping-by-sequencing applications. For re-sequencing, pooled sequencing, RNA-Seq, or large (to very large)-scale reduced representation designs, the HiSeq 4,000 can generate up to 1.5 Tb of data from an eight-lane flow cell in just under 4 days, and the emerging system NovaSeq promises to deliver even more ca. 6 Tb in just 3 days. In addition to these systems, in 2012, Illumina acquired a synthetic long-read technology, Moleculo – now called TruSeq Synthetic Long Read. In this library preparation approach, fragments up to 10 kb are sheared, cloned, and uniquely barcoded for short-read sequencing. The short reads are assembled into synthetic long reads or fragments based on barcode clustering. The diversity of systems, input flexibility, throughput, and very low error rate make Illumina the most cost-effective and widely used sequencing platform (Reuter et al. 2015).

6.2 *Pacific Biosciences*

The single molecule real-time (SMRT) sequencing strategy was one of the first long-read platforms commercialized (Pacific Biosciences) and exploits a strand displacing polymerase to sequence the same molecule multiple times generating a clone-free, circular consensus sequence (Travers et al. 2010). This approach improves accuracy and reduces biases associated with cloning-based approaches (e.g., GC bias in Illumina-derived data); thus PacBio error is random. The heart of this platform is the SMRT cell which contains the 1×10–21 L (zeptoliter) zero-mode waveguide (ZMW) wells (Goodwin et al. 2016). On the RSII system, each SMRT cell contains 150,000 wells generating ca. 1 Gb per cell, whereas cells for the recently introduced Sequel system contain 1,000,000 wells generating ca. 8 Gb per cell. Both systems produce mean read lengths that are typically >14 kb, but read length is a log normal distribution such that there are few very long reads (up to 100 kbp) and many reads ≪14 kb. This read distribution and a high error rate relative to Illumina (Berlin et al. 2015) require relatively high coverage for accurate genotyping (>25X). As more Sequel systems become available, the cost of generating sufficient coverage with only PacBio reads is falling; however, hybrid approaches using Illumina short reads for error correction are quite cost-effective (Lee et al. 2016). The power of using long reads to uncover important variation was recently demonstrated via exome sequencing of uncharacterized regions of the pine genome (Neves et al. 2013). In this study, long reads were used to improve the accuracy of de novo assemblies for the targeted exome sequences, resulting in better capture of full-length regions, and reduce complexity resulting from high levels of heterozygosity.

6.3 *Oxford Nanopore*

Since introducing the MinION in early 2014, Oxford Nanopore Technologies has emerged as a leader in commercializing nanopore sequencing. Like the SMRT cells for PacBio, nanopore sequencing relies on cells with hundreds of microscopic wells; only at the center of these wells are synthetic bilayers with enmeshed biologic pores (Wang et al. 1998). Also like PacBio, the reads are "single molecule" so a distribution of read lengths is generated from 6 kb up to >60 kb and error rates are high (reported rates range from 4 to 15%). The MinION system is a very compact unit (about the size of an eyeglass case) with a USB adaptor that connects to a laptop making the unit extremely portable; however throughput remains low at ca. 5–10 Gb flow cell. This low throughput, combined with high run failure rate and high error rate, has limited adoption of this platform. The development of a benchtop system, PromethION, which is a cluster of up to 48 flow cells, can generate 240–480 Gb, in a footprint no larger than a business class desktop. The small footprint, library simplicity, read lengths, and speed of data acquisition

suggest that this platform will find a niche. Indeed, the portability of this unit was instrumental in determining population structure among Ebola strains during an outbreak (Quick et al. 2016).

6.4 10X Genomics

A synthetic long-read system like TruSeq from Illumina, 10X Genomics, adds an encapsulating system (GEM) and retains very large fragments (100 kb) within a micelle for barcoding. This approach however requires an additional microfluidic device in addition to the underlying sequencing (for a review, see Goodwin et al. 2016). While this approach generates long reads with very high accuracy, it is with a higher cost, one that is commensurate with that of true long reads at sufficient depth for similar accuracy (Lee et al. 2016). This system has already been used in a hybrid approach combining with a true long-read platform for error correction and novel variant discovery (Mostovoy et al. 2016). A distinct advantage of this system is the potential for cell sorting and single-cell sequencing, particularly single-cell RNA-Seq via the Chromium™ Single Cell Controller, enabling the comparison of populations of cells.

7 Future Perspectives

With an expanding repertoire of sequencing platforms and the precipitous decline in per-base costs (exceeding Moore's law), the ability to generate sufficient data for any given population genomic question is quickly becoming trivial. Whether the protocol requires identification of structural variants, sequencing 1,000 whole genomes to a depth of 10X, or a nimble sequencer that can be used in remote areas, there is a platform available. The ability to combine platforms or tailor inputs to specific needs further amplifies this flexibility. In spite of the increase in throughput with each new chemistry and platform, the per-base cost of data generated on Illumina systems has flattened somewhat in recent years. The advent of technologies that deliver more data at lower costs (e.g., NovaSeq, Sequel, and PromethION) brings the promise of a second revolution in sequencing/genotyping. Nevertheless, the goal of characterizing genome diversity "telomere to telomere" (Shendure et al. 2017) remains elusive, particularly for species with complex genomes containing abundant repetitive elements. Emerging long-read sequencers (Nanopore, PacBio) have begun to address the assembly problem associated with short reads, and we expect future technological developments to further advance this objective. This increased competition in the sequencing market should allow for generation of comprehensive genomic datasets for non-model species, which comprise the vast majority of species of interest in population genomics, at a depth and quality once reserved for model systems.

References

Avise JC. Molecular markers, natural history, and evolution. 2nd ed. Sunderland: Sinauer Associates; 2004.

Backert S, Nielsen BL, Börner T. The mystery of the rings: structure and replication of mitochondrial genomes from higher plants. Trends Plant Sci. 1997;2:477–83.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 2008;3:e3376.

Barendse W, Armitage SM, Kossarek LM, Shalom A, Kirkpatrick BW, Ryan AM, Clayton D, Li L, Neibergs HL, Zhang N, Grosse WM. A genetic linkage map of the bovine genome. Nat Genet. 1994;6:227–35.

Barrett JW, Rajora OP, Yeh FCH, Dancik BP, Strobeck C. Mitochondrial-DNA variation and genetic-relationships of Populus species. Genome. 1993;36:87–93.

Beckmann JS, Kashi Y, Hallerman EM, Nave A, Soller M. Restriction fragment length polymorphism among Israeli Holstein-Friesian dairy bulls. Anim Genet. 1986;17:25–38.

Beismann H, Barker JH, Karp A, Speck T. AFLP analysis sheds light on distribution of two Salix species and their hybrid along a natural gradient. Mol Ecol. 1997;6:989–93.

Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40:e72.

Bensch S, Åkesson M. Ten years of AFLP in ecology and evolution: why so few animals? Mol Ecol. 2005;14:2899–914.

Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53–9.

Berlin K, Koren S, Chin CS, et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015;33:623.

Bernatchez L, Guyomard R, Bonhomme F. DNA sequence variation of the mitochondrial control region among geographically and morphologically remote European brown trout Salmo trutta populations. Mol Ecol. 1992;1:161–73.

Birky CW Jr. Transmission genetics of mitochondria and chloroplasts. Annu Rev Genet. 1978;12:471–512.

Birky CW Jr. Evolution and variation in plant chloroplast and mitochondrial genomes. In: Gottlieb L, editor. Plant evolutionary biology. Netherlands: Springer; 1988. p. 23–53.

Birol I, Raymond A, Jackman SD, et al. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. Bioinformatics. 2013;29:1492–7.

Bishop MD, Kappes SM, Keele JW, Stone RT, Sunden SL, Hawkins GA, Toldo SS, Fries R, Grosz MD, Yoo J. A genetic linkage map for cattle. Genetics. 1994;136:619–39.

Blouin MS, Parsons M, Lacaille V, Lotz S. Use of microsatellite loci to classify individuals by relatedness. Mol Ecol. 1996;5:393–401.

Boitard S, Schlotterer C, Nolte V, Pandey RV, Futschik A. Detecting selective sweeps from pooled next-generation sequencing samples. Mol Biol Evol. 2012;29:2177–86.

BOL. 2016. http://www.barcodeoflife.org

Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980;32:314–31.

Browning BL, Browning SR. Genotype imputation with millions of reference samples. Am J Hum Genet. 2016;98:116–26.

Buchert GP, Rajora OP, Hood JV, Dancik BP. Effects of harvesting on genetic diversity in old growth eastern white pine in Ontario, Canada. Conserv Biol. 1997;11:747–58.

Buerkle CA, Gompert Z. Population genomics based on low coverage sequencing: how low should we go? Mol Ecol. 2013;22:3028–35.

Buth DG. Genetic principles and the interpretation of electrophoretic data. In: Whitmore DH, editor. Electrophoretic and isoelectric focusing techniques in fishery management. Boca Raton: CRC Press; 1990. p. 1–22.

Callen DF, Thompson AD, Shen Y, Phillips HA, Richards RI, Mulley JC, Sutherland GR. Incidence and origin of "null" alleles in the (AC)n microsatellite markers. Am J Hum Genet. 1993;52:922–7.

Camper SA, Luck DN, Yao Y, Woychik RP, Goodwin RG, Lyons RH Jr, Rottman FM. Characterization of the bovine prolactin gene. DNA. 1984;3:237–49.

Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. Nature. 1987;325:31–6.

Cavers S, Navarro C, Lowe AJ. Chloroplast DNA phylogeography reveals colonization history of a Neotropical tree, Cedrela odorata L., in Mesoamerica. Mol Ecol. 2003;12:1451–60.

Cervera MT, Storme V, Soto A, Ivens B, Van Montagu M, Rajora OP, Boerjan W. Intraspecific and interspecific genetic and phylogenetic relationships in the genus *Populus* based on AFLP markers. Theor Appl Genet. 2005;111:1440–56.

Chamberlain JS, Gibbs RA, Ranier JE, Nguyen PN, Caskey CT. Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. Nucleic Acids Res. 1988;16:11141–56.

Chat J, Chalak L, Petit RJ. Strict paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in intraspecific crosses of kiwifruit. Theor Appl Genet. 1999;99:314–22.

Chhatre VE, Rajora OP. Genetic divergence and signatures of natural selection in marginal populations of a keystone, long-lived conifer, eastern white pine (*Pinus strobus*) from northern Ontario. PLoS One. 2014;9:e97291.

Christe C, Stölting KN, Paris M, et al. Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. Mol Ecol. 2016;26:59. https://doi.org/10.1111/mec.13765.

Clegg MT, Brown AH, Whitfeld PR. Chloroplast DNA diversity in wild and cultivated barley: implications for genetic conservation. Genet Res. 1984;43:339–43.

Clegg MT, Ritland K, Zurawski G. Processes of chloroplast DNA evolution. In: Karlin S, Nevo E, editors. Evolutionary processes and theory. New York: Academic Press; 1986. p. 275–94.

Craft KJ, Owens JD, Ashley MV. Application of plant DNA markers in forensic botany: genetic comparison of Quercus evidence leaves to crime scene trees using microsatellites. Forensic Sci Int. 2007;165:64–70.

Curtis SE, Clegg MT. Molecular evolution of chloroplast DNA sequences. Mol Biol Evol. 1984;1:291–301.

Dayanandan S, Bawa KS, Kesseli R. Conservation of microsatellites among tropical trees (Leguminosae). Am J Bot. 1997;84:1658–63.

De Wit P, Pespeni MH, Ladner JT, et al. The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. Mol Ecol Resour. 2012;12:1058–67.

Demesure B, Sodzi N, Petit RJ. A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. Mol Ecol. 1995;4:129–31.

DeWoody JA, Avise JC. Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. J Fish Biol. 2000;56:461–73.

Doebley J, Renfroe W, Blanton A. Restriction site variation in the zea chloroplast genome. Genetics. 1987;117:139–47.

Dowling TE, Moritz C, Palmer JD, Rieseberg LH. Nucleic acids III: analysis of fragments and restriction sites. In: Hillis DM, Moritz C, Mable BK, editors. Molecular systematics. 2nd ed. Sunderland: Sinauer Associates; 1996. p. 249–320.

Duminil J, Pemonge MH, Petit RJ. A set of 35 consensus primer pairs amplifying genes and introns of plant mitochondrial DNA. Mol Ecol Notes. 2002;2:428–30.

Dumolin-Lapegue S, Pemonge M-H, Petit RJ. An enlarged set of consensus primers for the study of organelle DNA in plants. Mol Ecol. 1997;6:393–7.

Eckert AJ, Pande B, Ersoz ES, et al. High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). Tree Genet Genomes. 2009;5:225–34.

Edelist C, Lexer C, Dillmann C, Sicard D, Rieseberg LH. Microsatellite signature of ecological selection for salt tolerance in a wild sunflower hybrid species, *Helianthus paradoxus*. Mol Ecol. 2006;15:4623–34.

Elshire RJ, Glaubitz JC, Sun Q, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6:e19379.

Estoup A, Rousset F, Michalakis Y, Cornuet JM, Adriamanga M, Guyomard R. Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (Salmo trutta). Mol Ecol. 1998;7:339–53.

Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W. Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theor Appl Genet. 2002;104:399–407.

Evans J, Kim J, Childs KL, et al. Nucleotide polymorphism and copy number variant detection using exome capture and next-generation sequencing in the polyploid grass *Panicum virgatum*. Plant J. 2014a;79:993–1008.

Evans LM, Slavov GT, Rodgers-Melnick E, et al. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. Nat Genet. 2014b;46:1089–96.

Fabian DK, Kapun M, Nolte V, et al. Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. Mol Ecol. 2012;21:4748–69.

Fageria MS, Rajora OP. Effects of harvesting of increasing intensities on genetic diversity and population structure of white spruce. Evol Appl. 2013;6:778–94.

Faivre-Rampant P, Zaina G, Jorge V, et al. New resources for genetic studies in Populus nigra: genome-wide SNP discovery and development of a 12k Infinium array. Mol Ecol Resour. 2016;16:1023–36.

Fan JB, Oliphant A, Shen R, et al. Highly parallel SNP genotyping. Cold Spring Harb Symp Quant Biol. 2003;68:69–78.

Fischer MC, Rellstab C, Tedder A, et al. Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. Mol Ecol. 2013;22:5594–607.

Fullwood MJ, Wei C-L, Liu ET, Ruan Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res. 2009;19:521–32.

Gagnaire P-A, Pavey SA, Normandeau E, Bernatchez L. The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. Evolution. 2013;67:2483–97.

Ghislain M, Spooner DM, Rodriguez F, et al. Selection of highly informative and user-friendly microsatellites (SSRs) for genotyping of cultivated potato. Theor Appl Genet. 2004;108:881–90.

Gill P, Jeffreys AJ, Werrett DJ. Forensic application of DNA 'fingerprints'. Nature. 1985;318:577–9.

Gillham NW. Organelle heredity. New York: Raven Press; 1978.

Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009;27:182–9.

Godoy JA, Jordano P. Seed dispersal by animals: exact identification of source trees with endocarp DNA microsatellites. Mol Ecol. 2001;10:2275–83.

Goncalves da Silva A, Barendse W, Kijas JW, et al. SNP discovery in nonmodel organisms: strand bias and base-substitution errors reduce conversion rates. Mol Ecol Resour. 2015;15:723–36.

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguichi AY, Young AB, Shoulson I, Bonilla E, Martin JB. A polymorphic DNA marker linked to Huntington's disease. Nature. 1983;306:238–44.

Hamilton MB. Four primer pairs for the amplification of chloroplast intergenic regions with intraspecific variation. Mol Ecol. 1999;8:521–2.
Harris SA, Ingram R. Chloroplast DNA and biosystematics: the effects of intraspecific diversity and plastid transmission. Taxon. 1991;1:393–412.
Hebert FO, Renaut S, Bernatchez L. Targeted sequence capture and resequencing implies a predominant role of regulatory regions in the divergence of a sympatric lake whitefish species pair (*Coregonus clupeaformis*). Mol Ecol. 2013;22:4896–914.
Helentjaris T, Slocum M, Wright S, Schaefer A, Nienhuis J. Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphisms. Theor Appl Genet. 1986;72:761–9.
Heuertz M, Fineschi S, Anzidei M, Pastorelli R, Salvini D, Paule L, Frascaria-Lacoste N, Hardy OJ, Vekemans X, Vendramin GG. Chloroplast DNA variation and postglacial recolonization of common ash (Fraxinus excelsior L.) in Europe. Mol Ecol. 2004;13:3437–52.
Hodges E, Xuan Z, Balija V, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet. 2007;39:1522–7.
Hoelzel AR. Molecular genetic analysis of populations: a practical approach. Oxford: IRL Press; 1992.
Holliday JA, Ritland K, Aitken SN. Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (*Picea sitchensis*). New Phytol. 2010;188:501–14.
Holliday JA, Zhou L, Bawa R, Zhang M, Oubida RW. Evidence for extensive parallelism but divergent genomic architecture of adaptation along altitudinal and latitudinal gradients in *Populus trichocarpa*. New Phytol. 2016;209:1240. https://doi.org/10.1111/nph.13643.
Hou ZG, Jiang P, Swanson SA, et al. A cost-effective RNA sequencing protocol for large-scale gene expression studies. Sci Rep. 2015;5:9570.
Hugot JP, Chamaillard M, Zouali H, Lesage S. Association of *NOD2* leucine-rich repeat variants with susceptibility to Crohn's disease. Nature. 2001;411:599.
Johns C, Lu M, Lyznik A, Mackenzie S. A mitochondrial DNA sequence is associated with abnormal pollen development in cytoplasmic male sterile bean plants. Plant Cell. 1992;4:435–49.
Johnston SE, Orell P, Pritchard VL, et al. Genome-wide SNP analysis reveals a genetic basis for sea-age variation in a wild population of Atlantic salmon (*Salmo salar*). Mol Ecol. 2014;23:3452–68.
Jones FC, Grabherr MG, Chan YF, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484:55–61.
Kang BY, Mann IK, Major JE, Rajora OP. Near-saturated and complete genetic linkage map of black spruce (*Picea mariana*). BMC Genomics. 2010;24:515.
Kang BY, Major JE, Rajora OP. A high-density genetic linkage map of a black spruce (*Picea mariana*) × red spruce (*Picea rubens*) interspecific hybrid. Genome. 2011;54:128–43.
Kashi Y, Hallerman E, Soller M. Marker-assisted selection of candidate bulls for progeny testing programmes. Anim Prod. 1990;51:63–74.
Kessler C. Class II restriction endonucleases. In: Obe G, Basler A, editors. Cytogenetics. Berlin: Springer Verlag; 1987. p. 225–79.
Kiialainen A, Karlberg O, Ahlford A, et al. Performance of microarray and liquid based capture methods for target enrichment for massively parallel sequencing and SNP discovery. PLoS One. 2011;6:e16486.
Kim S, Misra A. SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng. 2007;9:289–320.
Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press; 1983.
King TL, Kalinowski ST, Schill WB, Spidle AP, Lubinski BA. Population structure of Atlantic salmon (Salmo salar L.): a range-wide perspective from microsatellite DNA variation. Mol Ecol. 2001;10:807–21.

Knoop V. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr Genet. 2004;46:123–39.
Kocher TD, Thomas WK, Meyer A, Edwards SV, Pääbo S, Villablanca FX, Wilson AC. Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proc Natl Acad Sci USA. 1989;86:6196–200.
Kofler R, Pandey RV, Schlotterer C. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics. 2011;27:3435–6.
Kofler R, Betancourt AJ, Schlotterer C. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. PLoS Genet. 2012;8:e1002487.
Kofler R, Gomez-Sanchez D, Schlotterer C. PoPoolationTE2: comparative population genomics of transposable elements using pool-seq. Mol Biol Evol. 2016a;33:2759–64.
Kofler R, Langmuller AM, Nouhaud P, Otte KA, Schlotterer C. Suitability of different mapping algorithms for genome-wide polymorphism scans with pool-seq data. G3 Genes Genomes Genet. 2016b;6:3507–15.
Kohn MH, York EC, Kamradt DA, Haught G, Sauvajot RM, Wayne RK. Estimating population size by genotyping faeces. Proc R Soc Lond B Biol Sci. 1999;266:657–63.
Korbel JO, Urban AE, Affourtit JP, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007;318:420–6.
Kozarewa I, Ning Z, Quail MA, et al. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G plus C)-biased genomes. Nat Methods. 2009;6:291–5.
Kress WJ, Erickson DL. DNA barcodes: methods and protocols. Methods Mol Biol. 2012;858:3–8.
Kwok PY. Methods for genotyping single nucleotide polymorphisms. Annu Rev Genomics Hum Genet. 2001;2:235–58.
Lansman RA, Shade RO, Shapira JF, Avise JC. The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. III. Techniques and potential applications. J Mol Evol. 1981;17:214–26.
Lansman RA, Avise JC, Aquadro CF, Shapira JF, Daniel SW. Extensive genetic variation in mitochondrial DNAs among geographic populations of the deer mouse, Peromyscus maniculatus. Evolution. 1983;37:1–16.
Lawson Handley LJ, Perrin N. Advances in our understanding of mammalian sex-biased dispersal. Mol Ecol. 2007;16:1559–78.
Lee H, Gurtowski J, Yoo S, et al. Third-generation sequencing and the future of genomics. bioRxiv. 2016. https://doi.org/10.1101/048603.
Lepoittevin C, Bodenes C, Chancerel E, et al. Single-nucleotide polymorphism discovery and validation in high-density SNP array for genetic analysis in European white oaks. Mol Ecol Resour. 2015;15:1446–59.
Lewontin RC, Hubby JT. A molecular approach to the study of genic heterozygosity in natural populations: amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics. 1966;54:595–609.
Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387–406.
Little P, Annison G, Darling S, Williamson R, Cambar T, Model B. Model for antenatal diagnosis of b-thalassemia and other monogenic disorders by molecular analysis of linked DNA polymorphisms. Nature. 1980;285:144–7.
Loftus RT, MacHugh DE, Bradley DG, Sharp PM, Cunningham P. Evidence for two independent domestications of cattle. Proc Natl Acad Sci USA. 1994;91:2757–61.
Lonsdale DM, Brears T, Hodge TP, Melville SE, Rottmann WH. The plant mitochondrial genome: homologous recombination as a mechanism for generating heterogeneity. Philos Trans R Soc B. 1988;319:149–63.

Loridon K, Burgarella C, Chantret N, et al. Single-nucleotide polymorphism discovery and diversity in the model legume *Medicago truncatula*. Mol Ecol Resour. 2013;13:84–95.
Malenfant RM, Coltman DW, Davis CS. Design of a 9K illumina beadchip for polar bears (*Ursus maritimus*) from RAD and transcriptome sequencing. Mol Ecol Resour. 2015;15:587–600.
Maxam AM, Gilbert W. New method for sequencing DNA. Proc Natl Acad Sci USA. 1977;74:560–4.
Maxam AM, Gilbert W. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. 1980;65:499–559.
May B. Starch gel electrophoresis of allozymes. In: Hoelzel AR, editor. Molecular genetic analysis of populations: a practical approach. 2nd ed. New York: Oxford University Press; 1998. p 1–28 and 371–378.
McKernan KJ, Peckham HE, Costa GL, et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. 2009;19:1527–41.
Metzlaff M, Börner T, Hagemann R. Variations of chloroplast DNAs in the genus Pelargonium and their biparental inheritance. Theor Appl Genet. 1981;60:37–41.
Meyer A. Molecular phylogenetic studies of fishes. In: Beaumont AR, editor. Evolution and genetics of aquatic organisms. New York: Chapman and Hall; 1993.
Meyer A, Kocher TD, Basasibwaki P, Wilson AC. Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial DNA sequences. Nature. 1990;347:550–3.
Mizuki N, Ota M, Kimura M, Ohno S, Ando H, Katsuyama Y, Yamazaki M, Watanabe K, Goto K, Nakamura S, Bahram S. Triplet repeat polymorphism in the transmembrane region of the *MICA* gene: a strong association of six GCT repetitions with Behcet disease. Proc Natl Acad Sci U S A. 1997;94:1298–303.
Moran Z, Orth DJ, Schmitt JD, Hallerman EM, Aguilar R. Effectiveness of DNA barcoding for identifying piscine prey items in stomach contents of piscivorous catfishes. Environ Biol Fish. 2015;99:161–7.
Morgante M, Olivieri AM. PCR-amplified microsatellites as markers in plant genetics. Plant J. 1993;3:175–82.
Morizot DC, Schmidt ME. Starch gel electrophoresis and histochemical visualization of proteins. In: Whitmore DH, editor. Electrophoretic and isoelectric focusing techniques in fishery management. Boca Raton: CRC Press; 1990. p. 23–80.
Mostovoy Y, Levy-Sakin M, Lam J, et al. A hybrid approach for de novo human genome sequence assembly and phasing. Nat Methods. 2016;13:587.
Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol. 1987;155:335–50.
Murphy RW, Sites JW, Buth DG, Haufler CH. Isozyme electrophoresis. In: Hillis DM, Moritz C, Mable BK, editors. Molecular systematics. 2nd ed. Sunderland: Sinauer Associates; 1996. p. 51–120.
Nadeau NJ, Whibley A, Jones RT, et al. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. Philos Trans R Soc B Biol Sci. 2012;367:343–53.
Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E. Variable number of tandem repeat (VNTR) markers for human gene mapping. Science. 1987;235:1616–22.
National Conservation Training Center (NCTC) 2017. https://nctc.fws.gov/courses/csp/csp3157/content/terms/microsatellite.html
Neale DB, Sederoff RR. Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in loblolly pine. Theor Appl Genet. 1989;77:212–6.
Neiman MR, Sundling S, Groenberg H, et al. Library preparation and multiplex capture for massive parallel sequencing applications made efficient and easy. PLoS One. 2012;7:e48616.
Neves LG, Davis JM, Barbazuk WB, Kirst M. Whole-exome targeted sequencing of the uncharacterized pine genome. Plant J. 2013;75:146–56.

Nielsen EE, Hansen MM, Loeschcke V. Analysis of microsatellite DNA from old scale samples of Atlantic salmon Salmo salar: a comparison of genetic composition over 60 years. Mol Ecol. 1997;6:487–92.

Normark BB, McCune AR, Harrison RG. Phylogenetic relationships of neopterygian fishes, inferred from mitochondrial DNA sequences. Mol Biol Evol. 1991;8:819–34.

Okou DT, Steinberg KM, Middle C, et al. Microarray-based genomic selection for high-throughput resequencing. Nat Methods. 2007;4:907–9.

Olson MS, McCauley DE. Mitochondrial DNA diversity, population structure, and gender association in the gynodioecious plant Silene vulgaris. Evolution. 2002;56:253–62.

Paetkau D, Calvert W, Stirling I, Strobeck C. Microsatellite analysis of population structure in Canadian polar bears. Mol Ecol. 1995;4:347–54.

Paetkau D, Waits LP, Clarkson PL, Craighead L, Vyse E, Ward R, Strobeck C. Variation in genetic diversity across the range of North American brown bears. Conserv Biol. 1998;12:418–29.

Palmé AE, Su Q, Rautenberg A, Manni F, Lascoux M. Postglacial recolonization and cpDNA variation of silver birch, Betula pendula. Mol Ecol. 2003;12(2):201–12.

Palmé AE, Su Q, Palsson S, Lascoux M. Extensive sharing of chloroplast haplotypes among European birches indicates hybridization among Betula pendula, B. pubescens and B. nana. Mol Ecol. 2004;13(1):167–78.

Palmer JD. Evolution of chloroplast and mitochondrial DNA in plants and algae. In: McIntyre RJ, editor. Molecular evolutionary genetics. New York: Plenum Press; 1985. p. 131–240.

Palmer JD. Mitochondrial DNA in plant systematics: applications and limitations. In: Soltis PS, Soltis DE, Doyle JJ, editors. Molecular systematics of plants. New York: Springer; 1992. p. 36–49.

Palmer JD, Herbon LA. Tricircular mitochondrial genomes of Brassica and Raphanus: reversal of repeat configurations by inversion. Nucleic Acids Res. 1986;14:9755–64.

Palmer JD, Herbon LA. Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. J Mol Evol. 1988;28:87–97.

Palmer JD, Thompson WF. Rearrangements in the chloroplast genomes of mung bean and pea. Proc Natl Acad Sci USA. 1981;78:5533–7.

Palmer JD, Zamir D. Chloroplast DNA evolution and phylogenetic relationships in Lycopersicon. Proc Natl Acad Sci USA. 1982;79:5006–10.

Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu YL, Song K. Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc Natl Acad Sci USA. 2000;97:6960–6.

Parker HG, Kim LV, Sutter NB, Carlson S, Lorentzen TD, Malek TB, Johnson GS, DeFrance HB, Ostrander EA, Kruglyak L. Genetic structure of the purebred domestic dog. Science. 2004;304:1160–4.

Pascoal S, Cezard T, Eik-Nes A, et al. Rapid convergent evolution in wild crickets. Curr Biol. 2014;24:1369–74.

Pavy N, Pelgas B, Beauseigle S, et al. Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. BMC Genomics. 2008;9:21.

Pavy N, Gagnon F, Rigault P, et al. Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. Mol Ecol Resour. 2013;13:324–36.

Pavy N, Gagnon F, Deschenes A, et al. Development of highly reliable in silico SNP resource and genotyping assay from exome capture and sequencing: an example from black spruce (*Picea mariana*). Mol Ecol Resour. 2016;16:588–98.

Peakall R, Gilmore S, Keys W, Morgante M, Rafalski A. Cross-species amplification of soybean (Glycine max) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. Mol Biol Evol. 1998;15:1275–87.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One. 2012;7:e37135.

Petit RJ, Csaikl UM, Bordács S, Burg K, Coart E, Cottrell J, van Dam B, Deans JD, Dumolin-Lapègue S, Fineschi S, Finkeldey R. Chloroplast DNA variation in European white oaks: phylogeography and patterns of diversity based on data from over 2600 populations. For Ecol Manag. 2002;156:5–26.

Phillips J, Panny S, Kazazian H, Bochun C, Scott C, Smith R. Prenatal diagnosis of sickle cell anemia by restriction endonuclease analysis: hindIII polymorphisms in v-globin genes extend applicability. Proc Natl Acad Sci USA. 1980;77:2853–6.

Picelli S, Bjorklund AK, Reinius B, et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 2014;24:2033–40.

Plant-Microbe Genomics Facility (PMGF). 2017. https://pmgf.osu.edu/services/genotyping/example

Plomion C, Bartholome J, Lesur I, et al. High-density SNP assay development for genetic analysis in maritime pine (*Pinus pinaster*). Mol Ecol Resour. 2016;16:574–87.

Pring DR, Lonsdale DM. Molecular biology of higher plant mitochondrial DNA. Int Rev Cytol. 1985;97:1–46.

Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science. 1987;238:336–41.

Quéméré E, Hibert F, Miquel C, Lhuillier E, Rasolondraibe E, Champeau J, Rabarivola C, Nusbaumer L, Chatelain C, Gautier L, Ranirison P. A DNA metabarcoding study of a primate dietary diversity and plasticity across its entire fragmented range. PLoS One. 2013;8:e58971.

Quick J, Loman NJ, Duraffour S, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530:228.

Rajora OP, Rahman MH, Buchert GP, Dancik BP. Microsatellite DNA analysis of genetic effects of harvesting in old-growth eastern white pine (*Pinus strobus*) in Ontario, Canada. Mol Ecol. 2000;9:339–48.

Remington DL, Whetten RW, Liu BH, O'Malley DM. Construction of an AFLP genetic map with nearly complete genome coverage in *Pinus taeda*. Theor Appl Genet. 1999;98:1279–92.

Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015;58:586–97.

Rheindt FE, Fujita MK, Wilton PR, Edwards SV. Introgression and phenotypic assimilation in Zimmerius flycatchers (Tyrannidae): population genetic and phylogenetic inferences from genome-wide SNPs. Syst Biol. 2014;63:134–52.

Ritter E, Gebhardt C, Salamini F. Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. Genetics. 1990;125:645–54.

Rivin CJ, Zimmer EA, Cullis CA, Walbot V, Huynh T, Davis RW. Evaluation of genomic variability at the nucleic acid level. Plant Mol Biol Report. 1983;1:9–16.

Roberts JR. Restriction and modification enzymes and their recognition sequences. Nucleic Acids Res. 1984;12:R167–204.

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem. 1996;242:84–9.

Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. Science. 1998;281:363.

Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW. Ribosomal DNA spacer length polymorphism in barley: mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci USA. 1984;81:8014–8.

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. 1985;230:1350–4.

Saiki RK, Gelfand DH, Stoffel S, Scharf ST, Higuchi R, Horn GT, Mullis KB, Ehrlich HA. Primer-directed enzymatic amplification of DNA. Science. 1988;239:487–91.
Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 1977;74:5463–7.
Schardl CL, Pring DR, Lonsdale DM. Mitochondrial DNA rearrangements associated with fertile revertants of S-type male-sterile maize. Cell. 1985;43:361–8.
Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006;78:629–44.
Schlotterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. Nat Rev Genet. 2014;15:749–63.
Shen R, Fan JB, Campbell D, et al. High-throughput SNP genotyping on universal bead arrays. Mutat Res Fundam Mol Mech Mutagen. 2005;573:70–82.
Shendure J, Balasubramanian S, Church GM, et al. DNA sequencing at 40: past, present and future. Nature. 2017;550:345. https://doi.org/10.1038/nature24286. Advance Online Publication.
Sinclair WT, Morman JD, Ennos RA. The postglacial history of Scots pine (Pinus sylvestris L.) in western Europe: evidence from mitochondrial DNA variation. Mol Ecol. 1999;8:83–8.
Sobel JM, Streisfeld MA. Strong premating reproductive isolation drives incipient speciation in Mimulus aurantiacus. Evolution. 2015;69:447–61.
Soria-Carrasco V, Gompert Z, Comeault AA, et al. Stick insect genomes reveal natural Selection's role in parallel speciation. Science. 2014;344:738–42.
Southern EM. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol. 1975;98:503–17.
Sperisen C, Büchler U, Gugerli F, Mátyás G, Geburek T, Vendramin GG. Tandem repeats in plant mitochondrial genomes: application to the analysis of population differentiation in the conifer Norway spruce. Mol Ecol. 2001;10:257–63.
Spooner DM, McLean K, Ramsay G, Waugh R, Bryan GJ. A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. Proc Natl Acad Sci U S A. 2005;102:14694–9.
Straub SCK, Parks M, Weitemier K, et al. Navigating the tip of the genomic icebeg: next-generation sequencing for plant systematics. Am J Bot. 2012;99:349–64.
Streiff RE, Labbe TH, Bacilieri RO, Steinkellner HE, Glossl JO, Kremer AN. Within-population genetic structure in Quercus robur L. and Quercus petraea (Matt.) Liebl. assessed with isozymes and microsatellites. Mol Ecol. 1998;7:317–28.
Suren H, Hodgins KA, Yeaman S, et al. Exome capture from the spruce and pine giga-genomes. Mol Ecol Resour. 2016;16:1136–46.
Sytsma KJ, Gottlieb LD. Chloroplast DNA evolution and phylogenetic relationships in Clarkia Sect. peripetasma (Onagraceae). Evolution. 1986;40:1248–61.
Taberlet P, Gielly L, Pautou G, Bouvet J. Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Mol Biol. 1991;17:1105–9.
Taberlet P, Camarra JJ, Griffin S, Uhres E, Hanotte O, Waits LP, Dubois-Paganon C, Burke T, Bouvet J. Noninvasive genetic tracking of the endangered Pyrenean brown bear population. Mol Ecol. 1997;6:869–76.
Tollefsrud MM, Kissling R, Gugerli F, Johnsen Ø, Skrøppa T, Cheddadi R, Van der Knaap WO, Latałowa M, TerHürne-Berson RU, Litt T, Geburek T. Genetic consequences of glacial survival and postglacial colonization in Norway spruce: combined analysis of mitochondrial DNA and fossil pollen. Mol Ecol. 2008;17:4134–50.
Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. 2010;38:e159.
Travis SE, Ritland K, Whitham TG, Keim P. A genetic linkage map of Pinyon pine (*Pinus edulis*) based on amplified fragment length polymorphisms. Theor Appl Genet. 1998;97:871–80.

Tuskan GA, DiFazio S, Jansson S, et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science. 2006;313:1596–604.
Van Oosterhout C, Hutchinson WF, Wills DP, Shipley P. Micro-Checker: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes. 2004;4:535–8.
Vigilant L, Pennington R, Harpending H, Kocher TD, Wilson AC. Mitochondrial DNA sequences in single hairs from a southern African population. Proc Natl Acad Sci USA. 1989;86:9350–4.
Vignal A, Milan D, San Cristobal M, Eggen A. A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol. 2002;34:275–305.
Vila C, Leonard JA, Gotherstrom A, et al. Widespread origins of domestic horse lineages. Science. 2001;291:474–7.
Vos P, Hogers R, Bleeker M, Reijans M, Van de Lee T, Hornes M, Friters A, Pot J, Paleman J, Kuiper M, Zabeau M. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 1995;23:4407–14.
Wagner DB, Furnier GR, Saghai-Maroof MA, Williams SM, Dancik BP, Allard RW. Chloroplast DNA polymorphisms in lodgepole and jack pines and their hybrids. Proc Natl Acad Sci USA. 1987;84:2097–100.
Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science. 1998;280:1077–82.
Ward BL, Anderson RS, Bendich AJ. The mitochondrial genome is large and variable in a family of plants (Cucurbitaceae). Cell. 1981;25:793–803.
Wendel JF. New World tetraploid cottons contain Old World cytoplasm. Proc Natl Acad Sci USA. 1989;86:4132–6.
Williams JGK, Kubelik AR, Livak KJ, Rafalski SA, Tingey SV. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. 1990;18:6531–5.
Wilson GA, Strobeck C, Wu L, Coffin JW. Characterization of microsatellite loci in caribou Rangifer tarandus, and their use in other artiodactyls. Mol Ecol. 1997;6:697–9.
Yanez JM, Naswa S, Lopez ME, et al. Genomewide single nucleotide polymorphism discovery in Atlantic salmon (*Salmo salar*): validation in wild and farmed American and European populations. Mol Ecol Resour. 2016;16:1002–11.
Zhou L, Holliday JA. Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. BMC Genomics. 2012;13:703.
Zhou L, Bawa R, Holliday JA. Exome resequencing reveals signatures of demographic and adaptive processes across the genome and range of black cottonwood (*Populus trichocarpa*). Mol Ecol. 2014;23:2486–99.
Zurawski G, Clegg MT. Evolution of higher-plant chloroplast DNA-encoded genes: implications for structure-function and phylogenetic studies. Annu Rev Plant Physiol. 1987;38:391–418.

Computational Tools for Population Genomics

Jarkko Salojärvi

Abstract With the rapidly dropping costs of sequencing, it is now possible to study the genomes and populations of any species to obtain precise evidence about their evolution and adaptation. Here, we will give an overview of software tools for processing raw sequencing reads into population-level data, and then go through the common population genomics analyses on these data and computational tools developed for them, as well as give insights into the computational solutions and their efficiency.

We first address the tools and pipelines for processing next-generation sequencing data from heterogeneous data sources into population-level data comprising single nucleotide polymorphisms or copy-number variants. After a brief discussion on all-purpose software tools for carrying out standard population genetic analyses, we provide a more detailed overview of different types of population genomics data analyses, loosely grouped under population genetics and demography, evolutionary population genomics, phylogenomics, and comparative genomics, as well as suggest current tools for the analyses. Under population genetics and demography analyses, we discuss methods for exploring population genomic diversity and genetic structure, population admixture, interspecific introgression events, and inferences about overall population history. The evolutionary genomics analyses include methods and tools for studying patterns of selection, such as hard and soft sweeps and population differentiation but also genome-wide association studies and pan-genomes between individuals and populations, as well as paleogenomics research. Under phylogenomics and comparative genomics, we provide an overview of the computational tools used for studies on polyploid species, phylogenomics, and comparative genomics of gene space evolution within and between species.

Keywords Admixture · Data analysis · Evolutionary population genomics · Introgression · Paleogenomics · Polyploidy · Population genetics · Population genomics · Single nucleotide polymorphisms · Software

J. Salojärvi (✉)
School of Biological Sciences, Nanyang Technological University, Singapore, Singapore
e-mail: jarkko@ntu.edu.sg

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_57,

1 Introduction

The influx of new genome data is bringing about a golden era for research on populations; the genomic footprints of demography, evolution, and adaptation of a wide variety of species can now be studied. Population genomics research is a field where population genetics studies are carried out with information derived from the whole genome data. This may sound like a trivial extension, but in fact the availability of full genomes transforms all analyses by introducing the well-known statistical challenge of high-dimensional data, the notorious "small n, large p" problem, into genetics. A second challenge comes from theory, since simplifying assumptions made by many methods, for example, independence of different genomic loci in many likelihood-based models, do not necessarily hold. On the other hand, analyzing whole genomes instead of a small set of markers opens up new opportunities for obtaining information on the populations and species, its demography, selection pressure, evolution, and the underlying causal variants.

In this chapter, we will go through different analyses that can be carried out with the population-level whole genome sequence (WGS) data. Many of the methods are implementations of theoretical research on population genetics and analyze single nucleotide polymorphism (SNP) data. Therefore, we start by introducing the standard methodology for obtaining SNPs from raw next-generation sequencing reads. Since the standard SNP calling software is not able to detect larger genomic insertions, deletions, and duplications, we address the software for detecting copy-number variation separately. We will then go through the general methods and software tools for analyzing aspects of population genetics and demography, evolutionary genomics, and further expand the discussion to other types of data sets such as pan-genomes and the challenges in analyzing more complex data, for example, sequencing reads from polyploid species or ancient DNA. We loosely group various population genomics analyses under three categories: population genetics and demography, evolutionary population genomics, and phylogenomics and comparative genomics.

2 Single Nucleotide Polymorphisms

Most of the theory of genetics has been derived under the assumption that genome evolves through random point mutations, single nucleotide polymorphisms (SNPs), which are inherited from parent to child. The coalescent model further assumes that all genome sequences trace back to a common ancestral sequence, and this time can be estimated given the number of mutations introduced to the genome per generation. Altogether, this means that it is possible to identify the relationship between samples from the proportions of shared SNPs and, on larger time scales, to look at admixture between subpopulations and introgression between species by analyzing common SNPs.

In this section, we discuss how to obtain the SNPs (SNP calling) from next-generation sequencing data. This process involves several steps, out of which the first one is perhaps the most important, quality control of the sequencing reads. Therefore, we will initiate this Section by discussing the general aspects that may compromise high-quality SNP calling and then go through the steps required for SNP calling. In case of an existing reference genome, the procedure is very similar for all data types, including whole genome sequencing, restriction site-associated sequencing (RADseq), RNA sequencing, or exome sequencing data. Protocols exist for SNP calling using specific set of tools (see, e.g., DePristo et al. 2011; Nielsen et al. 2011; Langmead and Salzberg 2012), but overall the steps for obtaining the SNPs is similar, no matter which set of tools are used (Olson et al. 2015). These general procedures and standard tools are discussed in Sect. 2.2.

Alternatively, a reference genome for the studied species may not be available. Nevertheless, populations can be analyzed from marker-based sequencing data, such as RADseq, or transcriptome sequencing. We will discuss this methodology in Sect. 2.3. Finally, after the SNP data has been obtained, it needs to be filtered for high-quality SNPs and annotated, if this information is available. Furthermore, more accurate information can be obtained by phasing the genotype data. These issues will be discussed in the concluding subsections.

2.1 Quality Control Is Essential for All Data

High-quality data is paramount for all data analyses, and population genomics is not an exception. With noisy data or data from poorly designed experiments, all inferences will be unreliable or possibly even false. An inherent problem in high-dimensional data is that even if the results are random, always something that "makes sense" can be found out by cherry-picking the data. To avoid these issues, careful preprocessing and appropriate filtering of the data is essential, as well as carrying out statistical tests of the claims and making sure that the number of individuals in the study is sufficient.

2.1.1 Issues Affecting the Quality of the SNP Calls

Below we address some of the quality issues affecting the reliability of the inference with population data.

Genome Quality

The quality of the reference assembly is essential to all genomics work. For population genomics, the reference gives information on the order of the SNP loci across the genome and provides a way to link and compare the SNP variation

between individuals. Except for genomes assembled from Sanger sequence data, the quality of older genome assemblies is generally worse than in more recent genomes assembled using a combination of long-read platforms (PacBio, Oxford Nanopore) and high-throughput short-read technology (i.e., Illumina, Ion Torrent, BGISEQ).

Many population genomic analyses, such as admixture or introgression analysis, assume independent polymorphisms, and they are not as heavily affected by poor reference quality. An exception is a so-called over-assembled genome, where the assembly size is larger than the true genome size (estimated by, e.g., flow cytometry). The most common reason for over-assembly is high heterozygosity, which results in different assembly paths for different haplotypes, and, therefore, different contigs can map to the same physical region in the genome. When the same genomic locus is present twice in the reference assembly, short reads from resequenced individuals are mapped to either of these regions, depending on which haplotype is more similar. This can cause artifacts, such as regions of low genomic variation in the population; these in turn could be falsely interpreted, for example, as selective sweeps.

When working with a non-model organism, it may be necessary to test the quality of the genome assembly used as reference. For example, Quast gives an overall summary of the assembly, its size, and length distribution of the contigs and scaffolds (Gurevich et al. 2013). More arduous but also comprehensive analysis tool is REAPR which also estimates the amounts of assembly errors in the genome by analyzing sequencing data mapped to the reference (Hunt et al. 2013).

Read Length

Resequencing is usually carried out using short-read sequencing platforms. For example, in Illumina HiSeq platforms, the resulting reads are typically paired-end, meaning that in an insert size library of ~500 bp, the 150 bp from the ends of the fragments are read by the platform. For short reads, highly repetitive sequences pose a problem, since the read length may not be enough to identify a unique region in the genome. Nonunique regions are typically outside of gene-coding regions (repetitive DNA), but ambiguous mapping is possible also in recently duplicated genes or segmental duplications. Similarly, non-ambiguous or false mappings may occur if the species has high nucleotide diversity and, therefore, high variation in the genome. Additionally, in case of autopolyploid species, the amount of ambiguously mapping reads can be huge.

Coverage

Even though the cost of sequencing has been dropping rapidly, it is still expensive to sequence populations of individuals. Therefore, in many population genomics studies, the coverage – how many times an individual locus has been sequenced with short reads – of resequencing is limited. This leads to problems in SNP calls.

For example, if a locus has been sequenced to a coverage of 4, and each read is from either one of the haplotypes with the probability of 0.5, there is overall a $2 \times 6.25\%$ chance that all the reads are from the same haplotype, meaning that a heterozygous call is missed with 12.5% probability. Luckily this problem will be quickly alleviated, since with the recent extremely high-throughput platforms, such as NovaSeq, the library construction will be the highest cost, not the sequencing itself.

Genome Annotation

High-quality genome annotation is essential when evaluating the SNP data, since the annotation information can be used to filter the data for neutral SNPs, such as fourfold degenerate or intergenic SNPs, or then to find SNPs with putative causal effect; SNPs that cause a non-synonymous mutation in the protein sequence encoded by the gene.

2.1.2 Quality Control Tools for Sequencing Data

Quality control of the sequencing data is the essential first step in data analysis. The most common tool for monitoring the quality of the sequenced library is FASTQC (Andrews 2010), which provides summaries of the library size and quality scores of the reads, as well as lists overrepresented sequences. The tool also gives the number of reads in the sequencing library, which can be used to calculate the expected average coverage along the reference genome. Based on this, it can be decided whether more sequencing is needed to obtain high enough coverage for data analyses.

Furthermore, the KmerGenie software can be used to give an estimate of the sequencing coverage, genome size, as well as best k-mer value (division of data into substrings of length k) for subsequent analyses (Chikhi and Medvedev 2014).

2.1.3 Read Trimming

The next step after ensuring the quality of the data is to trim it by removing low-quality data and adapters used in sequencing. Based on the output from the quality control, the parameters for read trimming software, such as Cutadapt (Martin 2011) or Trimmomatic (Bolger et al. 2014) can be tuned to filter out low-quality data. The tools also remove the adapters used in sequencing and therefore prepare a set of reads that are ready to be mapped to the genome. In Cutadapt, the adapter sequences are defined by the user, whereas Trimmomatic has a library of the common adapters used in Illumina TruSeq protocols. However, in exotic cases, the adapters need to be defined by the software user.

In addition to specific tools developed for read trimming, some analysis pipelines, such as Stacks, developed for RADseq data (Rochette and Catchen 2017), incorporate their own read trimming modules.

2.2 *Reference-Based SNP Calling*

SNP calling methodology can be split into two general categories based on the availability of the reference genome. In case of reference-based SNP calling, all of the reads are first mapped to the genome using a fast alignment method, and then a SNP caller software is invoked.

2.2.1 Read Alignment

Over the years, several tools have been developed for read alignment. In order to carry out fast searches through the genome, the first N bases (typically 15–25 bases) of the reads are used as a seed region for narrowing down the searches. First-generation tools, such as Bowtie (Langmead et al. 2009) and MAQ (Li et al. 2008), allowed no mismatches in the seed region and therefore the first 15–25 bases of the reads had to align to the reference without any gaps or errors. The second-generation methods allow a few gaps and errors (or, variation) also in the beginning of the reads. With the second-generation methods, read mapping has become a standard procedure. Current state-of-the-art tools include bwa-mem (Li 2013), Bowtie2 (Langmead and Salzberg 2012), and HISAT2 (Kim et al. 2015); all use similar algorithms, such as suffix tree and Burrows-Wheeler transform, to carry out a fast search for the matching locus (Canzar and Salzberg 2017). An outcome of the methods is a Sequence Alignment Map (SAM) or Binary Alignment Map (BAM) file.

More recently, a third generation of ultrafast alignment methods using pseudoalignment have emerged. For example, minimap2 (Li 2018) can obtain accuracy similar to second-generation mapping tools in standard alignment tasks and is superior in aligning long reads, for example, from PacBio sequencing. However, in some cases, the method can also be slower than second-generation methods, for example, with short-read data having a high error rate or short reads from Hi-C data.

2.2.2 SNP Calling

After the sequence alignment to the reference has been carried out, the actual SNP calling can be done. A standard procedure requires several steps, including adding read group information, sorting BAM files according to genomic coordinates of the

mapped reads, removing or marking PCR duplicates, and the eventual SNP calling, resulting in a Variant Call Format (VCF) file (Auwera et al. 2018).

Since sequencing coverage for some individuals can be low, current methods for SNP calling carry out this step by genotyping whole populations together instead of analyzing one individual at a time. This helps in genotyping loci with low sequence coverage, since information from other individuals can be used to infer the most likely allele configuration. We will next go through five alternative tools for SNP calling.

Genome Analysis Toolkit (GATK)

Perhaps the current de facto standard for SNP calling, the genome analysis toolkit, GATK (McKenna et al. 2010; DePristo et al. 2011; Auwera et al. 2018), is a set of tools developed by Broad Institute. Even though not always the best one in comparisons (Hwang et al. 2015; Sandmann et al. 2017), it is the common benchmark for all methods. The reason for popularity is that the software is well maintained due to resources available from a large institute, and the user community and developers are very active on the web. Therefore, in case of problems, help is always available.

The GATK SNP calling proceeds through a pipeline where the first step is to estimate a so-called general VCF (gVCF) file containing genotype likelihoods for a single individual at every site. The idea is that the file summarizes all the necessary information for subsequent SNP calls and is much faster to handle than large BAM files. In the second stage, joint genotype calling is carried out for a population, each individual represented by its own gVCF file. The two-stage process makes it faster to carry out SNP calling for different populations, since in a gVCF file, much of the necessary preprocessing is already carried out and only joint genotyping needs to be done.

SAMtools

Perhaps the simplest and fastest SNP calling software pipeline, SAMtools calls the SNPs from a BAM file by forming a pileup of the reads, filtering them by mapping quality, and then performing SNP calling using bcftools (Li et al. 2009; Li 2011). Although typically used for SNP calling for single individuals, also multiple individual SNP calls are possible.

Freebayes

Perhaps the best competitor to GATK, Freebayes uses Bayesian inference to determine the genotype configuration at a given locus by using Ewens sampling formula as the prior (Garrison and Gabor 2012). A particular strength of Freebayes

is its haplotyping which is obtained by read-backed phasing, identifying reads that span several SNPs at the same time.

ANGSD

Developed especially for low-coverage sequencing data, the benefit of Analysis of Next-Generation Sequencing Data, ANGSD, is that the whole implementation of the analysis pipeline is probabilistic (Korneliussen et al. 2014). Calculations are carried out using phenotype likelihoods and probabilistic models, instead of reverting to explicit SNP calls that lose information. This makes the tool the best option for low-coverage sequencing data, since uncertainty regarding the SNP calls can be handled optimally. The downside of this tool is that the manual is not very detailed and the methods are described only in the scientific publications, making it difficult to link the processing options with the methodology used.

DeepVariant

DeepVariant is a recent method that uses deep belief networks implemented in Google TensorFlow machine learning library to call SNPs (Poplin et al. 2018). In the original publication, the method performed significantly better than comparison methods GATK and Freebayes. The method is computationally considerably more demanding but is the first SNP caller able to use graphical processing units (GPUs) to parallelize the SNP calls and, therefore, accelerate calling.

2.2.3 SNP Annotation

After obtaining a VCF file, the SNPs can be annotated based on the genome information. This means identifying the locations of the detected SNPs inside the gene models and possible regulatory elements and subsequently assigning a possible functional impact for the SNPs. Annotation tools include SnpEff (Cingolani et al. 2012a, b), Annovar (Wang et al. 2010), and the Ensembl Variant Effect Predictor (McLaren et al. 2016).

2.3 De Novo SNP Calling

At the time of writing this chapter, the NCBI genome database listed reference genomes for 1,739 animals, 639 plants, and 3,456 fungal species. This is a minute amount of total life diversity on earth, and therefore, for most species the reference genome is not available. In this case, population analyses have been carried out

mostly using marker-based analyses. Restriction site-associated DNA sequencing (RADseq) and its variants are the prevailing method; see Andrews et al. (2016) for a review of different RADseq technologies. Other methods include transcriptome-based analyses, such as RNA sequencing and exome capture sequencing. Compared to RADseq, there are considerably less transcriptome-based population genomics studies, mainly because of the higher costs; RNA sequencing is currently considerably more expensive than whole genome sequencing or RADseq. Other issues include the low stability of the RNA molecules and the fact that the expression profile depends on tissue, time of the day, and environmental conditions. In transcriptome-based analyses, first a complete transcriptome is constructed by de novo assembly using software such as Trinity (Grabherr et al. 2011) or Oases (Schulz et al. 2012); see Geniza and Jaiswal (2017) for a review of different tools. Once assembled, the transcriptome can be used as a reference for variant calling. For example, Trinity includes a script for running GATK software using STAR aligner (Dobin et al. 2013).

In case of RADseq, the current industry standard tool is Stacks (Rochette and Catchen 2017). An alternative tool for this is PyRAD (Eaton 2014), which is better able to tolerate indels, making it a preferable choice when analyzing more divergent species. The SNP calling procedure in both methods is similar, albeit the actual methodological implementations differ. In Stacks, the RADseq analysis is initiated by clustering the reads, first into putative alleles and then putative loci within a sample. Stacks has parameters controlling the number of identified alleles, such as minimum read coverage and number of mismatches allowed in a read; similar parameters exist for further clustering of the alleles into genomic loci. After the within-sample loci are identified, they are matched between samples to find homologous sites, all together forming a catalog of shared loci. Again, a certain number of mismatches are allowed for homologous loci.

The deficiency of RADseq approaches is obviously the lack of reference genome, which makes it difficult to choose "correct" parameters for clustering. The robustness of the clustering parameters has been explored, and there exists a rule of thumb (Paris et al. 2017). However, the rule was obtained by analyzing species with low effective population size and, therefore, low heterozygosity. In plants, high heterozygosity is not uncommon, and therefore having several SNPs per one RADseq read is highly likely, suggesting more loose clustering parameters. Additionally, in reality, the SNP density varies by genomic region, and therefore, uniform clustering parameters could introduce a bias in the data.

2.4 SNP Filtering

The aim of SNP calling tools is to detect variants and assign a quality score to assess the reliability of the call. At the next stage, the data will then be filtered to select high-quality SNP calls. The selection of filtration parameters is specific to the data

set and the analysis task at hand and depends on the overall mapping quality and SNP call accuracy observed for the population. In this step, data is usually analyzed by developing summaries of the quality values present in the variant call format (VCF) file, for example, by reading the VCF file into R and visualizing the distribution of mapping quality values and read coverage on the SNPs, for example, by density plots. After identifying the proper filtration parameters for removing SNPs with low quality or coverage, there are several software tools to carry out filtration, such as the tools implemented in GATK (Auwera et al. 2018), SnpEff/SnpSift (Cingolani et al. 2012a, b), or vcftools (Danecek et al. 2011).

During SNP filtering, it is typical to filter out also rare SNPs. However, this should be done with careful consideration since the filtering affects all subsequent analyses; model-based admixture methods such as STRUCTURE have been reported to be sensitive to MAF threshold (Linck and Battey 2017), and even principal component analysis produces varying results with different MAF thresholds (De la Cruz and Raska 2014), albeit to a lesser extent (Linck and Battey 2017). Naturally, also, methods analyzing site frequency spectrum and rare alleles will be affected. Additionally, the proportion of rare SNPs that significantly contribute to phenotypic traits is large. For example, in *Arabidopsis* GWAS, 35% of Bonferroni-corrected significant associations were observed with SNPs having MAF less than 5%, and further 28% had intermediate MAF of 5–10% (Togninalli et al. 2018).

2.5 *Phasing*

More accurate genetic analyses can be carried out if the genome data can be phased, that is, to be able to produce SNP data where the haplotypes have been identified. There are essentially three methodologies for obtaining phased data. First approach is read-backed phasing, which detects so-called haplotype blocks by identifying reads that span several SNPs and, therefore, find linked alleles. Using mapped reads, these blocks can be extended until an ambiguous region is encountered. The end result is a genome with haplotype blocks, regions where phasing has been obtained, spanned by unphased regions. A common problem in these approaches is that finding the relative phasing of different haplotype blocks is not possible based on short-read data. Read-backed phasing is implemented, for example, in GATK and FreeBayes software.

Phasing can also be carried out with trios, data consisting of parents and their progeny. In this case phasing is obtained by looking at recombination and SNP patterns observed in progeny. One such tool for phasing is whatsHap (Patterson et al. 2015). Finally, computational phasing can be carried out in large populations using software such as Beagle (Browning and Browning 2007) and Eagle2 (Loh et al. 2016).

3 Copy-Number Variation

In addition to SNPs, genomes contain a large number of structural variants (SVs): insertions, deletions, inversions, or copy-number variation. This genomic variation is an important mechanism for evolution and adaptation (Iskow et al. 2012). The standard SNP calling software processes only reads that are mapped to the reference with high confidence and therefore is able to detect only very short, few base pair insertions or deletions. In order to detect large indels and copy-number variation, specific tools for detecting structural variants have been developed. The methods can be categorized into two groups based on the data analysis type. In the first category, the methods analyze the paired-end and split reads to detect anomalies such as paired-end reads where only one end maps to the reference genome, or reads where the insert size based on mapping deviates from the library insert size, or reads where the orientation of the different ends is altered. An example of such software is DELLY (Rausch et al. 2012).

The second category of methods for detecting SVs is to identify regions where read coverage deviates from the average read coverage across the genome. For example, CNVnator (Abyzov et al. 2011) monitors the read-depth along the genome and identifies regions that deviate from the mean read depth. Methods also exist which combine the two sources, such as LUMPY (Layer et al. 2014), which combines split-read analysis and read coverage analysis using a probabilistic model to make a combined prediction of the SVs.

Most user-friendly tool for SV analysis is SpeedSeq (Chiang et al. 2015), which implements a full pipeline for SV calling. It uses LUMPY to initially detect SVs and then read-depth analysis by CNVnator to detect SVs that were not detected by LUMPY because of unmappable or repetitive sequence. Finally, it uses SVTyper for refining the SV breakpoints.

Genome STRucture in Populations (Genome STRiP) is a 12-stage SV discovery pipeline developed by Broad institute (Handsaker et al. 2015). For a given uniquely alignable genomic region, Genome STRiP models the distribution of read depths observed in the individuals using constrained Gaussian mixture models. The model-based approach makes it possible to estimate the most likely copy number for each genome and the confidence of the assignment. Mainly developed for humans, the software utilizes pre-computed metadata identifying the uniquely alignable regions from reference genome. However, for other species it is possible to set up the necessary metadata files using the tools in the Genome STRiP package. An outcome of the SV calling software is a file following the VCF format.

4 Population Genomic Analyses Using SNP or Structural Variant Data

After SNP calling, annotation, and filtering steps, the genotyping of the individuals has been completed, and the data is ready to be analyzed. In this section, we will go through some of the most common analyses for population genomic data and the

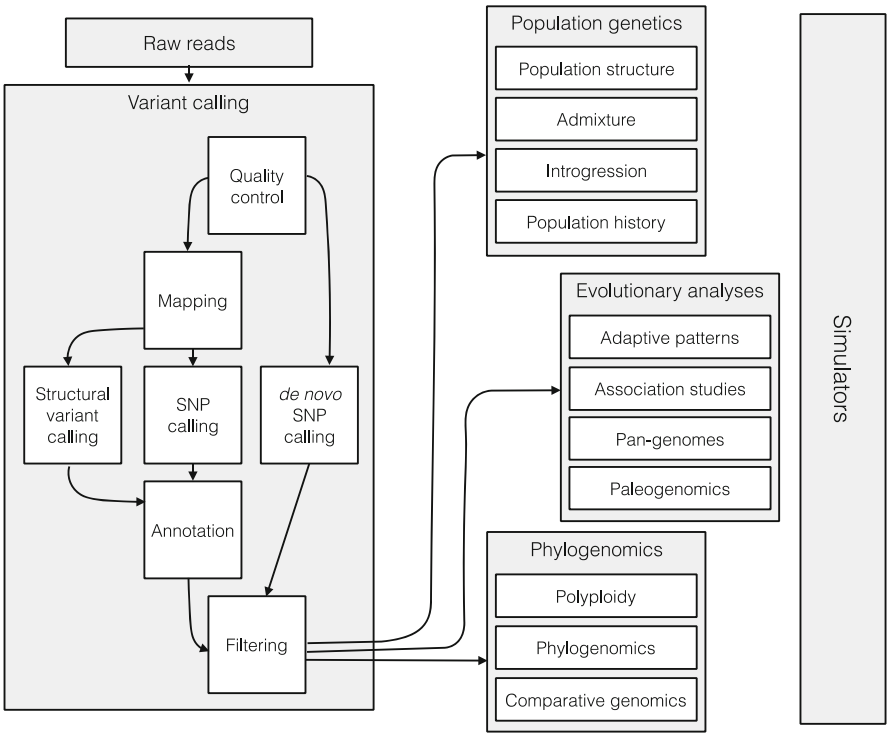

Fig. 1 In general, population genomic analyses can be split into research on population genetics and demography, evolutionary population genomics analyses, or phylogenomics and comparative genomics. Simulations can be run in parallel to all the analyses in order to test different evolutionary scenarios or the robustness and performance of new methods

available software tools. In addition to obtaining overall population genetic parameters that characterize the population, the population genomic analyses carried out with the data could be loosely grouped into the following three not so distinct categories: population genetics and demography, evolutionary population genomics, and phylogenomics and comparative population genomics. These are illustrated in Fig. 1.

4.1 All-Purpose Tools for Common Population Genetic Analyses

General-purpose analysis tools for population data are implemented in vcftools (Danecek et al. 2011), PLINK (Purcell et al. 2007, Chang et al. 2015), various R packages, and other software.

4.1.1 Vcftools

The vcftools is a simple tool for filtering the population genotyping data for certain sites, individuals, or allele frequency. Additionally, it can estimate standard population genetic parameters, such as linkage disequilibrium, population differentiation in terms of F_{ST}, and heterozygosity. The downside of the tool is that parallelization is not implemented.

4.1.2 PLINK

PLINK is a highly efficient tool for population genetic analyses, incorporating several different analyses, such as kinship estimation with identity-by-descent (IBD) method. The software can read-in VCF files, but after initial import, it uses its own data format, ped. The benefit of the file format is that analysis is extremely fast, but the encoding restricts the analysis to biallelic SNPs.

4.1.3 R

The number of R packages developed for analyzing population genomics data is rapidly expanding. Here, we suggest the reader to look into the functionalities in Pegas (Paradis 2010), PopGenome (Pfeifer et al. 2014), evobiR (Blackmon and Adams 2015), SNPRelate (Zheng et al. 2012), phangorn (Schliep 2011), and APE (Paradis et al. 2004) packages.

4.1.4 ANGSD

Designed for low-coverage sequencing data, the Analysis of Next-Generation Sequencing Data, ANGSD, has implementations for estimating the general population genetic parameters but also more advanced analyses such as admixture analysis though NGSadmix (Skotte et al. 2013) and estimation of IBD probabilities with NGSRelate (Korneliussen and Moltke 2015) are available.

4.2 Population Genetics and Demography

The first category of analyses, population genetics and demography, incorporates the overall analysis of the population genomic diversity, population genetic structure, and population history and demography. The current genetic diversity of the population results from a complex history of alternating population size, gene flow between populations, possible introgression from other species, as well as selection

and evolution. The general descriptive population genetic statistics such as nucleotide diversity will give an overall characterization of the population, but in order to properly analyze the different historical events, it is important to first obtain an understanding of the current population structure.

The key rule in all data analysis is to initiate it with models, which make the least amount of assumptions on the data, and proceed incrementally to more advanced models that make stronger modeling assumptions as more understanding of the data is obtained. Similarly, also the analysis of SNP data from populations is initiated by visualizing the overall variance of the data with principal component analysis (PCA).

4.2.1 Population Structure

Since most of the SNPs in the genomes are likely neutral, the global pattern of SNPs is largely due to drift processes and genetic relationships between the samples. This so-called population structure, differences in genetic ancestry of the sampled individuals, explains also much of the phenotypic variation observed in populations. Population structure can be used to estimate the relationships between samples and also as a null hypothesis where the observed phenotype is explained by drift processes.

The best tool for providing an initial view of the data and the population structure is principal component analysis (PCA), since it makes the least number of assumptions about the data. Perhaps the most common tool for PCA is EIGENSTRAT (Price et al. 2006), part of the EIGENSOFT package (Patterson et al. 2006). The downside of the tool is that the VCF file containing the SNPs needs to be converted to its own internal format before analysis. The package has tools for the conversion (convertf), but it doesn't have direct converter from VCF. Therefore, a VCF file needs to be first converted to PLINK format and that in turn to EIGENSTRAT. In addition to EIGENSTRAT, principal component analysis can be carried out in PLINK (Chang et al. 2015), and various packages in R programming language, for example, SNPRelate (Zheng et al. 2012), PCAdapt (Luu et al. 2016), and a combination of Adegenet (Jombart 2008) and ade4 (Dray and Dufour 2007). A complementary approach to PCA is a nonlinear PCA, known also as principal coordinate analysis or multidimensional scaling (MDS). The MDS is implemented, for example, in PLINK and Adegenet/ade4.

Instead of measuring purely the Euclidean distance between samples, there exist also genetically motivated ways of estimating relatedness, such as identity-by-state (IBS) and identity-by-descent (IBD) analyses. IBS analysis estimates the proportion of shared SNPs, whereas IBD estimates the proportion of haplotype blocks inherited by descent. Different approaches for IBD have been implemented, for example, in PLINK, NGSRelate (Korneliussen and Moltke 2015), and RELATE (Albrechtsen et al. 2008). In R, the SNPRelate package (Zheng et al. 2012) incorporating many IBD estimating methods is recommended. Finally, the refined IBD uses phased haplotype data for the inference (Browning and Browning 2013).

Finally, so-called admixture modeling is a model-based method for estimating ancestral populations and their admixture proportions in the individuals under study.

Since its introduction in the beginning of the 2000s (Pritchard et al. 2000), the analysis of admixture has become a common tool for all population analyses, and it could be viewed as an alternative way of estimating the population structure. However, compared to PCA, the admixture analyses make much stronger modeling assumptions on the data, and therefore also the results depend on how well the data fits these assumptions. We will therefore discuss admixture modeling separately in Sect. 4.2.2.

4.2.2 Admixture

Admixture analyses divide the genomes into ancestral non-admixed populations and estimate their relative proportions in each individual. The computational model behind admixture analysis is known as latent Dirichlet allocation (Blei et al. 2003), or discrete PCA (Buntine and Jakulin 2004), where for each locus, the allele is probabilistically generated from a set of ancestral populations. The method is completely data-driven, as the parameters of the model – proportions of ancestral populations in each individual – are estimated from the data using posterior sampling (Markov chain Monte Carlo-based methods), variational approximation, or maximum likelihood fitting. The difference between the implementations is that MCMC-based methods sample from the exact posterior, giving accurate results, but on the other hand they are slower to run. Variational methods estimate an approximate posterior distribution of the model parameters making heavy independence assumptions, and, thus, they give less accurate results but the execution is much faster. Finally, maximum likelihood solutions fit the model parameters to the likelihood without prior (thus assuming a uniform prior for the parameters). Most common methods include MCMC-based STRUCTURE (Pritchard et al. 2000), fastStructure using variational approximation (Raj et al. 2014), or maximum likelihood-based methods FRAPPE (Tang et al. 2005), ADMIXTURE (Alexander et al. 2009), and NGSadmix (Skotte et al. 2013). Recent developments include a fine-scale method that uses phased haplotype data to identify admixture. The fineSTRUCTURE (Lawson et al. 2012) first uses ChromoPainter to identify shared haplotypes in the population and then estimates their admixture; for a practical application of the software, see, for example, Kerminen et al. (2017).

The generative model underlying the admixture model also has its deficiencies, such as sensitivity to uneven sample sizes (Puechmaille 2016). Additionally, very different demographic scenarios can result in similar admixture compositions and, therefore, the results may be subject to over-interpretation (Lawson et al. 2018). In order to help in the interpretation and in identifying the most likely underlying scenario, a set of complementary analyses have recently been suggested, the so-called badMIXTURE which analyzes the goodness of fit of the admixture model (Lawson et al. 2018).

4.2.3 Introgression

In addition to admixture between populations, there can be gene flow between species through introgression. This may happen in cases where species may have already been split, but divergence has not yet resulted into a complete reproductive isolation barrier. Genomic research on introgression was heavily affected by the introduction of F3 and F4 statistics, developed to detect introgression with Neanderthals and humans (Green et al. 2010). These methods are implemented in the package Admixtools, which in addition to the formal test for introgression (F3) and test for the directionality of the introgression (F4) also contains methods to estimate the timing of the introgression and an implementation for comparing possible scenarios of introgression in studies concerning several species (Patterson et al. 2012). Since then, a revised statistic, the D statistic (also commonly known as ABBA-BABA statistic), was introduced as an improvement (Durand et al. 2011). The D statistic is implemented, for example, in PopGenome R package (Pfeifer et al. 2014). When trying to identify the regions under gene flow, the statistic has been applied to smaller genomic windows. However, it was recently shown that the D statistic produces inflated values when effective population size is low, and as a result, regions with low genomic diversity result in false positives, high D values (Martin et al. 2015). As a correction, a combination of f_d and d_{XY} statistics has been suggested, the first one to identify introgression and second one to identify regions of low diversity (Martin et al. 2015).

Even though formal tests for introgression exist, determining the directionality and proportions in the case of several populations and species is still very much manual work. Search for a solution that incorporates admixture events increases exponentially with the number of populations, and therefore, a global optimal solution is practically impossible to identify. However, greedy solutions exist, and they are implemented in

TreeMix (Pickrell and Pritchard 2012), Ohana (Cheng et al. 2017), and Admixturegraph R package (Leppälä et al. 2017). In Admixtools, it is possible to compare the model fit given different admixture solutions using qpGraph (Patterson et al. 2012).

4.2.4 Population History

Population history, the historical changes in effective population size, is of fundamental interest in population genetics. Several methods exist for estimating population history, all derived using different assumptions and summary statistics. The first set of methods estimate effective population size from the number of recombination events observed in a single individual or a small set of genomes. First, such method was pairwise sequentially Markovian coalescent model, PSMC (Li and Durbin 2011), which is still the method of comparison in several studies. However, PSMC is sensitive to population structure and can give false results, for example, in case of

population bottlenecks (Mazet et al. 2015). The MSMC2 (Schiffels and Durbin 2014) uses phased whole genome data and improves on PSMC by extending inference to several haplotypes and accelerates exact calculations, such that some of the approximations used in PSMC are not needed. The demographic inference using composite approximate likelihood (diCal) achieves similar improvements (Sheehan et al. 2013). The second method category estimates effective population sizes from haplotype lengths (Harris and Nielsen 2013).

Third methodology estimates population history from a site frequency spectrum (SFS). The methods are most accurate if an unfolded site frequency spectrum is used, and for this means the ancestral allele state needs to be identified first. State-of-the-art methods use several species and phylogenetic associations between them to estimate most likely ancestral state; see, e.g., phangorn package (Schliep 2011) in R. After inferring ancestral state and SFS, methods such as Stairway plots (Liu and Fu 2015) can be used to estimate a mixture model for the SFS data, where the mixture proportions are the effective population sizes at different times. Momi2 is a more recent method using similar strategy (Kamm et al. 2018). Benefit of SFS-based methods is that they do not suffer as heavily from population structure. Finally, the SMC++ integrates two methodologies by pairing coalescent HMM with site frequency spectrum estimation from a larger set of samples (Terhorst et al. 2017).

4.2.5 Mutation Rate

Correct estimate of mutation rate is essential for many analyses, since it helps in dating the divergence times in phylogenetic trees and major events in population history, such as population bottlenecks. In general, researchers are using mutation rates estimated in model species, since only a few studies exist on this subject in other species. One possibility is to obtain an indirect estimate of the mutation rate by comparing the divergence of orthologs between species and identify the amount of neutral mutations in the genes. The mutation rate can then be calibrated if the time of the species split can be estimated, for example, from fossil evidence.

An alternative method is the direct estimation of mutation rate from parent-progeny trios. Given parent-child relationships, the de novo mutations are identified by first estimating SNPs between the father-mother-child trios and then using trio calling software such as DeNovoGear (Ramu et al. 2013). In humans, the mutation rate estimates in different populations are converging to similar values with the direct method (Campbell and Eichler 2013). Interestingly, the indirect estimation gives twice as high mutation rate than indirect method, creating a conundrum (Moorjani et al. 2016).

4.3 Evolutionary Population Genomics Analyses

Genomic adaptation to the prevailing environmental conditions is a fundamental research question in ecology and evolutionary biology. Population genomics

addresses this question by looking at specific signatures in genome-level data or by seeking for association between genomic loci and phenotypic traits. Additionally, the analysis includes large-scale genomic variation between species and populations, such as copy-number variation and pan-genomes.

4.3.1 Genomic Patterns of Selection

A large body of population genetics research is devoted to studying strong positive selection for certain alleles. Hard selective sweep patterns appear under strong positive selection where the frequency of the favored allele rapidly increases and eventually reaches fixation in the population. During this process, genomic hitchhiking occurs where also the neutral alleles in linkage disequilibrium with the beneficial allele are inherited as well and reach fixation. As a result, the underlying genomic region is swept from variation. After reaching fixation, the region again accumulates random mutations. Their mutations are more recent than the sweep, and, therefore, the local site frequency spectrum, a histogram showing the number of SNPs shared by 1..N individuals, shows an overrepresentation of recently derived alleles.

The software for detecting signatures of selective sweeps looks for regions of reduced variation, a site frequency spectrum that is skewed toward recent alleles, or specific linkage disequilibrium patterns. The simplest method is to calculate statistics, such as Tajima's D, Fay and Wu's H, or similar. The ANGSD estimates many of these statistics as a part of the pipeline, whereas R packages Pegas (Paradis 2010) and PopGenome (Pfeifer et al. 2014) have functions for estimating these from VCF data. Finally, vcftools (Danecek et al. 2011) is able to calculate the basic statistics, Tajima's D, heterozygosity, and runs of homozygosity (ROH) from VCF-formatted data.

More advanced statistical approaches are implemented in specific software. Tools analyzing changes in site frequency spectrum include Sweepfinder2 (DeGiorgio et al. 2015) and SweeD (Pavlidis et al. 2013). An alternative approach, OmegaPlus, implements the omega statistic to detect anomalies in linkage disequilibrium (Alachiotis et al. 2012). Finally, Sweepy (Druet et al. 2013) identifies regions with reduced heterozygosity with a hidden Markov model having three states: neutral, intermediate, and sweep. Instead of training, the model parameters were fixed by the authors based on cattle data.

In most of the population genetics/genomics studies, the effect sizes of loci that have been significantly associated with the traits are very small, implying that most of the traits are polygenic. Under this scenario, it is very unlikely that a beneficial mutation at a single locus would provide a remarkable fitness advantage. Indeed, hard selective sweeps have turned out to be quite rare in nature. In contrast, soft sweeps occur in cases where several mutations in a genomic region have a fitness advantage, and, therefore, a palette of haplotypes is under selection in the same region. Soft sweeps and ongoing strong positive selection are currently detected with methods that use phased SNP data to identify alleles on their way to fixation or under balancing selection. The integrated haplotype score (iHS) statistic (Voight et al.

2006) seeks alleles driven to intermediate frequency by measuring the decay of haplotype homozygosity for a given derived allele, compared to the decay observed for the respective ancestral allele. Further refinements of this method are the "number of segregating sites by length" statistic, n_{SL} (Ferrer-Admetlla et al. 2014), and H12 and H2/H1 statistics (Garud et al. 2015), which are more robust to fluctuations in recombination and mutation rate.

Due to the polygenic nature of most traits, a major proportion of evolution occurs through long-term local adaptation to environmental conditions. Differentiation between populations can be measured by carrying out a genomic scan with the Wright fixation index (F_{ST}). The F_{ST} essentially implements the famous analysis of variance criterion to genetic data by comparing between population variance and within population variance. When estimated using the neutrally evolving loci, its average over the genome gives the overall differentiation between populations, and significant deviations from this average score in specific genomic regions will identify loci potentially under selection. The F_{ST} statistic is implemented in vcftools (Danecek et al. 2011) and R packages PopGenome (Pfeifer et al. 2014), Pegas (Paradis 2010), StAMPP (Pembleton et al. 2013), and HIERFSTAT (Goudet 2004). A similar measure, Q_{ST}, measuring the genetic diversity of different phenotypic grouping can be implemented using the same functions.

However, in addition to selective processes, the genetic structure of a population is dictated by random genetic drift processes, such as drift due to founder effects, and population bottlenecks. For example, a founder effect may occur following the establishment of a new population in a new environment. If the population is small, the population will differ from other populations only because of limited genetic variation present in the founding individuals. Population bottlenecks, for example, due to harsh environmental conditions, can produce similar artifacts, whereas migration introduces new alleles to the population and reduces the levels of population differentiation. For these reasons, a plain F_{ST} measure is being replaced by methods which attempt to decouple the drift processes and selection. A standard methodology is to use the population structure as the null model for drift and then detect loci where the allele distributions cannot be explained by the null model. Models such as FDIST2 (Beaumont and Nichols 1996) and BayeScan (Foll and Gaggiotti 2008) assume independent samples and simulate a null model under specific population history scenario, whereas FLK (Bonhomme et al. 2010) and BayEnv2 (Günther and Coop 2013) estimate population structure from data and use this as the null model.

Finally, one emerging trend to tackle the polygenicity of complex traits is to use epistatic models, such as population graphs and redundancy analysis, which analyze multilocus data (see, e.g., Legendre and Fortin 2010; Rajora et al. 2016; Salojärvi et al. 2017).

4.3.2 Genome-Wide Association Studies

One of the fundamental questions in population genomics is how variation in different loci is linked with the observed phenotypes. These data are analyzed in genome-wide association studies (GWAS), where phenotype and genotype information is collected from cohorts having sizes between hundreds to hundreds of thousands of individuals.

The simplest methods for analyzing GWAS data compute the correlation or estimate a linear model for the allele frequencies and trait values (in case of continuous trait) or the difference between allele distributions (in case of a categorical trait). After site-wise analysis over all genomic loci, multiple testing correction of the *p*-values obtained for the individual loci is carried out using either the conservative Bonferroni correction or the more loose false discovery rate correction methods (e.g., Benjamini-Hochberg correction).

Similar to all population genomic studies, population structure is the major factor contributing many false positives. More sophisticated models implement linear mixed models (LMM), which take the population structure into account by introducing covariates that model their contribution. Perhaps the first such model was Efficient Mixed Model Association (EMMA) software (Kang 2008). It models population structure with a random effect where the variance structure is obtained from a kinship matrix describing the relationship between samples. However, the time required for computation scaled cubically with the number of individuals.

Speed and, therefore, scalability can be improved by approximate methods. The genome-wide rapid association using mixed model and regression (GRAMMAR) set up a two-stage process, where in the first stage the observed phenotypes were modeled with a linear model using the kinship information (Aulchenko et al. 2007a, b). Residuals from this analysis were then used as input to the association analyses incorporating genomic data. The method is implemented in the GenABEL package in R (Aulchenko et al. 2007a, b). More sophisticated methods such as P3D, Population Parameters Previously Determined (Zhang 2010) in TASSEL (Bradbury et al. 2007), and EMMAX (Kang 2010) take similar approach by using kinship to fix some of the parameters in the linear mixed model; the P3D uses the null model with only kinship data to fix the variance components in the linear mixed model used for estimating associations with SNPs, whereas in EMMAX the kinship matrix is assumed to contribute to the noise covariance.

Instead of making compromises in accuracy, speedups can be obtained also by careful analysis of the original exact method. The GEMMA carries out a single Eigen decomposition of the relatedness matrix and uses this to replace several computationally demanding Eigen decomposition steps in EMMA, thus reducing the time complexity to quadratic in terms of the number of individuals (Zhou and Stephens 2012).

Further extension to linear mixed model approach is to model several correlated phenotypic variables at once. The so-called multi-trait mixed model (MTMM) uses

similar approximation to EMMAX and P3D to estimate and fix the covariance matrix before estimating the associations (Korte et al. 2012).

With whole genome sequencing data, genomic loci are not independent. The so-called epistatic models attempt to identify genetic dependencies between loci. However, with immensely many different SNP combinations, evaluating all of the different SNP combinations quickly becomes computationally prohibitive. A proper solution for this problem has not been found yet. Various different approaches have been proposed, many of them based on a combination of exhaustive searches and greedy optimization; see Niel et al. (2015) for a review of the current status.

4.3.3 Pan-Genomes

The pan-genome defines the entire genomic repertoire of a given species or, in microbiology, the phylogenetic clade (Vernikos et al. 2015). The concept originated from microbiology where species borders are notoriously difficult to specify and is presently well established (Vernikos et al. 2015) with a large palette of analysis software for bacterial pan-genomes (Xiao et al. 2015). Beyond microbial research, pan-genome analysis has recently gained attention in plant genomics (Golicz et al. 2015), although still relatively few pan-genome studies have been published (see, e.g., Cao et al. 2011; Li et al. 2014; Wang et al. 2018). The aim in pan-genome analysis is to divide gene space to so-called core, cloud, and shell genomes where the split is made according to the prevalence; the core genes are present in all individuals, shell genes in at least two individuals, and cloud genes in only one. The genes in the different categories appear to differ by their function, for example, in Wang et al. (2018), the core was observed to be enriched for GO terms related to growth, development, and reproduction, whereas shell and cloud genomes were enriched for regulation of immune and defense responses and ethylene metabolism.

One method implementing the pan-genome analysis is eukaryotic pan-genome analysis toolkit (EUPAN), a software pipeline implemented to detect presence/absence variation among the genomes of many individuals (Hu et al. 2017). The method uses a "map-to-pan" strategy, where each of the individual genomes are first assembled de novo. After this the pan-genome is constructed by mapping the contigs to a reference genome and identifying non-redundant novel sequences. After ab initio gene prediction, the presence/absence variation is determined based on reads mapped against pan-genome sequences. This strategy was used in the rice (*Oryza sativa*) pan-genome (Wang et al. 2018).

4.3.4 Ancient DNA and Paleogenomics

Ancient samples and sample collections maintained by natural historical museums provide invaluable information about ancestral populations. Since they give information about the genome up to 1 million years ago, they can be used to study migration patterns, species evolution, and adaptation. The analysis of ancient DNA

was initiated in human studies (Green et al. 2010; Rasmussen et al. 2010; Meyer et al. 2012; Slon et al. 2018), and the research has made it possible to track down human ancestry across different time scales (Llamas et al. 2017). The main obstacle in ancient DNA analysis is the sample quality. After the death of an organism, the DNA molecules get fragmented and degraded over time. The level of degradation varies across samples and environments but also within the specific sample. So far the oldest samples where DNA has been sequenced currently date back to around a million year (Orlando et al. 2015).

Depending on the level of DNA conservation, sequencing typically produces very short reads, and if assembled also the contigs are very short. These contigs can be organized by comparing against a modern genome, but for extinct species, this does not provide a reliable view of the genome structure. Methods such as FPSAC (Rajaraman et al. 2013) and the further development EWRA (Luhmann et al. 2018) attempt to solve this problem by estimating the genome structure by comparing several related species.

The ancient DNA typically contains contamination by modern DNA molecules, which have limited degradation and fragmentation. These patterns can be used to remove the contamination in the sample. Implementations include PMDtools (Skoglund et al. 2014), mapDamage software (Jónsson et al. 2013), as well at AtLAS, a toolbox for SNP calling in ancient DNA (Link et al. 2017), taking into account degradation due to postmortem damage (Kousathanas et al. 2016). The methods analyze reads by looking for hallmarks of DNA degradation and either remove them (PMDtools) or in the more recent methods recalibrate the base quality scores according to their probability of being damaged (mapDamage and AtLAS).

In terms of alignment, ancient DNA sequences contain a considerable amount of damaged bases, which typically accumulate toward read ends. Therefore, for best quality alignment results, the whole read length should be used for identifying the mapping region, instead of using so-called seed regions for fast alignment. Additionally, the phylogenetic distance to the reference genome affects alignment and should be taken into account (Schubert et al. 2012). Several probabilistic alignment methods have been developed to take these effects into account, such as BWA PSSM (Kerpedjiev et al. 2014), sesam (Rasmussen et al. 2010), and Anfo (Briggs et al. 2007).

4.4 Phylogenomics and Comparative Genomics Analyses

How species are born was the fundamental question by Darwin already in the nineteenth century, and the answer is still very much unknown. However, in plants, a common pattern of speciation is the formation of polyploids (Soltis and Soltis 2009). After formation of polyploids, the duplicated genomes start to lose genes in a so-called fractionation stage, which eventually results in a diploid species. These phenomena can be studied using phylogenomics and comparative genomics.

4.4.1 Polyploids

Polyploidy is highly prevalent among plant and fish species. For example, roughly 60–70% of flowering plant lineages have polyploid ancestry; new polyploids are formed at a frequency of 1 per 100,000 individuals, and approximately 2–4% of speciation events involve polyploidization (Van de Peer et al. 2009). Compared to this, the amount of published polyploid reference genomes is relatively small. This is for several reasons; genetics of polyploid species is more difficult, and genome assemblies are harder to carry out due to sequence similarity among subgenomes.

Besides reference genomes, high sequence similarity is also a problem in resequencing because even with a high-quality genome assembly, short reads may not be long enough to identify the correct subgenome. In case of allopolyploid species, the genomes of ancestral diploid parents may be used for identifying the subgenome where the read originates from. For example, in PolyCat software (Page et al. 2013), a SNP index of homeologous loci between cotton (*Gossypium* spp.) subgenomes was used for RNA sequencing reads to identify the subgenome where they originated from. The same authors have also developed a PolyDog software (Page and Udall 2015) which, given a reference assembly of an allopolyploid species, identifies reads that map uniquely to only one of the subgenomes. Both tools are implemented in the bambam software package (Page et al. 2014).

However, in some cases large effective population size and the resulting high heterozygosity may help in genome assembly. For example, in the hexaploid genome of sweet potato (*Ipomoea batatas*), the high average density of 1 SNP per 58 bp made it possible to phase 30% of the genome into six haplotypes by read-backed phasing that extended seed regions based on read support (Yang et al. 2017). Additionally, biological variants can be exploited to obtain good genome assembly such as aneuploidy (International Wheat Genome Sequencing Consortium 2014) or doubled haploid (Garcia-Mas et al. 2012; Zhang et al. 2014) individuals.

Overall, the development of tools for population genomic analysis of polyploid species is still in the very beginning. SuperMASSA software was developed for SNP genotyping populations where the ploidy level can be unknown (Serang et al. 2012), whereas the recent R package updog genotypes polyploids by accounting for allelic bias, over-dispersion, and sequencing errors with an empirical Bayes approach (Gerard et al. 2018). Further analyses can be carried out, for example, with StAMPP software, developed for analyzing genetic differentiation and structure of populations with mixed ploidy levels (Pembleton et al. 2013).

4.4.2 Phylogenomics

SNP calling can be carried out also using several species. In this case, the SNPs represent nucleotide differences between species. For longer time scales, the SNPs may be flipping back and forth, especially with gene-coding genomic regions where the number of neutrally evolving sites is limited due to functional constraints

imposed by the encoded protein. However, the effect is reduced when a large set of intergenic SNPs and neutrally evolving SNPs from gene-coding regions are collected. The SNPhylo software estimates phylogeny by first converting the SNPs into a FASTA-formatted file and then estimating the phylogeny using DNAML from PHYLIP package (Lee et al. 2014). Alternatively the produced FASTA files can be used as input to other software for estimating phylogeny, such as RAxML (Stamatakis 2014).

4.4.3 Comparative Genomics

The evolution of the number and size of gene families is in a key role when studying the adaptation and evolution of species (Demuth and Hahn 2009), and the variation also closely ties with population genetics and genomics. Tandemly duplicated genes may have more relaxed selection pressure (Salojärvi et al. 2017), and in humans, copy-number variation has been found to be associated with tandem duplication regions, resulting also in gene duplications (Sudmant et al. 2010).

The general computational methodology to study gene family evolution is fairly well established. Methods such as OrthoMCL (Li et al. 2003; Chen et al. 2006) and OrthoFinder (Emms and Kelly 2015) first carry out all-vs-all BLAST using amino acid sequences from the species under study and then cluster the pairwise similarity matrix using Markov clustering (Enright et al. 2002). The clusters, orthogroups, form a set of genes with common ancestry, putative orthologs, and paralogs. It is worth noting that the grouping is merely computational and possibly mostly represents the overall gene family behavior. If one would inspect properly validated gene families with common ancestry and domain composition, they may be split into several orthogroups or be incorporated into large orthogroups with many more genes. In Salojärvi et al. (2017), the proper clustering coefficient was searched by analyzing how well the computational clustering matched with known gene family splits.

In order to aid downstream analyses, OrthoFinder is also able to infer gene trees for each orthogroup as well as estimate a rooting for the species tree based on gene duplication events (Emms and Kelly 2017). The method also reconciles the gene trees and produces estimates of gene loss, gain, birth events, or incomplete lineage sorting using DLCpar (Wu et al. 2014) or its own internal method Recon, making it possible to directly analyze gene family evolution. The downside of most reconciliation methods is that the models are not implemented to take into account whole genome duplications, a feature that is very common in, e.g., plant evolution. However, for example, Phyldog can model these events (Boussau et al. 2013). Finally, the software tool ANGES reconstructs ancestral genome maps by analyzing the syntenic organization of extant related genomes (Jones et al. 2012) and is thus an alternative method for identifying gene duplication events in the extant species.

A complementary approach to gene tree reconciliation methods is to estimate a birth-death rate model of gene families. Several probabilistic implementations exist, such as CAFÉ (De Bie et al. 2006) and Badirate (Librado et al. 2012). Probabilistic

implementation makes it possible to identify gene families which expand or contract significantly more than expected based on general behavior.

5 Simulation

Since the evolutionary history of a species can be very complex, involving bottlenecks, isolated populations, migration, admixture, and introgression events, it is important to be able to estimate what types of footprints these different events leave in the population. For this means, several simulators have been developed which can generate genomic data under different evolutionary scenarios. When comparing the simulated data to observed population data, it is then possible to identify the likely population history or simulate the future behavior of a population. Another possible use for the simulated data is to test the robustness and performance of new statistical methods under different scenarios. For a thorough review of 42 different simulators, we refer the reader to Hoban et al. (2012). In addition to these, fastSimCoal2 is a more recent, highly versatile simulator (Excoffier et al. 2013). Current state-of-the-art methodology uses approximate Bayesian computation (ABC) to facilitate inference (Sunnåker et al. 2013).

6 Future Perspectives and Conclusion

In this chapter, we provided a brief overview of different computational tools available for analyzing population genomics data. The set of tools and the analysis types listed are by no means comprehensive, since we are missing many important new and rising fields. For example, the dropping cost of bisulfite sequencing makes it possible to estimate methylation status of the genome for populations. In humans, epigenetic variation contributes to the natural variation between populations (Heyn et al. 2013). A second interesting research field is the estimation of ultrahigh density linkage maps. With low-cost sequencing, it is viable to sequence whole genomes in parent-progeny experiments. This produces millions of markers, which is too much for standard methods that analyze linkage between genomic loci. Recently developed Lep-MAP3 software is able to manage the additional complexity and provides reliable estimates even with low-coverage sequencing (Rastas 2017). However, also metagenomics, population-level RNA sequencing and expression QTLs are emerging fields within the scope of population genomics.

In summary, the genomics research field is expanding rapidly and will eventually encompass all research where biological data is produced by sequencing. Reference genomes for new species are emerging at an increasing rate, and with the dropping cost of sequencing, whole genome sequencing will be the method of choice for all analyses. WGS makes it possible to accumulate large population genomic data sets which can be analyzed for any given purpose beyond the original study. With the

increasing amounts of data, method development for population genomics is currently flourishing, with a huge number of different solutions developed for each task. In time, some of these will prevail and will be incorporated into standard analyses pipelines; which ones, only time will tell.

References

Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974–84.

Alachiotis N, Stamatakis A, Pavlidis P. OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. Bioinformatics. 2012;28(17):2274–5.

Albrechtsen A, Sand Korneliussen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genet Epidemiol. 2008;33(3):266–74.

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64.

Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016;17:81.

Aulchenko YS, de Koning D-J, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics. 2007a;177(1):577–85.

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007b;23(10):1294–6.

Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2018;43(1):11.10.11–33.

Beaumont MA, Nichols RA. Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond Ser B Biol Sci. 1996;263(1377):1619.

Blackmon H, Adams RA. EvobiR: tools for comparative analyses and teaching evolutionary biology. 2015. http://coleoguy.github.io/.

Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, SanCristobal M. Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics. 2010;186(1):241–62.

Boussau B, Szöllősi GJ, Duret L, Gouy M, Tannier E, Daubin V. Genome-scale coestimation of species and gene trees. Genome Res. 2013;23(2):323–30.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23(19):2633–5.

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S. Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci. 2007;104(37):14616.

Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084–97.

Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 2013;194(2):459.

Buntine W, Jakulin A. Applying discrete PCA in data analysis. Proceedings of the 20th conference on uncertainty in artificial intelligence. Banff, Canada: AUAI Press; 2004. p. 59–66.

Campbell CD, Eichler EE. Properties and rates of germline mutations in humans. Trends Genet. 2013;29(10):575–84.

Canzar S, Salzberg SL. Short read mapping: an algorithmic tour. Proc IEEE Inst Electr Electron Eng. 2017;105(3):436–58.

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Muller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet. 2011;43(10):956–63.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):7–7.

Chen F, Mackey AJ, Stoeckert CJ, Roos DS. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 2006;34(Database issue):D363–8.

Cheng JY, Mailund T, Nielsen R. Fast admixture analysis and population tree estimation for SNP and NGS data. Bioinformatics. 2017;33(14):2148–55.

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. 2015;12:966.

Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2014;30(1):31–7.

Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front Genet. 2012a;3:35.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. Fly. 2012b;6 (2):80–92.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.

De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006;22(10):1269–71.

De la Cruz O, Raska P. Population structure at different minor allele frequency levels. BMC Proc. 2014;8(Suppl 1):S55.

DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. SWEEPFINDER2: increased sensitivity, robustness, and flexibility. arXiv. 2015:2–7.

Demuth JP, Hahn MW. The life and death of gene families. Bioessays. 2009;31(1):29–39.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.

Dray S, Dufour A-B. The ade4 package: implementing the duality diagram for ecologists. J Stat Software. 2007;1(4).

Druet T, Pérez-Pardal L, Charlier C, Gautier M. Identification of large selective sweeps associated with major genes in cattle. Anim Genet. 2013;44(6):758–62.

Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Mol Biol Evol. 2011;28(8):2239–52.

Eaton DAR. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics. 2014;30(13):1844–9.
Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16(1):157.
Emms DM, Kelly S. STRIDE: species tree root inference from gene duplication events. Mol Biol Evol. 2017;34(12):3267–78.
Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30(7):1575–84.
Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS Genet. 2013;9.
Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol. 2014;31.
Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics. 2008;180(2):977.
Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, González VM, Hénaff E, Câmara F, Cozzuto L, Lowy E, Alioto T, Capella-Gutiérrez S, Blanca J, Cañizares J, Ziarsolo P, Gonzalez-Ibeas D, Rodríguez-Moreno L, Droege M, Du L, Alvarez-Tejado M, Lorente-Galdos B, Melé M, Yang L, Weng Y, Navarro A, Marques-Bonet T, Aranda MA, Nuez F, Picó B, Gabaldón T, Roma G, Guigó R, Casacuberta JM, Arús P, Puigdomènech P. The genome of melon (Cucumis melo L.). Proc Natl Acad Sci U S A. 2012;109(29):11872–7.
Garrison EM, Gabor M. Haplotype-based variant detection from short-read sequencing. ArXiv. 2012. https://arxiv.org/abs/1207.3907.
Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS Genet. 2015;11:e1005004.
Geniza M, Jaiswal P. Tools for building de novo transcriptome assembly. Curr Plant Biol. 2017;11–12:41–5.
Gerard D, Ferrão LFV, Garcia AAF, Stephens M. Genotyping polyploids from messy sequencing data. bioRxiv. 2018.
Golicz AA, Batley J, Edwards D. Towards plant pangenomics. Plant Biotechnol J. 2015;14(4):1099–105.
Goudet J. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. Mol Ecol Notes. 2004;5(1):184–6.
Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2011;29(7):644–52.
Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, Hansen NF, Durand EY, Malaspinas A-S, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Ž, Gušic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710–22.
Günther T, Coop G. Robust identification of local adaptation from allele frequencies. Genetics. 2013;195(1):205–20.
Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.
Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA. Large multiallelic copy number variations in humans. Nat Genet. 2015;47:296.
Harris K, Nielsen R. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 2013;9(6):e1003521.

Heyn H, Moran S, Hernando-Herraez I, Sayols S, Gomez A, Sandoval J, Monk D, Hata K, Marques-Bonet T, Wang L, Esteller M. DNA methylation contributes to natural human variation. Genome Res. 2013;23(9):1363–72.
Hoban S, Bertorelle G, Gaggiotti OE. Computer simulations: tools for population and evolutionary genetics. Nat Rev Genet. 2012;13:110.
Hu Z, Sun C, Lu K-c, Chu X, Zhao Y, Lu J, Shi J, Wei C. EUPAN enables pan-genome studies of a large number of eukaryotic genomes. Bioinformatics. 2017;33(15):2408–9.
Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 2013;14(5):R47.
Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5:17875.
International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science. 2014;345(6194):1251788.
Iskow RC, Gokcumen O, Lee C. Exploring the role of copy number variants in human adaptation. Trends Genet. 2012;28(6):245–57.
Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008;24(11):1403–5.
Jones BR, Rajaraman A, Tannier E, Chauve C. ANGES: reconstructing ANcestral GEnomeS maps. Bioinformatics. 2012;28(18):2388–90.
Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 2013;29(13):1682–4.
Kamm JA, Terhorst J, Durbin R, Song YS. Efficiently inferring the demographic history of many populations with allele count data. bioRxiv. 2018.
Kang HM. Efficient control of population structure in model organism association mapping. Genetics. 2008;178:1709–23.
Kang HM. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42:348–54.
Kerminen S, Havulinna AS, Hellenthal G, Martin AR, Sarin A-P, Perola M, Palotie A, Salomaa V, Daly MJ, Ripatti S, Pirinen M. Fine-scale genetic structure in Finland. G3. 2017;7(10):3459.
Kerpedjiev P, Frellsen J, Lindgreen S, Krogh A. Adaptable probabilistic mapping of short reads using position specific scoring matrices. BMC Bioinformatics. 2014;15(1):100.
Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
Korneliussen TS, Moltke I. NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. Bioinformatics. 2015;31(24):4009–11.
Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014;15(1):356.
Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet. 2012;44(9):1066–71.
Kousathanas A, Leuenberger C, Link V, Sell C, Burger J, Wegmann D. Inferring heterozygosity from ancient and low coverage genomes. Genetics. 2016.
Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8(1):e1002453.
Lawson DJ, van Dorp L, Falush D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nat Commun. 2018;9(1):3258.
Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15(6):R84.

Lee T-H, Guo H, Wang X, Kim C, Paterson AH. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics. 2014;15(1):162.
Legendre P, Fortin M-J. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Mol Ecol Resour. 2010;10(5):831–44.
Leppälä K, Nielsen SV, Mailund T. admixturegraph: an R package for admixture graph manipulation and fitting. Bioinformatics. 2017;33(11):1738–40.
Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93.
Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv. 2013. e-prints.
Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018:bty191.
Li H, Durbin R. Inference of human population history from whole genome sequence of a single individual. Nature. 2011;475(7357):493–6.
Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89.
Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18(11):1851–8.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
Li Y-H, Zhou G, Ma J, Jiang W, Jin L-G, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L, Zhang S-S, Zuo Q, Shi X-H, Li Y-F, Zhang W-K, Hu Y, Kong G, Hong H-L, Tan B, Song J, Liu Z-X, Wang Y, Ruan H, Yeung CKL, Liu J, Wang H, Zhang L-J, Guan R-X, Wang K-J, Li W-B, Chen S-Y, Chang R-Z, Jiang Z, Jackson SA, Li R, Qiu L-J. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol. 2014;32:1045.
Librado P, Vieira FG, Rozas J. BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics. 2012;28(2):279–81.
Linck EB, Battey CJ. Minor allele frequency thresholds strongly affect population structure inference with genomic datasets. bioRxiv. 2017.
Link V, Kousathanas A, Veeramah K, Sell C, Scheu A, Wegmann D. ATLAS: analysis tools for low-depth and ancient samples. bioRxiv. 2017. https://doi.org/10.1101/105346.
Liu X, Fu Y-X. Exploring population size changes using SNP frequency spectra. Nat Genet. 2015;47(5):555–9.
Llamas B, Willerslev E, Orlando L. Human evolution: a tale from ancient genomes. Philos Trans R Soc B Biol Sci. 2017;372(1713):20150484.
Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, Schoenherr S, Forer L, McCarthy S, Abecasis GR, Durbin R, Price AL. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48:1443.
Luhmann N, Chauve C, Stoye J, Wittler R. Scaffolding of ancient contigs and ancestral reconstruction in a phylogenetic framework. IEEE/ACM Trans Comput Biol Bioinform. 2018. https://doi.org/10.1109/TCBB.2018.2816034.
Luu K, Bazin E, Blum MGB. pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resour. 2016;17(1):67–77.
Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10.
Martin SH, Davey JW, Jiggins CD. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. Mol Biol Evol. 2015;32(1):244–57.
Mazet O, Rodríguez W, Chikhi L. Demographic inference using genetic data from a single individual: separating population size variation from population structure. Theor Popul Biol. 2015;104:46–58.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20 (9):1297–303.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The ensembl variant effect predictor. Genome Biol. 2016;17(1):122.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andrés AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338 (6104):222.

Moorjani P, Gao Z, Przeworski M. Human germline mutation and the erratic evolutionary clock. PLoS Biol. 2016;14(10):e2000744.

Niel C, Sinoquet C, Dina C, Rocheleau G. A survey about methods dedicated to epistasis detection. Front Genet. 2015;6(285).

Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12(6):443–51.

Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML, Zook JM. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet. 2015;6:235.

Orlando L, Gilbert MTP, Willerslev E. Reconstructing ancient genomes and epigenomes. Nat Rev Genet. 2015;16:395.

Page JT, Udall JA. Methods for mapping and categorization of DNA sequence reads from allopolyploid organisms. BMC Genet. 2015;16(2):S4.

Page JT, Gingle AR, Udall JA. PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. G3. 2013;3(3):517.

Page JT, Liechty ZS, Huynh MD, Udall JA. BamBam: genome sequence analysis tools for biologists. BMC Res Notes. 2014;7(1):829.

Paradis E. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics. 2010;26(3):419–20.

Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20(2):289–90.

Paris JR, Stevens JR, Catchen JM. Lost in parameter space: a road map for stacks. Meth Ecol Evol. 2017;8(10):1360–73.

Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12): e190.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. Ancient admixture in human history. Genetics. 2012;192(3):1065–93.

Patterson M, Marschall T, Pisanti N, Van Iersel L, Stougie L, Klau GW, Schönhuth A. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. J Comput Biol. 2015;22 (6):498–509.

Pavlidis P, Živković D, Stamatakis A, Alachiotis N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol Biol Evol. 2013;30(9):2224–34.

Pembleton LW, Cogan NOI, Forster JW. StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. Mol Ecol Resour. 2013;13 (5):946–52.

Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol Biol Evol. 2014;31(7):1929–36.

Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8(11):e1002967.

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, Gross SS, Dorfman L, McLean CY, DePristo MA. Creating a universal

SNP and small indel variant caller with deep neural networks. bioRxiv. 2018. https://doi.org/10.1101/092890.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904.

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945.

Puechmaille SJ. The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. Mol Ecol Resour. 2016;16(3):608–27.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 2014;197(2):573.

Rajaraman A, Tannier E, Chauve C. FPSAC: fast phylogenetic scaffolding of ancient contigs. Bioinformatics. 2013;29(23):2987–94.

Rajora OP, Eckert AJ, Zinck JWR. Single-locus versus multilocus patterns of local adaptation to climate in eastern white pine (Pinus strobus, Pinaceae). PLoS One. 2016;11(7):e0158691.

Ramu A, Noordam MJ, Schwartz RS, Wuster A, Hurles ME, Cartwright RA, Conrad DF. DeNovoGear: de novo indel and point mutation discovery and phasing. Nat Methods. 2013;10:985.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, Bertalan M, Nielsen K, Gilbert MTP, Wang Y, Raghavan M, Campos PF, Kamp HM, Wilson AS, Gledhill A, Tridico S, Bunce M, Lorenzen ED, Binladen J, Guo X, Zhao J, Zhang X, Zhang H, Li Z, Chen M, Orlando L, Kristiansen K, Bak M, Tommerup N, Bendixen C, Pierre TL, Grønnow B, Meldgaard M, Andreasen C, Fedorova SA, Osipova LP, Higham TFG, Ramsey CB, Hansen TVO, Nielsen FC, Crawford MH, Brunak S, Sicheritz-Pontén T, Villems R, Nielsen R, Krogh A, Wang J, Willerslev E. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010;463:757.

Rastas P. Lep-MAP 3: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics. 2017;33(23):3726–32.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28(18):i333–9.

Rochette NC, Catchen JM. Deriving genotypes from RAD-seq short-read data using Stacks. Nat Protoc. 2017;12:2640.

Salojärvi J, Smolander O-P, Nieminen K, Rajaraman S, Safronov O, Safdari P, Lamminmäki A, Immanen J, Lan T, Tanskanen J, Rastas P, Amiryousefi A, Jayaprakash B, Kammonen JI, Hagqvist R, Eswaran G, Ahonen VH, Serra JA, Asiegbu FO, de Dios Barajas-Lopez J, Blande D, Blokhina O, Blomster T, Broholm S, Brosché M, Cui F, Dardick C, Ehonen SE, Elomaa P, Escamez S, Fagerstedt KV, Fujii H, Gauthier A, Gollan PJ, Halimaa P, Heino PI, Himanen K, Hollender C, Kangasjärvi S, Kauppinen L, Kelleher CT, Kontunen-Soppela S, Koskinen JP, Kovalchuk A, Kärenlampi SO, Kärkönen AK, Lim K-J, Leppälä J, Macpherson L, Mikola J, Mouhu K, Mähönen AP, Niinemets Ü, Oksanen E, Overmyer K, Palva ET, Pazouki L, Pennanen V, Puhakainen T, Poczai P, Possen BJHM, Punkkinen M, Rahikainen MM, Rousi M, Ruonala R, van der Schoot C, Shapiguzov A, Sierla M, Sipilä TP, Sutela S, Teeri TH, Tervahauta AI, Vaattovaara A, Vahala J, Vetchinnikova L, Welling A, Wrzaczek M, Xu E, Paulin LG, Schulman AH, Lascoux M, Albert VA, Auvinen P, Helariutta Y, Kangasjärvi J. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. Nat Genet. 2017;49:904.

Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH, Dugas M. Evaluating variant calling tools for non-matched next-generation sequencing data. Sci Rep. 2017;7:43169.

Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014;46(8):919–25.
Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592–3.
Schubert M, Ginolhac A, Lindgreen S, Thompson JF, Al-Rasheid KAS, Willerslev E, Krogh A, Orlando L. Improving ancient DNA read mapping against modern reference genomes. BMC Genomics. 2012;13(1):178.
Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28(8):1086–92.
Serang O, Mollinari M, Garcia AAF. Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. PLoS One. 2012;7(2):e30906.
Sheehan S, Harris K, Song YS. Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. Genetics. 2013;194(3):647–62.
Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Pääbo S, Krause J, Jakobsson M. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc Natl Acad Sci. 2014;111(6):2229.
Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. Genetics. 2013;195(3):693–702.
Slon V, Mafessoni F, Vernot B, de Filippo C, Grote S, Viola B, Hajdinjak M, Peyrégne S, Nagel S, Brown S, Douka K, Higham T, Kozlikin MB, Shunkov MV, Derevianko AP, Kelso J, Meyer M, Prüfer K, Pääbo S. The genome of the offspring of a Neanderthal mother and a Denisovan father. Nature. 2018;561(7721):113–6.
Soltis PS, Soltis DE. The role of hybridization in plant speciation. Annu Rev Plant Biol. 2009;60(1):561–88.
Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Genomes P, Eichler EE. Diversity of human copy number variation and multicopy genes. Science. 2010;330(6004):641–6.
Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C. Approximate bayesian computation. PLoS Comput Biol. 2013;9(1):e1002803.
Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol. 2005;28(4):289–301.
Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole-genomes. Nat Genet. 2017;49(2):303–9.
Togninalli M, Seren Ü, Meng D, Fitz J, Nordborg M, Weigel D, Borgwardt K, Korte A, Grimm DG. The AraGWAS Catalog: a curated and standardized Arabidopsis thaliana GWAS catalog. Nucleic Acids Res. 2018;46(D1):D1150–6.
Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 2009;10:725.
Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin Microbiol. 2015;23:148–54.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4(4):e154.
Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, Shen F, Cui X, Yu H, Li Z, Chen M, Detras J, Zhou Y, Zhang X, Zhao Y, Kudrna D, Wang C, Li R, Jia B, Lu J, He X, Dong Z, Xu J, Li Y, Wang M, Shi J, Li J, Zhang D, Lee S, Hu W, Poliakov A, Dubchak I, Ulat VJ, Borja FN, Mendoza JR, Ali J, Li J, Gao Q, Niu Y, Yue Z, Naredo MEB, Talag J, Wang X, Li J, Fang X, Yin Y, Glaszmann J-C, Zhang J, Li J, Hamilton

RS, Wing RA, Ruan J, Zhang G, Wei C, Alexandrov N, McNally KL, Li Z, Leung H. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018;557(7703):43–9.

Wu Y-C, Rasmussen MD, Bansal MS, Kellis M. Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. Genome Res. 2014;24:475–86.

Xiao J, Zhang Z, Wu J, Yu J. A brief review of software tools for pangenomics. Genomics Proteomics Bioinformatics. 2015;13(1):73–6.

Yang J, Moeinzadeh MH, Kuhl H, Helmuth J, Xiao P, Haas S, Liu G, Zheng J, Sun Z, Fan W, Deng G, Wang H, Hu F, Zhao S, Fernie AR, Boerno S, Timmermann B, Zhang P, Vingron M. Haplotype-resolved sweet potato genome traces back its hexaploidization history. Nat Plants. 2017;3(9):696–703.

Zhang Z. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2010;42:355–60.

Zhang H, Tan E, Suzuki Y, Hirose Y, Kinoshita S, Okano H, Kudoh J, Shimizu A, Saito K, Watabe S, Asakawa S. Dramatic improvement in genome assembly achieved using doubled-haploid genomes. Sci Rep. 2014;4:6780.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28 (24):3326–8.

Zhou X, Stephens M. Genome-wide efficient mixed model analysis for association studies. Nat Genet. 2012;44(7):821–4.

Population and Evolutionary Genetic Inferences in the Whole-Genome Era: Software Challenges

Alexandros Stamatakis

Abstract The continuous advances in DNA sequencing technologies are driving a constantly accelerating accumulation of nucleotide sequence data at the whole-genome scale. As a consequence, evolutionary biology researchers have to rely on a growing number of increasingly complex software. All widely used tools in the field have grown considerably, in terms of the number of features as well as lines of code and consequently also with respect to software complexity. Complexity is further increased by exploiting parallelism on multi-core and hardware accelerator architectures. Moreover, typical analysis pipelines now include a substantially larger number of components than 5–10 years ago. A topic that has received little attention in this context is that of code quality and verification of widely used data analysis software. Unfortunately, the majority of users still tend to blindly trust the software and the results it produces. To this end, we assessed the software quality of three highly cited tools in population genetics (Genepop, Migrate, Structure) that are being routinely used in current data analysis pipelines and studies. We also review widely unknown problems associated with floating-point arithmetics in conjunction with parallel processing. Since the software quality of the tools we analyzed is rather mediocre, we provide a list of best practices for improving the quality of existing tools but also list techniques that can be deployed for developing reliable, high-quality scientific software from scratch. Finally, we also discuss some general policy issues that need to be addressed for improving software quality as well as ensuring support for developing new and maintaining existing software.

Keywords Numerical stability · Parallel computing · Reproducibility · Software quality · Software verification

A. Stamatakis (✉)
Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany
e-mail: Alexandros.Stamatakis@h-its.org

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*,
Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_42,

1 Introduction

With next-generation sequencing (NGS) data coming off age and being routinely used by now, evolutionary biology is becoming an increasingly quantitative and computational discipline (see also Barone et al. 2017). These massive amounts of data have also triggered a paradigm shift from a hypothesis-driven toward a more data-driven science.

One can also observe a gradual transformation into a true computational science as evolutionary biology increasingly relies on supercomputers (e.g., Misof et al. 2014 or Jarvis et al. 2014) as well as multi-core servers and accelerator architectures. This is a transition other disciplines such as astrophysics, geophysics, or fluid dynamics accomplished decades ago. This transition is challenging because it requires nontrivial parallel programming techniques (e.g., Alachiotis et al. 2012) and introduces additional reproducibility issues as well as error sources (i.e., nondeterministic program behavior) which we will also briefly discuss in this chapter.

Apart from the increasing use of parallelism with all its associated complications, researchers also have to rely on a substantially larger number of increasingly complex core software components. These core components are mostly written in `C` or `C++` because they are typically highly compute- and floating point intensive. By software complexity we refer to the fact that widely used tools in the broad field of evolutionary biology have grown considerably, in terms of the number of features, models, and lines of code. For instance, the Bayesian phylogenetic inference tool MrBayes (Ronquist et al. 2012) had approximately 49,000 lines of code in 2005 and already about 94,000 in 2014. Furthermore, evolutionary analysis software now supports a substantially larger set of models (e.g., substitution models, demographic scenarios, variants of the coalescent, approximate Bayesian computation approaches), hardware platforms (e.g., GPUs, FPGAs, etc.), and types of parallelism (e.g., embarrassingly parallel, fine-grain, coarse-grain, multigrain, hybrid approaches) than a decade ago.

Another challenge is that constantly growing datasets also induce increased numerical difficulties, as most population genetics codes calculate probabilities for some quantity and are thus prone to either exhibit numerical underflows (e.g., Pavlidis et al. 2013) or yield inaccurate results because of roundoff error propagation. The deployment of more complex and parameter-rich models further complicates matters, since it is often difficult to devise, for instance, numerically stable maximum likelihood parameter optimization procedures as these increasingly complex and parameter-rich models may exhibit several local maxima, for instance.

We do not only have to handle the software complexity of stand-alone core components but also need to consider the increasing number of core components in current analysis pipelines. In the "Sanger days," the analysis pipeline for evolutionary analyses used to be straightforward, once the sequences were available. For a phylogenetic study, it merely consisted of three steps: align → infer tree → visualize tree. For NGS data and huge phylogenomic datasets, such as the insect transcriptome

(Misof et al. 2014) or bird genome evolution (Jarvis et al. 2014) projects, the data analysis pipelines have become substantially longer and more complex. They also require user expertise in an increasing number of bioinformatics areas (e.g., SNP calling, orthology assignment, NGS error correction, read assembly, dataset assembly, partitioning of datasets, divergence times inference, etc.). In addition, these pipelines require a plethora of scripts to transform formats and to assemble workflows. Even format transformation is not trivial in case of badly specified or simply inappropriately used data formats which can lead to the incorrect presentation of results. Such a behavior was recently demonstrated for the widely used Newick phylogenetic tree file format (Czech et al. 2017). Moreover, helper scripts are typically written in languages such as `perl`, a language that is highly susceptible to coding errors due to lack of typing or `python` that uses dynamic typing and can thus not be subjected to a comprehensive type-check either. The term "typing" refers to the data types of variables (e.g., integer or floating point) that are passed to, and returned by, functions. Without strict typing a function expecting an integer argument can be invoked with a floating point value as an argument and exhibit undefined or unexpected behavior. Thus, programming languages with stricter type control reduce the potential for errors. Ideally the languages used should be fully type-safe. Our main concern is that, if each core software component (henceforth, we use code as synonym for software) or script component i used in such a pipeline has a probability of being "buggy" P_i, the probability that there is a bug in the pipeline increases dramatically with the number of components. If detected too late, errors in the early data analysis pipeline stages (e.g., NGS assembly, SNP calling, alignment) for large-scale data analysis projects can have a dramatic impact on downstream analyses such as coalescent simulations or phylogenetic inferences as they will all have to be repeated. In fact, this *has* happened in every large-scale data analysis project we have been involved in thus far. Given that our field needs to compete with established computational sciences for scarce supercomputing or cloud resources, repeating large evolutionary analyses can result in a substantial waste of computational resources. Current large-scale phylogenomic analysis projects can require up to 75 million processor hours on supercomputers.

Algorithmic problems that might generally be perceived as "being solved" such as the alignment of closely related sequences of individuals from a single population may also exhibit methodological pitfalls. It was shown that errors in multiple sequence alignments can yield a dramatically increased false-positive rate in tests for positive selection (Fletcher and Yang 2010) which can, however, be alleviated by taking alignment uncertainty into account (Redelings 2014). Thus, the alignment problem is generally not solved. This is true even for apparently simple cases such as pair-wise sequence alignment algorithms with affine gap penalties. Here, an error in the initial formulation of the algorithm (Gotoh 1982) has propagated into several textbooks, university lecture slides, and, more importantly, widely used implementations (Flouri et al. 2015). While we will not discuss methodological pitfalls here, we wish to emphasize that they exist and may also lead to incorrect inferences. In the following we will only focus on software quality and verification issues in population genetics software.

Based on the prolegomena, our goals in this chapter are to (1) assess the quality of current population genetics software and (2) to propose potential solutions, including software analysis tools, for improving the quality of population genetics software. We wish to emphasize that the quality measures we deploy only represent one possible approach to assessing software and reflect a soft probabilistic notion that "something might perhaps go wrong." Software quality is not necessarily an indicator for software correctness, but, as demonstrated repeatedly in software engineering research, a strong correlation *does* exist (e.g., Briand et al. 1999, 2000; Casalnuovo et al. 2015).

For assessing software quality, we downloaded and scrutinized – using a common set of criteria – three widely used and highly cited population genetics codes. An analogous study has been conducted for a broader range of evolutionary biology software in Darriba et al. (2018). Based on the software analysis results, we assemble a list of best practices and discuss some possible policy changes that might contribute to improving software quality.

It is absolutely not our intention to criticize the developers of the tools we assessed since they have made major contributions to the field. Instead, our goal is to emphasize that users should be aware of the fact that software is imperfect and that software quality should also constitute a criterion for selecting the most appropriate tool for conducting population genetics analyses.

The remainder of this chapter is organized as follows. In Sect. 2 we assess the software quality of three widely used population genetics tools. Then, we discuss some more general issues and additional error sources that are induced by deploying parallelism for large-scale data analyses in Sect. 4. We conclude our chapter with a suggestions for best practices in software development in Sect. 5 and discuss possible policy changes for improving software quality in Sect. 6.

2 Software Quality Analysis of Three Population Genetics Codes

In our software quality analysis, we focus on core tools that are typically open source, easy to obtain, and written in `C` or `C++` for computational efficiency. Note that, it is generally much harder to obtain the scripts used for large-scale empirical data analysis pipelines deployed for empirical population genetics studies as they are not always available and generally poorly documented.

While one might expect at least the core tools to exhibit a high software quality since "they are being used by everyone" and "they yield reasonable results," this is, as we will show, not the case for some exemplary standard tools. Note that, in the following, we only assess the software quality of these core tools using some rather straightforward yet informative criteria. As stated before, our findings do not imply that the codes do not work correctly. However, since there exists a strong correlation between code quality and correctness (Briand et al. 1999, 2000), software of bad

quality is substantially more likely to yield incorrect results. Our software analysis results allow for identifying potential weaknesses of the tools which *do* allow to deliberately make them fail under specific settings. What we intend to emphasize is that not enough attention and funding are spent on analyzing as well as improving the software quality of widely used tools with tens of thousands of citations, since potential and substantial bugs in these tools can have a dramatic impact on published research, including the worst-case scenario: the withdrawal of hundreds of papers due to bugs in one of the core tools.

We want to emphasize that the issue of software verification and correctness should receive substantially more attention from the application developers but also from pure computer science that needs to develop novel tools for automatic or at least semiautomatic verification of complex numerical codes in population genetics and other areas of bioinformatics.

To explore the software quality and probability of potential software issues, we analyzed the following three standard population genetics toolkits following a similar approach as in Darriba et al. (2018): Genepop (two main publications Raymond and Rousset (1995) and Rousset (2008) have over 19,000 citations; Google scholar, accessed January 25, 2017), Migrate (the four main publications Beerli and Felsenstein (1999, 2001), Beerli (2006) and Beerli and Palczewski (2010) have over 3000 citations; Google scholar, accessed January 27, 2017), and Structure (over 24,000 citations for the four main papers Pritchard et al. (2000), Falush et al. (2003, 2007) and Hubisz et al. (2009); Google scholar accessed January 25, 2017).

Note that the results of our analyses merely provide an intuition about what and how much could potentially go wrong. A detailed study of the warnings and detection of potential bugs for only one of these programs would require more than half a year of work for a programmer who is not familiar with the software which is beyond the scope of this chapter. Our main intention is to assess the current state, increase awareness about the issue, and provide some simple techniques and suggestions for improving code quality.

2.1 Experimental Setup

All three codes are written in `C/C++` and we analyzed them as follows. Initially, we simply counted the lines of code (excluding comments) and conducted a so-called cyclomatic code complexity (McCabe 1976) analysis using the lizard tool (https://github.com/terryyin/lizard). The cyclomatic complexity provides a measure for quantifying the control flow complexity in software (for a brief description, see https://en.wikipedia.org/wiki/Cyclomatic_complexity). Typically, functions with a complexity exceeding 10 or 15 are judged as being too complex. They should thus, ideally, be redesigned and restructured such as to increase modularization. Thereafter, we assessed the amount of code duplication using the simian tool (http://www.harukizaemon.com/simian/). Then, we deployed the `clang/clang++` compiler and enabled the following warning flags `-Weverything -Wno-padded -Wno-float-`

`equal -Wno-vla` to assess how many warnings the codes generate. Note that the `clang` compiler generates a substantially higher number of warnings than the `gcc` compiler suite, because it entails a static code analysis tool. Based on our experience, it reliably detects a significantly higher number of type mismatches in function calls and variable assignments than `gcc`. Subsequently, we applied the Linux command `grep assert` to all source files to determine if assertions were used. Assertions provide a means of verifying that the code, and more specifically its variables, is in the expected state, for instance, before a function call returns. The use of assertions essentially allows for implementing a, at least partial, correctness verification mechanism based on Hoare Logic (Hoare 1969) which is a formal framework to prove the correctness of code (see https://en.wikipedia.org/wiki/Hoare_logic). Our working hypothesis is that the more assertions a software author has inserted, the more he/she has attempted to reason about the correctness of the code. There also exists a recent software engineering study using a large collection of `C/C++` codes obtained from `github` which suggests that functions *with* assertions *do* have significantly fewer defects (Casalnuovo et al. 2015). Finally, we assessed the memory management of the three softwares via the standard `valgrind` tool using the `--leak-check=full` and `--show-reachable=yes` flags. In the following three sections, we discuss our findings for the three population genetics tools we assessed.

2.2 Genepop (V4)

The cyclomatic complexity analysis revealed that there are 53 functions with a cyclomatic complexity greater than 15, the three highest complexity values being 166, 128, and 74. Thus, given those very high numbers, there might be a need to simplify and modularize several functions. The code duplication analysis revealed that there are 2252 duplicate lines of code (LoC) in a total of 167 blocks of code, while the total LoC number without comments is 10, 583. Thus, there are approximately 20% of code duplication that could be avoided to improve maintainability of the software. Compiling the code with `clang` generated 1585 warnings, that is, approximately 1 warning per 7 LoC. We further found that not a single assertion is used in the entire code. Finally, the memory management analysis with `valgrind` (using `./Genepop settingsFile=sampleSettings.txt Mode=Batch`) indicated that 984 bytes of RAM are possibly lost and that, at program termination, 19, 056 bytes of RAM are still reachable. In other words, the program does not properly de-allocate the memory it used, at program termination. This can become problematic especially if the Genepop `main()` function is integrated and called by some larger surrounding `C` or `C++` code several times, because this will generate memory leaks. However, other memory issues are regularly being fixed by using `valgrind` during Genepop development.

At the time of writing this, F. Rousset, the main author of Genepop, was aware of the above issues. In the meantime, most issues have been taken into account in the

context of the redesign of Genepop using the `R` programming language (pers. comm., June 15, 2017, see https://cran.r-project.org/web/packages/genepop/index.html).

2.3 Migrate (Version 3.6.11)

For migrate, we only assessed the main source file directory and not the sub-directories, since the software also includes a plethora of third-party libraries for random number generation, assembling PDF documents, or compressing files. In total we detected 148 functions with a cyclomatic complexity greater than 15, the three highest values being 239, 126, and 114, that is, the top candidates for restructuring. We further detected 10, 145 duplicate lines in 867 code blocks for 83, 860 LoC in the main source directory (12% of duplicated code). Compiling the code with `clang` generated 1818 `clang` warnings, that is, about 1 warning per 46 LoC. No assertions are used in the source files of the main code directory. However, some of the third-party libraries such as `zlib` (compression) or `SFMT` (random number generator) *do* use assertions. Finally, we analyzed the memory behavior by executing `migrate-n` with the default test files in the `example/` directory of the distribution. We only modified the parameter file as follows (`long-sample=100` and `burn-in=1000`) to obtain reasonable execution times in conjunction with executing `valgrind` on top of migrate, since `valgrind` substantially increases run-times (typically the factor ranges between 5 and 100 according to the official `valgrind` documentation). The `valgrind` tool reports that 18, 008 bytes of allocated RAM are definitely lost, 7, 903, 896 bytes of allocated RAM are indirectly lost, 2, 910, 391 bytes of allocated RAM are possibly lost, and 240, 661 bytes of allocated RAM are still reachable at program termination.

P. Beerli, the main author of migrate, is aware of the above issues and is currently working on fixing them in the planned release v3.7.1 of his code (pers. comm., Jan 27, 2017).

2.4 Structure (Version 2.3.4)

Overall, we detected 31 functions with a cyclomatic complexity exceeding 15, the three highest being 86, 82, and 60. We detected 280 duplicate lines of code in 29 code blocks for a total of 6060 LoC amounting to about 5% of code duplication. Compiling the code with `clang` yielded 600 warnings, that is, roughly 1 warning per 10 LoC. We found that only one assertion is being used in source file `mymath.c`. We provide the corresponding code snippet below:

```
if(z <= 0.)
  {
   fprintf(stderr,
           "lgamma function failed with wrong input (%f)\n",z);
   assert(0);
   exit(-1);
   }
```

Here we want to show that the assertion is, in fact, not used as intended, because it will always fail since the Boolean expression it should evaluate is, in fact, a constant that is set to 0 (i.e., always evaluates to `FALSE`). Here, the assertion is merely used to exit the program, and the subsequent command `exit(-1)` will never be executed since the code will fail with the assertion prior to executing `exit(-1)`.

The memory check with `valgrind` and using simulated microsatellite test data from http://pritchardlab.stanford.edu/software/structure-data_v.2.3.1.html (with the following settings `LABEL=1`, `POPDATA=1`, `POPFLAG=1`, `NUMLOCI=5`, `PLOIDY=2`, `MISSING=-999`, `ONEROWPERIND=0`) generated no errors whatsoever. This is not surprising, since the original `Makefile` written by the authors already includes some commented out lines for using `valgrind`, that is, the tool was apparently used by the authors to generate code without memory leaks.

J. Pritchard, the main author of Structure, is aware of the above analyses and issues we detected and has no objections about them (pers. comm., Jan 29, 2017). The issues we found cannot be fixed in Structure due to lack of manpower.

The usage of `valgrind` and the lack of manpower for sustainable code maintenance lead to the conclusion of this section: if some straightforward standard tools for improving software quality are routinely used (e.g., `valgrind` for Structure), code quality can already be substantially improved. Furthermore, increased long-term funding for maintaining and occasionally redesigning such important tools from scratch is required.

3 Impact

The simple code quality metrics deployed in the preceding section only serve as proxies for software quality. Note that software quality and the probability of program faults, that is, either a crash of the program or incorrect behavior, are indeed correlated (see, e.g., Khoshgoftaar and Seliya 2003; Nagappan and Ball 2005). Thus, these analyses should provide sufficient evidence that additional measures to enhance software quality are required in the tested tools. This will increase confidence that they *do* work correctly.

We outline a simple example of how the Structure tool can be made to fail with an uninformative error message because of a programming error. If we analyze the warnings produced by the clang compiler, we observe the following programming error:

```
structure.c:3136:17: warning: implicit conversion
  changes signedness: 'int' to 'unsigned long'
  [-Wsign-conversion]
  lambda=calloc(MAXPOPS, sizeof (double));
```

This warning indicates that `MAXPOPS` has been defined as a 32-bit signed integer variable. However, the C function `calloc` is expecting an *unsigned*, typically 64-bit integer value of type `size_t`. Thus, if `MAXPOPS` is set to a value exceeding the signed 32-bit integer number range, we expect a failure to occur. To test this, Structure was executed under the following, admittedly rather unrealistic, setting:

```
structure -K 3000000000 infile
```

and yielded the following error message:

```
Error in assigning memory (not enough space?)
Exiting the program due to error(s) listed above.
```

This error message is misleading since the actual value of `MAXPOPS` before the problematic memory allocation is `-1294967296` (value obtained via code instrumentation). In fact, the failure occurs because (1) `calloc()` is invoked with a negative value since an inadequate integer type is used for allocating memory, (2) no range check for the command line input parameters is deployed (i.e., the value of `-K 3000000000` exceeds the signed 32-bit integer range), and (3) no assertions to verify the allowed value range of this variable are used (e.g., `assert (MAXPOPS > 0)`).

While this evidently represents a constructed example, this type of programming error in memory allocations is present in all three tools assessed here. As a consequence, they are all prone to yield analogous program failures. In fact, compiling Structure with `clang` yielded 88 cases where either `calloc()` or `malloc()` is invoked with incorrect integer arguments. While this type of programming error will not constitute a problem for the average use case, it is likely to emerge when analyzing large datasets. In other words, it limits the scalability of the tools.

As this type of exemplary errors might not affect the correctness of the tools, but merely their stability, the extremely frequent occurrence of dangerous implicit type conversions as in the above example is also likely to affect program correctness. Assessing the correctness of the tools is beyond the scope of this chapter as an in-depth study of only one tool would require at least a year of work.

Here, our intention is to show that it is relatively straightforward to construct examples for which the tools will fail and that, given the insights from the area of empirical software engineering, it is likely that they contain errors.

4 Numerical and Parallel Computing Challenges

Recent years have witnessed a substantial paradigm change in computer hardware and programming approaches with the introduction of multi-core architectures and accelerator systems such as graphics processing units (GPUs) and the Intel Xeon PHI many-core system. Such architecture-level advancements also have an impact on code verification, on debugging, and on the reproducibility of results. This is because the complexity of software development for parallel architectures requires an additional set of programming skills and also a distinct way of approaching algorithm design. For instance, parallel computing introduces an additional class of bugs, so-called race conditions. Race conditions are bugs that only occasionally appear in a nondeterministic fashion due to varying execution speeds among concurrent threads of execution, yielding parallel software harder to develop, test, and verify.

In addition, parallelization introduces serious complications with respect to the reproducibility of analyses. The main problem here is a numerical one. Suppose one intends to compute a sum over some floating point values f_i as $f = \sum_{i=1}^{n} f_i$. Further assume that the data for calculating these individual values f_i is distributed to a certain number of processors $p \leq n$. Then, the value f will typically be computed via a so-called parallel reduction operation as implemented, for instance, in the `MPI_Reduce()` collective communication routine of the Message Passing Interface (MPI) that still *is* the de facto standard for massively parallel computing. Now assume that $p := n/2$. In such a case, each processor will first add two values $f_i + f_{i+1}$ locally and then invoke `MPI_Reduce()` to communicate this partial result and calculate the overall sum f. If we now assume that $p := n/4$, each processor will initially add four values locally and subsequently invoke `MPI_Reduce()`. For $n := 8$ and $p := 4$, the sum might be computed as $\{\{(f_1 + f_2) + (f_3 + f_4)\} + \{(f_5 + f_6) + (f_7 + f_8)\}\}$ where the placement of the curly brackets {} denoting the parallel reduction depends, in fact, on the specific implementation of the `MPI_Reduce()` operation. Thus, the addition order induced by {} may also vary between different MPI implementations. As a consequence, even using different MPI implementations may induce distinct addition orders for the above sum. Hence, because of roundoff error propagation, one may obtain different results when executing code on the same computer system with the same number of processors, compiled with distinct MPI implementations. Furthermore, if $n := 8$ and $p := 2$, the sum is guaranteed to be executed in a distinct order since the partial sums will become larger: $\{(f_1 + f_2 + f_3 + f_4) + (f_5 + f_6 + f_7 + f_8)\}$. Thus, if the calculations carried out are numerically sensitive, as they mostly are in population genetics, since we typically operate with probabilities, we might obtain a signal for positive selection with two processors but not with four processors due to round-off error propagation. Note that this type of parallel reduction operation is very common in parallel codes and that analogous phenomena can be observed for multi-core parallel programming frameworks such as OpenMP (Open Memory Programming).

The problem of lacking associativity in floating point operations does not only apply to parallel programs though. Different compilers such as, for instance, `icc` and `gcc` may choose to reorder instructions (always assuming associativity) in different ways when code optimization flags (`-O` switch) are enabled which is typically the case. Thus, one might obtain different results across compilers or even for one and the same compiler when distinct optimization levels have been chosen via `-O`. We have, in fact, observed both phenomena (deviations within and across compilers) with RAxML (Stamatakis 2014). Thus, numerical deviations can also be easily observed for sequential codes, and caution is advised since such apparently small deviations may lead to a substantial divergence in the final result (see example provided in Darriba et al. 2018).

In conjunction with the above deviations, it also becomes extremely difficult to assess the correctness of numerical codes. If such a numerical deviation is detected, it is often unclear if it is a bug or indeed just a numerical deviation. While one can conduct a formal roundoff error analysis for an analytical mathematical equation, this is almost impossible as soon as numerical optimization routines are being used to optimize the value of that function. Thus, while one could determine a sufficiently exact value of a function by using arbitrary numerical precision libraries, as soon as this function needs to be optimized, there is no means to determine the expected or allowed variance/deviation of the optimized value.

5 Best Practices

There exist several ways in which software quality can be improved. The code analysis tools and criteria we have deployed in our analysis in Sect. 2 can and *should* be applied to all new software being developed. Also, software quality aspects as well as software analysis tools should receive more attention in programming courses for undergraduate and graduate students. In the programming courses we teach at the computer science department of the Karlsruhe Institute for Technology, we regularly apply the above criteria (usage of assertions, `valgrind`, `clang` compiler warnings, cyclomatic complexity analysis) for grading. In addition, as a community, we need to interact more intensively with software engineering researchers at computer science departments, since, after all, we are developing production level tools. Beyond the simple tools we have analyzed, there exists a plethora of more advanced software analysis tools such FindBugs for `JAVA` programs (http://findbugs.sourceforge.net/) or Cppcheck for `C++` codes (http://cppcheck.sourceforge.net/) that attempt to identify spurious code at a higher level that can be achieved by compilers. Some of these tools can also be integrated with `github`. Developers should keep in mind that investing some effort during initial program development will reduce the subsequent maintenance load. The main problem with this is that it is entirely unpredictable whether a prototype software one has developed will become a widely used bioinformatics tool or not.

Thus far, we have only discussed code quality, but not addressed code verification. A code quality assessment may only provide a notion as to whether a software tool is more or less likely to exhibit defects. Hence, the question arises how the actual code verification process could be improved. Firstly, standardized testing procedures should be applied. Secondly, the results of the tool should be, if possible, compared with competing codes implementing the same function, provided that such codes exist. Thirdly, in an ideal world, two independent teams should be working on developing two independent implementations simultaneously based upon the same specification whose outputs will then be systematically compared. This is how aircraft autopilots (usually with three independent teams using at least two distinct programming languages) are being developed. Evidently, we lack the time and funding for being able to apply most of the aforementioned techniques for software verification.

We believe that this is the main problem of bioinformatics software development. There is insufficient funding for sustainable development, reengineering, and maintenance of widely used software tools, given the tremendous citation impact such tools have but also the harm that can be done (including paper retractions) by software bugs. Thus, as a community, we need to (1) adopt a standard discipline of using software analysis tools, (2) put more emphasis on testing and verifying software, and (3) increase the pressure on funding agencies to implement actions for sustainable software development and maintenance.

Another important factor to consider is that, while the code is correct, the actual specification might be incorrect or incomplete. Examples for this are the aforementioned mathematical issues in Gotoh's pair-wise alignment algorithm (Flouri et al. 2015) or the erroneous Hastings correction for a widely used topological proposal mechanism in MCMC-based Bayesian inference of phylogenies (Holder et al. 2005) that was being used for several years until finally detected and corrected.

Another source of errors that might at least induce reproducibility problems is the usage of external libraries. Here, we consider library version management as being the main problem. If a code relies on some external libraries, it might yield distinct results depending on which version of the library happens to be installed. Thus, explicit library version management needs to be integrated into our tools to prevent this. When preparing a study on the impact of false positives for positive selection in population genetics (Pavlidis et al. 2012), we were, initially, not able to reproduce our own results. An intense search for the source of the problem revealed that a simulation tool we were using relied on a random number generator implemented in the widely used `boost C++` library. It turned out that a different version of `boost` was installed on the Linux system where we attempted to reproduce our results which generated a distinct sequence of random numbers for the same random number seed.

6 Future Perspectives and Conclusion

We have analyzed three highly cited population genetics tools and assessed their software quality which tends to be comparatively mediocre. While this does by no means imply that these three tools work incorrectly, the probability that they *do* contain bugs is high based on results from the field of empirical software engineering. In addition, even if correct, the tools might nonetheless experience failures, in particular on very large datasets, as we highlight by an appropriate, yet admittedly constructed, example.

One possible future direction for improving code quality is to make reviewers and editors of journals that have dedicated software tracks, for instance, *systematic biology, bioinformatics application notes, molecular ecology resources*, more aware of this issue. Thus, reviewers could be asked to conduct analyses similar to ours when reviewing software papers. We try to already apply this when reviewing such papers. Alternatively, authors could be asked to submit a code analysis report (including code duplication, warnings, results of `valgrind` analyses) together with their software papers. For standard programming languages such as `C/C++` or `JAVA`, such tests could, to a large extent, also be automated. Such a policy change would substantially improve awareness about code quality issues. In fact, we are currently working on developing an open-source tool for code quality checking that could be used for this purpose.

Another future direction is to emphasize the importance of code quality in graduate and undergraduate teaching. In conjunction with this, we also need to raise the awareness about software quality in the general user community, as we do with this chapter.

Finally, there is a substantial lack of funding for code development, despite the fact that widely used software packages contribute enormously to the citation records of entire departments. Hence, funding agencies should initiate additional and substantially more funding schemes for software development, redesign, and verification.

Acknowledgements This work was financially supported by the Klaus Tschira Foundation.

References

Alachiotis N et al. OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. Bioinformatics 2012;28(17):2274–5.

Barone L, Williams J, Micklos D. Unmet needs for analyzing biological big data: a survey of 704 NSF principal investigators. bioRxiv 2017. https://doi.org/10.1101/108555. http://biorxiv.org/content/early/2017/02/15/108555

Beerli P. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics 2006;22(3):341–5.

Beerli P, Felsenstein J. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 1999;152(2):763–73.

Beerli P, Felsenstein J. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. Proc Natl Acad Sci 2001;98(8):4563–8.
Beerli P, Palczewski M. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. Genetics 2010;185(1):313–26.
Briand LC, Wüst J, Ikonomovski SV, Lounis H. Investigating quality factors in object-oriented designs: an industrial case study. In: Proceedings of the 21st international conference on software engineering. New York: ACM; 1999. p. 345–54.
Briand LC, Wüst J, Daly JW, Porter DV. Exploring the relationships between design measures and software quality in object-oriented systems. J Syst Softw 2000;51(3):245–73.
Casalnuovo C, Devanbu P, Oliveira A, Filkov V, Ray B. Assert use in GitHub projects. In: Proceedings of the 37th international conference on software engineering - volume 1, ICSE '15. Piscataway: IEEE Press; 2015. p. 755–66. http://dl.acm.org/citation.cfm?id=2818754.2818846
Czech L, Huerta-Cepas J, Stamatakis A. A critical review on the use of support values in tree viewers and bioinformatics toolkits. Mol Biol Evol 2017;34(6):1535.
Darriba D, Flouri T, Stamatakis A. The state of software for evolutionary biology. Mol Biol Evol 2018;35(5):1037–46.
Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 2003;164(4):1567.
Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes 2007;7(4):574–8.
Fletcher W, Yang Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol Biol Evol 2010;27(10):2257.
Flouri T, Kobert K, Rognes T, Stamatakis A. Are all global alignment algorithms and implementations correct? bioRxiv (2015). https://doi.org/10.1101/031500. http://biorxiv.org/content/early/2015/11/12/031500
Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol 1982;162(3):705–8. https://doi.org/10.1016/0022-2836(82)90398-9. http://www.sciencedirect.com/science/article/pii/0022283682903989
Hoare CAR. An axiomatic basis for computer programming. Commun ACM 1969;12(10):576–80
Holder MT, Lewis PO, Swofford DL, Larget B. Hastings ratio of the LOCAL proposal used in Bayesian phylogenetics. Syst Biol 2005;54(6):961–5.
Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 2009;9(5):1322–32.
Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 2014;346(6215):1320–31.
Khoshgoftaar TM, Seliya N. Fault prediction modeling for software quality estimation: comparing commonly used techniques. Empir Softw Eng 2003;8(3):255–83.
McCabe TJ. A complexity measure. IEEE Trans Softw Eng 1976;SE-2(4):308–20.
Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science 2014;346(6210):763–7.
Nagappan N, Ball T. Static analysis tools as early indicators of pre-release defect density. In: Proceedings of the 27th international conference on software engineering, ICSE '05. New York: ACM; 2005. p. 580–6.
Pavlidis P, Jensen JD, Stephan W, Stamatakis A. A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. Mol Biol Evol 2012;29(10):3237–48.
Pavlidis P, Živkovic D, Stamatakis A, Alachiotis N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol Biol Evol 2013;30(9):2224.
Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155(2):945.

Raymond M, Rousset F. Genepop (version 1.2): population genetics software for exact tests and ecumenicism. J Hered 1995;86(3):248–9.

Redelings B. Erasing errors due to alignment ambiguity when estimating positive selection. Mol Biol Evol 2014;31(8):1979.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 2012;61(3):539–42.https://doi.org/10.1093/sysbio/sys029. http://sysbio.oxfordjournals.org/content/61/3/539.abstract

Rousset F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Resour 2008;8(1):103–6.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30(9):1312–3.

Part III
Concepts and Approaches

Population Epigenomics: Advancing Understanding of Phenotypic Plasticity, Acclimation, Adaptation and Diseases

Ehren R. V. Moler, Abdulkadir Abakir, Maria Eleftheriou, Jeremy S. Johnson, Konstantin V. Krutovsky, Lara C. Lewis, Alexey Ruzov, Amy V. Whipple, and Om P. Rajora

All authors have contributed significantly to the writing of this chapter and approve its final version. All authors except for the first and last are listed alphabetically by their last names.

E. R. V. Moler
Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA
e-mail: erm287@nau.edu

A. Abakir · M. Eleftheriou · L. C. Lewis · A. Ruzov
Wolfson Centre for Stem Cells, Tissue Engineering and Modelling (STEM), Division of Cancer and Stem Cells, School of Medicine, Centre for Biomolecular Sciences, University of Nottingham, Nottingham, UK
e-mail: Abdulkadir.abakir@nottingham.ac.uk; Maria.Eleftheriou@nottingham.ac.uk; Lara.Lewis@nottingham.ac.uk; Alexey.Ruzov@nottingham.ac.uk

J. S. Johnson
School of Forestry, Northern Arizona University, Flagstaff, AZ, USA
e-mail: jeremy.johnson@nau.edu

K. V. Krutovsky
Department of Forest Genetics and Forest Tree Breeding, Georg-August University of Göttingen, Göttingen, Germany

Department of Ecosystem Science and Management, Texas A&M University, College Station, TX, USA

Laboratory of Population Genetics, N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia

Genome Research and Education Center, Siberian Federal University, Krasnoyarsk, Russia
e-mail: konstantin.krutovsky@forst.uni-goettingen.de

A. V. Whipple
Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA
e-mail: amy.whipple@nau.edu

O. P. Rajora (✉)
Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, NB, Canada
e-mail: Om.Rajora@unb.ca

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_59,

Abstract Advances in chromatin state mapping, high-throughput DNA sequencing, and bioinformatics have revolutionized the study and interpretability of epigenomic variation. The increasing feasibility of obtaining and analyzing detailed information on epigenetic mechanisms across many individuals and populations has enabled the study of epigenomic variation at the population level and its contributions to phenotypic variation, acclimation, ecological adaptation, and disease traits. Over the past decade, researchers from disparate life sciences ranging from epidemiology to marine conservation have begun approaching their subjects through the lens of population epigenomics. Epigenetic mechanisms involve molecular alterations in chromatin through DNA methylation and histone modifications, as well as complex non-coding RNAs and enzyme machinery, all leading to altered transcription and post-transcriptional RNA processing resulting in changes in gene expression. Genetic and environmental variation and stochastic epimutations give rise to epigenomic variation. Notably, some forms of epigenomic variation are quite stable and in some instances may be transmitted through one or more rounds of meiosis. Epigenomic variation can contribute significantly to phenotypic plasticity, stress responses, disease conditions, and acclimation and adaptation to habitat conditions across a wide variety of organisms during their lifetime but also across multiple generations. The purpose of this chapter is to provide an overview of population epigenomics concepts, approaches, challenges, and applications. We discuss the molecular basis of epigenetic mechanisms and their variation and heritability across diverse tissues and taxa. We then discuss the sources of epigenomic variation, within – and among – population epigenomic variation in plants and animals, and the evolutionary context of epigenomic variation before reviewing current molecular and bioinformatics methods for screening epigenomic variation. We then explore the contribution and association of epigenomic variation with phenotypic and ecological adaptation traits in plants and common disease conditions in humans and pharmacoepigenomics, as well as the main challenges and future research directions in population epigenomics.

We emphasize challenges and potential solutions unique to the study of epigenomes and how those challenges are amplified by the diversity of pathways by which genes and environments can affect gene expression. With proper application and interpretation, the field of population epigenomics will continue to yield profound insights toward a better understanding of phenotypic plasticity, acclimation, ecological adaptation, heritability, human diseases, and pharmacogenomics.

Keywords DNA methylation · Epigenome-wide association study (EWAS) · Evolution · Histone modifications · Missing heritability · Non-coding RNAs · Pharmacoepigenomics · Phenotypic plasticity · Population epigenomics · Source and heritability of epigenomic variation

1 Introduction

Epigenetics is the study of potentially heritable changes in gene expression that are not strictly due to nucleotide changes such as substitutions, insertions and deletions (indels), or other rearrangements of the underlying DNA sequence. Diverse epigenetic mechanisms detected across all three domains of life are characterized by genetically, environmentally, and developmentally mediated molecular phenotypes that may trigger or result from cell differentiation and development, and which demonstrate varying degrees of heritability through mitosis and meiosis (Cortijo et al. 2014; Heard and Martienssen 2014). Epigenetic mechanisms involve molecular alterations in chromatin through DNA methylation and histone modifications and transcriptional and translational interference via non-coding RNAs (Fig. 1; Johnson and Tricker 2010), leading to altered transcription and post-transcriptional RNA processing resulting in changes in gene expression. Epigenomics is the investigation of the interactions among multiple epigenetic mechanisms at the genome-wide level and how they interact with the genome to influence chromatin

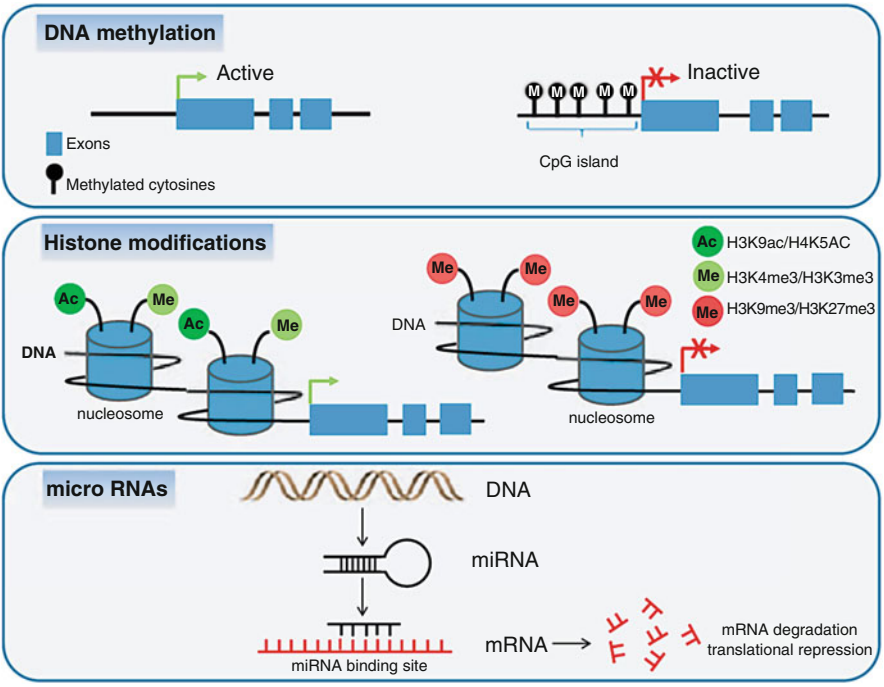

Fig. 1 The three most commonly investigated epigenetic mechanisms affecting gene expression are DNA methylation, histone modifications, and non-coding RNA. Reproduced with permission from D'addario et al. (2013)

function and gene expression. Epigenomic variation can result from genetic and environmental factors, as well as from stochastic epimutations (see review in Taudt et al. 2016).

In a series of papers published in 1942, Conrad Waddington presented the concept of the "epigenotype" to describe processes of gene regulation suspected to influence cell differentiation and phenotypic plasticity (Jamniczky et al. 2010; Waddington 2012; Deans and Maggert 2015). While definitions of epigenetics now often involve a heritability component following the popularization of that association by Holliday in 1994, there is still no ultimate consensus of what constitutes, and thus how to study, epigenetic phenomena (Richards 2006; Deans and Maggert 2015) nor to what degree epigenetic mechanisms and their effects are heritable (Pecinka and Scheid 2012; Furrow 2014; Whipple and Holeski 2016). Irrespective of their heritability, epigenetic mechanisms are indispensable for the development and survival of most organisms (Zemach and Zilberman 2010).

Population epigenetics was described by Richards (2008) as "emerging as an active subfield at the interface of molecular genetics, genomics, and population biology, [that] addresses questions concerning the prevalence and importance of epigenetic variation in the natural world." With the development of massive high-throughput parallel sequencing techniques to assay genome-wide epigenetic marks, such as bisulfite DNA sequencing, epigenomics has progressed from investigating individual epigenomes to studying epigenomic variation across populations and species, leading to the research field of population epigenomics, which is now a rapidly growing field of basic and applied research.

By distinguishing the contribution of the epigenome to the variation in traits and gene expression in and among populations, the field of population epigenomics is unravelling the complexities of the evolutionary process and revolutionizing biotechnological approaches for improving human health and the environment. The broad utility of these methods for interrogating non-genetic sources of phenotypic variation draws researchers from across the life sciences to consider the role of the epigenome in their respective study systems. This has resulted in notable outcomes, such as important discoveries in human disease processes (Ling and Groop 2009; Rodríguez-Paredes and Esteller 2011), environmental toxicology (Birney et al. 2016; Martin and Fry 2018), novel advances in stem cell therapy (Lunyak and Rosenfeld 2008; Atlasi and Stunnenberg 2017), new approaches in molecular breeding for the improvement of agronomic crops (King et al. 2010; Zheng et al. 2017), and ambitious concepts and biotechnologies for the conservation of species and ecosystems in the face of a rapidly warming global climate (Sáez-Laguna et al. 2014; Van Oppen et al. 2017). These advances will undoubtedly reveal new challenges in the study of epigenomes, as will studying the role of epigenomes at increasingly complex levels of biological organization.

The objective of this chapter is to provide a discussion of population epigenomics concepts, methods, challenges, and applications. First, we discuss the molecular basis of epigenetic phenomena and their taxonomic diversity, tissue specificity, and heritability. We then examine the evolution and sources of epigenomic variation before discussing epigenomic variation within and among populations of plants and

animals. Thereafter, we provide an overview of methods used to measure epigenomic variation and bioinformatics methods for analyzing population epigenomics data. Subsequently, we review the influence and association of epigenomic variation with phenotypic and ecological acclimation and adaptation traits in plants, common disease conditions in humans, and pharmacoepigenomics. Lastly, we discuss challenges, research needs, and future directions in population epigenomics.

2 The Molecular Basis of Epigenetic Phenomena

2.1 Epigenetic Mechanisms

DNA methylation, histone modifications, and non-coding RNA are the most well-studied epigenetic mechanisms. Chromatin is thought to be at the core of epigenetic gene regulation, affecting gene expression patterns and ultimately the phenotype via changes in accessibility of the DNA to transcription factors (Chen et al. 2017). The nucleosome, the basic building block of chromatin, is comprised of approximately 147 bp of negatively charged DNA wound twice around a histone octamer consisting of heterodimers of H3/H4 and H2A/H2B histones (Hansen 2002). The N-terminus of a histone molecule is positively charged and contains numerous lysine and arginine residues that interact with negatively charged DNA, limiting its accessibility to transcription factors (Peterson and Laniel 2004). The bulk of genomic DNA is incorporated into the nucleosome with around 10–60 residues acting as a linker region connecting subsequent nucleosomes together (Hansen 2002). Compaction of these nucleosome units produces structures of approximately 10 nm in diameter known as chromatin fibers (Hansen 2002). Like origami, these chromatin fibers are condensed further, firstly into 30 nm fibers, then into 100–400 nm interphase filaments, and finally into chromosomes (Peterson and Laniel 2004). Organization of these chromatin structures can be altered by DNA methylation, histone modifications, and non-coding RNAs that collectively define chromatin states allowing for either expression (euchromatin) or repression (heterochromatin) of different genes (Allis and Jenuwein 2016). DNA methylation can result in the compaction of chromatin, and small RNA can direct DNA methylation to a specific genomic region via RNA-directed DNA methylation (RdDM) (review in Bernstein and Allis 2005). Chromatin compaction is known to suppress gene expression by inhibiting the accessibility to DNA by transcription machinery. Importantly, most chromatin modifications are reversible (Allis and Jenuwein 2016). The dynamics of the patterns of chromatin modifications enables biological processes, such as development, differentiation, acclimation and adaptation (Taudt et al. 2016). Organisms from different branches of the tree of life can vary in these mechanisms, some lineages having lost entire pathways (Zemach and Zilberman 2010).

2.1.1 DNA Methylation

One of the most frequently studied chromatin modifications in plant, animal, and fungal genomes is the covalent addition of a methyl group to the fifth carbon of the cytosine pyrimidine ring, leading to the generation of 5-methylcytosine (5mC) (Holliday and Pugh 1975; Riggs 1975; Law and Jacobsen 2010). The human genome contains approximately 28 million CpG dinucleotides, of which 60–80% are methylated (Taudt et al. 2016). Although different tissues are characterized by various levels of CpG methylation, CpG islands located in the proximal regions of gene promoters generally remain unmethylated in animal genomes (Ehrlich et al. 1982).

In mammalian genomes, there are three DNA methyltransferase (DNMT) enzymes, which add methyl groups to DNA (Edwards et al. 2017). DNMT1, the maintenance methyltransferase shared across numerous lineages, has an affinity for hemimethylated CpGs that are generated following DNA replication or during DNA damage repair (Bostick et al. 2007). Ubiquitin-like containing PHD and RING finger domains 1 (UHRF1) recruits DNMT1 to hemimethylated DNA, where this enzyme reproduces the pattern of DNA methylation present on the original strand of DNA onto the newly synthesized strand (Bostick et al. 2007). DNMT3a and DNMT3b are de novo DNA methyltransferases that, along with catalytically inactive DNMT3L, methylate cytosine residues in "naked" unmethylated DNA (Okano et al. 1999). Although DNMT1 and DNMT3a/b are, respectively, designated as maintenance and de novo methyltransferases, these functions are not mutually exclusive (Okano et al. 1999; Fatemi et al. 2002).

Global levels of DNA methylation are relatively static in most tissues. However, during cellular differentiation the DNA methylation status of a fraction of all CpGs in the genome exhibits dynamic changes that modulate tissue-specific gene expression (Gifford et al. 2013; Ziller et al. 2013). These alterations in the patterns of DNA methylation are influenced by the chromatin state through a cross-talk between methylation/demethylation machinery and histone modifications including H3K9me and H3K4me (Cedar and Bergman 2009; Du et al. 2015). The removal of 5mC from DNA can occur either through passive or active demethylation (Smith and Meissner 2013). Passive demethylation is replication-dependent. During this process, 5mC is being "diluted out" with each successive round of replication in the absence of DNMT1 and/or UHRF1 (Wu and Zhang 2014). Active DNA demethylation was initially described in plants where Demeter (DME)/repressor of silencing 1 (ROS1) family of DNA glycosylases mediate removal of 5mC when coupled with base excision repair (BER) machinery. In addition to this, a direct removal of the methyl group from 5mC has also been reported in plants; however, this process is thought to be thermodynamically unfavorable in mammalian cells (Zhu 2009).

5-Hydroxymethylcytosine (5hmC), an oxidized derivative of 5mC initially identified in bacteriophages (Wyatt and Cohen 1952), was later found in non-negligible quantities in the mouse genome (Kriaucionis and Heintz 2009). The ten-eleven translocation (TET) family of DNA dioxygenase enzymes (TET1–3) was shown to

catalyze the conversion of 5mC into 5hmC (Tahiliani et al. 2009). Further oxidation of 5hmC then forms 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) by TET proteins (Ito et al. 2011; He et al. 2011). 5hmC may facilitate passive dilution of 5mC by impairing binding of DNMT1/URHF1 to the hemi-modified DNA (Wu and Zhang 2014). Moreover, thymine-DNA glycosylase (TDG) can recognize and excise 5caC and 5fC from DNA (Maiti and Drohat 2011; He et al. 2011). Analogously to DME/ROS1-mediated demethylation in plants, TDG-driven excision of 5caC and 5fC produces an abasic site that can be repaired by BER machinery resulting in regeneration of non-modified cytosine (Chen and Riggs 2011). In addition to their roles as intermediates in the processes of active and passive DNA demethylation, according to multiple studies, the oxidized forms of 5mC (5hmC, 5caC, and 5fC) may have their own functional epigenetic significance (Song and He 2013). Thus, accumulation of 5fC and 5caC at cell-type-specific promoters, which correlates with transcriptional activity of the corresponding genes, has been observed during glial/neural and hepatic differentiation, implying a potential role of these modifications in regulation of gene expression (Wheldon et al. 2014; Lewis et al. 2017). The TET/TDG/BER-dependent pathway of active demethylation is most documented for mammalian systems to date.

DNA methylation is usually associated with transcriptional repression and has been linked to a plethora of biological processes, including X chromosome inactivation, genomic imprinting, heritable repression of retrotransposons, pluripotency regulation, and gene silencing in development and disease (Edwards et al. 2017; Iurlaro et al. 2017). In addition to silencing of coding genes, transposon-derived sequences, such as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs), are often heavily methylated in mammals (Edwards et al. 2017). Interestingly, in plants, DNA methylation also occurs predominantly on repetitive sequences, including transposons (Zhang et al. 2006). Given that transposable elements are a significant threat to the genome integrity due to their ability to replicate and integrate randomly throughout the genome, their tight regulation is of particular importance for the heritable transfer of genetic information (Fedoroff 2012). Correspondingly, DNA methylation represents one of the main mechanisms allowing to maintain repetitive elements in a silenced state in both plants and animals (Law and Jacobsen 2010).

Genomic imprinting is an epigenetic process defined by the expression of genes in a parental-origin-specific manner (Ferguson-Smith 2011). Imprinting was initially discovered while studying the inheritance of maize (*Zea mays*) kernel coloration, when specific phenotypes were attributed to the parental germline environment of a gene instead of differences in its DNA sequence (Kermicle 1970). Imprinting has been reported for mammals, plants, and insects (Kermicle 1970).

X-inactivation is an example of whole chromosome imprinting whereby one of the female X chromosomes is silenced to equalize its transcriptional output to the male XY (Plath et al. 2002). Although X chromosome inactivation is instigated by the ncRNA *Xist*, DNA methylation is central to maintaining its inactive state (Csankovszki et al. 2001). In addition to the continued expression of *Xist* and

deacetylation of histones, de novo methylation of CpG islands is required for permanent silencing of the X chromosome (Bird 2002).

Apart from locus-specific changes in 5mC content, two events of genome-wide DNA demethylation and remethylation occur in mammalian development (Edwards et al. 2017). One wave of global genome demethylation is observed upon migration of dividing primordial germ cells toward developing gonads, and the second demethylation event occurs in cleavage stage embryos soon after fertilization (Monk et al. 1987; Edwards et al. 2017). These waves of genome-wide demethylation and subsequent remethylation are currently understood as reprogramming events and are correlated with the loss of cellular memory and the resetting of cellular potency (Iurlaro et al. 2017). In contrast, for plants, studies in *Arabidopsis* have shown that in pollen, the germline cells do not undergo erasure of DNA methylation (Slotkin et al. 2009). Rather DNA methylation is lost in the vegetative nucleus, resulting in re-expression of transposons and the production of small RNAs. It has been demonstrated that these small RNAs can travel to the germline cells, where it is suggested they reinforce methylation states (Slotkin et al. 2009).

Although methylation of cytosine is the most abundant and well-studied DNA modification, adenine within DNA has been shown to be methylated in some instances (N^6-methyladenine, 6mA) (O'Brown and Greer 2016). Until recently, the presence of 6mA in DNA had been described only for prokaryotes in the context of host defense mechanisms (Vanyushin et al. 1968). However, since 2015, a number of studies have documented the presence of this mark in plants, insects, and mammals (Fu et al. 2015; Greer et al. 2015; Zhang et al. 2015a, b; Liu et al. 2016; Wang et al. 2017; Xiao et al. 2017). Despite indications of its possible involvement in the regulation of transcription, activity of transposable elements, embryo development, and inheritance in these systems, potential functional roles of this DNA modification in eukaryotes remain to be elucidated (Luo et al. 2015; Sun et al. 2015; Luo and He 2017).

2.1.2 Non-coding RNAs (ncRNAs)

A majority of the non-protein-coding transcripts produced from a genome are functionally active as RNA molecules and play numerous regulatory roles in the cell (Uchida and Dimmeler 2015). These RNAs, known as non-coding RNAs (ncRNAs), are functional transcripts that are not translated into proteins. Classification of ncRNAs is often based on size, dividing them into small (<30 nt) and long (>200 nt) transcripts (Uchida and Dimmeler 2015). Small ncRNAs (sncRNAs) include microRNAs (miRNAs), small interfering RNAs (siRNAs), Piwi-interacting RNAs (piRNAs), transfer RNAs (tRNAs), and small nucleolar RNAs (snRNAs). Ribosomal RNAs (rRNAs) and natural antisense transcripts (NATs) are within the scope of long ncRNAs (lncRNAs); however numerous other lncRNAs exist (Chen et al. 2017).

Micro RNAs (miRNAs)

Almost 40 thousand types of miRNAs have been discovered, and these evolutionarily conserved single-stranded RNAs (20–24 nucleotides long) are thought to be involved in many important biological processes by regulating the expression of approximately half of all genes in a cell post-transcriptionally (Kaikkonen et al. 2011). Although some miRNAs are transcribed from independent loci, most miRNAs are clustered and are transcribed as a part of a single polycistronic unit, most commonly from intergenic regions of the genome (Karius et al. 2012; Uchida and Dimmeler 2015).

Mature miRNA integrates with the RNA-induced silencing complex (RISC) in order to guide its binding to the 3′ untranslated region (UTR) of target mRNA (Bartel 2004). The degree of base pairing between the mature miRNA seed sequence and the target mRNA 3′UTR determines either repression or degradation of the corresponding mRNA (Li et al. 2010). In case of perfect complementarity between the seed sequence and mRNA 3′UTR, the Argonaute protein cleaves the resulting complex, whereas non-perfect complementarity usually leads to translational inhibition of mRNAs (Kaikkonen et al. 2011). Thus, via mRNA targeting, miRNAs can modulate the gene expression patterns for hundreds of different targets and, consequently, influence many biological processes, such as proliferation, differentiation, and metabolism (Uchida and Dimmeler 2015).

Small Interfering RNAs (siRNAs)

miRNAs and siRNAs are similar in many aspects, including their size (20–24 nt) and ability to associate with the RISC complex to silence gene function. However, they have divergent origins and biogenesis pathways (Kaikkonen et al. 2011). Both require Dicer for processing and the Argonaute family of proteins to support their silencing abilities, but siRNAs do not rely on Drosha, a class 2 ribonuclease III enzyme, and are mainly processed from long, linear, fully complementary dsRNAs as opposed to the stem-loop precursors described for miRNAs (Carthew and Sontheimer 2009; Kim et al. 2009). Analogously to miRNAs, the extent of complementarity between siRNA and its target determines the particular mode of siRNA-dependent silencing, but most siRNAs almost exclusively mediate cleavage and degradation of their target mRNAs (Kaikkonen et al. 2011).

Initially, only exogenous siRNAs were considered as a primitive form of genome defense that act in response to foreign nucleic acids including viruses, transposons, and transgenes (Kaikkonen et al. 2011); however, it soon became apparent that endogenous siRNAs transcribed from loci containing transposons and repetitive elements could, similar to exogenous siRNA and piRNAs, contribute to the suppression of transposon activity (Carthew and Sontheimer 2009). Interestingly, siRNAs have also been associated with sequence-specific silencing through the upregulation of epigenetic marks that induce formation of heterochromatin (Kaikkonen et al. 2011).

PIWI-Interacting RNAs (piRNAs)

piRNAs are 24–31 nucleotides long and are characterized by a 2′-O-methyl modification at the 3′, as well as a preference for uridine at the 5′ (Siomi and Siomi 2009). Unlike miRNAs and siRNAs, piRNAs are processed from a single-stranded precursor transcript (Vagin et al. 2006). piRNAs form effector complexes known as piRNA-induced silencing complexes (piRISCs) with PIWI proteins belonging to a germline-specific subclass of the Argonaute family (Iwasaki et al. 2015). piRNAs are transcribed from piRNA clusters, intergenic regions containing large numbers of different transposons (Kaikkonen et al. 2011; Iwasaki et al. 2015). Initially, piRNAs were identified in *Drosophila*, where they are complementary to numerous transposable and repetitive elements (Aravin et al. 2003). Correspondingly, piRNAs act primarily as the essential regulators of transposon activity within the genome during germline development (Iwasaki et al. 2015). Interestingly, piRNAs have also been linked to transposon regulation in somatic cells (Li et al. 2009a; Malone et al. 2009). As piRNAs are transcribed from loci that are similar to their targets, to successfully regulate transposition, they need to recognize their "self" genes from "non-self" transposable elements that are to be targeted (Malone and Hannon 2009). A combination of diversity in the sequences for target transposons and in piRNA processing mechanisms makes these RNAs one of the most diverse and the largest subgroups of ncRNAs (Siomi et al. 2011).

Long Non-coding RNAs (lncRNAs)

Unlike highly conserved sncRNAs that regulate gene silencing through specific base pairing, long non-coding RNAs (lncRNAs) have low-level sequence conservation and use diverse mechanisms of regulation which are not yet fully characterized. Similar to protein-translating mRNA, lncRNAs are transcribed by RNA polymerase II (Wang and Chang 2011). They are often 5′ capped and spliced and contain a 3′-polyadenylated tail (Chen et al. 2017). Unlike protein-coding genes, lncRNAs lack open reading frames (ORF); their encoded RNA sequences are shorter, and the abundance of the expressed transcripts is lower compared with mRNAs (Wang and Chang 2011). lncRNAs are enriched in the nucleus compared to the cytoplasm, and their expression is highly cell type, tissue type, and developmental stage-specific (Chen et al. 2017).

lncRNAs are commonly classified according to genomic location as sense, antisense, intronic, intergenic, enhancer, and circular RNAs (Uchida and Dimmeler 2015). Sense lncRNAs usually share the same promoter and overlap with a protein-coding transcript, whereas antisense lncRNAs are present in the strand opposite to a protein-coding gene (Uchida and Dimmeler 2015). Intronic lncRNAs are transcribed from the introns of a translated gene, and long intergenic non-coding RNAs (lincRNAs) can be found between two transcribed genes. Enhancer RNAs (eRNAs) are produced from enhancer regions of protein-coding genes, and circular RNAs are usually formed following the splicing of a protein-coding gene whereby

the product covalently binds to itself (Uchida and Dimmeler 2015; Chen 2016). lncRNAs have also been grouped according to their function as imprinting-related, scaffolds, enhancer activation, and molecular sponges (Uchida and Dimmeler 2015) but as individual lncRNAs may fulfill several biological roles; therefore, these groups are not mutually exclusive (Wang and Chang 2011).

The lncRNA *Xist*, the first functionally characterized lncRNA involved in imprinting, silences one of two X chromosomes (Brown et al. 1991; Herzing et al. 1997). *Xist* silences one of the female XX chromosomes to equalize its transcriptional output compared to male XY (Plath et al. 2002). On the active X chromosome, *Xist* is silenced in the *cis* position (self-inactivation); however, on the inactive X chromosome, *Xist* is activated both in *cis* and *trans* positions (non-self-inactivation) (Chen et al. 2017).

When a lncRNA acts as a scaffold, it directs different biological activities through the recruitment of additional functional proteins (Uchida and Dimmeler 2015). Scaffold lncRNAs represent the most abundant subgroup of these RNAs, a type of lincRNA consisting of more than 10,000 molecular species (Chen 2016). Unlike most lncRNAs, lincRNAs are highly evolutionarily conserved across different species (Guttman et al. 2009). lincRNAs have a distinctive chromatin signature; their promoter and transcribed regions are marked by trimethylated lysines 4 (H3K4me3) and 36 (H3K36me3) of histone 3, and both are associated with actively transcribed genes (Khalil et al. 2009; Guttman et al. 2009). Current experimental evidence suggests that lincRNAs act as flexible scaffolds, guiding chromatin-modifying complexes to particular loci within the genome, enabling the creation of cell-type-specific epigenetic states and instigating different transcriptional programs (Tsai et al. 2010; Guttman et al. 2011).

The eRNAs are a group of lncRNAs transcribed from enhancers (Uchida and Dimmeler 2015). eRNAs modulate enhancer activation and range in size from 0.1 to 9 kb (Kim et al. 2010; Kaikkonen et al. 2011). Similar to other lincRNAs, eRNAs are evolutionarily conserved and have a distinct chromatin signature (Heintzman et al. 2007). eRNA-producing regions are usually characterized by high enrichment of monomethylated (H3K4me1) and low content of trimethylated lysine 4 on histone 3 (H3K4me3) (Heintzman et al. 2009). As the initiation of eRNA transcription occurs from RNA polymerase II binding sites, followed by bidirectional elongation of the transcript, eRNA expression levels positively correlate with those of nearby mRNAs (Kim et al. 2010; Chen et al. 2017).

Unlike eRNAs, molecular sponges regulate gene expression via sequestering molecules that interact with a particular region of the genome (Chen et al. 2017). Circular RNA (circRNA) arising from introns or protein-coding exons via linking their 3′ and 5′ ends commonly acts as molecular sponges (Zhang et al. 2013a, b; Jeck et al. 2013). Although the overall range of biological roles of circRNAs is still rather unclear, one of their known functions is the sequestration of miRNAs (Uchida and Dimmeler 2015). A circRNA containing more than 70 conserved miRNA target sites known as ciRS-7 was shown to act as a sponge for miR-7 in both human and mouse brain (Hansen et al. 2013). ciRS-7 is strongly associated with Argonaute proteins in

a miR-7-dependent manner and upregulates miR-7 target levels by suppressing miR-7 activity (Hansen et al. 2012, 2013).

2.1.3 Histone Modifications

Histones are the core components of the nucleosome. They are of fundamental importance for the epigenetic regulation of chromatin structure and undergo a large number of chemical modifications that are considered epigenetic (Hansen 2002). Although the post-translational modifications of histones were known since the early 1960s (Allfrey et al. 1964), histone modifications were functionally linked with chromatin structure only in 1997 after achieving a high-resolution X-ray structural determination of the nucleosome (Luger et al. 1997). The 20–35 residue amino-N-terminal histone tail extends from the nucleosome unit enabling its interaction with neighboring nucleosomes and is instrumental in the folding of nucleosomes into higher-order chromatin fibers (Peterson and Laniel 2004; Bannister and Kouzarides 2011). Histones possess more than 130 post-translational modifications (PTMs) that include acetylation, methylation, phosphorylation, sumoylation, ubiquitination, deamination, beta-*N*-acetylglucosamine, ADP ribosylation, histone tail clipping, and histone proline isomerization (Bannister and Kouzarides 2011; Rivera and Ren 2013). Most of these PTMs are observed in both the amino and carboxyl terminal tails of histones; however, central histone domains can also be modified (Bannister and Kouzarides 2011). In 2000, the "histone code" hypothesis, stating that the combined nature of different histone modifications defines different combinatorial chromatin states, was proposed in several studies (Strahl and Allis 2000; Jenuwein and Allis 2001). According to this hypothesis, the patterns of histone modifications present at defined locations in the genome can be interpreted by other proteins resulting in a specific downstream event (Strahl and Allis 2000). Although consensus has not been achieved for what the "histone code" actually means, it is generally assumed that histone modifications contribute to control of gene expression via either structural changes of chromatin or recruitment of transcription factors, coactivators, and suppressors in order to achieve active, poised, or silenced transcriptional states of the corresponding genes (Peterson and Laniel 2004; Bannister and Kouzarides 2011; Chen et al. 2017).

2.2 Taxonomic Diversity of Epigenetic Patterns

Although epigenetic mechanisms play a key role in the evolution of phenotypic and functional biological diversity in myriad animal and plant taxa, and are conserved across a wide range of species, most fungi and invertebrate animals investigated so far appear to make less use of DNA methylation than plants and animals (Zemach and Zilberman 2010; Zhong 2016; Yung and Elsässer 2017). Importantly, many post-translational modifications originated in prokaryotes as metabolic intermediates

and acquired an "epigenetic" role only in multicellular organisms (Yung and Elsässer 2017). Thus, comparative analysis of epigenetic modifications between different species may provide an insight into both biological roles of epigenetic marks and the evolution of specific developmental processes (Xiao et al. 2014; Roadmap Epigenomics Consortium et al. 2015; Zhong 2016; Hardcastle et al. 2018).

Cytosine DNA methylation (5mC) is a major epigenetic modification commonly found in plants, animals, and fungi (Yung and Elsässer 2017). Its global levels vary across different eukaryotes, with the amount of cytosine residues that are methylated representing 0–3% of all the cytosine residues in the genome of insects, 5% in mammals and birds, 10% in fish and amphibians, and sometimes more than 30% in the genomes of certain plants (Field et al. 2004). Unlike in mammals, in insects, 5mC is enriched in the gene bodies of actively transcribed loci, where it is involved in the control of gene expression (Yan et al. 2015; Jaenisch and Bird 2003). In the honeybee (*Apis* spp.), gene body methylation has also been linked with alternative splicing (Wedd and Maleszka 2016), and an intriguing though disputed finding from studies of eusocial insects linked 5mC variation with development of different castes and behavioral patterns (Yan et al. 2014, 2015).

Extensive 5mC variation exists in plants, mediated by a suite of plant-specific methyltransferases, and correlates strongly with the distribution of transposable elements across a given genome (Niederhuth et al. 2016; Bewick et al. 2017). Plants with a more complex genome have a wide distribution of 5mC that, like in mammals, contributes to preventing transposition of repetitive elements (Zemach et al. 2010). Similar to insects, plants contain 5mC in a CG context, but also in CHG and CHH contexts (where H is any base), within approximately a third of gene bodies of actively transcribed protein-coding genes (Cokus et al. 2008; Lister et al. 2008). CHG methylation is specifically mediated by CHROMOMETHYLASE 3, while CHH methylation is specifically mediated by CHROMOMETHYLASE 2. CG, CHG, and CHH methylation also share a common mediator: DOMAINS REARRANGED METHYLASE 2. 5mC content and distribution vary significantly across plant species and have been lost altogether within some algal species (Bewick et al. 2016, 2017). Thus, *Chlorella* possesses a highly methylated genome, whereas *Volvox* contains only low levels of DNA methylation (Lister et al. 2008).

Importantly, all of the DNMTs in eukaryotes are highly homologous to bacterial DNA methyltransferases (Goll and Bestor 2005). While in mammals, de novo methylation is mediated by the DNMT3 class of enzymes, their plant counterparts belong to the DOMAIN REARRANGED METHYLTRANSFERASE 2 (DRM2) protein family (Law and Jacobsen 2010; Zhang et al. 2018). Targeting of these enzymes to DNA significantly differs between these organisms. Mammalian DNMT3 is recruited to chromatin through its association with histones; however, DRM2 is targeted to the DNA via siRNAs through RdDM (Law and Jacobsen 2010; Zhang et al. 2018). RdDM is a major mechanism of DNA methylation in plants that is generated via a different pathway in plants than in other eukaryotes, consisting of 24 nt small RNAs produced by two RNA polymerases specific to plants: Pol IV interacting with non-coding RNA produced by Pol V to target DOMAINS REARRANGED METHYLASE 2 (Matzke and Mosher 2014). There is strong

evidence of genetic variants in these plant-specific pathways associated with population level differences in DNA methylation (Schmitz et al. 2013b).

The most common DNA modification in prokaryotes is adenine methylation (6mA), while the role of 5mC in bacteria is rather poorly understood (Vanyushin et al. 1968; Breiling and Lyko 2015). There are two types of 6mA methyltransferases in bacteria: restriction-modification systems, which protect the prokaryotic host from the invasion of foreign (phage) DNA, and solitary methyltransferases, e.g., *Dam* (Wion and Casadesús 2006). In bacterial genomes, 6mA, in combination with solitary methylases, is implicated in influencing virulence of diverse human and animal pathogens as well as providing signals for DNA-protein interactions (Vanyushin et al. 1968; Low et al. 2001; Casadesús and Low 2006; Kahramanoglou et al. 2012).

In addition to bacterial genomes, 6mA has also recently been identified in a wide range of multicellular organisms including *Arabidopsis*, *Chlamydomonas*, *Drosophila*, *C. elegans*, *Tetrahymena thermophila*, rice (*Oryza sativa*), zebrafish (*Danio rerio*), pig (*Sus scrofa*), and *Homo sapiens* (Fu et al. 2015; Greer et al. 2015; Zhang et al. 2015a, b; Liu et al. 2016; Wang et al. 2017; Xiao et al. 2018; Zhang et al. 2018). Unlike that of 5mC, the function of 6mA in these organisms is currently largely unclear. However, it does seem to correlate with activation and/or silencing of genes in certain biological systems studied to date (Luo et al. 2015; Sun et al. 2015; Luo and He 2017).

Several recent reports have provided experimental evidence for epigenetic-like functional roles of the active demethylation intermediates 5hmC, 5caC, and 5fC (Song and He 2013). The generation of 5caC and 5fC in plant genomes in response to environmental stresses has also been reported (Tang et al. 2014), but not 5hmC due to the lack of TET enzymes in plants. Moreover, TET homologues and the oxidized derivatives of 5mC have also been detected in fungi *Coprinopsis cinerea*, but their biological roles in these organisms are yet to be elucidated (Zhang et al. 2014).

ncRNAs are present not only in eukaryotes but also in bacteria, archaea, and viruses (Storz 2002). Numerous early ncRNA studies were carried out in unicellular eukaryotes (Volpe et al. 2002; Mochizuki et al. 2002). These sncRNAs are among the most highly conserved sequences in vertebrate genomes, whereas the lncRNAs have limited evolutionary conservation (Pang et al. 2006; Pollard et al. 2006). One of the most ancient sncRNAs is the hammerhead ribozyme. This catalytic RNA was discovered in subviral plant pathogens in addition to archaea, bacteria, and eukaryotic genomes (Przybilski et al. 2005; Seehafer et al. 2011). Endogenous siRNAs have also been observed in different species such as plants, worms, flies, and mammals. However, the complexity of the siRNA biogenesis pathway is not equal in all organisms (Kim et al. 2009). Within humans, identification of endogenous siRNAs is limited (Xia et al. 2013). Bacterial siRNAs are not related to eukaryotic small RNAs. Unlike bacterial siRNAs, endogenous siRNA of eukaryotes is 20–30 nucleotides long and specifically associates with the Argonaute family of proteins (Kim et al. 2009). Interestingly, the high degree of base pairing between miRNAs and their target mRNAs leading to target degradation is more commonly observed in

plants, whereas miRNA-dependent translational repression most commonly occurs in humans due to imperfect complementarity (He and Hannon 2004).

Unlike eukaryotes, archaea do not contain complex nucleosomal units, but prototypic histones are associated with DNA packaging in these prokaryotes (Mattiroli et al. 2017). Moreover, the amino acid sequences forming the contacts between nucleosomal histones and DNA are conserved between archaea and eukaryotes indicating an important functional role for these interactions (Mattiroli et al. 2017). The core histone fold, regulatory sites on the histone tail, and histone PTMs are evolutionarily conserved in protozoans, and most of the components of the machinery involved in the regulation of PTMs also appear to possess some degree of evolutionary conservation (Postberg et al. 2010; Talbert et al. 2012). Lysine acetylation is present in all kingdoms of life; sirtuin deacetylases are conserved between eukaryotes, archaea, and bacteria; and the catalytic core of the SET domain of lysine methyltransferases can be found in prokaryotic (bacterial) proteins (Soppa 2010; Alvarez-Venegas 2014; Yung and Elsässer 2017). Furthermore, the donor molecules employed by the eukaryotes for the generation of some PTMs (e.g., acetyl CoA and ATP) serve as intermediates involved in metabolic feedback regulation in prokaryotes (Sharma and Rando 2017).

Thus, despite the diverse roles of specific epigenetic marks in various species, most of the basic epigenetic mechanisms are of very ancient origin, and, therefore, elucidating their roles in different contexts should be of immense interest for understanding the most fundamental principles of the homeostasis and development of biological systems.

2.3 Cell and Tissue Specificity of Epigenetic Patterns

Epigenetic memory refers to the transmission of gene expression states through multiple generations of a cell line, independent of initiation signals or genetic variation (Ng and Gurdon 2008). With few exceptions, the hundreds of cell types present in a multicellular eukaryote contain identical genomes, yet the functions they perform differ substantially (Watanabe et al. 2013). This functional variation is facilitated by changes in gene expression resulting from enhancer-promoter interactions, chromatin assembly, transcription factors, transposable element mobilizations, and attendant epigenomic modifications (Li et al. 2016). Mechanisms of gene regulation both reflect changes in cellular environments and ontogeny, and collectively act to progressively silence transcriptionally active gene regions as cell differentiation proceeds. Once a developing cell becomes committed to a cell fate, it cannot switch to another cell fate, in part due to accumulated epigenetic modifications that buffer cell differentiation (Hochedlinger and Plath 2009; Takahashi et al. 2018). Elimination of the epigenetic memory accumulated during cellular differentiation, exposure to conducive cellular environments, and induction of embryonic transcription factor network expression have allowed stem cell researchers to reprogram a wide range of differentiated somatic cells to a pluripotent state (Lunyak and Rosenfeld 2008;

Watanabe et al. 2013). Conversely, epigenetic alterations accrued during the cell differentiation process result in cell populations consisting of an epigenome mosaic. Therefore, any multi-cell epigenomic study of tissue samples composed of multi-aged cells, let alone different cell types, constitutes a collection of epigenomes (Jaffe and Irizarry 2014; Wijetunga et al. 2014). Mixed epigenome samples present challenges for the analysis and interpretation of epigenomic data, particularly in epigenome-wide association studies (EWAS), of which investigators are often not aware (Greally 2017). A study using five publicly available datasets from epigenome-wide associations between human disease and DNA methylation content of whole blood found that blood DNA methylation levels explain over 19% of variation in blood cell-type composition present among study samples (Jaffe and Irizarry 2014). Similarly, studies of stem cells from diverse cell sources have found that induced pluripotent stem cells maintain characteristic epigenetic profiles depending on cell origin even prior to cell reprogramming (Shiota et al. 2002; Watanabe et al. 2013). Studies of the DNA methylome of multiple tissues from *Arabidopsis thaliana* found divergence in methylation profiles across tissue types that also varied in their degree of divergence based upon sequence context (Fig. 2; Calarco et al. 2012; Kawakatsu et al. 2016a, b). Together, this suggests a strong potential for spurious epigenomic associations to confound epigenomic studies any time when samples represent heterogeneous cell compositions.

Mixed-cell sample deconvolution strategies for simplifying heterogeneous samples include methods of cell-type sorting via flow cytometric approaches such as fluorescence-activated cell sorting, immunomagnetic separation, or microfluidic microchips (Jaffe and Irizarry 2014; Wijetunga et al. 2014), single-cell genome-wide bisulfite sequencing (Smallwood et al. 2014), and bioinformatics approaches (Teschendorff and Zheng 2017). Naïve assessment of heterogeneous tissue- and cell-specific epigenetic profiles clearly presents a significant source of spurious epigenomic variation, i.e., false detection of epigenomic variation among samples. However, leverage of cell- or tissue-specific epigenomic variation allows dissection of the epigenetic contribution to gene expression differences underlying differential tissue and cell development. For instance, conspicuous tissue-specific DNA methylation patterns offer useful biomarkers of various human diseases (Hewitt et al. 2017; Keller et al. 2017; Yang et al. 2017), targets for the identification of loci related to DNA methylation-associated phenotypic variation in response to imposed environmental stress (Alonso et al. 2017), and methods for quantitative determination of cell-type proportions present in heterogeneous samples (Baron et al. 2006). Additionally, cell-type-specific DNA methylation changes in the development of certain cell lines, such as male and female plant germ cells, have important consequences for the propagation of accumulated epigenetic modifications to daughter cells via signaling factors, often in the form of small RNAs present in plant sperm nuclei and seed endosperm (Calarco et a. 2012; Springer and Schmitz 2017).

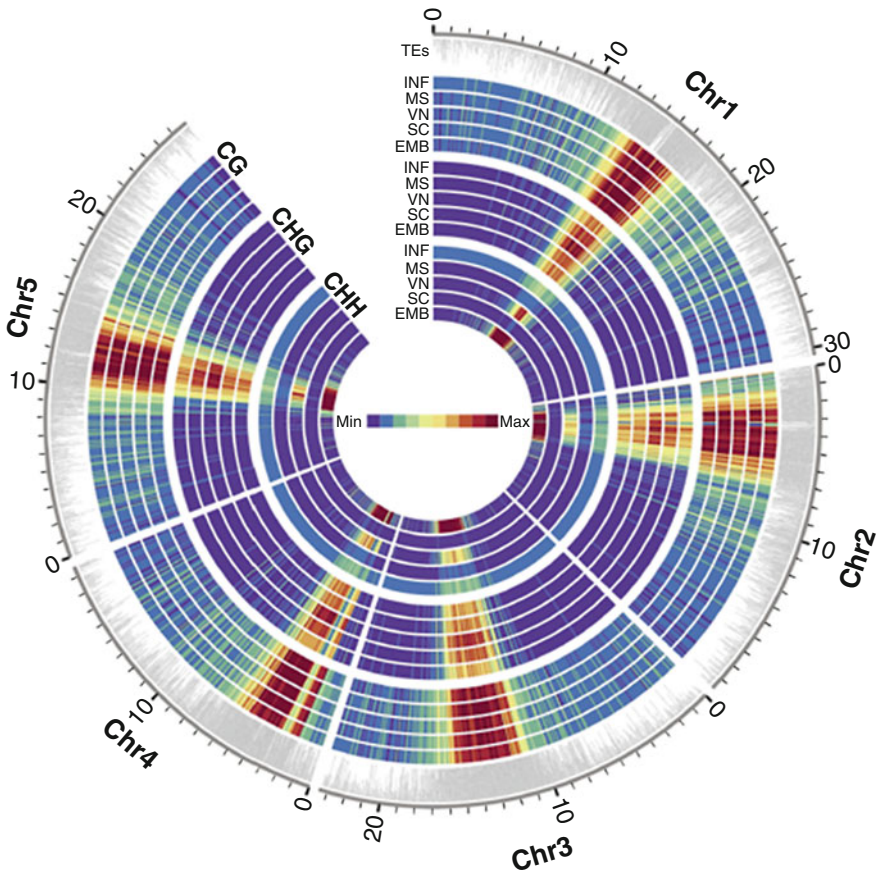

Fig. 2 Heat map of DNA methylation of the primary methylation contexts in different tissues of *Arabidopsis thaliana*: *INF* inflorescence, *MS* microspore, *VN* vegetative nucleus, *SC* sperm cell, *EMB* embryo. Reproduced with permission from Calarco et al. (2012)

2.4 Heritability of Epigenetic Patterns

Characteristic patterns of epigenomic variation occur within subspecies, species, and populations (Verhoeven et al. 2010; Zemach and Zilberman 2010; Schmitz et al. 2011). The resetting of most epigenetic patterns that occurs during gamete and zygote formation is necessary to enable zygote cell totipotency, yet epigenetic patterns are largely conserved within lineages and populations, suggesting the activity of mechanisms for their transmission through mitosis and meiosis, as well as their reestablishment from environmental and genetic cues (Schmitz et al. 2011; Calarco et al. 2012). Interestingly, transmission of gene-independent epialleles (loci differing in chromatin states among cells or organisms) through meiosis is well documented in plants, but less so in animals, although plants maintain totipotent

cells well beyond embryogenesis while animals do not (Calarco 2012). This difference relates to the tendency of the epigenomes of animals, mammals in particular, to be erased and then reestablished during zygote formation, while plant epigenomes undergo far less loss or reprogramming and are instead reinforced during gamete formation (Heard and Martienssen 2014).

The inheritance of silent gene expression states appears to mostly involve the transmission of DNA methylation profiles, as evidenced by the finding that qualitative "on/off" effects on gene expression are commonly associated with DNA methylation (Springer and Schmitz 2017). Meanwhile, the transmission of active and quantitative gene expression states commonly involves histone variants, especially the histone variant H3.3. Along with multiple histone modifications, H3.3 is often enriched at active chromatin sites (Ng and Gurdon 2008). The mechanism enabling maintenance of a pure epigenetic state (sensu Richards 2006) related to DNA methylation through mitosis is facilitated by the semiconservative nature of DNA replication. After mitosis, each daughter cell has one parental DNA strand with a methylation pattern matching that of the parental cell and one newly synthesized strand lacking methylation. The resulting hemimethylated state is the preferred substrate of certain DNA methyltransferases, which preferentially methylate CG or CHG nucleotides on the new strand paired with methylated complementary sequences on the parental strand (Adams and Burdon 1985; Ng and Gurdon 2008). Transmission of the histone variant H3.3 depends upon the synthesis and deposition of H3.3 near chromatin sites already enriched in H3.3, which may occur in most phases of the cell cycle (Ng and Gurdon 2008). Enrichment of H3.3 near sites in the mother cell enriched in H3.3 increases the ratio of this histone variant to the typical H3.1 form, improving the likelihood of H3.3 recruitment into newly replicated chromosomes, thus maintaining H3.3 density and position along the chromatin across cell generations. It is misleading, however, to discuss the inheritance of DNA methylation or histone variants in isolation from the influence of small RNAs, which play a role in orchestrating many instances of the former. For instance, patterns of CHH methylation, once lost in the sperm cells and microspores of *A. thaliana*, are restored by small interfering RNA (siRNA) and RdDM pathways associated with regions of active CG demethylation of transposable elements (TEs) flanking imprinted genes (Calarco et al. 2012).

Transgenerational epigenetic inheritance, i.e., transmission of epigenetic states through meiosis, requires the transfer of epigenetic phenotypes through the germ line. Determining that a putatively epigenetic trait with a known source is transgenerationally heritable requires, at a minimum, observing the persistence of the trait across generations in the absence of the source (Mirbahai and Chipman 2014). Correlated gene expression and epigenetic profile alterations following intragenerational epigenome alterations are frequently short-lived in plants, decaying after a small number of cell cycles, much less transmitting through meiosis. These short-term responses may cause beneficial phenotypic plasticity, as in the salt-stress exposure of *A. thaliana* described by Wibowo et al. (2016) and the multi-generation drought stress of *A. thaliana* reported by Van Dooren et al. (2018). Correlated gene expression and epigenetic profile alterations meeting the conditions of epigenetic transgenerational inheritance have been demonstrated only for

intergenerationally accumulated changes in plants, such as the B' epiallele in maize that has remained stable through thousands of cell cycles (Richards 2006), maintenance of epigenetic variation across many generations in *A. thaliana* (Schmitz et al. 2011; Hagmann et al. 2015), and in DNA methylation changes in rice after consistent multi-generational stress exposure (Zheng et al. 2017). No evidence yet exists to suggest that coupled gene expression-epigenetic alterations resulting from intragenerational sources of epimutations are transmitted through meiosis in plants (Pecinka and Scheid 2012), unlike for animal studies such as the study of Nilsson et al. (2012) describing the transmission of DNA methylation states correlated to the induction of ovarian disease in rats (*Rattus* spp.). While environmentally-induced transgenerational epigenetic inheritance of disease reportedly occurs in rats (Maamar et al. 2018), more studies are needed to conclude that transgenerational epigenetic inheritance occurs in mammals (Horsthemke 2018). Nevertheless, there is evidence for heritable changes in DNA methylation in response to environmental stresses in plants, but the strength of inheritance depends upon environmental conditions, and DNA methylation changes could persist through clonal propagation (review in Richards et al. 2017). In general, the extent to which environmentally-induced transgenerational epigenetic inheritance occurs and what role it plays in adaptive evolution remains inconclusive and controversial (Luikart et al. 2018).

However, as described next, a significant aspect of epigenomic variation with respect to evolution may relate to the fact that phenotypic changes associated with inheritance of epigenetic alterations have often been found to outpace changes associated with genetic alterations (Rando and Verstrepen 2007). Two possible mechanisms for increased rates of phenotypic change related to epigenetic alterations include (1) the effect of 5mC on facilitating cytosine deamination, thereby increasing rates of point mutations (Feinberg and Irizarry 2010), and (2) increased TE transposition rates (with attendant altered DNA methylation patterns) following stress exposure or hybridization – observed to increase the rate of TE insertion $\times$ (TE copy)$^{-1}$ $\times$ (generation)$^{-1}$ from 10^{-5} to 1, and thereby contributing to rapid chromosomal rearrangement (Bonchev and Parisod 2013).

3 Sources and Evolution of Epigenomic Variation

Epigenomes provide mechanisms for modifying gene expression according to environmental and developmental contexts. Sources of epigenomic variation may be due to epigenetic variants arising purely from genetic variation, the interaction of genetic and environmental variation, purely environmental variation, or stochastic epimutation events. Richards (2006) categorized types of epigenomic variation based upon the source of variation, as follows: *obligatory* (a strictly genetic source), *facilitated* (the genome facilitates or potentiates an epigenomic state), and *pure* (epigenomic variation is due to the environment or stochastic epimutations, not the underlying genome). The degree to which an epigenome may add to the phenotypic variation in a population above that already afforded by population genetic variation depends upon the extent to which an epigenome is pure (Klironomos et al. 2013),

and whether an epigenomic mechanism is retained by natural selection probably depends in part upon the costs and benefits of maintaining phenotypic variation.

Retention of an alternate source of phenotypic variation, such as an epigenome, entails metabolic costs that are outweighed by associated benefits conferred to cells, organisms, and populations that retain the system. For example, DNA methylation of cytosine utilizes S-adenosyl-L-methionine ($C_{15}H_{23}N_6O_5S^+$; abbr. SAM), one of the most metabolically expensive compounds that cells construct, as the donor of methyl groups (CH_3). SAM has a metabolic cost of 12 ATP equivalents per carbon atom, compared to the cost of 6 ATP equivalents per carbon atom in a glucose molecule (Adams and Burdon 1985). The high cost of SAM implies that there must be both sufficient selective advantage for organisms to direct large amounts of metabolic energy toward SAM biogenesis and strong selective pressure for the parsimonious consumption of CH_3 from SAM. Two crucial functions of DNA methylation and related epigenetic mechanisms would appear to justify the retention of such a metabolically expensive system: (1) enabling the regulation of proliferation of transposable elements and (2) facilitating the generation of myriad cell types, developmental changes, and phenotypic variation, all from the same underlying genetic information. It follows then that the epigenetic machinery responsible for these functions should be retained only to the extent that the machinery contributes to the adaptation of organisms to their environments without undue costs. For instance, in stable environments, the benefits of higher variation that may enable survival in rapidly changing environments may not outweigh the metabolic cost associated with maintaining high phenotypic variation via pure epigenetic means (Relyea 2002).

For epigenetic modifications to influence adaptation, these modifications must be subject to natural selection, which has the following requirements: (1) epigenetic modifications produce phenotypic *variation*, (2) related phenotypic variants contribute to *differential fitness*, and (3) phenotypic variation generated by epigenetic modifications are *heritable* (Darwin 1859). For natural selection to act upon epigenetic variation similarly to genetic variation, epigenetic variation and its effects must be sufficiently stable to allow time for multiple selective forces to act upon the differential fitness it produces (Rahavi and Kovalchuk 2013; Iglesias and Cerdán 2016). Because the duration of an epigenetic change influences its relevance to adaptation and different sources of epigenetic change have different durations of influence, understanding the potential for an epigenetically-mediated trait to influence adaptation requires consideration and investigation of the potential sources of observed epigenetic variation (Chadha and Sharma 2014).

3.1 Genetic Sources of Epigenomic Variation

Genetic variants associated with epialleles/epigenetic variants and activity of transposable elements (TEs) constitute major genetic sources of epigenetic variation

(Suzuki and Bird 2008; Taudt et al. 2016; Springer and Schmitz 2017). Whole genome duplication events (polyploidization) also strongly influence epigenomic variation, and though the precise role of polyploidization in the generation of epigenomic variation remains speculative, it is likely to be heavily dependent upon TE activity (Zhang et al. 2015a, b). See Taudt et al. (2016) for a review of genetic sources of population epigenomic variation.

Genetic variants strongly associated with variation in methylated DNA loci operate as distinct quantitative trait loci for epigenetic molecular phenotypes and are identified through linkage mapping between genetic variants and methylated DNA loci (Denker and de Laat 2015; Chen et al. 2016; Taudt et al. 2016). QTL associated with DNA methylation (meQTL), histone variants (hQTL), and large-scale patterning of histone variants forming variable chromatin domains (chQTL) may reside near to (*cis* <50 kb) or hundreds of kb from (*trans*) an associated epigenetic phenotype (Taudt et al. 2016). Regardless of the proximity of genetically determined epigenetic marks, such epigenetic variants follow the same strict Mendelian patterns of inheritance as the genes they relate to.

Both the *cis* and *trans* regulatory genetic mechanisms conditioning both population epigenomic variation and individual epigenetic marks have been identified in a wide variety of species, and numerous meQTL and hQTL have been identified (Taudt et al. 2016). For example, 15% of the >3 million genome-wide CpG sites in humans were found to be associated with meQTL (McClay et al. 2015), and nearly all meQTL were found in *cis* configuration (Taudt et al. 2016). All hQTL detected in the human study were also in *cis* configuration (review in Taudt et al. 2016). In a study of the genome-wide differentially methylated regions (DMRs) and whole genome DNA sequences in worldwide natural accessions of *Arabidopsis*, Schmitz et al. (2013a) found that 35% of the DMRs could be associated with meQTL, with 26% of the associations mapped in *cis* and 74% in *trans* configuration. In another study, Schmitz et al. (2013b), using recombinant inbred lines of *A. thaliana* for examining the inheritance of DNA methylation, reported that >90% of DMRs mapped to a meQTL, implying that up to 10% of DMRs detected in the study may relate to non-genetic factors. A study of maize showed that nearly half of the DMRs identified across 51 genotypes were significantly associated with meQTL in *cis* configuration with or within DMRs (Eichten et al. 2013). Furthermore, many of the DMRs identified in this study occurred near TEs, a common finding in epigenome-wide association studies reflecting the often-cited suspected origin of DNA methylation as a means to regulate TE mobilization (Fedoroff 2012). Many *cis*-acting meQTLs in plants are thought to be due to SNP alleles tagging nearby structural variants, such as TEs, that spread methylation into flanking regions (see Taudt et al. 2016). A study of the extent of *trans*-acting hQTL in mice showed that roughly 25% of histone variants are under genetic control in *trans* configuration to the histone variants detected (Baker 2018).

The stochastic transposition of repetitive gene sequences, resulting from the mobilization of transposable elements, is both a result and a source of epigenetic variation (Fedoroff 2012; Liang et al. 2014). Mobilization of transposable elements is expected to be followed by a feedback cascade that results in TE remethylation

either rapidly in cells without compromised methylation pathways or after multiple rounds of DNA replication, leading to eventual TE silencing (Bousios and Gaut 2016). If TEs escape silencing due to environmentally or developmentally induced disruption of the DNA methylation machinery, TEs may proliferate within the genome and spread TE DNA fragments (i.e., targets of regulation by DNA methylation) and genetic regulatory networks throughout the genome (Rey et al. 2016). This was shown to occur in the genome of *Populus* spp., where cytosine methylation patterns were altered up to 2 kb upstream and downstream of TE insertion sites, which strongly correlated to altered gene expression resulting from the methylation of transcription factor binding sites located within the region of modified methylation near TE insertions (Liang et al. 2014). The coupled activities of DNA methylation, TEs, and altered gene expression via modified transcription factor activity suggest that DNA methylation not only regulates but also provides a mechanism for realizing selective advantages from retaining TEs. Otherwise, TEs likely would not persist within genomes since there would be sufficient selective pressure to eliminate deleterious, unregulated stochastic TE insertions resulting in chromosomal disruptions, with potentially negative consequences (Kazazian 1998), through selection or homologous recombination (Fedoroff 2012; Bonchev and Parisod 2013). Differences in numbers and sites of TE insertion likely relate to more than merely the specific genomic site into which a TE is inserted, as TE insertion typically elicits a flood of gene-silencing DNA methylation proximal and even distal to the insertion site (Fedoroff 2012; Rey et al. 2016).

3.1.1 Environmental Sources of Epigenomic Variation

Though genetic mutations that affect epigenetic variation occur at a relatively constant rate, exposure to environmental stresses, such as temperature extremes, drought, and toxins can lead to altered epigenetic states through numerous mechanisms. The best understood mechanisms include TE activation or mobilization and accelerated rates of somatic mutation of genes linked to epialleles, such as epigenetic modifier genes (Fedoroff 2012; Bonchev and Parisod 2013; Liang et al. 2014; Greenblatt and Nimer 2014; Weng et al. 2014; Rey et al. 2016). Epigenetic alterations resulting from environmental perturbation is a stochastic process with the fundamental result of increasing phenotypic variation, which may by chance result in adaptive changes in gene expression along with less beneficial outcomes (Feinberg and Irizarry 2010; Becker and Weigel 2012; Calarco et al. 2012). For instance, random TE insertions at particular genomic sites were found to increase the capacity of wild populations of *Drosophila melanogaster* to adapt to temperature and precipitation regimes (Bonchev and Parisod 2013). A population of wild barley (*Hordeum* spp.) growing vigorously on a dry site was found to have more full-length TE insertions and less truncated LTR retrotransposon insertions than a nearby population occupying a more favorable, moist habitat (Bonchev and Parisod 2013). On the other hand, dozens of mammalian diseases are known to result from retrotransposon insertions, demonstrating the stochastic and unbiased nature of TE

transposition (Kazazian 1998). In another *A. thaliana* study, mobilization of the *ONSEN* retrotransposon via heat shock stress was observed in plants with an impaired siRNA pathway but not in wild-type plants or those not exposed to the heat stress, resulting in the formation of a "stress memory" in siRNA-impaired plants (Ito et al. 2011). While it is inaccurate to consider the stress memory as a Lamarckian "trigger" capable of inducing a specific phenotypic change yielding more heat-tolerant plants, the stress memory is associated with the activation of novel regulatory genetic networks that by chance may result in the generation of heat tolerance upon which natural selection could act.

The well-documented environmentally-responsive quality of TEs represents a mechanism through which the environment may indirectly influence epigenomic variation. However, there are numerous examples of environmental influences directly altering epigenetic phenotypes without apparent genetic influences, though in most cases it is still possible that undetected *trans*-acting genetic variants are involved (Greally 2017; Richards et al. 2017). Studies have, however, demonstrated that patterns of epigenetic variation may be substantially altered quite independently of genetic variation in plants shown to have minimum genetic variation after growing for a number of generations (Schmitz et al. 2011, 2013b) and after controlled environmental exposures to stresses such as vernalization (He et al. 2003), drought (Zheng et al. 2017; Van Dooren et al. 2018), tissue culturing (Stroud et al. 2013), vegetation density reduction in a natural setting (Herrera and Bazaga 2016), exposure to plant defense hormones (Verhoeven et al. 2010), and nutrient withholding (Verhoeven et al. 2010). Divergent epigenetic variation was also detected in a monozygotic human twin study of DNA methylation profiles correlated to diabetes (Zhao et al. 2011) and lifestyle differences (Fraga et al. 2005a). Numerous studies taking a survey approach have also reported putative differences in epigenomic profiles independent of genetic variation detected through various genotyping assays in plants (Lira-Medeiros et al. 2010; Herrera et al. 2013; Latzel et al. 2013), fungi (Zimmerman et al. 2016), fish (Mirbahai and Chipman 2014), mice (Wilson and Sengoku 2013), and human toxicology studies (summarized by Martin and Fry 2018).

3.2 *Evolution of Epigenomic Variation Within Populations*

A means for achieving greater phenotypic diversity and plasticity (phenotypic diversity arising in the same genotype in response to different environments) was implicit in Waddington's original formulation of the concept of epigenetics. By providing a source of phenotypic variation independent of population-level changes in allele frequencies, and by influencing the extent of phenotypic variation possible via a given gene × environment interaction, epigenetic phenomena may provide a critical stepping stone between phenotypic plasticity and the stabilization of

expression of facultatively plastic (canalized) phenotypic responses (Johnson and Tricker 2010; Grativol et al. 2012; Schlichting and Wund 2014; Richards et al. 2017). Given that environmental variation will inevitably favor the ability of a population to adapt to new conditions, the ability to activate mechanisms for enhancing phenotypic diversity could be profoundly beneficial (Sultan 2000; Nicotra et al. 2010; Baythavong 2011). A brief review of the role of phenotypic diversity and plasticity in the process of evolution will help to clarify their interaction and the potential relevance of epigenetics to evolution.

Microevolution is a stochastic process that occurs within and among populations, whereby drift and selection act upon mutations (Hendry and Kinnison 2001), which may eventually lead to conspicuous trait and species divergences (Dobzhansky 1937). Phenotypic plasticity refers to the capacity of a genotype to express different behavioral, morphological, and physiological responses depending on the environment in which it occurs (Price et al. 2003; Schlichting and Wund 2014). Increased variation in a phenotypic trait due to phenotypic plasticity may soften the impact of an environmental stressor on a population before natural selection acts upon population gene frequencies (West-Eberhard 2005). The degree and rate of persistence or reduction in the phenotypic plasticity of a population depend heavily upon the range of environmental variation to which a population must continually adapt and the adaptive landscape of a population (Price et al. 2003). Such plastic developmental responses may represent past functionalities or functions produced by de novo changes in gene regulation that reveal previously hidden portions of a reaction norm to natural selection, such as previously silenced genes (Sultan 2003; West-Eberhard 2005; Rey et al. 2016), which may result either from coordinated cellular responses to specific environmental stimuli or arise through pure chance (Price et al. 2003). Enhanced trait variation produced by epigenetic variation could enhance the efficiency of genetic accommodation by increasing the variation in phenotypes that arise from a given genotype, thereby potentially increasing the frequency of phenotypes that confer selective advantages under a given environmental regime (Schlichting and Wund 2014). No matter what their source, plastic phenotypic responses to stress can delay the trait-purifying step of selection that eventually shifts a population's fitness peak toward a new adaptive peak or, alternatively, eliminates a population altogether due to a lack of short-term capacity to adjust to a stressor (Fig. 3; Price et al. 2003).

The duration of the delay between initial phenotypic responses and selection for adaptive traits that track the new adaptive peak is proportional to the level of plasticity for adaptive traits in a population and the nature of the adaptive landscape (Price et al. 2003). Multiple adaptive peaks in an evolutionary landscape, due to high environmental heterogeneity, for instance, may select for increased phenotypic plasticity, while high plasticity is likely to delay adaptation and any concomitant *reductions* in plasticity that must precede canalization of an adaptive phenotype, possibly leading to subsequent genetic assimilation (Price et al. 2003; Baythavong 2011; Schlichting and Wund 2014). Accordingly, a population with low to moderate plasticity, yet sufficient plasticity to move a population's fitness peak toward a new adaptive peak, may adapt to the new environment more rapidly than a highly plastic

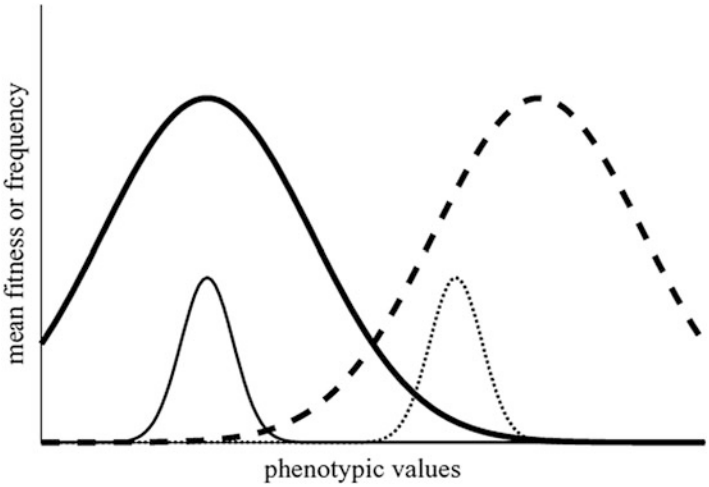

Fig. 3 Plasticity contributes to a peak shift in changing environments; bold line shows mean fitness; bold dashed line shows the shift of mean fitness in a new environment; thin solid line shows trait distribution in the old environment; thin dotted line shows trait distribution after plastic response to new environment. Reproduced with permission from Price et al. (2003)

population. Meanwhile, populations with a history of rapid genetic assimilation due to the stability or homogeneity of their environments are less likely to survive stochastic, extreme environmental stresses (Price et al. 2003; Aitken et al. 2008).

In a controlled study of two genetically uniform inbred lines of *A. thaliana* with different levels of variation in DNA methylation patterns (i.e., epigenetic recombinant inbred lines, abbr. epiRILs; Johannes et al. 2009), one line with highly variable genome-wide DNA methylation patterns (epigenetically diverse) and one line with low variation (epigenetically uniform), it was concluded that epigenetic diversity accounted for enhanced resilience and growth (Latzel et al. 2013). The two epiRILs were studied for their responses to both a common pathogen and interspecific competition. These studies revealed that the epigenetically diverse genetic line produced 40% more biomass than the epigenetically uniform line, and morphological differences between the lines were more pronounced when plants were under biotic stress. In another epiRIL study of *A. thaliana*, Cortijo et al. (2014) described the phenotypic outcomes of induced methylation of numerous genomic regions and found that inducing methylation accounted for 60–90% of the heritability for flowering time and primary root length through the F_3 generation. Importantly, the study showed that these traits can be propagated through artificial selection and that the methylated regions related to the traits of interest were also variable within natural populations of *A. thaliana*, suggesting that natural selection would likely act upon these epigenetic traits in the same manner as strictly genetically based evolution (Cortijo et al. 2014). Epigenetic mechanisms, especially DNA methylation, may, thus, facilitate the acclimation response of organisms to a variety of

abiotic and biotic stresses through phenotypic plasticity (see also Richards et al. 2017). Hence, epigenomic variation has been described as a potentially significant source of variation in plants and animals in response to climate change (Bräutigam et al. 2013).

3.3 *Epigenomic Variation Within and Among Populations and Species*

The description of epigenomic variation within and among populations and species is in its early stages. The epigenome is more complex than the genome in that it is subject to direct genetic, developmental, and environmental influences at short time scales and consists of diverse mechanisms. The presence and prevalence of epigenetic mechanisms, especially the machinery involved in DNA methylation, differ across taxa as shown by comparative epigenomics studies across phyla and within angiosperms (Niederhuth et al. 2016) with some model species like *Drosophila* spp. and *Caenorhabditis* spp. lacking some methylation mechanisms found in other organisms. Additional differences, such as germline differentiation occurring later in development for plants than animals (Sharma 2013) and less complete demethylation during reproduction in plants (Heard and Martienssen 2014), also suggest substantial differences in the distribution and type of population epigenomic variation across taxa. Furthermore, the epigenome is a phenotype as well as in some instances having the potential to facilitate the inheritance of other phenotypes (Greally 2017). Thus, epigenomic variation within and among populations and species is logistically more difficult to ascertain and describe than genomic variation, and thus underlying theory for the former has been slower to develop (Banta and Richards 2018). The population and species distribution of epigenomic variation is just beginning to be explored in natural populations (Richards et al. 2017), with studies on a small number of model systems, particularly maize and *A. thaliana*, providing more detailed insight. Nevertheless, evidence so far suggests that epigenomic variation is widespread in wild populations of plants (Schmitz et al. 2013b; Niederhuth et al. 2016) and animals (review in Hu and Barrett 2017).

3.3.1 Sex Differences and the Epigenome

Based on biological mechanisms and empirical results from quantitative genetics, patterns and frequencies of epigenetic inheritance are expected to vary between sexes. There are differences between sexes in epigenome reprograming, and sexes differ in the opportunity for transmission of cytoplasmic signaling molecules (Calarco et al. 2012; Jiang et al. 2013; Heard and Martienssen 2014). Meanwhile, quantitative genetic studies of maternal and paternal effects have found variable

effects of the sexes (Roach and Wulff 1987; Galloway and Etterson 2007; Bonduriansky and Head 2007). Parental effects are not equivalent to epigenetics, but epigenetics is one mechanism by which parental effects can be transmitted (Vogt 2017). In *Mimulus guttatus*, crosses in controlled breeding and a demethylation treatment together demonstrated that both male and female parents demonstrated transgenerational induction of increased glandular trichome production in response to simulated insect damage, and the paternal effect persisted after demethylation treatment, whereas the maternal effect did not (Akkerman et al. 2016). This is suggestive of a methylation-dependent transmission in the maternal line and some other mechanism, such as siRNA, meditating the paternal effect (Akkerman et al. 2016). In zebrafish (*Danio rerio*), the sperm methylome is the one passed to early embryos (Jiang et al. 2013). In non-model species, sex differences in epigenetic inheritance have been detected by methylation-sensitive amplified polymorphism (MSAP) data. For instance, a methylome fragment analysis assay in a conifer full-sib family found greater inheritance of fragments from the maternal line than the paternal line (Avramidou et al. 2015).

Gender-specific methylation patterns in some species (Janoušek et al. 1996; Piferrer 2013) suggest that methylation may influence sex determination or sex-related trait expression (Chatterjee et al. 2016). Plant taxa that have more recently evolved dioecy from monoecious ancestors are useful models for investigating the evolution of sex-determining mechanisms (Bräutigam et al. 2017). In *Populus balsamifera*, the genome region associated with sex determination includes a gene encoding methyltransferase. There are also overall sex differences in methylation levels across the *P. balsamifera* genome, and methylation difference between the sexes is greatest at one gene within the region associated with sex determination (Bräutigam et al. 2017). In sea bass (*Centropristis striata*), with mixed genetic and environmental sex determination, temperature variation early in the development alters methylation patterns suggesting a role for epigenetics in environmental sex determination (Piferrer 2013).

3.3.2 Epigenomic Variation Within and Among Plant Populations

A relatively detailed picture of methylome variation among and within populations of the model plant *A. thaliana* has emerged from a series of studies examining natural variation and offspring of multi-generation crosses. Methylation of TE-rich regions of the genome is high and relatively consistent across individuals and time (Vaughn et al. 2007; Becker et al. 2011). Gene regions are less methylated but are the source of a large fraction of DMRs detected. Instability of gene region methylation status generates novel variation (Schmitz et al. 2011). These DMRs can be inherited but experience high rates of loss or back-mutation in contrast with gene sequence variation (Becker et al. 2011; Van der Graaf et al. 2015). This combination of phenomena is probably the reason for the observation that after 31 generations, the accumulation of differences in the epigenome is not different from the accumulation of differences in the genome (Becker et al. 2011). Intriguingly, DMRs show less

association with two potential promoters of methylation, transposable elements and small interfering RNAs.

A. thaliana accessions vary substantially in TE composition, and this variation coincides and interacts with methylation variation to impact gene expression (Underwood et al. 2017). Geographic surveys of *A. thaliana* suggest an association of low temperature with lower methylation of TEs that may result from temperature-related natural selection acting on genetic control of methylation levels (Underwood et al. 2017). A genome-wide association study (GWAS) analysis of methylation found substantial genetic control of methylation (Kawakatsu et al. 2016a, b) and that the epigenomic changes were particularly associated with immunity genes.

Research in crop plants, which has provided a large proportion of insights into population epigenomics, highlights some potential differences between *A. thaliana* and most other plants. *Arabidopsis* houses a small genome with low levels of methylation and TEs relative to other plants. Thus, epigenomes of other plant taxa may serve more important roles in generating phenotypic variation within and among populations (Kawakatsu et al. 2016a, b; Song and Cao 2017). On the other hand, comparative analysis of angiosperm methylomes suggests that clonal propagation and other common crop production methods may lead crops to have a different distribution of epigenomic variation than plants that are not cultivated (Niederhuth et al. 2016). For instance, a greater proportion of the rice genome consists of TEs, with correspondingly higher levels of methylation as compared to *A. thaliana* (Song and Cao 2017). Furthermore, differences in the types of TEs occurring in the genomes of Asian and African rice species account for most observed variation in those genomes (Wang et al. 2015). Among rice subspecies, differences in TEs predominate (Song and Cao 2017). Epigenetic differences between subspecies were also well correlated with gene expression differences (He et al. 2010), and hybrids showed high levels of non-additive epigenetic variation, especially for methylation as well as for transcription differences (He et al. 2010).

Maize has long been a model species for the study of TEs due to the pioneering work of McClintock (1951). A selection experiment in rice found extensive methylation and morphological changes accompanying an artificial selection experiment (Zheng et al. 2017). Approximately 30% of the methylation variants were inherited by additional generations after the removal of selective pressure. A study comparing a Chilean land race of maize to a maize reference genome found increases in long non-coding RNAs that respond to salt and boron stress in the landrace, demonstrating within-species variation is yet another potential epigenetic regulatory mechanism (Huanca-Mamani et al. 2018). In summarizing studies that included epigenome profiling across diverse accessions in five plant species, mostly crops, Springer and Schmitz (2017) noted that substantial variation occurs despite high conservation of methylomes and that this variation can potentially be harnessed for crop improvement both in cases of environmental induction and transgenerational inheritance.

Various *Populus* species have served as model tree species for genomics studies, and they are also well suited to epigenomics work because of the ease of clonal propagation. Common garden tests containing clonal replicates, implemented in

multiple habitats, allows the separation of the effect of genomes and environments on epigenomes and outward phenotypes, enabling some insight into epigenomic variation and its consequences (Whipple and Holeski 2016). Two studies of clonally reproduced *Populus* spp. cuttings and cultivars grown in distinct environments, and then clonally reproduced again for growing a second clonal "generation" in another common environment, provide insight into environmental sources of epigenomic variation (Raj et al. 2011; Schönberger et al. 2016). Differences in the previous growing conditions of clones resulted in differentially methylated regions and differences in miRNA expression among clonal individuals currently growing in the same environment. Some of these epigenetic changes were also related to changes in gene expression.

For most non-model plant systems, population epigenomics work is in its infancy and consists of assays on anonymous surveys of methylated restriction sites, as recently reviewed by Richards et al. (2017). Other studies may not encompass the entire genome but are still important in this young field for investigating associations between traits and stably-inherited epialleles (Richards 2006; Jablonka and Raz 2009; Richards et al. 2017). Common findings presented in papers on non-model plant species reviewed in Richards et al. (2017) demonstrate that often, but not always, observed epigenomic variation is greater than genomic variation and that epigenomic variation is frequently correlated with environmental or trait variation. For example, Gugger et al. (2016) studied DNA methylation variation in 58 natural populations of *Quercus lobata* sampled across the species' range and found significant associations of 43 single methylation variations (SMVs) with each of the four climate variables. One recent addition to these types of studies is a counter example where genomic variation was found to be greater than epigenomic variation in a deciduous shrub *Vitex negundo* var. *heterophylla* (Chinese chastetree) (Lele et al. 2018). In another recent study, by combining field and common garden trait measures as well as epigenomic assays, Groot et al. (2018) showed that some of the epigenomic variation in the shrub species *Scabiosa columbaria* was inherited and appeared to be environmentally induced. The epigenomes and genomes of plants grown in the field were correlated, but not for garden-grown plants. Additional work is needed to determine the mechanism of inheritance of epigenome variation and to what degree observed variation is genetically determined, facilitated, or pure.

Additional insights into natural variation in epigenomes have come from studies of forest trees, especially Norway spruce (*Picea abies*). Long-running provenance trials established across wide environmental gradients with the same source populations enabled the detection of not only genetic source but also temperature during seed development (embryogenesis), to offspring phenological traits in this species (Kvaalen and Johnsen 2008, Johnsen et al. 2009). The epigenetic memory of the environmental temperature during Norway spruce embryogenesis was shown to consistently and reproducibly affect phenology of the resulting trees in a manner often ascribed to ecotypes and gradual phenotypic differences across environmental clines (Yakovlev et al. 2012; Carneros et al. 2017). Further study found an association of miRNAs and transcriptome variation with seed development temperature

(Yakovlev et al. 2016; Yakovlev and Fossdal 2017). The latest efforts to characterize the miRNA expression in this species identified over 1,000 highly expressed miRNAs, more than half of which were differentially expressed across temperature treatments. Many of these miRNAs themselves target genes involved in epigenetic regulation. Studies that do not explicitly assay the epigenome but make inventive use of biological systems allowing inference of potential epigenetic phenomena have also provided insights into epigenomic variation among populations. For instance, Dewan et al. (2018) used grafted clonal *Populus nigra* to show that there are similar temperature effects during seed production on offspring traits. And clonal propagation in contrasting environments with *Pinus pinaster* has been used to infer the likelihood of an epigenetic component to seedling performance (Zas et al. 2013).

Hybridization and polyploidization can increase the occurrence of epigenetic alterations, which may serve as mechanisms to cope with genomic instability resulting from hybridization or generate novel phenotypic variation (Paun et al. 2010; Jackson 2017). Two orchid species have gone through independent allopolyploid hybridization events across multiple sites and climates, providing a model system for surveying epialleles associated with hybridization and ecotypic differentiation. Paun et al. (2010) detected epialleles that showed patterns of selection and strong association with climatic variation. Additional studies could elucidate whether the epialleles are induced by the environment each generation, stably inherited, or the result of an undetected genetic variant (Balao et al. 2016).

3.3.3 Epigenomic Variation Within and Among Animal Populations

Despite the attention to epigenetic mechanisms involved in human diseases, little attention has been devoted to human population epigenomic variation (Kelly et al. 2017). In a worldwide survey of human epigenomic variation, Carja et al. (2017) found evidence that patterns of variation closely follow genetic variation and are likely largely under genetic control, with much greater stability than is generally found in plants. Similarly, other authors found population variation in methylation within genic regions with genetic control likely, but with the genes involved also varying across human populations (Fraser et al. 2012). A third study, which more closely mimics designs seen in non-model plant species, made comparisons of geographically distinct populations where genetic difference predominated, versus genetically similar populations occupying recently divergent environments (Fagny et al. 2015). For geographically dispersed populations, most methylome variation seemed to be attributable to genetic loci controlling methylation states. In addition, the methylation differences were predominately located at metabolic and developmental genes. For genetically similar populations in contrasting environments, there was less evidence of genetic control of the epigenome, and the loci with methylation differences were co-located with immune system genes (Fagny et al. 2015). Finally, in a comparative study of human and mouse DNA methylation, differentially methylated regions among individuals were disproportionally found in developmental genes for both species (Feinberg and Irizarry 2010).

In general, animals have not been investigated for natural variation in the epigenome to the same extent as plants (Vogt 2017). There are, however, examples in animal studies of greater epigenome diversity than genome diversity (as reviewed in Vogt 2017). Invasions by genetically uniform individuals, as may occur in asexual animal species, can be useful for understanding associations of the epigenome with environmental variation and/or trait variation (Vogt 2017). For instance, in an introduced, parthenogenic snail (*Cornu* spp.), greater epigenomic and trait variation were associated with differences between habitats (Thorson et al. 2017).

4 Methods for Screening Population Epigenomic Variation

DNA methylation and chromatin modification states are the most heavily studied of all epigenetic marks due to the accessibility of the associated assays, with most studies conducted on DNA methylation. Variation in DNA methylation among individuals and populations is usually determined by examining DMRs and differentially methylated positions (DMPs, SMVs, or single methylation polymorphisms – SMPs) – akin to single nucleotide polymorphisms (SNPs). These data are then used for downstream analyses for determining various population epigenomic parameters and associations with various phenotypic, disease, and adaptive traits. Here, we first briefly describe the molecular methods used for assaying the epigenomic marks/variants and then bioinformatics methods for determining epigenomic states and epigenotypes.

4.1 Molecular Methods

4.1.1 Global Methylation and Methylation-Sensitive Marks

Global DNA methylation analyses are mostly used to address questions regarding the extent and proportion of methylation present in a genome and do not enable high-resolution detection of the sequence context in which methylation occurs. Since the discovery of DNA methylation using paper chromatography (Hotchkiss 1948), chromatography-based methods have been regarded as a gold standard approach for the analysis of global DNA methylation due to their accuracy and reproducibility (Ettre 2001). Thin-layer chromatography (TLC) and high-performance liquid chromatography (HPLC) (Friso et al. 2002; Magaña et al. 2008) are currently the most extensively used chromatographic methods for assessing the global levels of 5-methylcytosine (5mC). In TLC-based techniques, purified DNA is digested to mononucleotides using nuclease P1 and then separated on a thin-layer chromatography plate based on their distinct size and mobility (Kuchino et al. 1987). The relative intensity of the spots (proportional to the amount of each analyte) can be visualized using various techniques (Oakeley 1999). To confirm separation of 5mC from unmethylated cytosine (C), 5mC monophosphate is usually run in parallel on a

control plate (Oakeley 1999). While highly accurate, TLC has a limited resolution compared to HPLC (Reich and Schibli 2007). First employed by Kuo and colleagues for the global analysis of methylated cytosine, HPLC is one of the oldest and most accurate tools for analysis of global DNA methylation (Kuo et al. 1980). In this approach, DNase I, nuclease P1, and alkaline phosphatase are used for hydrolysis of DNA into individual deoxyribonucleosides that are then separated by reverse-phase high-pressure liquid chromatography (Kuo et al. 1980). The separated analytes obtained from as little as 3 μg of the total DNA can be subsequently detected using ultraviolet absorption at 254 and 280 nm (Kuo et al. 1980; Armstrong et al. 2011).

Further development of chromatographic methods led to their coupling with mass spectrometry that provided a unique advantage for understanding the chemical composition of separated analytes. Most widely used variants of these techniques include thin-layer chromatography mass spectrometry (TLC-MS) (Song et al. 2005) and HPLC-MS (Friso et al. 2002; Le et al. 2011). During TLC-MS and HPLC-MS, analytes separated using chromatographic methods are passed through the mass spectrometer for confirmation of known chemical species, identification of novel bases, and quantitative measurement of the analytes (Song et al. 2016; Chowdhury et al. 2017). This coupling of chromatographic methods with mass spectrometry recently resulted in the identification of oxidized forms of 5mC, 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fc), and 5-carboxylcytosine (5caC) in mammalian genomes (Kriaucionis and Heintz 2009; Tahiliani et al. 2009; Ito et al. 2011; He et al. 2011). The combination of chromatographic and MS methods is currently extensively used for the identification and analysis of the 5mC oxidative derivatives (e.g., Zhang et al. 2012a, b).

Although the chromatography- and spectrometry-based methods offer an unparalleled accuracy and precision for detection of DNA modifications, these approaches do not provide any spatial information necessary to understand the biological functions of DNA methylation in tissue and organs with cell-type-specific DNA methylation patterns (Abakir et al. 2016). Generation of the 5mC-specific antibody allowed the development of immunochemical techniques that offer a robust and rapid analysis of global levels of DNA methylation as well as other DNA modifications in individual cells of different tissues (Santos and Dean 2006; Abakir et al. 2016). Although the immunochemical techniques currently employed for the detection of global DNA methylation in mammalian systems may vary in detail, they essentially involve the same steps. Among these are fixation of the cells or tissue sections in 4% paraformaldehyde (PFA), their permeabilization with a detergent (e.g., 0.1% Triton X-100), and depurination of the DNA using hydrochloric acid (HCl) to facilitate antibody access to DNA (Santos and Dean 2006; Abakir et al. 2016). Next, the samples are treated with specific anti-5mC antibody, and, finally, a fluorescent-labelled secondary antibody is used to visualize the 5mC signal by conventional or confocal microscopy (Santos and Dean 2006; Kremer et al. 2012). Importantly, the immunochemical techniques can also be employed for co-detection of 5mC with other DNA modifications (e.g., 5hmC, 5fC, or 5caC) (Ruzov et al. 2011; Almeida et al. 2012). Moreover, this protocol has recently been modified by

incorporating peroxidase-conjugated secondary antibody and tyramide signal amplification step that adds in an unprecedented sensitivity to the immunochemical detection of DNA modifications (Globisch et al. 2010; Wheldon et al. 2014; Abakir et al. 2016).

Importantly, the techniques described above do not provide information regarding the sequence specificity of observed DNA methylation. In contrast, employing methylation-sensitive isoschizomers of restriction enzymes (e.g., *Hpa*II sensitive to DNA methylation and a methylation-insensitive *Msp*I that both can recognize the same DNA restriction site depending on whether it is methylated or not) can lead to determining the global sequence-specific patterns of DNA methylation based on their fingerprints (Waalwijk and Flavell 1978; Lindsay and Bird 1987). This technique and its variants, such as MSAP and methylation-sensitive amplified fragment length polymorphism (MS-AFLP), though not quantitative, are particularly appealing in ecological and evolutionary studies where reference genomes are often not available (Reyna-Lopez et al. 1997; Yaish et al. 2014; Alonso et al. 2015). However, as the ability of isoschizomers to differentiate 5mC from unmethylated cytosine is restricted to sites of recognition of the corresponding enzymes, these methods have limited resolution (Yaish et al. 2014; Richards et al. 2017). To improve the resolution of the isoschizomer-based analysis, a modification of these techniques, EpiRAD, has recently been developed (Peterson et al. 2012; Schield et al. 2016). This technique is based on the use of barcoded adaptors that allow fragmentation of the samples by different pairs of restriction enzymes before size selection, amplification, and sequencing of the fragments (Peterson et al. 2012; Schield et al. 2016). The MSAP, MS-AFLP, and EpiRAD methods yield information on polymorphic DNA methylation loci and DNA methylation epigenotypes.

4.1.2 Bisulfite Sequencing

As both 5mC and C have the same base-pairing characteristics, identification of the methylation status of individual nucleotides had been a major hurdle for DNA methylation analysis prior to the advent of bisulfite sequencing. Although sodium bisulfite deaminates cytosine bases to uracil, thus changing the genomic DNA sequence, such treatment does not affect methylated cytosine (Clark et al. 2006). Subsequent PCR amplification of the bisulfite-treated DNA fragments leads to incorporation of unmethylated cytosines in place of the 5mC, while bisulfite-modified unmethylated cytosine (C) is being amplified as thymine (T); therefore, these bases can be discriminated from each other by standard sequencing techniques (Sanger and Coulson 1975; Sanger et al. 1977). Direct bisulfite sequencing based on the use of strand-specific PCR primers for amplification of the bisulfite-converted DNA followed by cloning of the amplified fragments into a vector and sequencing of the corresponding insert was developed by Frommer and colleagues in 1992 (Frommer et al. 1992; Clark et al. 2006). Since then, bisulfite sequencing has been widely used for determining the DNA methylation status of individual CpGs at both the level of single loci and genome-wide (Eckhardt et al. 2006; Kawakatsu et al.

2016a, b). As conventional bisulfite sequencing requires a cloning step, it is a low-throughput and time-consuming method. These limitations of conventional bisulfite sequencing were overcome by the introduction of pyrosequencing, where the ratio of C (methylated cytosine) and T in bisulfite-treated PCR fragments is determined by the amounts of C and T incorporated by DNA polymerase during sequencing reaction (Wong et al. 2006; Yaish et al. 2014). Unfortunately, use of pyrosequencing for the analysis of DNA methylation patterns in eukaryotes with large genomes is limited due to its prohibitively high cost (Zilberman and Henikoff 2007; Kacmarczyk et al. 2018). Thus, Meissner and colleagues introduced the reduced representation bisulfite sequencing (RRBS) as a more affordable alternative to pyrosequencing (Meissner et al. 2005). RRBS combines restriction digestion (to enrich for CpG-containing regions) and bisulfite sequencing to provide methylation analysis at single base resolution (Meissner et al. 2005). Although the technique is popular among epigeneticists due to its relatively low cost, its limitations consist of low sequencing coverage of some genomic regions that may originate from the incomplete digestion of methylated CpGs by restriction enzymes (Gu et al. 2011). More recently, employing the next-generation "sequencing by synthesis" approach together with bisulfite sequencing, numerous studies have provided genome-wide DNA methylation map (Cokus et al. 2008; Lister et al. 2008). Typically, genomic DNA is fragmented and ligated with Illumina adapters in which all cytosines are methylated and then bisulfite converted; the sites of methylated cytosines in the genome are revealed by deep sequencing (Cokus et al. 2008; Lister et al. 2008; Ziller et al. 2013). Bisulfite-converted DNA sequences can be processed for identifying DMRs, DMPs, SMVs, and/or SMPs.

Despite extensive use of bisulfite sequencing for analysis of DNA methylation patterns, it has several important limitations (Kacmarczyk et al. 2018). Specifically, incomplete bisulfite conversion may result in false detection of unmethylated cytosines as methylated, though standard bisulfite conversion kits typically achieve high conversion rates (Kurdyukov and Bullock 2016). Although prolonged bisulfite treatment has been shown to reduce such false positives, it can also result in degradation of the DNA (Grunau et al. 2001). These limitations highlight the delicate balance between achieving full conversion of unmethylated cytosines and retaining DNA integrity (Kurdyukov and Bullock 2016). Another major limitation of conventional bisulfite sequencing is its inability to discriminate between 5mC and its oxidized derivative, 5hmC (Nestor et al. 2010). To investigate the function of the relatively large quantities of 5hmC in mammalian genomes, recently, there has been several modifications of conventional bisulfite sequencing that allow mapping of 5hmC have been developed. One of them is based on chemical conversion of 5hmC to 5fC, which is called as unmethylated cytosine during sequencing (Booth et al. 2012, 2013). Yu and colleagues described an alternative technique that involves conversion of 5hmC to β-glucosyl-5-hydroxymethylcytosine (5gmC) protecting 5hmC from further oxidation (Yu et al. 2012). After this conversion, all 5mC bases are oxidized to 5caC by recombinant TET1 protein followed by the bisulfite treatment that, subsequently, converts 5caC to uracil leaving the original 5hmC (transformed to 5gmC) unaffected and called as C in sequencing reaction

(Yu et al. 2012). Moreover, Neri and colleagues developed the methylation-assisted bisulfite sequencing (MAB-seq) that allows mapping of 5fC and 5caC distribution patterns (Neri et al. 2016). In this method, bacterial CpG methyltransferase M.SssI converts unmethylated cytosines to 5mC (called as cytosine after bisulfite treatment) discriminating it from 5fC and 5caC transformed to uracil by bisulfite. Collectively, these methods can be used to determine the patterns of 5mC and all of its oxidation derivatives at single base resolution.

More recently, single molecule sequencing technologies (e.g., MinION, Oxford Nanopore Technologies) were employed for the discrimination of 5mC from non-methylated cytosine based on their distinct ionic currents (Rand et al. 2017). Although still in development, this approach looks very promising as it overcomes the need for chemical treatment of DNA for mapping the methylation patterns (Simpson et al. 2017).

4.1.3 NGS ChIP Sequencing

The composition and chemical nature of proteins interacting with DNA define chromatin states in eukaryotic genomes (Ren et al. 2000). Chromatin immunoprecipitation (ChIP) is a method that allows mapping of the sites of protein-DNA interactions using antibodies raised against specific chromatin-associated proteins or histone modifications (Jackson and Chalkley 1981). Most of the ChIP protocols typically include the following steps: cross-linking of DNA/chromatin-associated proteins using formaldehyde, sonication of recovered chromatin into shorter fragments, selective pulldown of the DNA bound by the protein of interest using specific antibodies, purification of the immunoprecipitated DNA fragments, and their analysis by qPCR or next-generation sequencing (NGS) (Buck and Lieb 2004). This approach has been extensively used for studying transcription (Weinmann and Farnham 2002; Valouev et al. 2008), DNA replication (Jackson and Chalkley 1981; Gadaleta et al. 2015), and cellular identity (Whyte et al. 2013; Rehimi et al. 2016).

Analogously to ChIP, specific antibodies raised against 5mC and its oxidized derivatives can also be used for immunoprecipitation of DNA fragments enriched in specific modifications in a technique termed DNA immunoprecipitation (DNA-IP or DIP; meDIP for 5mC-DNA-IP) that, in combination with high-throughput sequencing (HTS), is instrumental in determining the genomic distribution of these epigenetic marks in different systems (Weber et al. 2005; Pomraning et al. 2009). Moreover, meDIP can also be combined with ChIP for mapping the patterns of 5mC on DNA bound by the protein of interest (Mikkelsen et al. 2007; Moison et al. 2015). These approaches allow determining and comparing genome-wide distributions of histone modifications, specific histone variants, transcription factors, and DNA modifications (Novak et al. 2006).

Importantly, there are several factors that may affect the reliability of ChIP and DIP datasets (Lentini et al. 2018). Most crucial of them is the sensitivity and specificity of the antibody used in the immunoprecipitation (Spencer et al. 2009).

Moreover, a recent study reported that the intrinsic affinity of IgG for short unmodified tandem repeats may affect DIP-based genome profiling resulting in false positive rate of 55 to 99% (Lentini et al. 2018). Thus, normalization of the DIP datasets not only for input but also for IgG controls seems important for the reliable immunoprecipitation-based analysis of DNA modifications. Another important limitation of these techniques is that, unlike bisulfite sequencing-based approaches, neither of them yields single base resolution data.

4.1.4 sRNA Sequencing

Primarily evolved as a key defense system for silencing of parasitic foreign genetic material (Storz 2002; Lu et al. 2005), small RNAs (sRNAs) have also been shown to play critical roles in gene regulation and post-transcriptional silencing of gene expression (Studholme 2012). As sRNAs are usually less than 40 nt in length, several specific approaches have been designed for their capture and analysis of their distribution (Lu et al. 2005; Hafner et al. 2008). Although these approaches differ in their throughput and amount of required input material, they all involve isolation of sRNA, sRNA enrichment by size selection, ligation of the 5′ adaptors to both ends of sRNAs, conversion of sRNA into cDNA, and amplification/sequencing of corresponding cDNA fragments (Shendure and Ji 2008). Until recently, sRNA sequencing studies were mainly using either pyrosequencing or ABI Solid sequencing platforms; however, polymerase-based sequencing by synthesis on the Illumina sequencing platform is currently the most popular approach for sRNA analysis (Creighton et al. 2009; Eminaga et al. 2013).

In summary, although none of the current methodologies allow complete deciphering of chromatin states across individuals, creative integration of the described methods should help in furthering our understanding of how epigenetic variation influences evolution.

4.2 Bioinformatics Methods

Bioinformatics is performed on sequence data and other information provided by molecular assays for epigenomic variant and epigenotype calling and downstream analyses. Since most of the downstream analyses after epigenomic variant and epigenotype calling are similar to those used in population genomics, here we focus only on the bioinformatics methods used for epigenetic variant and epigenotype calling. An overview of bioinformatics methods used in population genomics is provided by Salojärvi (2018) in this book.

Bioinformatics analysis has become the rate-limiting step in all epigenomics analyses. Decreasing costs coupled with greater ease and speed of sequence data generation has resulted in the need to process increasingly large and complex

datasets. As HTS replaces arrays, individual samples are themselves larger and more complex, while multiplex sequencing has provided a means for the simultaneous sequencing of up to thousands of different samples. Bioinformatic analysis is now the most time-consuming step of most epigenomics studies. The type of data that are produced can also determine what downstream analyses are possible. It is, therefore, necessary for any epigenomics experiment to carefully consider how the data are to be managed and analyzed at the outset of a project. The number and type of bioinformatics programs and methods have proliferated alongside the data being produced. As a result, a fully comprehensive review of bioinformatics methods is beyond the scope of this chapter. Instead, we will highlight important aspects of bioinformatics analyses and common themes and emphasize key downstream steps unique to each method.

4.2.1 Microarray Data

Microarray data can provide information on single base methylation status and differentially methylated sites. Analysis of microarray data is more straightforward and computationally less intensive than that of sequencing data. Despite platform specifics, the basic output of microarrays consists of a measurement of fluorophore intensity from hybridization. Each signal of the array is derived from a set of probes of known sequence and typically known location in the genome. This prevents the need for subsequent mapping to the genome and simplifies downstream analyses. It also limits the dynamic range of measurement based on the number of available probes and often limits the resolution of the data and the potential for conducting further analyses on the dataset.

For both arrays and sequencing, the "raw" data which researchers usually work with have already undergone initial processing by proprietary software specific to each platform. This step, known as "feature extraction" for microarrays, converts scanned images into formats which may be specific to that platform (e.g., CEL for Affymetrix) (Grant et al. 2007). The first step in bioinformatics analysis for most researchers is to import these data formats and apply quality control (Fig. 4). This involves the creation of a variety of diagnostic plots for the identification of problematic arrays, missing data, etc. and filtering of these data (Grant et al. 2007).

Normalization is a critical step in analysis, used to reduce technical variation between arrays (Fig. 4). Proper normalization and the method used are dependent upon a number of factors, from overall experimental design to the specific platform used (Grant et al. 2007). Normalization methods and differential analyses commonly used in expression analyses may not be appropriate for ChIP-chip or other such assays (Buck and Lieb 2004). A special note should be made about the Illumina Infinium BeadChip technology (Illumina Inc.), whose HumanMethylation450 array is one of the most widely used for population-scale epigenomics in humans. A number of specialized programs have been developed for its analysis (see Morris and Beck 2015).

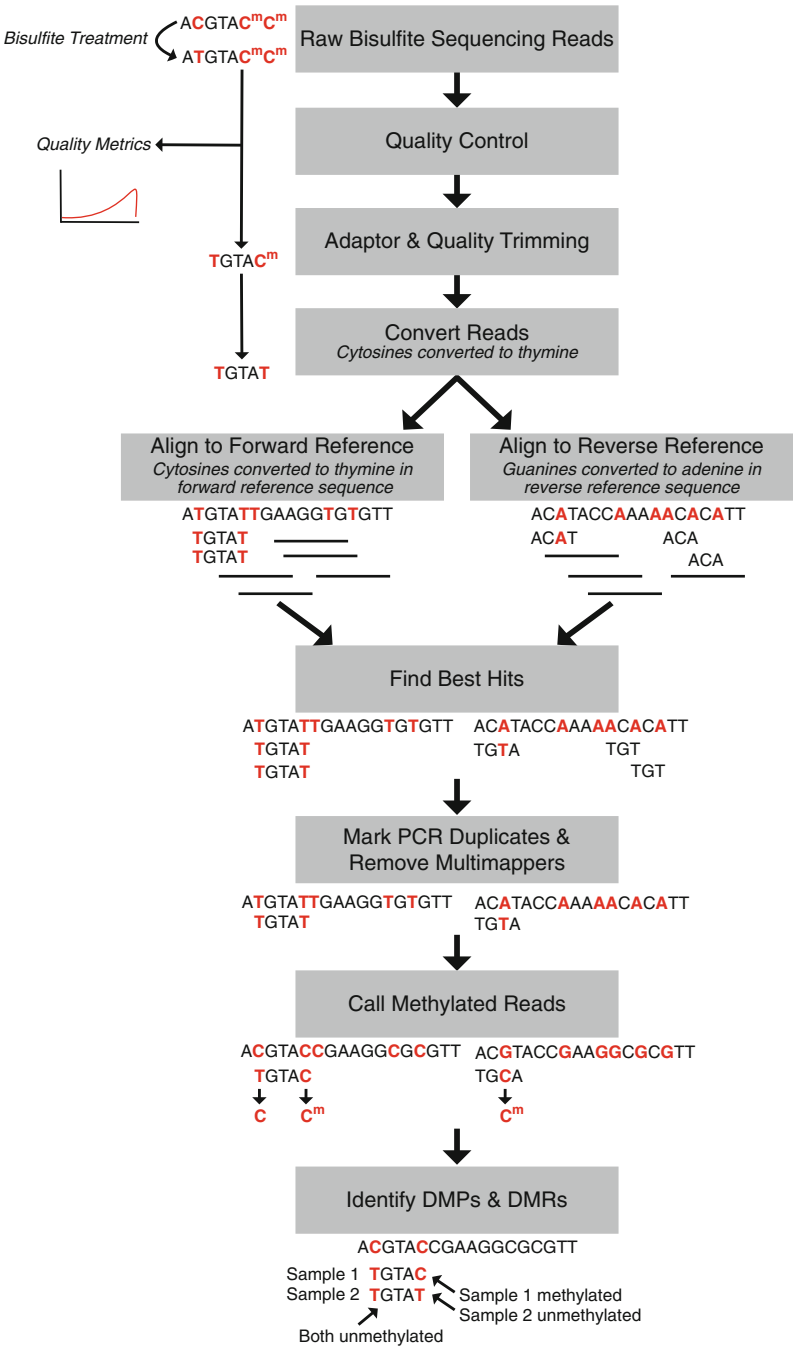

Fig. 4 A typical workflow for bioinformatics analysis of whole genome bisulfite sequencing

4.2.2 Sequencing Data

Declining costs, higher resolution, increased dynamic range, and the availability of more downstream analytical approaches together make HTS an increasingly attractive assay for population epigenomics. The downside of HTS is the greater challenge of bioinformatics analysis. For any sequencing-based epigenomics assay for which a reference genome is available, the key first step typically involves mapping sequenced reads onto the genome. This requires the following steps, which are common in the analysis of other types of HTS data: quality control, adapter and quality trimming, alignment, and post-alignment filtering (Figs. 4 and 5). Commonly used tools can be used for each of these steps. Alternatively, pipelines have been developed for many applications that combine all steps.

As for microarrays, quality control is an initial assessment of the raw data based on sets of common metrics, such as base quality, adapter contamination, overrepresented sequences, GC content, etc. This step itself usually does not involve any filtering of the data, and perhaps the most commonly used tool is FastQC. Read trimming is then used to improve mappability of the raw data to the genome. This is done by removing low-quality bases and contamination from adapter sequences. Versatile tools like Cutadapt (Martin 2011) and Trimmomatic (Bolger et al. 2014) can perform all these steps (and more) at once.

Next, reads are aligned or mapped back to a reference genome, a task for which a number of excellent programs are available. Several of the most commonly used programs include Bowtie (Langmead et al. 2009), its successor Bowtie 2 (Langmead and Salzberg 2012), BWA-MEM (Li 2013), and its predecessors BWA (Li and Durbin 2009) and BWA-SW (Li and Durbin 2010). Duplicated reads arising from library construction bias downstream analyses, and it is typically not possible to accurately analyze reads that map to multiple locations. For most analyses, tools such as SAMtools (Li et al. 2009b) and Picard (http://broadinstitute.github.io/picard/) can be used to mark or remove these.

4.2.3 Enrichment-Based Sequencing

Most sequencing-based chromatin assays work by enriching for DNA from regions marked by a particular chromatin modification. These include methods specific to certain histones and histone modifications (ChIP-seq), methylated DNA (MeDIP-seq, MBD-seq, etc.), and DNA accessibility (DNase-seq, ATAC-seq, FAIRE-seq, MNase-seq, etc.). When mapped back to a reference genome, sequencing reads should stack up over these regions (Fig. 4). This creates "peaks" which can be distinguished from the background and could identify patterns of chromatin modifications among individuals (Fig. 5). It is often recommended to sequence unbound input DNA or DNA from a non-specific antibody to use as a control for background noise. Depending on the assay, these peaks represent where a particular chromatin state is found or not found. Numerous peak calling programs have been developed,

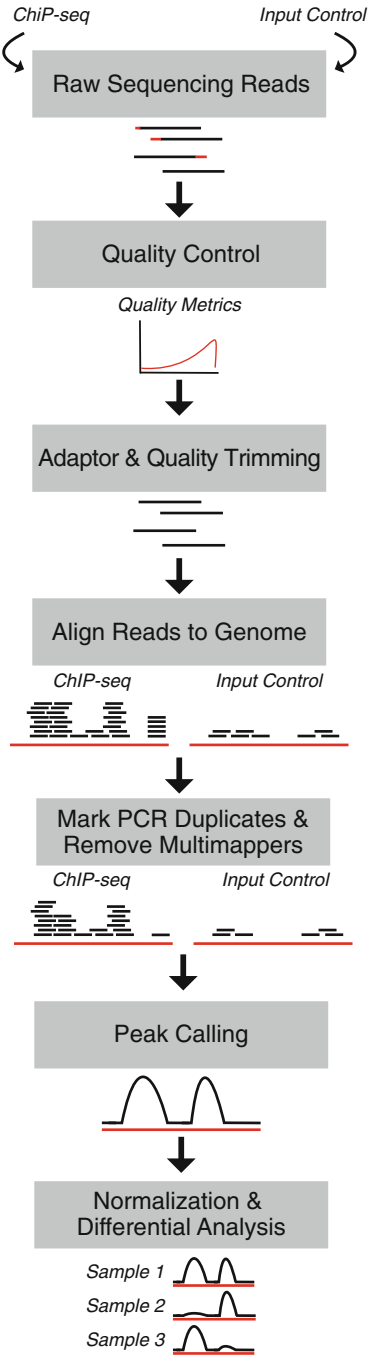

Fig. 5 A typical workflow for bioinformatics analysis of ChIP-sequencing data

as detailed by Bailey et al. (2013). An important consideration in choosing a program is whether or not the chromatin modification being assayed will create "broad" or "narrow" peaks. Transcription factors and other DNA-binding proteins are often limited to specific sequences, creating "narrow" peaks that are more easily detected computationally. However, some chromatin modifications often span large regions with no central peak, posing a greater computational challenge. For example, MACS is one of the most widely used programs for peak calling, applicable to both narrow and wide peaks (Zhang et al. 2008), while programs like SICER (Xu et al. 2014) were developed specifically for application to broad peaks.

When comparing samples, differential peak calling is used to identify differences in the amplitude of peaks and then infer differences in chromatin or protein binding from differential peak amplitudes (Fig. 5). Normalization between samples is an important consideration and remains one of the more challenging aspects of such analyses. One of the most straightforward and commonly used approaches is to normalize by sequencing depth, but more sophisticated methods have also been developed. One approach to differential peak calling is to apply methods originally developed for RNA-seq such as DESeq/DESeq2 (Love et al. 2014; Anders and Huber 2010) or edgeR (Robinson et al. 2010) to identify differences in read counts at previously determined peaks. This strategy is incorporated in the DiffBind program (Stark and Brown 2011; Ross et al. 2010). Other approaches are used in programs like MAnorm (Shao et al. 2012). Comprehensive comparisons of differential peak calling methods are available (Bailey et al. 2013; Steinhauser et al. 2016).

4.2.4 Bisulfite Sequencing

While the processing and mapping of sequencing data described above applies to most sequencing applications, bisulfite sequencing methods require additional considerations (Fig. 4). These include whole genome bisulfite sequencing (WGBS), RRBS, and variations such as Tet-assisted bisulfite sequencing (TAB-seq). This is due in part because the conversion of unmethylated *C*s to *U*s during bisulfite treatment, and then to *T*s during subsequent PCR, artificially introduces additional sequence variation between the sequencing reads and the reference genome. This reduces mappability, causing certain quality metrics to diverge from expected values. Furthermore, the methylation data from each read is "strand-specific," and so the specific strand from which a read originates from must also be identified. To address these issues, several mapping strategies have been devised. Figure 4 demonstrates one such general strategy as used by pipelines like Methylpy (Schultz et al. 2015), Bismark (Krueger and Andrews 2011), BS Seeker and BS Seeker2 (Chen et al. 2010; Guo et al. 2013), Bison (Ryan and Ehninger 2014), and the BSmooth algorithm (implemented in R package bsseq) (Hansen et al. 2012). This mapping scheme works by temporarily converting all *C*s in reads to *T* (*G* to *A* for reverse reads) in silico, which are then mapped to a "forward reference" where all *C*s in the genome have been converted to *T* and a "reverse reference" where all *G*s have been converted to *A*. Best hits can then be identified, data merged, and the original read

sequence restored. Other strategies from the one just described have also been developed and implemented in programs such as BSMAP (Xi and Li 2009).

Ultimately, it is typically desirable to use bisulfite sequencing data to identify DMPs, SMVs, or DMRs between individuals (Fig. 4). Calling individual sites as methylated or unmethylated remains challenging and is highly dependent upon factors such as sequencing coverage and sources of error. Pipelines like Methylpy (Schultz et al. 2015) incorporate statistical methods of determining the methylation status of each individual site from a single sample. Multiple programs are available for the determination of DMPs and DMRs between samples. Some pipelines, such as Methylpy (Schultz et al. 2015) and bsseq (BSmooth) (Hansen et al. 2012) incorporate mapping and DMP/DMR calling in the same package. Several stand-alone programs are also available, such as DSS (Feng et al. 2014). Bioinformatics software developers have also recently begun incorporating machine learning algorithms, such as in the program HOME (Srivastava et al. 2017). Some programs, like methylKit (Akalin et al. 2012), were developed with additional considerations for RRBS data. See Shafi et al. (2018) for a review of DMP and DMR calling software.

5 Association of Epigenomic Variation with Phenotypes and Ecological Acclimation and Adaptation

Individuals within populations often exhibit broad phenotypic variation in morphological, physiological, and behavioral traits. This variation partially reflects differential selection upon phenotypes by environmental conditions particular to a given location, which is essentially the basis of local adaptation. In part, this is because individuals with heritable phenotypes that are well-adapted to their environments are likely to display fitness advantages over local conspecifics and thereby increase the frequency of these phenotypes within the local population. Population and landscape genomics studies frequently use correlative or association approaches to link variable genomic regions or specific alleles and haplotypes to environmental conditions or phenotypes (Balkenhol et al. 2017 – see the chapter in this book). However, genetic variation, such as SNPs, indels, and copy number variants (CNVs), may not be the only heritable variation contributing to local adaptation. Epigenetic modifications that affect adaptive traits may also be important for local adaptation in response to rapidly changing environments (Richards et al. 2010). Epigenetic modifications display a range of stability and heritable duration depending on their type and the function they serve and in some cases persist through meiosis. In contrast to genetic variation, the universality and role for such transgenerational epigenetic variation in natural plant systems remain incompletely known and controversial. However, examples from controlled plant studies demonstrating transgenerational epigenetic inheritance have generated an increase in research aimed at understanding

this departure from typical Mendelian inheritance. The implications of transgenerational epigenetic inheritance for evolutionary processes are essentially unexplored (Cushman 2014).

The forms of epigenetic regulation that are most ecologically important, and most well-described, separate into two groups including (1) enzymatically controlled, reversible covalent modifications of DNA and histone proteins, most often methylation, and (2) the activity of siRNAs which influence and in some cases dictate the former group (Nicotra et al. 2010). However, the relative ease of laboratory procedures for the study of 5mC have resulted in strong representation of this mechanism in studies of epigenetic variation (Birney et al. 2016; Richards et al. 2017). 5mC in CG, CHG, and CHH (where H can be A, C, or T) nucleotide contexts silences transposable elements and influences the expression of genes related to numerous ecologically important traits (Law and Jacobsen 2010). Meanwhile, the heritability of histone modifications remains less well-explored (Eichten et al. 2014; Verhoeven et al. 2016), but see Nightingale et al. (2006). Here, we provide an overview how epigenetic variation (epialleles) influences phenotypes and ecological acclimation and adaptation by exploring what is known about epigenetic control of ecologically and environmentally important traits and some approaches that can be employed to investigate the possible role of epigenetic mechanisms in phenotypic variation and ecological acclimation and adaptation.

Recent studies suggest that epigenetic mechanisms and epigenomic variation contribute significantly to phenotypes, abiotic and biotic stress responses, disease conditions, adaptation to habitat, and range distributions in a variety of organisms (review in Richards et al. 2017). However, most related studies have so far focused on plants and humans. Therefore, in this section, we briefly discuss the association of epigenomic variation with phenotypic, growth, disease, and ecological acclimation and adaptive traits in plants. The next section of this chapter explores this aspect in human disease conditions.

5.1 Phenotypic Traits

The adaptive genetic variation found in plants is commonly studied using GWAS that investigate how variable genomic markers (mainly SNPs) associate with variation of adaptive phenotypic traits (Korte and Farlow 2013) and environmental factors (such as climatic factors) using environmental association analysis (EAA) (Rellstab et al. 2015), often employing common garden approaches, which control for environment variation, to sort out genetic variation (Ingvarsson and Street 2011). Common garden and reciprocal transplant studies are useful for assessing phenotypic variation, partitioning variation into genetic and environmental components, and assessing the heritability of phenotypic variation. However, even with very large SNP datasets, these approaches often fail to identify SNPs associated with phenotypes, and the problem of the so-called missing or unexplained heritability persists (Talbot et al. 2017). This phenomenon occurs where variation in measured

phenotype is explained only by a small portion of genetic and environmental variation, leaving much of the variation unaccounted for. This unaccounted-for variation could in part be due to epigenetic mechanisms and epigenomic variation.

In the absence of genetic variation (Schmitz et al. 2013a), the same approaches used in population genomics, such as GWAS and EAA (Fig. 6), can be applied to investigate epigenetic associations with environmental heterogeneity, phenotypic variation, or both (Verhoeven et al. 2016). The presence of genetic and environmental variation drastically complicates the interpretation of epigenomic investigations due to interactions among the epigenome, the genome, and the environment in which they occur. Until recently, most studies of epigenetic variation have been based on methylation-sensitive amplified fragment length polymorphisms (MS-AFLPs) and related MS-RFLPs. While such non-sequencing-based approaches provided initial insights into patterns of genome methylation, they do not allow interrogation of epigenomic variation across the entire genome, which invites the strong possibility of falsely detecting positive associations to methylation variants (Verhoeven et al. 2016; Greally 2017). While these low-resolution approaches are very informative for certain questions and study systems, whole epigenome or RRBS does a better job of detecting epigenetic variation across an entire genome and improves insight into the proportion of the epigenome under genetic control (Eichten et al. 2013). RRBS is particularly effective for assessing numerous samples of non-model species where de novo assembly of the epigenome will be required to understand patterns of methylation.

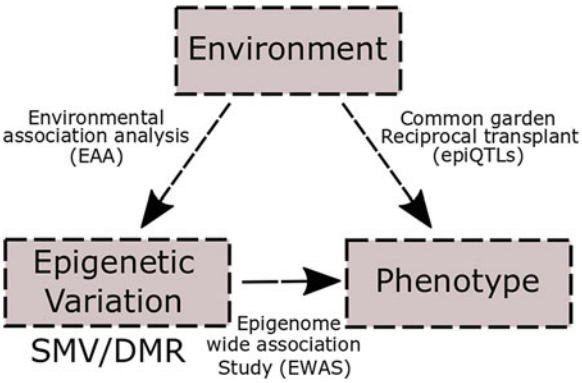

Fig. 6 After disentangling the confounding effects of genetic variation, relationship between epigenetic variation and local adaptation can be investigated through different approaches. EAA can be used to associate epigenetic variation with environmental heterogeneity such as climate or soil factors. EWAS can be used to associate epigenetic variation with variation in phenotypes either in the wild or in designed common gardens. epiQTLs can be used to assess patterns of heritability and variation in phenotypes with different environmental conditions in common gardens. A combination of analysis is likely to be the most informative. *SMV* single methylation variant, *DMRs* differentially methylated regions. Modified from Rellstab et al. (2015)

5.1.1 Environment Association Analysis: Outlier Detection

Outlier analysis is also used in population genomics to identify loci potentially involved in adaptive traits. Outlier tests assess patterns of allele frequencies to detect those that do not match the general patterns found throughout the rest of the genome, that is, they deviate from selectively neutral expectations (Luikart et al. 2003). In general, whole genome or reduced representation genomics facilitate the identification of thousands to millions of SNPs that can be screened for these outlier patterns, and then the loci can be correlated with environmental heterogeneity (Balkenhol et al. 2017). In epigenetic environmental association analysis (epiEAA), patterns of genome-wide or targeted 5mC can be correlated with environmental heterogeneity. Research on the role of DNA methylation in plants to variable environmental conditions has been demonstrated in the model plant *A. thaliana* (Dowen et al. 2012), for which stress-induced regulation of immune response was associated with 5mC variants, suggesting the importance of environmental context for epigenetic regulation of pathogen defense.

In wild plant populations, epiEAA has been used to assess 5mC variation in relation to habitat type and climate. For example, using RRBS, Gugger et al. (2016) associated SMVs found in *Quercus lobata* with climate variables across the species range in California. They found 43 SMVs associated with climate variables, notably maximum temperature. Interestingly, the 43 SMVs tended to occur in or near genes that were known to be involved with plant response to environment. Another EAA, using MS-AFLP, found that patterns of 5mC in *Spartina alterniflora* and *Borrichia frutescens* were significantly associated with salinity gradients representing different habitat types (low, moderate, and high salinity) (Foust et al. 2016). The association of epigenetic variation was interesting because it differed from genetic variation, where *S. alterniflora* had no association with genetic variation and habitat type and *Borrichia frutescens* had a genetic association in addition to the epigenetic association suggesting that environmental factors had at least a role in DNA methylation (Foust et al. 2016). More studies are needed to assess the extent to which genetics, pure epigenetic traits, and the environment drive these patterns of epigenetic variation in wild populations.

5.1.2 Epigenome-Wide Association Studies

In contrast to EAA approaches, which link epigenomic variation to environmental conditions, phenotype approaches can also be used to relate epigenetic/epigenomic variation to evolutionarily important or adaptive phenotypes. Almost any plant phenotype exhibiting variation or plasticity may be epigenetically controlled. Epigenetic variation may be genetically controlled, but environmental variation may also result in heritable DNA modifications, resulting in epigenetically-regulated plasticity (Feinberg and Irizarry 2010). In order for transgenerational epigenetic plasticity to be evolutionarily important, the variation needs to be adaptive and

heritable (Whipple and Holeski 2016). In EWAS, whole methylome or reduced representation methylomes can be scanned for DMRs and SMVs that can then be statistically associated with adaptive phenotypes. Controlled, replicated studies involving common garden and epiRILs will continue to serve a critical role in determining the extent to which genotypes explain observed correlations between sample phenotypes and epigenomes (Fig. 7); see also Richards et al. (2017).

5.1.3 Untangling Genetic vs. Epigenetic Control of Phenotypes

Several studies have addressed the correlation between genetic variation and epigenetic variation with differences in environment (Dubin et al. 2015; Meng et al. 2016; Taudt et al. 2016). For example, high epigenetic differentiation among populations of *Fallopia* spp. and epigenetic loci associated with microhabitat conditions suggest that these loci could contribute to phenotypic variation in genetically depauperate populations (Richards et al. 2012). In another case, *Helleborus foetidus* plants had higher epigenetic variation compared to genetic variation, and epigenetic markers

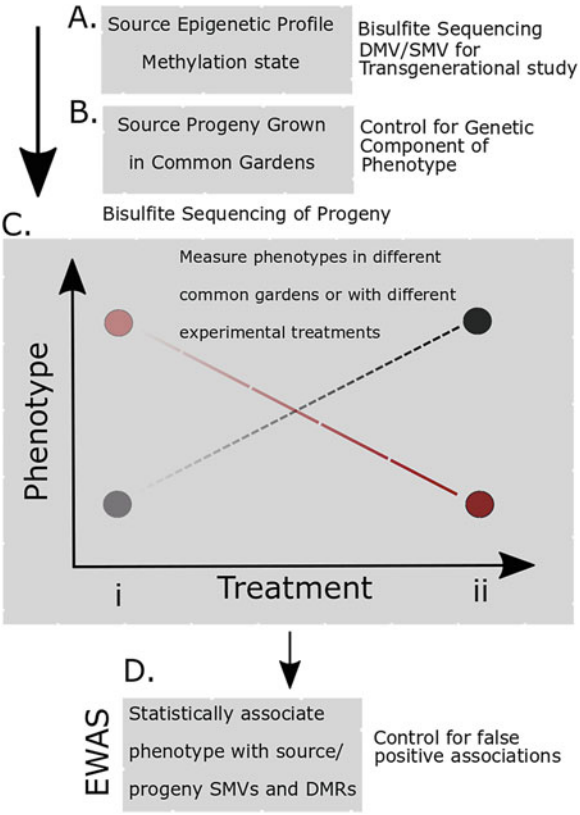

Fig. 7 Conceptual workflow for EWAS. Source populations are sampled. Either (**a**) wild populations are bisulfite sequenced and phenotypes are measured or (**b**, **c**) seeds are collected and grown in common gardens where controlled crosses can be made or known pedigrees can be used. They can then be bisulfite sequenced, and phenotypes can be measured under different environmental conditions or experimental treatments. Finally, identified variable epigenomic marks or regions can be statistically associated with variation in measured phenotypes (**d**). Different sources or different family lines can be tested under different environment or experimental conditions, and changes in phenotypes can be measured (**c**). By controlling for genotype, epigenetic contribution can be inferred

were significantly associated with whole plant functional traits suggesting that epigenetic diversity allows plants to exploit a more broad range of ecological conditions (Medrano et al. 2014). Given evidence of heritability of beneficial epigenetic variation, it can be argued that epigenetic variation is acclimative/adaptive. For example, epigenetic variation within and between populations of *Viola cazorlensis* was correlated with adaptive divergence (Herrera and Bazaga 2010).

5.1.4 epiQTL

Epigenomic variation at a single locus can be assessed by treating it as a quantitative trait (QTL) and calculating its heritability (Taudt et al. 2016). If epigenetic inheritance is absent and calculated heritability estimates are greater than zero, this suggests that epigenomic variation is under genetic control. However, if heritability estimates are zero, then variation may be a product of variation in environmental factors and epigenetically controlled plasticity. The use of EWAS can be used to detect variants (see Box 2 in Taudt et al. 2016), but to date most EWAS have been conducted in human disease studies or in model plant systems (e.g., Birney et al. 2016; Chen et al. 2016).

5.2 Ecologically and Environmentally Relevant Traits

When multiple replicated genotypes are grown together in the same environment (common garden), the genetic component of the phenotype can be identified providing insight into ecologically and environmentally relevant traits in plants. One approach is to use epiRILs to identify individuals with known epialleles, and then using reciprocal inbred lines/clones assesses variation in phenotype in different environments. Epigenetic recombinant inbred lines have been developed in model plants, for example, *A. thaliana*, and are a powerful tool that may help assess the role of transgenerational epigenetic variation on adaptive phenotypes (Johannes et al. 2009; Reinders et al. 2009). Using high-resolution sequencing approaches, thus, allows the identification of epiQTLs (here epiQTLs are quantitative trait loci that are pure epigenetic variants, rather than obligatory or facilitated) and can aid in the quantification of the role of epigenetic variation and selection for evolutionarily important traits.

In *A. thaliana*, EWAS and high-throughput phenotyping were used to identify DMRs in epiRILs, and identified epiQTL were then associated with defense compounds and flowering time (Aller et al. 2018). This study concluded that variation in DNA methylation accounted for less phenotypic variation than that accounted for by genetic variation. Nonetheless, analysis of epigenetic variation in plants and their associated phenotypes have provided important insights into the role of epigenetic regulation of acclimative and adaptive traits. For example, studies of *A. thaliana* have

shown heritable epigenetic variation in ecologically important traits, such as flower symmetry (Cubas and Vincent 1999), flowering time (Johannes et al. 2009), and methyltransferase regulation (Dubin et al. 2015; Kawakatsu et al. 2016a, b). Related studies in non-model organisms remain less common, and studies in wild populations are even rarer.

Several studies have found correlations between ecologically and environmentally relevant plant traits and epigenetic variation in methylation state (Table 1). For example, levels of cytosine methylation have been correlated with leaf shape (Herrera and Bazaga 2013), floral symmetry (Cubas and Vincent 1999), whole plant functional traits (Medrano et al. 2014; Lele et al. 2018), flowering time (Johannes et al. 2009), cold tolerance (Xie et al. 2015), salinity tolerance (Foust et al. 2016), disturbance (Herrera et al. 2016), disease susceptibility (Dowen et al. 2012; Sollars and Buggs 2018), and climate (Platt et al. 2015; Gugger et al. 2016) (Table 1). DNA methylation patterns have also been correlated with shifts in species ranges (Richards et al. 2012; Xie et al. 2015), functional diversity in terms of productivity and stability of plant populations (Latzel et al. 2013), and inbreeding depression in plants (Vergeer and Ouborg 2012).

6 Association of Epigenetic Mechanism and Epigenomic Variation with Human Diseases, and Pharmacoepigenomics

Epigenetic mechanisms have been implicated in several biological processes in humans, such as development, cell differentiation, X chromosome inactivation, and genomic imprinting (Kiefer 2007). During the last decades, a large number of studies have focused on elucidating the role of epigenetics in pathogenesis, progression, and response to treatment of different diseases (Murrell et al. 2013). Recently, an increasing number of population epigenomics studies in humans have focused on investigating associations between epigenomic variation and disease conditions through EWAS. Here, we provide an overview of association of epigenetic mechanism and epigenomic variation with human diseases.

6.1 Genomic Imprinting Diseases

Initial evidence for the importance of epigenetics in the pathogenesis of human disease came from studies of genomic imprinting, a mechanism underlying the monoallelic expression of a gene according to its parental origin (Reik 1989). Certain imprinting-related diseases, e.g., Beckwith-Wiedemann syndrome (BWS), lead not only to developmental abnormalities but also to somatic overgrowth and predisposition to childhood malignancies, highlighting the importance of epigenetic

Table 1 Selected studies showing associations of epigenomic variation with phenotypic, disease, and adaptive traits in plants using epiEAA or EWAS

Traits	Species	Method	Results	Reference
Floral symmetry	*Linaria vulgaris*	RFLP/cDNA	Methylation of Lcyc gene alters floral symmetry from bilateral to radial	Cubas and Vincent (1999)
Methylation state	*Taraxacum officinale*	Multivariate analysis	Progeny of abiotically stressed parent plants was raised in a common garden and screened for changes in methylation state. Environmental stress increased epigenetic variation in offspring	Verhoeven et al. (2010)
Epigenetic variation	*Viola cazorlensis*	MS-AFLPs epigenetic vs. genetic variation (correlation)	Epigenetic structure within and between populations found to correlate to adaptive phenotypes	Herrera and Bazaga (2010)
Methylation state, height	*Oryza sativa* L.	RR Methylome/MSAP	Nutrient deficiency (Nitrogen) alters heritable methylation profiles	Kou et al. (2011)
Pathogen defense genes	*Arabidopsis thaliana*	Whole genome methylation	Variation in DNA methylation associated with abiotic stress alters gene expression	Dowen et al. (2012)
Methylation state	Japanese knotweed (*Fallopia* spp.)	MS-AFLPs epigenetic vs. genetic variation (GenAlEx)	High epigenetic differentiation among sites with evidence of some epigenetic loci responding to local microhabitat conditions, suggesting that epigenetics may contribute to phenotypic variation in genetically depauperate populations	Richards et al. (2012)
Methylation profiles, leaf type	*Ilex aquifolium* (Aquifoliaceae)	MSAP LME	MSAP markers were significantly associated with leaf shape (prickly vs. non-prickly). DNA from prickly leaves had significantly less methylation	Herrera and Bazaga (2013)
Whole plant, leaves, regenerative traits	*Helleborus foetidus*	LME genetic vs. epigenetic variation	Plants had higher epigenetic vs. genetic variation. 13% of MSAP markers were significantly associated with plant traits. Epigenetic diversity may allow plants to exploit a broad range of ecological conditions	Medrano et al. (2014)

(continued)

Table 1 (continued)

Traits	Species	Method	Results	Reference
CpG methylation polymorphisms	*Quercus lobata*	EWAS	CpG methylation polymorphisms correlated to local adaptation to climate	Platt et al. (2015)
Cold tolerance	*Ageratina adenophora*	Whole genome methylation	A decrease in methylated sites was associated with an increase in freezing tolerance and linked to the ICE1 epiallele for demethylation. Epigenetics correlates with cold tolerance and is associated with invasibility in crofton weed in China	Xie et al. (2015)
Climate, maximum temperature	*Quercus lobata*	EAA/RRBS	43 SMVs significantly associated with four climate variables. CG methylation is important in locally adaptive evolution or plasticity in plants	Gugger et al. (2016)
Salt tolerance, habitat type	*Spartina alterniflora; Borrichia frutescens*	EAA/MS-AFLP	Sampled two salt marsh perennials along high, mid, and low salinity gradients in GA. Found evidence of epigenetic variation associated with habitat (salinity), suggesting epigenetic control of salt tolerance	Foust et al. (2016)
Leaf shape and photosynthetic traits	*Populus simonii*	MS-AFLPs/ EWAS Population epigenomic variation	413 methylation sites were used to partition *P. simonii* into three distinct populations. Epigenetic subpopulations were associated with environmental variation. Association analysis identified methylated regions that may influence photosynthesis and leaf development	Ci et al. (2016)
Whole plant, leaves, regenerative traits	*Vitex negundo* var. *heterophylla*	MS-AFLPs epigenetic vs. genetic variation (correlation/Mantel)	Both genetic and epigenetic diversity were low between habitats within sites, suggesting a genetic basis of adaptation to habitats and significant correlation between epigenetic (but not genetic) variation and plant phenotypes	Lee et al. (2017)

(continued)

Table 1 (continued)

Traits	Species	Method	Results	Reference
Flowering time, defense compounds	*Arabidopsis thaliana*	epiQTL	Variation in DNA methylation was lower than that of genetic variation	Aller et al. (2018)
Disease susceptibility	*Fraxinus excelsior*	Whole genome methylation	Genes associated with disease susceptibility were differentially methylated	Sollars and Buggs (2018)

mechanisms for the pathogenesis of cell proliferation-related defects and tumorigenesis (Weksberg et al. 2003). Correspondingly, several studies show preferential loss of maternal alleles in the *11p15.5* region of the BWS-associated imprinted chromosomal area in BWS-related tumors (Soejima and Higashimoto 2013), providing evidence that epigenetics and imprinting contribute to etiology and development of BWS (Zoghbi and Beaudet 2016). There is currently a large body of experimental evidence on the contribution of epigenetic mechanisms to pathogenesis and development of a wide range of human diseases, including cardiovascular (Martinez et al. 2015), kidney-related (Reddy and Natarajan 2011), and neurological diseases (Landgrave-Gómez et al. 2015) and cancer (Bennett and Licht 2018).

6.2 Cardiovascular Diseases

Cardiovascular diseases include hypertension, cardiomyopathy, arrhythmias, atherosclerosis, and myocardial infarction (Duygu et al. 2013). Epigenomic profiling of DNA methylation revealed hypomethylation of promoter regions and hypermethylation of gene bodies in patients with primary and secondary cardiomyopathies (Movassagh et al. 2011). In addition to changes in the patterns of DNA methylation, elevated levels of *miR-499* have also been associated with heart failure in both hypertrophy and cardiomyopathy (Matkovich et al. 2012), and increased levels of *mir-208* have been shown to induce hypertrophy and fibrosis (Van Rooij et al. 2006). Importantly, overexpression of *mir-208* serves as a clinical diagnostic predictor of heart failure (Satoh et al. 2010). Moreover, DNA methylation, histone modifications, and miR-mediated regulation have all been implicated in hypertension development (Papait et al. 2013). Thus, decreased global 5-methylcytosine (5mC) content correlates with hypertension suggesting that global DNA hypomethylation is indicative of the progression of this disease (Smolarek et al. 2010). In contrast to global hypomethylation of the genome, methylation of the *11βHSD2* promoter region is associated with the disruption of *11βHSD2*-mediated conversion of cortisol into cortisone in hypertension (Friso et al. 2008; Udali et al. 2013). The subsequent imbalance of cortisol and cortisone promotes the pathogenesis of hypertension (Friso et al. 2008) that, apart from impacting

the cardiovascular system, leads to the development of kidney-related diseases (Quinkler and Stewart 2003).

6.3 Kidney-Related Diseases

The kidney is a complex multicellular organ that can be affected by various environmental signals (Adli et al. 2015). Different kidney cell types acquire distinct profiles of epigenetic marks during renal development and nephrogenesis (Patel and Dressler 2013; Adli et al. 2015). During the last decade, a number of studies have focused on the role of DNA methylation in chronic kidney diseases including diabetic nephropathy (Reddy and Natarajan 2011). Specific changes of DNA methylation in the diabetes susceptibility genes (e.g., *UNC13B*) were reported for patients with diabetes (Bell et al. 2010). Moreover, genome-wide profiling of 5mC revealed that DNA methylation of enhancers of the key fibrotic genes is altered in patients with chronic kidney disease leading to overexpression of the corresponding genes (Ko et al. 2013). Specific histone modifications were also implicated in the alterations of expression of fibrotic, cell cycle, and inflammatory genes causing chronic kidney diseases (Sun et al. 2010). Correspondingly, both hyperacetylation of histones and increased H2K4me were associated with changes in insulin expression leading to the development of diabetes (Ling and Groop 2009).

6.4 Neurodegenerative Diseases

Epigenetic mechanisms are the critical determinants of cell-type-specific gene expression taking place during development of the brain (Keverne 2014), a highly complex and specialized organ, where most cell types originate from neural stem cells (Leto et al. 2016). Deregulation of epigenetic marks contributes to the development of several neurological diseases that are characterized by damage of neurons in specific brain regions (Landgrave-Gómez et al. 2015). For example, transcriptional repression of a number of genes involved in the pathogenesis of X syndrome and schizophrenia is associated with hypermethylation of the corresponding loci (Rangasamy et al. 2013). Similarly, several studies carried out in post-mortem brain samples of patients with Alzheimer's disease showed hypermethylation of genes associated with disease progression (Lu et al. 2013) and global DNA hypomethylation (Mastroeni et al. 2009). Parkinson's disease is another neurodegenerative disease that is correlated with impaired DNA methylation (Miranda-Morales et al. 2017). Hypomethylation of the *SNCA* gene, which encodes α-synuclein, contributes to the pathogenesis of this disease via structural changes or overexpression of this protein leading to its aggregation and abnormal gene expression (Kaidery et al. 2013). Moreover, microRNAs (miRNAs), such as miR-133b and miR-34b/c that have critical roles in maturation of dopaminergic neurons and packaging/trafficking of α-synuclein, are dysregulated in

the brain of Parkinson's patients (Kim et al. 2009; Miñones-Moyano et al. 2011). Several miRNAs are also deregulated in Alzheimer's disease (Landgrave-Gómez et al. 2015). Thus, miRNAs involved in the processing of amyloid precursor protein (APP) (e.g., miR-124) are downregulated in Alzheimer patients, resulting in impaired splicing of APP (Smith et al. 2011) followed by aggregate generation of small β-amyloid (Aβ) peptides which are the main cause in the development of this disease (Zhang et al. 2011). Importantly, the ectopic expression of miR-124 can rescue this splicing defect (Smith et al. 2011).

6.5 Cancer

A large number of recent studies focus on both the role of epigenetic mechanisms in cancer progression and the potential for targeted manipulation of epigenetic readers, writers, and erasers for cancer treatment (Bennett and Licht 2018). Most population epigenomics studies also target cancer. Specific alterations in the DNA methylation patterns are associated with both the initiation and progression of cancer (Wajed et al. 2001). Both global hypomethylation and hypermethylation of CpG islands of tumor suppressor genes represent characteristic features of cancer epigenomes (Feinberg and Vogelstein 1983). For instance, the *CDKN2A* gene encodes a cyclin-dependent kinase inhibitor $p16^{INK4A}$ that is crucial for cell cycle regulation (Li et al. 2012). Hypermethylation of its promoter has been found in a wide range of tumors, is thought to lead to uncontrolled cell cycle progression, and correlates with poor overall survival in patients suffering from non-small cell lung and colorectal cancers (Xing et al. 2013). Furthermore, *p73,* a gene closely related to *p53*, was also found to be hypermethylated in lymphomas (Pei et al. 2011). Similar to promoters of tumor suppressor genes, promoter regions of DNA repair genes are often hypermethylated in cancer (Lahtz and Pfeifer 2011). This is usually associated with gene repression that, in turn, leads to cancer initiation (Lahtz and Pfeifer 2011). Thus, hypermethylation of the *BRCA1* promoter results in reduced overall survival in breast cancer (Zhu et al. 2015), and hypermethylation of the *MLH1* promoter is commonly found in endometrial carcinomas and is associated with microsatellite instability (Esteller et al. 1998).

Although specific genomic regions are hypermethylated in many tumors, the key feature of cancer genomes is their global hypomethylation (Rodriguez et al. 2006). Importantly, the overexpression of proto-oncogenes and growth factors that are critical for cancer progression and development is often associated with hypomethylation of their promoters (Szyf et al. 2004). Moreover, genomic instability has also been reported for different tumors, possibly related to hypomethylation of retrotransposable elements and their subsequent aberrant mobilization (Daskalos et al. 2009; Ross et al. 2010).

The oxidized forms of 5mC are also of high importance for tumor pathogenesis according to a number of recent studies (Ficz and Gribben 2014), with 5-hydroxymethylcytosine (5hmC) levels significantly reduced in many cancers

(e.g., breast, colon, and lung cancers) compared with the levels of this mark in normal tissues (Yu et al. 2012). This reduction in 5hmC content correlates with tumorigenesis and tumor progression in hematopoietic and solid tumors (Ficz and Gribben 2014). In myeloid malignancies, reduced 5hmC levels often correlate with *TET2* mutations (Abdel-Wahab et al. 2009; Ko et al. 2010), but, interestingly, the reduction of 5hmC levels in gliomas does not appear to be linked to mutations in TET proteins (Kraus et al. 2015). Surprisingly, unlike those of 5hmC, the levels of 5-carboxylcytosine (5caC) are elevated in some breast cancers, gliomas, and pediatric brain tumors suggesting that low 5hmC does not necessarily correlate with inactivation of TET-dependent oxidation of 5mC (Eleftheriou et al. 2015; Ramsawhook et al. 2017). Thus, it is likely that TET proteins may mediate preferential oxidation of 5mC to 5caC in some tumors which is in line with the reported significance of TET proteins and TET/TDG/5caC-dependent DNA demethylation for specification of glial and hepatic lineages (Wheldon et al. 2014; Lewis et al. 2017). Therefore, the oxidation of 5mC to 5caC in adult and pediatric brain tumors may represent an epigenetic signature of these cancers reflecting their likely neural progenitor/stem cell origin (Ramsawhook et al. 2018).

In addition to DNA methylation, the levels and distribution of histone modifications and miRNAs are also altered in cancers (Kunej et al. 2011; Sawan and Herceg 2010). Thus, the balance between histone acetylation and deacetylation was reported for many different tumors (Ropero and Esteller 2007). Increased histone acetyltransferase activity leads to hyperacetylation and the activation of proto-oncogenes in certain cancer types, whereas hypoacetylation results in silencing of tumor suppressor genes (Biswas and Rao 2017). Specifically, loss of lysine 16 acetylation (H4K16ac) and trimethylation of lysine 20 (H4K20me3) of histone H4 have been linked to a variety of cancers (Fraga et al. 2005b). Lysine methyltransferase SUV39H1 maintains genome stability; however, in acute myeloid leukemia, it causes abnormal methylation of H3K9 resulting in silencing of the $p15^{INK4B}$ tumor suppressor gene (Lakshmikuttyamma et al. 2010).

MiRNAs can contribute to the pathogenesis of cancer either directly or via the control of expression of epigenetic writers and readers (Kunej et al. 2011). Indeed, hypermethylation of miR-148 in breast cancer cells leads to tumor growth and metastasis, whereas its reactivation rescues the phenotype (Lujambio et al. 2008). Moreover, miR-124, the most prevalent miRNA in the brain, was found to be abnormally regulated in glioblastoma (GBM) (Karsy et al. 2012). This is particularly interesting as recent studies suggest that miR-124 may act as a tumor suppressor gene and could be useful in treating human GBM (Qiao et al. 2017).

6.6 *Epigenome-Wide Association Studies in Human Disease*

Recent advances in next-generation sequencing and bioinformatics allowed performing large-scale studies of human disease-associated epigenetic variation (Rakyan et al. 2011). A number of such epigenome-wide association studies focused

on genome-wide profiling of DNA methylation in different pathological conditions have been published over the last few years, and the corresponding datasets are currently available for the community (Rakyan et al. 2011). An EWAS that investigated potential links between metabolome and DNA methylation patterns identified an interesting association of differential DNA methylation with diabetes and smoking (Petersen et al. 2014). Another EWAS demonstrated association of high-density lipoprotein cholesterol (HPL-C) with the methylation status of a CpG site localized near the *DHCR24* gene, which is involved in cholesterol biosynthesis and is associated with metabolic traits (Braun et al. 2017). EWAS have also revealed a potential association of age-independent cardiovascular risk with DNA methylation (Fernández-Sanlés et al. 2018) and potential relationship between DNA methylation status of genes encoding liver enzymes and hepatic steatosis (Nano et al. 2017). Furthermore, EWAS have revealed significant associations of DNA methylation with kidney function (Chu et al. 2017), type 2 diabetes (Meeks et al. 2017), panic disorder (Shimada-Sugimoto et al. 2017), cardiovascular diseases (Nakatochi et al. 2017), cancer (Xu et al. 2013), chronic obstructive pulmonary disease and lung function (Lee et al. 2017), and other conditions.

Despite the majority of EWAS are purely correlational, they provide invaluable information that allows identification of potential targets for future in-depth mechanistic analysis and, thus, are of high importance for development of future specific epi-drugs. Also, caution should be exercised in the experimental design of EWAS to minimize or eliminate spurious associations (see Birney et al. 2016).

6.7 Epigenetic Biomarkers of Disease

Despite the large volume of information on epigenetic mechanisms of human diseases, there are currently no Food and Drug Administration (FDA) approved diagnostic tests relying solely on targeting epigenetic marks or the corresponding writer/reader proteins. However, there are several disease diagnostic tests incorporating epigenetic determinants (Kronfol et al. 2017). One of them is Cologuard, a screening test that examines the DNA methylation levels of the *BMP3* and *NDRG4* genes in combination with *KRAS* mutations and hemoglobin analysis (Kronfol et al. 2017). The methylation status of the *MGMT* promoter is used to predict the response to the temozolomide therapy in glioblastoma multiforme (Thon et al. 2013). Temozolomide alkylates or methylates DNA at the N-7 or O-6 positions of guanine residues causing DNA damage and, eventually, tumor cell death (Zhang et al. 2012a). Hypomethylation at the *MGMT* promoter leads to expression of the O^6-alkylguanine DNA acetyltransferase resulting in DNA repair inhibiting cell death caused by temozolomide (Esteller et al. 2000).

6.8 Epigenetic Drugs

While research on epigenome therapeutics is still in its infancy, the development of drugs targeting the epigenome attracts considerable attention from researchers working in the fields of epigenetics and biomedical science (Zhang et al. 2012b). These drugs are referred to as "epi-drugs" and are classified according to their targets as DNA methyltransferase inhibitors (DNMTi), histone acetyltransferase inhibitors (HATi), histone methyltransferase inhibitors (HMTi), histone demethylase inhibitors (HDMi), and histone deacetylase inhibitors (HDACi) (Kronfol et al. 2017).

DNMTi drugs inhibit DNA methylation and are divided into two groups: (1) nucleoside analogues/DNA binders and (2) antisense oligonucleotides (Singh et al. 2018). The first group of DNMTis integrates into the DNA and forms complexes with DNMTs promoting their degradation resulting in global reduction of DNA methylation (Stresemann and Lyko 2008). The first FDA-approved epi-drug targeting DNMTs was 5-azacytidine (Vidaza), an analogue of cytidine with the substitution of a nitrogen atom to a carbon atom in the 5-position of the heterocyclic ring (Jones and Taylor 1980). Vidaza has been used for clinical treatment of chronic myelomonocytic leukemia and myelodysplastic syndrome (Santi et al. 1984). As there are several limitations of Vidaza use, such as metabolic instability, low specificity, and induction of several side effects (Yoo and Jones 2006), other 5-azacytidine analogues including 5-aza-deoxycytidine, 5-fluoro-deoxycytidine (5-FC), zebularine, and S110 have been recently developed (Derissen et al. 2013). 5-FC has already been enrolled in clinical trials for the treatment of advanced solid tumors and acute myelomonocytic leukemia (Newman et al. 2015).

Another group of DNMTs is represented by molecules that do not integrate into the DNA but directly bind DNMTs and inhibit their activity (Singh et al. 2018). MG98, an antisense oligonucleotide, interacts with the 3′UTR of DNMT1 inhibiting methylation and promoting the expression of tumor suppressor genes (Amato 2007). However, due to poor efficacy and increased toxicity, it failed in clinical trials (Amato 2007).

Since patterns of histone deacetylation are often altered in cancer and metabolic diseases, HDACi epi-drugs are under development to inhibit the removal of the acetyl group from acetylated histones (Valente and Mai 2014). Several HDACis have been approved by the FDA (Eckschlager et al. 2017). One of them, Vorinostat, a histone deacetylase inhibitor that suppresses cell proliferation and promotes cell cycle arrest, is used for the treatment of cutaneous T-cell lymphoma (Xue et al. 2016). HMT inhibitors are the molecules that specifically target histone lysine methyl transferases (HKMTs) and protein arginine methyl transferases (PRMTs) competing with either their substrate or their cofactor S-adenosyl-L-methionine (SAM) for binding sites in the enzyme (Kronfol et al. 2017). There are currently three HMTi molecules at phases I/II of clinical trials: GSK126, EPZ6438 (Tazemetostat), and CPI-1209, which are all SAM inhibitors of enhancer of zeste homologue 2 (EZH2) (Castillo-Aguilera et al. 2017). These compounds are supposed to be used for the treatment of lymphomas and advanced solid tumors

(GSK126 and EPZ6438) as well as metastatic castration-resistant prostate cancer (CPI-1209) (Castillo-Aguilera et al. 2017). Another group of epi-drugs that target histone demethylases are the HDMis which are divided into two groups: the amine oxidases/histone demethylases (Lysine-specific demethylases LSD1/2) and the Jumonji C (JmjC) domain-containing histone demethylases (Morera et al. 2016). ORY-1001, a molecule that inhibits LSD1, is currently at phases I/II of clinical trials for the treatment of relapsed acute leukemia (Maiques-Diaz and Somervaille 2016). In preclinical studies, ORY-1001 has been found to reduce progression of acute myeloid leukemia (Maes et al. 2018). Additionally, a number of HATis are currently in preclinical studies for the treatment of several hematopoietic and solid tumors; however these studies are mainly restricted to in vitro experiments (Biswas and Rao 2017). Specifically, these are the compound C646, a small molecule inhibitor of the p300/CBP HAT family, and PU139 that inhibits several HAT subfamilies (Bowers et al. 2010). C646 was shown to inhibit cell growth in lung and prostate cancer cell lines and to cause growth arrest in melanoma lines, whereas PU139 was able to block neuroblastoma xenograft growth in mice (Gajer et al. 2015).

In summary, there is a large volume of experimental evidence supporting the association of abnormal regulation of epigenetic pathways with the development and progression of several diseases including cancer. Large-scale epigenetic profiling of different epigenetic-related diseases is warranted in order to target particular epigenetic regulators and develop highly specific epi-drugs.

7 Future Perspectives and Needs

Population epigenomics is a rapidly developing discipline at the intersections of molecular biology, physiology, population genetics and genomics. Variation in phenotypes, ecological acclimation and adaptation, and disease conditions represent just a few areas of research that have realized substantial advances by adopting population epigenomics methods. The recent development of RRBS sequencing has enabled high-resolution investigation of epigenomes of non-model organisms and across wild populations. However, numerous fundamental questions and challenges remain to be addressed in the field of population epigenomics, including:

- What is the extent of epigenomic variation within and between populations, and how closely does it follow patterns of genomic variation?
- To what extent is epigenomic variation influenced by genetic elements, environmental factors, and stochastic epimutations, and how do interactions among those sources of variation influence epigenomes? To what extent is genetically and environmentally induced epigenomic variation inherited transgenerationally? What conditions influence the number of mitotic and meiotic events through which an epigenetic phenotype may persist?
- How might epigenomic variation contribute to evolutionary processes? Does epigenomic variation facilitate strictly acclimation or both acclimation and

adaptation? Is epigenomic variation by itself sufficient to provide acclimation and adaptation under changing climate and environment conditions? How best can population epigenomics be applied to conserve or improve the sustainable use of natural resources and to treat human diseases?

In addition to the above challenges, there is an urgent need to improve cost-effective molecular and bioinformatics methods and to develop general principles in the field of epigenomics to generate robust experimental designs that will enable the accurate interpretation of results from population epigenomics studies. The current dearth of general principles and evolutionary theory concerning the role of the epigenome presents serious challenges to research in population epigenomics. This lack is due in part to the wide variety of epigenetic mechanisms that exist and their divergence across evolutionary histories, which often precludes the generalization of principles drawn from studies of model organisms to studies of wild systems (Richards et al. 2017). Additionally, epigenomes are far more complex than genomes and require association testing against the entire underlying genome in order to determine the extent to which the two are intertwined.

The bulk of our discussion has centered upon the three most commonly studied epigenetic mechanisms included in the epigenome (DNA methylation, histone modifications and variants, and non-coding RNAs). Other mechanisms also exhibit epigenetic properties, such as post-translational propagation of altered protein structure via prions (Halfmann and Lindquist 2010), ultrastructural chromatin compartmentalization producing topologically associated genomic domains (Jost et al. 2014), and microbial symbionts (Woodward et al. 2012). The epigenome, if limited to the three most commonly studied mechanisms, may largely constitute a molecular phenotype of an underlying genotype (Eichten et al. 2013; Greally 2017). How closely an epigenome adheres to the underlying genotype depends upon numerous factors and is especially dependent upon which mechanism is under consideration. There is a need for improved attribution of epigenetic states to genetic variation distant from or near to epialleles, such as *trans*-acting and *cis*-acting genetic elements, respectively. Disentangling genetic from non-genetic controls of observed epigenetic variation, in particular – the identification of false positives due to reverse causation – remains one of the largest challenges to the accurate interpretation of epigenomic data (Birney et al. 2016).

As our knowledge of the diversity of epigenetic marks continually expands (Stöger and Ruzov 2018), so does our need for new bioinformatics tools. Most of the bioinformatics tools that are used in population epigenomics research have been developed for population genomics analyses, and there is a need to develop bioinformatics methods, tools, and pipelines that are specific to the analysis of epigenomic data. Furthermore, there is a need to improve upon sequencing and genotyping technologies for population epigenomics studies to enhance the precision and cost-effectiveness of these technologies.

In contrast to genetic or genomic variation, epigenomic variation shows tissue and cell-specific patterns. Sampling a pool of cells and tissues from an individual can bias results and can yield erroneous conclusions, including spurious associations

with traits (see Birney et al. 2016). This results in an urgent need for quality control over experimental designs. To do so, individual tissues and cell types can be sorted and analyzed separately. Studies of mixed-cell samples should adopt strict bioinformatics measures to account for tissue and cell-type heterogeneity. Development of standards regarding sequencing read lengths and bioinformatics quality control is also needed to aid in the detection of epigenomic variation attributable to repetitive genomic regions and transcription factor binding sites.

Two major obstacles to the accurate interpretation of epigenomics studies discussed by Greally (2017) and Birney et al. (2016) are reverse causation and spurious associations. The problem of reverse causation in epigenomics is exemplified by interactions among methylation of the 5-carbon position of the DNA nucleotide cytosine (5mC), transcription factors, and transposable elements. For example, many instances of variation in 5mC may indicate, rather than cause, transcriptional regulation due to transcription factor binding prior to transcription, which alters local patterns of the enzyme-mediated covalent bonding of a methyl group (CH3) to cytosines (Greally 2017). Transposable element mobilization, which often alters gene expression (Agrawal 2001), may also represent both a cause and an outcome of 5mC dynamics (Fedoroff 2012). Spurious associations are likely to occur when a genetic basis for polymorphic chromatin-modifying molecular phenotypes evade detection, leading to the false conclusion that correlations between phenotypic traits and epigenetic phenomena are gene-independent. For instance, it has been estimated that approximately 22–80% of 5mC variation among the genomes of humans (Birney et al. 2016), and half of DMRs in maize, are significantly associated with genomic SNPs (Eichten et al. 2013). Further, 5mC varies not only among individual genomes but also across tissue and cell types within a single genome and even within a single chromosome over time. This flexibility of 5mC patterning results from the differential activity of multiple enzymes that are responsible for de novo methylation, the maintenance of methylation states across cell cycles, organismal ontogeny, as well as parental imprinting. However, depending on whether loci that encode enzymes associated with altered DNA methylation states occur locally (i.e., *cis*-acting methyl-QTLs, $<$50 kb from altered methylation states) or remotely (i.e., *trans*-acting methyl-QTLs, $>$50 kb and $<$2 Mb from altered methylation states) to a site with polymorphic methylation, genome-epigenome association assays will likely fail to detect causal genetic polymorphisms unless conducted using high-resolution epigenome profiles from samples representing cells of the same type and developmental stage. Fortunately, analytical tools and experimental approaches for averting the problems of reverse causation and spurious associations exist and are continually improving.

Overcoming these challenges will yield tremendous advances in fundamental and applied fields of research, such as contributing to the solution of the missing heritability problem (Whipple and Holeski 2016), developing new ways to combat the ravages of climate change (Nicotra et al. 2010; Bräutigam et al. 2013), and providing new approaches to some of the most widespread human diseases (Ling and Groop 2009; Bell et al. 2010; Bennett and Licht 2018).

8 Conclusion

Technical and conceptual advances have enabled the application of epigenomics tools to the interrogation of population processes, yielding profound advances in diverse fields of life science research. Awareness of the challenges to the accurate interpretation of epigenomic data will improve the ability to conduct cross-study comparisons in epigenomics and enable the development of general principles regarding the current role of epigenomic patterns across living systems. This will open new avenues for the application of epigenomics tools to environmental and medical subjects. While model systems will be instrumental in the development of general principles in epigenomics research, parallel studies in wild systems are likely to remain crucial as a reference for the relevance of findings gained from controlled studies. Ultimately, epigenomics tools and principles will be most useful when considered in light of the context in which the mechanisms evolved, within populations.

Acknowledgments E.R.V.M. and A.V.W. are supported by NSF Macrosystems grant no. EF-1442597. A.R.'s lab (A.A., M.E., L.C.L., and A.R.) is supported by Biotechnology and Biological Sciences Research Council [grant number BB/N005759/1] to A.R. A.A. is supported by Medical Research Council IMPACT DTP PhD Studentship [grant number MR/N013913/1] to A.A. OPR was supported by a Natural Sciences and Engineering Research Council of Canada Discovery Grant RGPIN 2017-04589. The authors thank Dr. Chad Niederhuth for helpful comments and contributions to writing the bioinformatics section, and Dr. Jesse Hollister for sharing his thoughts about the preliminary outline of the chapter.

References

Abakir A, Wheldon L, Johnson AD, Laurent P, Ruzov A. Detection of modified forms of cytosine using sensitive immunohistochemistry. J Vis Exp. 2016;16(114).

Abdel-Wahab O, Mullally A, Hedvat C, Garcia-Manero G, Patel J, Wadleigh M, et al. Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies. Blood. 2009;114(1):144–7.

Adams RL, Burdon RH. DNA methylation in the cell. In: Molecular biology of DNA methylation. New York: Springer; 1985. p. 9–18.

Adli M, Parlak M, Li Y, Eldahr S. Epigenetic states of nephron progenitors and epithelial differentiation. J Cell Biochem. 2015;116(6):893–902.

Agrawal AA. Phenotypic plasticity in the interactions and evolution of species. Science. 2001;294 (5541):321–6.

Aitken SN, Yeaman S, Holliday JA, Wang T, Curtis-McLane S. Adaptation, migration or extirpation: climate change outcomes for tree populations. Evol Appl. 2008;1(1):95–111.

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. methyl Kit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 2012;13(10):R87.

Akkerman KC, Sattarin A, Kelly JK, Scoville AG. Transgenerational plasticity is sex-dependent and persistent in yellow monkeyflower (Mimulus guttatus). Environ Epigenet. 2016;2(2): dvw003.

Aller EST, Jagd LM, Kliebenstein DJ, Burow M. Comparison of the relative potential for epigenetic and genetic variation to contribute to trait stability. G3. 2018. http://www.g3journal.org/content/early/2018/03/21/g3.118.200127.abstract.

Allfrey VG, Faulkner R, Mirsky AE. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. Proc Natl Acad Sci U S A. 1964;51:786–94.

Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nat Rev Genet. 2016;17(8):487–500.

Almeida RD, Loose M, Sottile V, Matsa E, Denning C, Young L, et al. 5-Hydroxymethyl-cytosine enrichment of non-committed cells is not a universal feature of vertebrate development. Epigenetics. 2012;7(4):383–9.

Alonso C, Pérez R, Bazaga P, Herrera CM. Global DNA cytosine methylation as an evolving trait: phylogenetic signal and correlated evolution with genome size in angiosperms. Front Genet. 2015;6(4):1–9.

Alonso C, Medrano M, Pérez R, Bazaga P, Herrera C, Alonso C, et al. Tissue-specific response to experimental demethylation at seed germination in the non-model herb *Erodium cicutarium*. Epigenomes. 2017;1(3):16.

Alvarez-Venegas R. Bacterial SET domain proteins and their role in eukaryotic chromatin modification. Front Genet. 2014;5:65.

Amato R. Inhibition of DNA methylation by antisense oligonucleotide MG98 as cancer therapy. Clin Genitourin Cancer. 2007;5(7):422–6.

Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.

Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, et al. The small RNA profile during Drosophila melanogaster development. Dev Cell. 2003;5(2):337–50.

Armstrong KM, Bermingham EN, Bassett SA, Treloar BP, Roy NC, Barnett MPG. Global DNA methylation measurement by HPLC using low amounts of DNA. Biotechnol J. 2011;6(1):113–7.

Atlasi Y, Stunnenberg HG. The interplay of epigenetic marks during stem cell differentiation and development. Nat Rev Genet. 2017;18(11):643–58.

Avramidou EV, Doulis AG, Aravanopoulos FA. Determination of epigenetic inheritance, genetic inheritance, and estimation of genome DNA methylation in a full-sib family of *Cupressus sempervirens* L. Gene. 2015;562(2):180–7.

Bailey T, Pawel K, Istvan L, Celine L, Qunhua L, Tao L, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol. 2013;9(11):e1003326.

Baker B. Context-dependent transgenerational plasticity in an annual plant: effects of parental shade versus sun on fitness and competitive performance. Masters thesis. 2018. https://wesscholar.wesleyan.edu/etd_mas_theses/189.

Balao F, Tannhäuser M, Lorenzo MT, Hedrén M, Paun O. Genetic differentiation and admixture between sibling allopolyploids in the *Dactylorhiza majalis* complex. Heredity. 2016;116(4):351–61.

Balkenhol N, Dudaniec RY, Krutovsky KV, Johnson JS, Cairns DM, Segelbacher G, et al. Landscape genomics: understanding relationships between environmental heterogeneity and genomic characteristics of populations. In: Rajora OP, editor. Population genomics concepts, strategies and approaches. Cham: Springer International Publishing AG; 2017. https://doi.org/10.1007/13836_2017_2.

Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res. 2011;21(3):381–95.

Banta JA, Richards CL. Quantitative epigenetics and evolution. Heredity. 2018;121:210–24.

Baron U, Turbachova I, Hellwag A, Eckhardt F, Berlin K, Hoffmüller U, et al. DNA methylation analysis as a tool for cell typing. Epigenetics. 2006;1(1):56–61.

Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281–97.

Baythavong BS. Linking the spatial scale of environmental variation and the evolution of phenotypic plasticity: selection favors adaptive plasticity in fine-grained environments. Am Nat. 2011;178(1):75–87.
Becker C, Weigel D. Epigenetic variation: origin and transgenerational inheritance. Curr Opin Plant Biol. 2012;15(5):562–7.
Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, et al. Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature. 2011;480(7376):245–9.
Bell CG, Teschendorff AE, Rakyan VK, Maxwell AP, Beck S, Savage DA. Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. BMC Med Genomics. 2010;3(1):33.
Bennett RL, Licht JD. Targeting epigenetics in cancer. Annu Rev Pharmacol Toxicol. 2018;58 (1):187–207.
Bernstein E, Allis CD. RNA meets chromatin. Genes Dev. 2005;19(14):1635–55.
Bewick AJ, Ji L, Niederhuth CE, Willing E-M, Hofmeister BT, Shi X, et al. On the origin and evolutionary consequences of gene body DNA methylation. Proc Natl Acad Sci U S A. 2016;113 (32):9111–6.
Bewick AJ, Niederhuth CE, Ji L, Rohr NA, Griffin PT, Leebens-Mack J, et al. The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. Genome Biol. 2017;18 (1):65.
Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002;16(1):6–21.
Birney E, Smith GD, Greally JM. Epigenome-wide association studies and the interpretation of disease -omics. PLoS Genet. 2016;12(6):e1006105.
Biswas S, Rao CM. Epigenetics in cancer: fundamentals and beyond. Pharmacol Therapeut. 2017;173:118–34.
Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
Bonchev G, Parisod C. Transposable elements and microevolutionary changes in natural populations. Mol Ecol Resour. 2013;13(5):765–75.
Bonduriansky R, Head M. Maternal and paternal condition effects on offspring phenotype in *Telostylinus angusticollis* (Diptera: Neriidae). J Evol Biol. 2007;20(6):2379–88.
Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science. 2012;336 (6083):934–7.
Booth MJ, Ost TWB, Beraldi D, Bell NM, Branco MR, Reik W, et al. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nat Protoc. 2013;8(10):1841–51.
Bostick M, Kim JK, Estève P-O, Clark A, Pradhan S, Jacobsen SE. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science. 2007;317(5845):1760–4.
Bousios A, Gaut BS. Mechanistic and evolutionary questions about epigenetic conflicts between transposable elements and their plant hosts. Curr Opin Plant Biol. 2016;30:123–33.
Bowers E, Yan G, Mukherjee C, Orry A, Wang L. Virtual ligand screening of the p300/CBP histone acetyltransferase: identification of a selective small molecule inhibitor. Chem Biol. 2010;17 (5):471–82.
Braun KVE, Dhana K, de Vries PS, Voortman T, van Meurs JBJ, Uitterlinden AG, et al. Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. Clin Epigenetics. 2017;9(1):15.
Bräutigam K, Vining KJ, Lafon-Placette C, Fossdal CG, Mirouze M, Marcos JG, et al. Epigenetic regulation of adaptive responses of forest tree species to the environment. Ecol Evol. 2013;3 (2):399–415.
Bräutigam K, Soolanayakanahally R, Champigny M, Mansfield S, Douglas C, Campbell MM, et al. Sexual epigenetics: gender-specific methylation of a gene in the sex determining region of *Populus balsamifera*. Sci Rep. 2017;7:45388.
Breiling A, Lyko F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. Epigenetics Chromatin. 2015;8:24.

Brown CJ, Lafreniere RG, Powers VE, Sebastio G, Ballabio A, Pettigrew AL, et al. Localization of the X inactivation centre on the human X chromosome in Xq13. Nature. 1991;349(6304):82–4.
Buck MJ, Lieb JD. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics. 2004;83(3):349–60.
Calarco JP, Borges F, Donoghue MTA, Van Ex F, Jullien PE, Lopes T, et al. Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. Cell. 2012;151 (1):194–205.
Carja O, MacIsaac JL, Mah SM, Henn BM, Kobor MS, Feldman MW, Fraser HB. Worldwide patterns of human epigenetic variation. Nat Ecol Evol. 2017;1(10):1577.
Carneros E, Yakovlev I, Viejo M, Olsen JE, Fossdal CG. The epigenetic memory of temperature during embryogenesis modifies the expression of bud burst-related genes in Norway spruce epitypes. Planta. 2017;246(3):553–66.
Carthew RW, Sontheimer EJ. Origins and mechanisms of miRNAs and siRNAs. Cell. 2009;136 (4):642–55.
Casadesús J, Low D. Epigenetic gene regulation in the bacterial world. Microbiol Mol Biol Rev. 2006;70(3):830–56.
Castillo-Aguilera O, Depreux P, Halby L, Arimondo P, Goossens L, Castillo-Aguilera O, et al. DNA methylation targeting: the DNMT/HMT crosstalk challenge. Biomolecules. 2017;7(1):3.
Cedar H, Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet. 2009;10(5):295–304.
Chadha S, Sharma M. Transposable elements as stress adaptive capacitors induce genomic instability in fungal pathogen Magnaporthe oryzae. PLoS One. 2014;9(4):e94415.
Chatterjee A, Lagisz M, Rodger EJ, Zhen L, Stockwell PA, Duncan EJ, Horsfield JA, Jeyakani J, Mathavan S, Ozaki Y, Nakagawa S. Sex differences in DNA methylation and expression in zebrafish brain: a test of an extended 'male sex drive' hypothesis. Gene. 2016;590(2):307–16.
Chen L-L. Linking long noncoding RNA localization and function. Trends Biochem Sci. 2016;41 (9):761–72.
Chen PY, Cokus SJ, Pellegrini M. BS Seeker: precise mapping for bisulfite sequencing. BMC Bioinformatics. 2010;11(1):203.
Chen Z, Riggs A. DNA methylation and demethylation in mammals. J Biol Chem. 2011;286 (21):18347–53.
Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell. 2016;167(5):1398–1414.e24.
Chen Z, Li S, Subramaniam S, Shyy JY-J, Chien S. Epigenetic regulation: a new frontier for biomedical engineers. Annu Rev Biomed Eng. 2017;19:195–219.
Chowdhury B, Cho I-H, Irudayaraj J. Technical advances in global DNA methylation analysis in human cancers. J Biol Eng. 2017;11(1):10.
Chu AY, Tin A, Schlosser P, Ko YA, Qiu C, Yao C, Joehanes R, Grams ME, Liang L, Gluck CA, Liu C. Epigenome-wide association studies identify DNA methylation associated with kidney function. Nat Commun. 2017;8(1):1286.
Ci D, Song Y, Du Q, Tian M, Han S, Zhang D. Variation in genomic methylation in natural populations of Populus simonii is associated with leaf shape and photosynthetic traits. J Exp Bot. 2016;67:723–37.
Clark SJ, Statham A, Stirzaker C, Molloy PL, Frommer M. DNA methylation: bisulphite modification and analysis. Nat Protoc. 2006;1(5):2353.
Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature. 2008;452 (7184):215–9.
Cortijo S, Wardenaar R, Colomé-Tatché M, Gilly A, Etcheverry M, Labadie K, et al. Mapping the epigenetic basis of complex traits. Science. 2014;343(6175):1145–8.
Creighton CJ, Reid JG, Gunaratne PH. Expression profiling of microRNAs by deep sequencing. Brief Bioinform. 2009;10(5):490–7.

Csankovszki G, Nagy A, Jaenisch R. Synergism of Xist Rna, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. J Cell Biol. 2001;153(4):773–84.
Cubas P, Vincent C. An epigenetic mutation responsible for natural variation in floral symmetry. Nature. 1999;401(6749):157.
Cushman SA. Grand challenges in evolutionary and population genetics: the importance of integrating epigenetics, genomics, modeling, and experimentation. Front Genet. 2014;5:197.
D'addario C, Francesco AD, Pucci M, Agrò AF, Maccarrone M. Epigenetic mechanisms and endocannabinoid signalling. FEBS J. 2013;280(9):1905–17. https://doi.org/10.1111/febs.12125.
Darwin C. On the origins of species by means of natural selection. London: Murray; 1859. p. 247.
Daskalos A, Nikolaidis G, Xinarianos G, Savvari P, Cassidy A, Zakopoulou R, et al. Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. Int J Cancer. 2009;124(1):81–7.
Deans C, Maggert KA. What do you mean, "epigenetic"? Genetics. 2015;199(4):887–96.
Denker A, de Laat W. A long-distance chromatin affair. Cell. 2015;162(5):942–3.
Derissen EJ, Beijnen JH, Schellens JH. Concise drug review: azacitidine and decitabine. Oncologist. 2013;18(5):619–24.
Dewan S, Vander Mijnsbrugge K, De Frenne P, Steenackers M, Michiels B, Verheyen K. Maternal temperature during seed maturation affects seed germination and timing of bud set in seedlings of European black poplar. Forest Ecol Manag. 2018;410:126–35.
Dobzhansky T. Genetics and the origin of species. New York: Columbia University Press; 1937.
Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR, et al. Widespread dynamic DNA methylation in response to biotic stress. Proc Natl Acad Sci U S A. 2012;109(32):E2183.
Du J, Johnson LM, Jacobsen SE, Patel DJ. DNA methylation pathways and their crosstalk with histone methylation. Nat Rev Mol Cell Biol. 2015;16(9):519–32.
Dubin MJ, Zhang P, Meng D, Remigereau MS, Osborne EJ, Casale FP, et al. DNA methylation in *Arabidopsis* has a genetic basis and shows evidence of local adaptation. Elife. 2015;4:1–23.
Duygu B, Poels EM, da Costa Martins PA. Genetics and epigenetics of arrhythmia and heart failure. Front Genet. 2013;4:219.
Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet. 2006;38(12):1378–85.
Eckschlager T, Plch J, Stiborova M, Hrabeta J. Histone deacetylase inhibitors as anticancer drugs. Int J Mol Sci. 2017;18(7).
Edwards DN, Ngwa VM, Wang S, Shiuan E, Brantley-Sieders DM, Kim LC, et al. The receptor tyrosine kinase EphA2 promotes glutamine metabolism in tumors by activating the transcriptional coactivators YAP and TAZ. Sci Signal. 2017;10(508).
Ehrlich M, Gama-Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA, et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic Acids Res. 1982;10(8):2709–21.
Eichten SR, Briskine R, Song J, Li Q, Swanson-Wagner R, Hermanson PJ, et al. Epigenetic and genetic influences on DNA methylation variation in maize populations. Plant Cell. 2013;25(8):2783–97.
Eichten SR, Schmitz RJ, Springer NM. Epigenetics: beyond chromatin modifications and complex genetic regulation. Plant Physiol. 2014;165(3):933.
Eleftheriou M, Pascual A, Wheldon L, Perry C, Abakir A. 5-Carboxylcytosine levels are elevated in human breast cancers and gliomas. Clin Epigenetics. 2015;7:88. https://doi.org/10.1186/s13148-015-0117-x.
Eminaga S, Christodoulou DC, Vigneault F, Church GM, Seidman JG. Quantification of microRNA expression with next-generation sequencing. Curr Protoc Mol Biol. 2013;103(1):4.17.1–4.17.14. https://doi.org/10.1002/0471142727.mb0417s103.
Esteller M, Levine R, Baylin SB, Ellenson LH, Herman JG. MLH1 promoter hypermethylation is associated with the microsatellite instability phenotype in sporadic endometrial carcinomas. Oncogene. 1998;17(18):2413–7.

Esteller M, Garcia-Foncillas J, Andion E, Goodman SN, Hidalgo OF, Vanaclocha V, et al. Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. N Engl J Med. 2000;343(19):1350–4.

Ettre LS. Milestones in chromatography: the birth of partition chromatography. LCGC. 2001;19 (5):506–12.

Fagny M, Patin E, Macisaac JL, Rotival M, Flutre T, Jones MJ, et al. The epigenomic landscape of African rainforest hunter-gatherers and farmers. Nat Commun. 2015;6:10047.

Fatemi M, Hermann A, Gowher H, Jeltsch A. Dnmt3a and Dnmt1 functionally cooperate during de novo methylation of DNA. Eur J Biochem. 2002;269(20):4981–4.

Fedoroff NV. Presidential address. Transposable elements, epigenetics, and genome evolution. Science. 2012;338(6108):758–67.

Feinberg A, Irizarry R. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc Natl Acad Sci U S A. 2010;107(Suppl 1):1757–64. https://doi.org/10.1073/pnas.0906183107.

Feinberg A, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature. 1983;301(5895):89–92.

Feng H, Conneely K, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. 2014;42(8):e69. https://doi.org/10.1093/nar/gku154.

Ferguson-Smith AC. Genomic imprinting: the emergence of an epigenetic paradigm. Nat Rev Genet. 2011;12(8):565–75.

Fernández-Sanlés A, Sayols-Baixeras S, Curcio S, Subirana I, Marrugat J, Elosua R. DNA methylation and age-independent cardiovascular risk, an epigenome-wide approach: the REGICOR study (REgistre GIroni del COR). Arterioscler Thromb Vasc Biol. 2018;38(3):645–52.

Ficz G, Gribben J. Loss of 5-hydroxymethylcytosine in cancer: cause or consequence? Genomics. 2014;104(5):352–7.

Field LM, Lyko F, Mandrioli M, Prantera G. DNA methylation in insects. Insect Mol Biol. 2004;13 (2):109–15.

Foust CM, Preite V, Schrey AW, Alvarez M, Robertson MH, Verhoeven KJF, et al. Genetic and epigenetic differences associated with environmental gradients in replicate populations of two salt marsh perennials. Mol Ecol. 2016;25(8):1639–52.

Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A. 2005a;102(30):10604–9.

Fraga MF, Ballestar E, Villar-Garea A, Boix-Chornet M, Espada J, Schotta G, et al. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. Nat Genet. 2005b;37(4):391–400.

Fraser H, Lam L, Neumann S, Kobor M. Population-specificity of human DNA methylation. Genome Biol. 2012;13(2):R8. https://doi.org/10.1186/gb-2012-13-2-r8.

Friso S, Choi S-W, Dolnikowski GG, Selhub J. A method to assess genomic DNA methylation using high-performance liquid chromatography/electrospray ionization mass spectrometry. Anal Chem. 2002;74(17):4526–31.

Friso S, Pizzolo F, Choi S-W, Guarini P, Castagna A, Ravagnani V, et al. Epigenetic control of 11 beta-hydroxysteroid dehydrogenase 2 gene promoter is related to human hypertension. Atherosclerosis. 2008;199(2):323–7.

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A. 1992;89(5):1827–31.

Fu Y, Luo G-Z, Chen K, Deng X, Yu M, Han D, et al. N6-methyldeoxyadenosine marks active transcription start sites in *Chlamydomonas*. Cell. 2015;161(4):879–92.

Furrow RE. Epigenetic inheritance, epimutation, and the response to selection. PLoS One. 2014;9 (7):e101559.

Gadaleta MC, Iwasaki O, Noguchi C, Noma K-I, Noguchi E. Chromatin immunoprecipitation to detect DNA replication and repair factors. Methods Mol Biol. 2015;1300:169–86.

Gajer J, Furdas S, Gründer A, Gothwal M, Heinicke U. Histone acetyltransferase inhibitors block neuroblastoma cell growth *in vivo*. Oncogenesis. 2015;4:e137. https://doi.org/10.1038/oncsis.2014.51.

Galloway LF, Etterson JR. Transgenerational plasticity is adaptive in the wild. Science. 2007;318 (5853):1134–6.

Gifford CA, Ziller MJ, Gu H, Trapnell C, Donaghey J, Tsankov A, et al. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell. 2013;153 (5):1149–63.

Globisch D, Münzel M, Müller M, Michalakis S, Wagner M, Koch S, et al. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. PLoS One. 2010;5 (12):e15367.

Goll MG, Bestor TH. Eukaryotic cytosine methyltransferases. Annu Rev Biochem. 2005;74:481–514.

Grant GR, Manduchi E, Stoeckert CJ. Analysis and management of microarray gene expression data. Curr Protoc Mol Biol. 2007;77(1):19.6.1–19.6.30.

Grativol C, Hemerly AS, Ferreira PCG. Genetic and epigenetic regulation of stress responses in natural plant populations. Biochim Biophys Acta. 2012;1819(2):176–85.

Greally JM. Population epigenetics. Curr Opin Syst Biol. 2017;1:84–9.

Greenblatt SM, Nimer SD. Chromatin modifiers and the promise of epigenetic therapy in acute leukemia. Leukemia. 2014;28(7):1396–406.

Greer EL, Blanco MA, Gu L, Sendinc E, Liu J, Aristizábal-Corrales D, et al. DNA methylation on N6-adenine in *C. elegans*. Cell. 2015;161(4):868–78.

Groot MP, Wagemaker N, Ouborg NJ, Verhoeven KJF, Vergeer P. Epigenetic population differentiation in field- and common garden-grown *Scabiosa columbaria* plants. Ecol Evol. 2018;8 (6):3505–17.

Grunau C, Clark SJ, Rosenthal A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. Nucleic Acids Res. 2001;29(13):e65.

Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc. 2011;6 (4):468.

Gugger PF, Fitz-Gibbon S, PellEgrini M, Sork VL. Species-wide patterns of DNA methylation variation in *Quercus lobata* and their association with climate gradients. Mol Ecol. 2016;25 (8):1665–80.

Guo W, Fiziev P, Yan W, Cokus S, Sun X. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics. 2013;14:774. https://doi.org/10.1186/1471-2164-14-774.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458 (7235):223–7.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature. 2011;477(7364):295–300.

Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, Lin C, et al. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods. 2008;44(1):3–12.

Hagmann J, Becker C, Müller J, Stegle O, Meyer RC, Wang G, et al. Century-scale methylome stability in a recently diverged *Arabidopsis thaliana* lineage. PLoS Genet. 2015;11(1): e1004920.

Halfmann R, Lindquist S. Epigenetics in the extreme: prions and the inheritance of environmentally acquired traits. Science. 2010;330(6004):629–32.

Hansen JC. Conformational dynamics of the chromatin fiber in solution: determinants, mechanisms, and functions. Annu Rev Biophys Biomol Struct. 2002;31:361–92.

Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 2012;13(10):R83.

Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, et al. Natural RNA circles function as efficient microRNA sponges. Nature. 2013;495(7441):384–8.

Hardcastle TJ, Müller SY, Baulcombe DC. Towards annotating the plant epigenome: the *Arabidopsis thaliana* small RNA locus map. Sci Rep. 2018;8(1):6338.
He Y, Michaels SD, Amasino RM. Regulation of flowering time by histone acetylation in *Arabidopsis*. Science. 2003;302(5651):1751–4.
He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet. 2004;5:522–31.
He G, Zhu X, Elling AA, Chen L, Wang X, Guo L, et al. Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. Plant Cell. 2010;22(1):17–33.
He Y-F, Li B-Z, Li Z, Liu P, Wang Y, Tang Q, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science. 2011;333(6047):1303–7.
Heard E, Martienssen RA. Transgenerational epigenetic inheritance: myths and mechanisms. Cell. 2014;157(1):95–109.
Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007;39(3):311–8.
Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009;459(7243):108–12.
Hendry A, Kinnison M. An introduction to microevolution: rate, pattern, process. Genetica. 2001;112–113(1):1–8.
Herrera CM, Bazaga P. Epigenetic differentiation and relationship to adaptive genetic divergence in discrete populations of the violet *Viola cazorlensis*. New Phytol. 2010;187(3):867–76.
Herrera CM, Bazaga P. Epigenetic correlates of plant phenotypic plasticity: DNA methylation differs between prickly and nonprickly leaves in heterophyllous *Ilex aquifolium* (Aquifoliaceae) trees. Bot J Linn Soc. 2013;171(3):441–52.
Herrera CM, Bazaga P. Genetic and epigenetic divergence between disturbed and undisturbed subpopulations of a Mediterranean shrub: a 20-year field experiment. Ecol Evol. 2016;6(11):3832–47.
Herrera CM, Medrano M, Bazaga P. Epigenetic differentiation persists after male gametogenesis in natural populations of the perennial herb *Helleborus foetidus* (Ranunculaceae). PLoS One. 2013;8(7):e70730.
Herrera CM, Medrano M, Bazaga P. Comparative spatial genetics and epigenetics of plant populations: heuristic value and a proof of concept. Mol Ecol. 2016;25(8):1653–64.
Herzing LB, Romer JT, Horn JM, Ashworth A. Xist has properties of the X-chromosome inactivation centre. Nature. 1997;386(6622):272–5.
Hewitt AW, Januar V, Sexton-Oates A, Joo JE, Franchina M, Wang JJ, et al. DNA methylation landscape of ocular tissue relative to matched peripheral blood. Sci Rep. 2017;7:46330.
Hochedlinger K, Plath K. Epigenetic reprogramming and induced pluripotency. Development. 2009;136(4):509–23.
Holliday R. Epigenetics: an overview. Dev Genet. 1994;15(6):453–7.
Holliday R, Pugh JE. DNA modification mechanisms and gene activity during development. Science. 1975;187(4173):226–32.
Horsthemke B. A critical view on transgenerational epigenetic inheritance in humans. Nat Commun. 2018;9(1):2973.
Hotchkiss RD. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. J Biol Chem. 1948;175(1):315–32.
Hu J, Barrett RD. Epigenetics in natural animal populations. J Evol Biol. 2017;30(9):1612–32.
Huanca-Mamani W, Arias-Carrasco R, Cárdenas-Ninasivincha S, Rojas-Herrera M, Sepúlveda-Hermosilla G, Caris-Maldonado JC, Bastías E, Maracaja-Coutinho V. Long non-coding RNAs responsive to salt and boron stress in the hyper-arid lluteño maize from atacama desert. Genes. 2018;9(3):170.
Iglesias FM, Cerdán PD. Maintaining epigenetic inheritance during DNA replication in plants. Front Plant Sci. 2016;7:38. https://doi.org/10.3389/fpls.2016.00038.

Ingvarsson PK, Street NR. Association genetics of complex traits in plants. New Phytol. 2011;189 (4):909–22.
Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science. 2011;333(6047):1300–3.
Iurlaro M, von Meyenn F, Reik W. DNA methylation homeostasis in human and mouse development. Curr Opin Genet Dev. 2017;43:101–9.
Iwasaki YW, Siomi MC, Siomi H. PIWI-interacting RNA: its biogenesis and functions. Annu Rev Biochem. 2015;84:405–33.
Jablonka E, Raz G. Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. Q Rev Biol. 2009;84(2):131–76.
Jackson SA. Epigenomics: dissecting hybridization and polyploidization. Genome Biol. 2017;18 (1):17–9.
Jackson V, Chalkley R. A new method for the isolation of replicative chromatin: selective deposition of histone on both new and old DNA. Cell. 1981;23(1):121–34.
Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003;33(Suppl):245–54.
Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 2014;15(2):R31.
Jamniczky HA, Boughner JC, Rolian C, Gonzalez PN, Powell CD, Schmidt EJ, et al. Rediscovering Waddington in the post-genomic age: operationalising Waddington's epigenetics reveals new ways to investigate the generation and modulation of phenotypic variation. Bioessays. 2010;32 (7):553–8.
Janoušek B, Široký J, Vyskot B. Epigenetic control of sexual phenotype in a dioecious plant, Melandrium album. Mol Gen Genet. 1996;250(4):483–90.
Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, et al. Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA. 2013;19(2):141–57.
Jenuwein T, Allis CD. Translating the histone code. Science. 2001;293(5532):1074–80.
Jiang L, Zhang J, Wang JJ, Wang L, Zhang L, Li G, Yang X, Ma X, Sun X, Cai J, Zhang J. Sperm, but not oocyte, DNA methylome is inherited by zebrafish early embryos. Cell. 2013;153 (4):773–84.
Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, et al. Assessing the impact of transgenerational epigenetic variation on complex traits. PLoS Genet. 2009;5(6): e1000530.
Johnsen Ø, Kvaalen H, Yakovlev IA, Dæhlen OG, Fossdal CG, Skrøppa T. An epigenetic memory from time of embryo development affects climatic adaptation in Norway spruce. Plant cold hardiness. From the laboratory to the field. Wallingford: CABI; 2009. p. 99–107.
Johnson LJ, Tricker PJ. Epigenomic plasticity within populations: its evolutionary significance and potential. Heredity. 2010;105(1):113–21.
Jones P, Taylor S. Cellular differentiation, cytidine analogs and DNA methylation. Cell. 1980;20 (1):85–93.
Jost D, Carrivain P, Cavalli G, Vaillant C. Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic Acids Res. 2014;42(15):9553–61.
Kacmarczyk TJ, Fall MP, Zhang X, Xin Y, Li Y, Alonso A, et al. "Same difference": comprehensive evaluation of four DNA methylation measurement platforms. Epigenetics Chromatin. 2018;11(1):21.
Kahramanoglou C, Prieto AI, Khedkar S, Haase B, Gupta A, Benes V, et al. Genomics of DNA cytosine methylation in Escherichia coli reveals its role in stationary phase transcription. Nat Commun. 2012;3:886.
Kaidery N, Tarannum S, Thomas B. Epigenetic landscape of Parkinson's disease: emerging role in disease mechanisms and therapeutic modalities. Neurotherapeutics. 2013;10(4):698–708.
Kaikkonen MU, Lam MTY, Glass CK. Non-coding RNAs as regulators of gene expression and epigenetics. Cardiovasc Res. 2011;90(3):430–40.

Karius T, Schnekenburger M, Dicato M, Diederich M. MicroRNAs in cancer management and their modulation by dietary agents. Biochem Pharmacol. 2012;83(12):1591–601.
Karsy M, Arslan E, Moy F. Current progress on understanding microRNAs in glioblastoma multiforme. Genes Cancer. 2012;3(1):3–15.
Kawakatsu T, Huang S-SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, et al. Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. Cell. 2016a;166(2):492–505.
Kawakatsu T, Stuart T, Valdes M, Breakfield N, Schmitz RJ, Nery JR, et al. Unique cell-type-specific patterns of DNA methylation in the root meristem. Nat Plants. 2016b;2(5):16058.
Kazazian HH. Mobile elements and disease. Curr Opin Genet Dev. 1998;8(3):343–50.
Keller M, Hopp L, Liu X, Wohland T, Rohde K, Cancello R, et al. Genome-wide DNA promoter methylation and transcriptome analysis in human adipose tissue unravels novel candidate genes for obesity. Mol Metabolism. 2017;6(1):86–100.
Kelly DE, Hansen MEB, Tishkoff SA. Global variation in gene expression and the value of diverse sampling. Curr Opin Syst Biol. 2017;1:102–8.
Kermicle JL. Dependence of the R-mottled aleurone phenotype in maize on mode of sexual transmission. Genetics. 1970;66(1):69–85.
Keverne EB. Significance of epigenetics for understanding brain development, brain evolution and behaviour. Neuroscience. 2014;264:207–17.
Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A. 2009;106(28):11667–72.
Kiefer JC. Epigenetics in development. Dev Dyn. 2007;236(4):1144–56.
Kim VN, Han J, Siomi MC. Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol. 2009;10(2):126–39.
Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465(7295):182–7.
King GJ, Amoah S, Kurup S. Exploring and exploiting epigenetic variation in crops. Genome. 2010;53(11):856–68.
Klironomos FD, Berg J, Collins S. How epigenetic mutations can affect genetic evolution: model and mechanism: problems & paradigms. Bioessays. 2013;35(6):571–8.
Ko M, Huang Y, Jankowska AM, Pape UJ, Tahiliani M, Bandukwala HS, et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. Nature. 2010;468(7325):839–43.
Ko Y, Mohtat D, Suzuki M, Park A, Izquierdo M. Cytosine methylation changes in enhancer regions of core pro-fibrotic genes characterize kidney fibrosis development. Genome Biol. 2013;14:R108. https://doi.org/10.1186/gb-2013-14-10-r108.
Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013;9(1):29.
Kou HP, Li Y, Song XX, Ou XF, Xing SC, Ma J, Von Wettstein D, Liu B. Heritable alteration in DNA methylation induced by nitrogen-deficiency stress accompanies enhanced tolerance by progenies to the stress in rice (Oryza sativa L.). J Plant Physiol. 2011;168(14):1685–93.
Kraus TFJ, Greiner A, Steinmaurer M, Dietinger V, Guibourt V, Kretzschmar HA. Genetic characterization of ten-eleven-translocation methylcytosine dioxygenase alterations in human glioma. J Cancer. 2015;6(9):832–42.
Kremer D, Metzger S, Kolb-Bachofen V. Quantitative measurement of genome-wide DNA methylation by a reliable and cost-efficient enzyme-linked immunosorbent assay technique. Anal Biochem. 2012;422(2):74–8.
Kriaucionis S, Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science. 2009;324(5929):929–30.
Kronfol MM, Dozmorov MG, Huang R, Slattum PW, McClay JL. The role of epigenomics in personalized medicine. Expert Rev Precis Med Drug Dev. 2017;2(1):33–45.
Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27(11):1571–2.

Kuchino Y, Hanyu N, Nishimura S. Analysis of modified nucleosides and nucleotide sequence of tRNA. Methods Enzymol. 1987;155:379–96.
Kunej T, Godnic I, Ferdin J, Horvat S, Dovc P, Calin GA. Epigenetic regulation of microRNAs in cancer: an integrated review of literature. Mutat Res. 2011;717(1–2):77–84.
Kuo KC, McCune RA, Gehrke CW, Midgett R, Ehrlich M. Quantitative reversed-phase high performance liquid chromatographic determination of major and modified deoxyribonucleosides in DNA. Nucleic Acids Res. 1980;8(20):4763–76.
Kurdyukov S, Bullock M. DNA methylation analysis: choosing the right method. Biology. 2016;5 (1):3. https://doi.org/10.3390/biology5010003.
Kvaalen H, Johnsen Ø. Timing of bud set in Picea abies is regulated by a memory of temperature during zygotic and somatic embryogenesis. New Phytol. 2008;177(1):49–59.
Lahtz C, Pfeifer G. Epigenetic changes of DNA repair genes in cancer. J Mol Cell Biol. 2011;3 (1):51–8.
Lakshmikuttyamma A, Scott SA, DeCoteau JF, Geyer CR. Reexpression of epigenetically silenced AML tumor suppressor genes by SUV39H1 inhibition. Oncogene. 2010;29(4):576–88.
Landgrave-Gómez J, Mercado-Gómez O, Guevara-Guzmán R. Epigenetic mechanisms in neurological and neurodegenerative diseases. Front Cell Neurosci. 2015;9:58.
Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9 (4):357–9.
Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
Latzel V, Allan E, Bortolini Silveira A, Colot V, Fischer M, Bossdorf O. Epigenetic diversity increases the productivity and stability of plant populations. Nat Commun. 2013;4:2875.
Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010;11(3):204–20.
Le T, Kim K-P, Fan G, Faull KF. A sensitive mass spectrometry method for simultaneous quantification of DNA methylation and hydroxymethylation levels in biological samples. Anal Biochem. 2011;412(2):203–9.
Lee MK, Hong Y, Kim SY, Kim WJ, London SJ. Epigenome-wide association study of chronic obstructive pulmonary disease and lung function in Koreans. Epigenomics. 2017;9(7):971–84.
Lele L, Ning D, Cuiping P, Xiao G, Weihua G. Genetic and epigenetic variations associated with adaptation to heterogeneous habitat conditions in a deciduous shrub. Ecol Evol. 2018;8 (5):2594–606.
Lentini A, Lagerwall C, Vikingsson S, Mjoseng HK, Douvlataniotis K, Vogt H, et al. A reassessment of DNA-immunoprecipitation-based genomic profiling. Nat Methods. 2018;15(7):499–504.
Leto K, Arancillo M, Becker E, Buffo A, Chiang C. Consensus paper: cerebellar development. Cerebellum. 2016;15(6):789–828.
Lewis LC, Lo PCK, Foster JM, Dai N, Corrêa IR, Durczak PM, et al. Dynamics of 5-carboxylcytosine during hepatic differentiation: potential general role for active demethylation by DNA repair in lineage specification. Epigenetics. 2017;12(4):277–86.
Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013. https://arxiv.org/pdf/1303.3997.pdf.
Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589–95.
Li J, Poi MJ, Tsai M-D. The regulatory mechanisms of tumor suppressor p16INK4 and relevance to cancer. Biochemistry. 2012;50(25):5566–82. https://doi.org/10.1021/bi200642e.
Li C, Vagin VV, Lee S, Xu J, Ma S, Xi H, et al. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. Cell. 2009a;137(3):509–21.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009b;25(1 Pt 2):1653–4.

Li Y, Kong D, Wang Z, Sarkar FH. Regulation of microRNAs by natural agents: an emerging field in chemoprevention and chemotherapy research. Pharm Res. 2010;27(6):1027–41.
Li H, Liu F, Ren C, Bo X, Shu W. Genome-wide identification and characterisation of HOT regions in the human genome. BMC Genomics. 2016;17(1):733.
Liang D, Zhang Z, Wu H, Huang C, Shuai P, Ye CY, et al. Single-base-resolution methylomes of populus trichocarpa reveal the association between DNA methylation and drought stress. BMC Genet. 2014;15(Suppl 1):1–11.
Lindsay S, Bird AP. Use of restriction enzymes to detect potential gene sequences in mammalian DNA. Nature. 1987;327(6120):336–8.
Ling C, Groop L. Epigenetics: a molecular link between environmental factors and type 2 diabetes. Diabetes. 2009;58(12):2718–25.
Lira-Medeiros CF, Parisod C, Fernandes RA, Mata CS, Cardoso MA, Ferreira PC. Epigenetic variation in mangrove plants occurring in contrasting natural environment. PLoS One. 2010;5 (4):e10326.
Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell. 2008;133 (3):523–36.
Liu J, Zhu Y, Luo G-Z, Wang X, Yue Y, Wang X, et al. Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. Nat Commun. 2016;7:13052.
Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
Low DA, Weyand NJ, Mahan MJ. Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. Infect Immun. 2001;69(12):7197–204.
Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC, Green PJ. Elucidation of the small RNA component of the transcriptome. Science. 2005;309(5740):1567–9.
Lu H, Liu X, Deng Y, Hong Q. DNA methylation, a hand behind neurodegenerative diseases. Front Aging Neurosci. 2013;5:85. https://doi.org/10.3389/fnagi.2013.00085.
Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature. 1997;389(6648):251–60.
Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet. 2003;4:981–94.
Luikart G, Kardos M, Hand BK, Rajora OP, Aitken SN, Hohenlohe PA. Population genomics: advancing understanding of nature. In: Rajora OP, editor. Population genomics: concepts, approaches and applications. Cham: Springer International Publishing AG, part of Springer Nature; 2018.
Lujambio A, Calin G, Villanueva A, Ropero S, Sánchez-Céspedes M. A microRNA DNA methylation signature for human cancer metastasis. Proc Natl Acad Sci U S A. 2008;105 (36):13556–1.
Lunyak VV, Rosenfeld MG. Epigenetic regulation of stem cell fate. Hum Mol Genet. 2008;17(R1): R28–36.
Luo G-Z, He C. DNA N6-methyladenine in metazoans: functional epigenetic mark or bystander? Nat Struct Mol Biol. 2017;24(6):503–6.
Luo G-Z, Blanco MA, Greer EL, He C, Shi Y. DNA N(6)-methyladenine: a new epigenetic mark in eukaryotes? Nat Rev Mol Cell Biol. 2015;16(12):705–10.
Maamar MB, Sadler-Riggleman I, Beck D, Skinner MK. Epigenetic transgenerational inheritance of altered sperm histone retention sites. Sci Rep. 2018;8(1):5308.
Maes T, Tirapu I, Estiarte A, Ciceri F, Lunardi S, Wiseman D. ORY-1001, a potent and selective covalent KDM1A inhibitor, for the treatment of acute leukemia. Cancer Cell. 2018;33 (3):495–511.
Magaña AA, Wrobel K, Caudillo YA, Zaina S, Lund G, Wrobel K. High-performance liquid chromatography determination of 5-methyl-2′-deoxycytidine, 2′-deoxycytidine, and other deoxynucleosides and nucleosides in DNA digests. Anal Biochem. 2008;374(2):378–85.

Maiques-Diaz A, Somervaille TC. LSD1: biologic roles and therapeutic targeting. Epigenomics. 2016;8(8):1103–16.
Maiti A, Drohat AC. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. J Biol Chem. 2011;286(41):35334–8.
Malone CD, Hannon GJ. Small RNAs as guardians of the genome. Cell. 2009;136(4):656–68.
Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, et al. Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. Cell. 2009;137(3):522–35.
Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.
Martin EM, Fry RC. Environmental influences on the epigenome: exposure-associated DNA methylation in human populations. Annu Rev Public Health. 2018;39:309–33.
Martinez SR, Gay MS, Zhang L. Epigenetic mechanisms in heart development and disease. Drug Discov Today. 2015;20(7):799–811.
Mastroeni D, McKee A, Grover A, Rogers J, Coleman PD. Epigenetic differences in cortical neurons from a pair of monozygotic twins discordant for Alzheimer's disease. PLoS One. 2009;4(8):e6617.
Matkovich SJ, Hu Y, Eschenbacher WH, Dorn LE, Dorn GW. Direct and indirect involvement of microRNA-499 in clinical and experimental cardiomyopathy. Circ Res. 2012;111(5):521–31.
Mattiroli F, Bhattacharyya S, Dyer PN, White AE, Sandman K, Burkhart BW, et al. Structure of histone-based chromatin in Archaea. Science. 2017;357(6351):609–12.
Matzke MA, Mosher RA. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nat Rev Genet. 2014;15(6):394–408.
McClay JL, Shabalin AA, Dozmorov MG, Adkins DE, Kumar G, Nerella S, Clark SL, Bergen SE, Hultman CM, Magnusson PK, Sullivan PF. High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. Genome Biol. 2015;16(1):291.
McClintock B. Chromosome organization and genic expression. In: Cold Spring Harbor symposia on quantitative biology, vol. 16. New York: Cold Spring Harbor Laboratory Press; 1951. p. 13–47.
Medrano M, Herrera CM, Bazaga P. Epigenetic variation predicts regional and local intraspecific functional diversity in a perennial herb. Mol Ecol. 2014;23(20):4926–38.
Meeks KA, Henneman P, Venema A, Burr T, Galbete C, Danquah I, Schulze MB, Mockenhaupt FP, Owusu-Dabo E, Rotimi CN, Addo J. An epigenome-wide association study in whole blood of measures of adiposity among Ghanaians: the RODAM study. Clin Epigenetics. 2017;9(1):103.
Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 2005;33(18):5868–77.
Meng D, Dubin M, Zhang P, Osborne EJ, Stegle O, Clark RM, et al. Limited contribution of DNA methylation variation to expression regulation in *Arabidopsis thaliana*. PLoS Genet. 2016;12(7):e1006141.
Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007;448(7153):553–60.
Miñones-Moyano E, Porta S, Escaramís G, Rabionet R, Iraola S, Kagerbauer B, et al. MicroRNA profiling of Parkinson's disease brains identifies early downregulation of miR-34b/c which modulate mitochondrial function. Hum Mol Genet. 2011;20(15):3067–78.
Miranda-Morales E, Meier K, Sandoval-Carrillo A, Salas-Pacheco J, Vázquez-Cárdenas P, Arias-Carrión O. Implications of DNA methylation in Parkinson's disease. Front Mol Neurosci. 2017;10:225. https://doi.org/10.3389/fnmol.2017.00225.
Mirbahai L, Chipman JK. Epigenetic memory of environmental organisms: a reflection of lifetime stressor exposures. Mutat Res. 2014;764–765:10–7.

Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. Cell. 2002;110(6):689–99.

Moison C, Assemat F, Daunay A, Arimondo PB, Tost J. DNA methylation analysis of ChIP products at single nucleotide resolution by Pyrosequencing®. In: Lehmann U, Tost J, editors. Pyrosequencing: methods and protocols. New York: Springer; 2015. p. 315–33. https://doi.org/10.1007/978-1-4939-2715-9_22.

Monk M, Boubelik M, Lehnert S. Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development. 1987;99(3):371–82.

Morera L, Lübbert M, Jung M. Targeting histone methyltransferases and demethylases in clinical trials for cancer therapy. Clin Epigenetics. 2016;8:57. https://doi.org/10.1186/s13148-016-0223-4

Morris T, Beck S. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. Methods. 2015;72(11):3–8.

Movassagh M, Choy M, Knowles D, Cordeddu L, Haider S. Distinct epigenomic features in end-stage failing human hearts. Circulation. 2011;124(22):2411–22.

Murrell A, Hurd PJ, Wood IC. Epigenetic mechanisms in development and disease. Biochem Soc Trans. 2013;41(3):697–9.

Nakatochi M, Ichihara S, Yamamoto K, Naruse K, Yokota S, Asano H, Matsubara T, Yokota M. Epigenome-wide association of myocardial infarction with DNA methylation sites at loci related to cardiovascular disease. Clin Epigenetics. 2017;9(1):54.

Nano J, Ghanbari M, Wang W, de Vries P, Dhana K. Epigenome-wide association study identifies methylation sites associated with liver enzymes and hepatic steatosis. Gastroenterology. 2017;153(4):1096–106. https://doi.org/10.1053/j.gastro.2017.06.003.

Neri F, Incarnato D, Krepelova A, Parlato C, Oliviero S. Methylation-assisted bisulfite sequencing to simultaneously map 5fC and 5caC on a genome-wide scale for DNA demethylation analysis. Nat Protoc. 2016;11(7):1191–205.

Nestor C, Ruzov A, Meehan R, Dunican D. Enzymatic approaches and bisulfite sequencing cannot distinguish between 5-methylcytosine and 5-hydroxymethylcytosine in DNA. Biotechniques. 2010;48(4):317–9.

Newman EM, Morgan RJ, Kummar S, Beumer JH, Blanchard MS, Ruel C, El-Khoueiry AB, Carroll MI, Hou JM, Li C, Lenz HJ. A phase I, pharmacokinetic, and pharmacodynamic evaluation of the DNA methyltransferase inhibitor 5-fluoro-2′-deoxycytidine, administered with tetrahydrouridine. Cancer Chemother Pharmacol. 2015;75(3):537–46.

Ng RK, Gurdon JB. Epigenetic inheritance of cell differentiation status. Cell Cycle. 2008;7(9):1173–7.

Nicotra AB, Atkin OK, Bonser SP, Davidson AM, Finnegan EJ, Mathesius U, et al. Plant phenotypic plasticity in a changing climate. Trends Plant Sci. 2010;15(12):684–92.

Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, et al. Widespread natural variation of DNA methylation within angiosperms. Genome Biol. 2016;17(1):1–19.

Nightingale KP, O'Neill LP, Turner BM. Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. Curr Opin Genet Dev. 2006;16(2):125–36.

Nilsson E, Larsen G, Manikkam M, Guerrero-Bosagna C, Savenkova MI, Skinner MK. Environmentally induced epigenetic transgenerational inheritance of ovarian disease. PLoS One. 2012;7(5):e36129.

Novak P, Jensen T, Oshiro MM, Wozniak RJ, Nouzova M, Watts GS, et al. Epigenetic inactivation of the HOXA gene cluster in breast cancer. Cancer Res. 2006;66(22):10664–70.

O'Brown ZK, Greer EL. N6-methyladenine: a conserved and dynamic DNA mark. In: Jeltsch A, Jurkowska RZ, editors. DNA methyltransferases – role and function. Cham: Springer International Publishing; 2016. p. 213–46. https://doi.org/10.1007/978-3-319-43624-1_10.

Oakeley EJ. DNA methylation analysis: a review of current methodologies. Pharmacol Ther. 1999;84(3):389–400. https://doi.org/10.1016/S0163-7258(99)00043-1.

Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 1999;99(3):247–57.
Pang KC, Frith MC, Mattick JS. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. 2006;22(1):1–5.
Papait R, Cattaneo P, Kunderfranco P, Greco C, Carullo P. Genome-wide analysis of histone marks identifying an epigenetic signature of promoters and enhancers underlying cardiac hypertrophy. Proc Natl Acad Sci U S A. 2013;110(50):20164–9.
Patel SR, Dressler GR. The genetics and epigenetics of kidney development. Semin Nephrol. 2013;33(4):314–26. https://doi.org/10.1016/j.semnephrol.2013.05.004.
Paun O, Bateman RM, Fay MF, Hedrén M, Civeyrel L, Chase MW. Stable epigenetic effects impact adaptation in allopolyploid orchids (Dactylorhiza: Orchidaceae). Mol Biol Evol. 2010;27(11):2465–73.
Pecinka A, Scheid OM. Stress-induced chromatin changes: a critical view on their heritability. Plant Cell Physiol. 2012;53(5):801–8.
Pei J-H, Luo S-Q, Zhong Y, Chen J-H, Xiao H-W, Hu W-X. The association between non-Hodgkin lymphoma and methylation of p73. Tumor Biol. 2011;32(6):1133.
Petersen AK, Zeilinger S, Kastenmüller G, Römisch-Margl W, Brugger M, Peters A, et al. Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. Hum Mol Genet. 2014;23(2):534–45.
Peterson CL, Laniel M-A. Histones and histone modifications. Curr Biol. 2004;14(14):R546–51.
Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One. 2012;7(5):e37135.
Piferrer F. Epigenetics of sex determination and gonadogenesis. Dev Dyn. 2013;242(4):360–70.
Plath K, Mlynarczyk-Evans S, Nusinow DA, Panning B. Xist RNA and the mechanism of X chromosome inactivation. Annu Rev Genet. 2002;36:233–78.
Platt A, Gugger PF, Pellegrini M, Sork VL. Genome-wide signature of local adaptation linked to variable CpG methylation in oak populations. Mol Ecol. 2015;24(15):3823–30.
Pollard KS, Salama SR, Lambert N, Lambot M-A, Coppens S, Pedersen JS, et al. An RNA gene expressed during cortical development evolved rapidly in humans. Nature. 2006;443(7108):167–72.
Pomraning KR, Smith KM, Freitag M. Genome-wide high throughput analysis of DNA methylation in eukaryotes. Methods. 2009;47(3):142–50.
Postberg J, Forcob S, Chang W-J, Lipps HJ. The evolutionary history of histone H3 suggests a deep eukaryotic root of chromatin modifying mechanisms. BMC Evol Biol. 2010;10:259.
Price TD, Qvarnström A, Irwin DE. The role of phenotypic plasticity in driving genetic evolution. Proc R Soc Lond B Biol Sci. 2003;270(1523):1433–40.
Przybilski R, Gräf S, Lescoute A, Nellen W, Westhof E. Functional hammerhead ribozymes naturally encoded in the genome of *Arabidopsis thaliana*. Plant Cell. 2005;17(7):1877–85.
Qiao W, Guo B, Zhou H, Xu W, Chen Y, Liang Y, et al. miR-124 suppresses glioblastoma growth and potentiates chemosensitivity by inhibiting AURKA. Biochem Biophys Res Commun. 2017;486(1):43–8.
Quinkler M, Stewart PM. Hypertension and the cortisol-cortisone shuttle. J Clin Endocrinol Metab. 2003;88(6):2384–92.
Rahavi SMR, Kovalchuk I. Changes in homologous recombination frequency in Arabidopsis thaliana plants exposed to stress depend on time of exposure during development and on duration of stress exposure. Physiol Mol Biol Plants. 2013;19(4):479–88.
Raj S, Bräutigam K, Hamanishi ET, Wilkins O, Thomas BR, Schroeder W, et al. Clone history shapes *Populus* drought responses. Proc Natl Acad Sci U S A. 2011;108(30):12521–6.
Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011;12(8):529–41.

Ramsawhook A, Lewis L, Coyle B, Ruzov A. Medulloblastoma and ependymoma cells display increased levels of 5-carboxylcytosine and elevated TET1 expression. Clin Epigenetics. 2017;9:18.
Ramsawhook A, Ruzov A, Coyle B. Wilms' tumor protein 1 and enzymatic oxidation of 5-methylcytosine in brain tumors: potential perspectives. Front Cell Dev Biol. 2018;6:26. https://doi.org/10.3389/fcell.2018.00026.
Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, et al. Mapping DNA methylation with high-throughput nanopore sequencing. Nat Methods. 2017;14(4):411–3.
Rando OJ, Verstrepen KJ. Timescales of genetic and epigenetic inheritance. Cell. 2007;128 (4):655–68.
Rangasamy S, D'Mello SR, Narayanan V. Epigenetics, autism spectrum, and neurodevelopmental disorders. Neurotherapeutics. 2013;10(4):742–56.
Reddy MA, Natarajan R. Epigenetics in diabetic kidney disease. J Am Soc Nephrol. 2011;22 (12):2182–5.
Rehimi R, Nikolic M, Cruz-Molina S, Tebartz C, Frommolt P, Mahabir E, et al. Epigenomics-based identification of major cell identity regulators within heterogeneous cell populations. Cell Rep. 2016;17(11):3062–76.
Reich E, Schibli A. High-performance thin-layer chromatography for the analysis of medicinal plants. Stuttgart: Thieme; 2007.
Reik W. Genomic imprinting and genetic disorders in man. Trends Genet. 1989;5(10):331–6.
Reinders J, Wulff BBH, Mirouze M, Mari-Ordonez A, Dapp M, Rozhon W, et al. Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. Genes Dev. 2009;23(8):939–50.
Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R. A practical guide to environmental association analysis in landscape genomics. Mol Ecol. 2015;24(17):4348–70.
Relyea RA. Costs of phenotypic plasticity. Am Nat. 2002;159(3):272–82.
Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL. Genome-wide location and function of DNA binding proteins. Science. 2000;290(5500):2306–9.
Rey T, Laporte P, Bonhomme M, Jardinaud M-F, Huguet S, Balzergue S, et al. MtNF-YA1, a central transcriptional regulator of symbiotic nodule development, is also a determinant of medicago truncatula susceptibility toward a root pathogen. Front Plant Sci. 2016;7:1837.
Reyna-Lopez G, Simpson J, Ruiz-Herrera J, Genetics M. Differences in DNA methylation patterns are detectable during the dimorphic transition of fungi by amplification of restriction polymorphisms. Mol Gen Genet. 1997;253(6):703–10.
Richards EJ. Inherited epigenetic variation – revisiting soft inheritance. Nat Rev Genet. 2006;7 (5):395–401.
Richards EJ. Population epigenetics. Curr Opin Genet Dev. 2008;18(2):221–6.
Richards CL, Bossdorf O, Verhoeven KJF. Understanding natural epigenetic variation. New Phytol. 2010;187(3):562–4.
Richards CL, Schrey AW, Pigliucci M. Invasion of diverse habitats by few Japanese knotweed genotypes is correlated with epigenetic differentiation. Ecol Lett. 2012;15(9):1016–25.
Richards CL, Alonso C, Becker C, Bossdorf O, Bucher E, Colomé-Tatché M, et al. Ecological plant epigenetics: evidence from model and non-model species, and the way forward. Ecol Lett. 2017;20(12):1576–90.
Riggs AD. X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genet. 1975;14 (1):9–25.
Rivera CM, Ren B. Mapping human epigenomes. Cell. 2013;155(1):39–55.
Roach DA, Wulff RD. Maternal effects in plants. Annu Rev Ecol Syst. 1987;18(1):209–35.
Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–30.
Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.

Rodriguez J, Frigola J, Vendrell E, Risques R-A, Fraga MF, Morales C, et al. Chromosomal instability correlates with genome-wide DNA demethylation in human primary colorectal cancers. Cancer Res. 2006;66(17):8462–8.

Rodríguez-Paredes M, Esteller M. Cancer epigenetics reaches mainstream oncology. Nat Med. 2011;17(3):330–9. https://doi.org/10.1038/nm.2305.

Ropero S, Esteller M. The role of histone deacetylases (HDACs) in human cancer. Mol Oncol. 2007;1(1):19–25.

Ross JP, Rand KN, Molloy PL. Hypomethylation of repeated DNA sequences in cancer. Epigenomics. 2010;2(2):245–69.

Ruzov A, Tsenkina Y, Serio A, Dudnakova T, Fletcher J, Bai Y, et al. Lineage-specific distribution of high levels of genomic 5-hydroxymethylcytosine in mammalian development. Cell Res. 2011;21(9):1332–42.

Ryan D, Ehninger D. Bison: bisulfite alignment on nodes of a cluster. BMC Bioinformatics. 2014;15:337. https://doi.org/10.1186/1471-2105-15-337.

Sáez-Laguna E, Guevara M-Á, Díaz L-M, Sánchez-Gómez D, Collada C, Aranda I, et al. Epigenetic variability in the genetically uniform forest tree species *Pinus pinea* L. PLoS One. 2014;9(8):e103145.

Salojärvi J. Computational tools for population genomics. In: Rajora OP, editor. Population genomics: concepts, approaches and applications. Cham: Springer International Publishing AG, part of Springer Nature; 2018.

Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94(3):441–8.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–7.

Santi DV, Norment A, Garrett CE. Covalent bond formation between a DNA-cytosine methyltransferase and DNA containing 5-azacytosine. Proc Natl Acad Sci U S A. 1984;81(22):6993–7.

Santos F, Dean W. Using immunofluorescence to observe methylation changes in mammalian preimplantation embryos. In: Nuclear reprogramming. Totowa, NJ: Humana Press; 2006. p. 129–38.

Satoh M, et al. Expression of microRNA-208 is associated with adverse clinical outcomes in human dilated cardiomyopathy. J Card Fail. 2010;16(5):404–10. https://doi.org/10.1016/j.cardfail.2010.01.002.

Sawan C, Herceg Z. 3-Histone modifications and cancer. Adv Genet. 2010;70(70):57–85.

Schield D, Walsh M, Card D, Andrew A, Adams R, Castoe T. EpiRADseq: scalable analysis of genomewide patterns of methylation using next-generation sequencing. Methods Ecol Evol. 2016;7(1):60–9.

Schlichting CD, Wund MA. Phenotypic plasticity and epigenetic marking: an assessment of evidence for genetic accommodation. Evolution. 2014;68(3):656–72.

Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, et al. Transgenerational epigenetic instability is a source of novel methylation variants. Science. 2011;334(6054):369–73.

Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, et al. Patterns of population epigenomic diversity. Nature. 2013a;495(7440):193–8.

Schmitz RJ, He Y, Valdes-Lopez O, Khan SM, Joshi T, Urich MA, et al. Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. Genome Res. 2013b;23(10):1663–74.

Schönberger B, Chen X, Mager S, Ludewig U. Site-dependent differences in DNA methylation and their impact on plant establishment and phosphorus nutrition in *Populus trichocarpa*. PLoS One. 2016;11(12):e0168623.

Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015;523(7559):212–6.

Seehafer C, Kalweit A, Steger G, Gräf S, Hammann C. From alpaca to zebrafish: hammerhead ribozymes wherever you look. RNA. 2011;17(1):21–6.
Shafi A, Mitrea C, Nguyen T, Draghici S. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. Brief Bioinform. 2018;19(5):737–53. https://doi.org/10.1093/bib/bbx013.
Shao Z, Zhang Y, Yuan G-C, Orkin SH, Waxman DJ. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biol. 2012;13(3):R16.
Sharma A. Transgenerational epigenetic inheritance: focus on soma to germline information transfer. Prog Biophys Mol Biol. 2013;113(3):439–46. https://doi.org/10.1016/j.pbiomolbio.2012.12.003.
Sharma U, Rando OJ. Metabolic inputs into the epigenome. Cell Metab. 2017;25(3):544–58.
Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26(10):1135–45.
Shimada-Sugimoto M, Otowa T, Miyagawa T, Umekage T, Kawamura Y, Bundo M, Iwamoto K, Tochigi M, Kasai K, Kaiya H, Tanii H. Epigenome-wide association study of DNA methylation in panic disorder. Clin Epigenetics. 2017;9(1):6.
Shiota K, Kogo Y, Ohgane J, Imamura T, Urano A, Nishino K, et al. Epigenetic marks by DNA methylation specific to stem, germ and somatic cells in mice. Genes Cells. 2002;7(9):961–9. https://doi.org/10.1046/j.1365-2443.2002.00574.x.
Simpson J, Workman R, Zuzarte P, David M, Dursi L, Detecting D, et al. Cytosine methylation using nanopore sequencing. Nat Methods. 2017;14(4):407–10. https://doi.org/10.1038/nmeth.4184.
Singh NN, Luo D, Singh RN. Pre-mRNA splicing modulation by antisense oligonucleotides. In: Exon skipping and inclusion therapies. New York, NY: Humana Press; 2018. p. 415–37.
Siomi H, Siomi MC. On the road to reading the RNA-interference code. Nature. 2009;457(7228):396–404.
Siomi MC, Sato K, Pezic D, Aravin AA. PIWI-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol. 2011;12(4):246–58.
Slotkin RK, Vaughn M, Borges F, Tanurdžić M, Becker JD, Feijó JA, Martienssen RA. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell. 2009;136(3):461–72.
Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014;11(8):817–20.
Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. 2013;14(3):204–20.
Smith P, Al H, Girard J, Delay C, Hébert S. In vivo regulation of amyloid precursor protein neuronal splicing by microRNAs. J Neurochem. 2011;116(2):240–7.
Smolarek I, Wyszko E, Barciszewska AM, Nowak S, Gawronska I, Jablecka A, et al. Global DNA methylation changes in blood of patients with essential hypertension. Med Sci Monit. 2010;16(3):CR149–55.
Soejima H, Higashimoto K. Epigenetic and genetic alterations of the imprinting disorder Beckwith-Wiedemann syndrome and related disorders. J Hum Genet. 2013;58(7):402–9.
Sollars ESA, Buggs RJA. Genome-wide epigenetic variation among ash trees differing in susceptibility to a fungal disease. BMC Genomics. 2018;19(1):502.
Song X, Cao X. Transposon-mediated epigenetic regulation contributes to phenotypic diversity and environmental adaptation in rice. Curr Opin Plant Biol. 2017;36:111–8.
Song C-X, He C. Potential functional roles of DNA demethylation intermediates. Trends Biochem Sci. 2013;38(10):480–4.
Song L, James SR, Kazim L, Karpf AR. Specific method for the determination of genomic DNA methylation by liquid chromatography-electrospray ionization tandem mass spectrometry. Anal Chem. 2005;77(2):504–10.
Song Y, Ci D, Tian M, Zhang D. Stable methylation of a non-coding RNA gene regulates gene expression in response to abiotic stress in *Populus simonii*. J Exp Bot. 2016;67(5):1477–92.

Soppa J. Protein acetylation in archaea, bacteria, and eukaryotes. Archaea. 2010. pii: 820681. https://doi.org/10.1155/2010/820681.
Spencer C, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 2009;5(5):e1000477. https://doi.org/10.1371/journal.pgen.1000477.
Springer NM, Schmitz RJ. Exploiting induced and natural epigenetic variation for crop improvement. Nat Rev Genet. 2017;18(9):563–75.
Srivastava A, Karpievitch Y, Eichten S, Borevitz J, Lister R. HOME: a histogram based machine learning approach for effective identification of differentially methylated regions. BioRxi. 2017. https://doi.org/10.1101/228221.
Stark R, Brown G. DiffBind: differential binding analysis of ChIP-Seq peak data. 2011. http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf.
Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. Brief Bioinform. 2016;17(6):953–66.
Stöger R, Ruzov A. Beyond CpG methylation: new modifications in eukaryotic DNA. Front Cell Dev Biol. 2018;6:87. https://doi.org/10.3389/fcell.2018.00087.
Storz G. An expanding universe of noncoding RNAs. Science. 2002;296(5571):1260–3.
Strahl BD, Allis CD. The language of covalent histone modifications. Nature. 2000;403(6765):41–5.
Stresemann C, Lyko F. Modes of action of the DNA methyltransferase inhibitors azacytidine and decitabine. Int J Cancer. 2008;123(1):8–13.
Stroud H, Ding B, Simon SA, Feng S, Bellizzi M, Pellegrini M, et al. Plants regenerated from tissue culture contain stable epigenome changes in rice. Elife. 2013;2:e00354. https://elifesciences.org/articles/00354
Studholme DJ. Deep sequencing of small RNAs in plants: applied bioinformatics. Brief Funct Genomics. 2012;11(1):71–85.
Sultan SE. Phenotypic plasticity for plant development, function and life history. Trends Plant Sci. 2000;5(12):537–42.
Sultan SE. Phenotypic plasticity in plants: a case study in ecological development. Evol Dev. 2003;5(1):25–33.
Sun G, Reddy MA, Yuan H, Lanting L, Kato M, Natarajan R. Epigenetic histone methylation modulates fibrotic gene expression. J Am Soc Nephrol. 2010;21(12):2069–80. https://doi.org/10.1681/ASN.2010060633.
Sun Q, Huang S, Wang X, Zhu Y, Chen Z, Chen D. N6-methyladenine functions as a potential epigenetic mark in eukaryotes. Bioessays. 2015;37(11):1155–62.
Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet. 2008;9(6):465–76.
Szyf M, Pakneshan P, Rabbani SA. DNA methylation and breast cancer. Biochem Pharmacol. 2004;68(6):1187–97.
Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science. 2009;324(5929):930–5.
Takahashi S, Osabe K, Fukushima N, Takuno S, Miyaji N, Shimizu M, et al. Genome-wide characterization of DNA methylation, small RNA expression, and histone H3 lysine nine di-methylation in *Brassica rapa* L. DNA Res. 2018; https://doi.org/10.1093/dnares/dsy021.
Talbert PB, Ahmad K, Almouzni G, Ausió J, Berger F, Bhalla PL, et al. A unified phylogeny-based nomenclature for histone variants. Epigenetics Chromatin. 2012;5:7.
Talbot B, Chen T-W, Zimmerman S, Joost S, Eckert AJ, Crow TM, et al. Combining genotype, phenotype, and environment to infer potential candidate genes. J Hered. 2017;108(2):207–16.
Tang Y, Xiong J, Jiang H-P, Zheng S-J, Feng Y-Q, Yuan B-F. Determination of oxidation products of 5-methylcytosine in plants by chemical derivatization coupled with liquid chromatography/tandem mass spectrometry analysis. Anal Chem. 2014;86(15):7764–72.

Taudt A, Colomé-Tatché M, Johannes F. Genetic sources of population epigenomic variation. Nat Rev Genet. 2016;17(6):319–32.
Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. Epigenomics. 2017;9(5):757–68.
Thon N, Kreth S, Kreth F. Personalized treatment strategies in glioblastoma: MGMT promoter methylation status. OncoTargets Ther. 2013;6:1363–72.
Thorson JLM, Smithson M, Beck D, Sadler-Riggleman I, Nilsson E, Dybdahl M, et al. Epigenetics and adaptive phenotypic variation between habitats in an asexual snail. Sci Rep. 2017;7 (1):1–11.
Tsai M, Manor O, Wan Y, Mosammaparast N, Wang J. Long noncoding RNA as modular scaffold of histone modification complexes. Science. 2010;329(5992):689–93.
Uchida S, Dimmeler S. Long noncoding RNAs in cardiovascular diseases. Circ Res. 2015;116 (4):737–50.
Udali S, Guarini P, Moruzzi S, Choi S, Friso S. Cardiovascular epigenetics: From DNA methylation to microRNAs. Mol Aspects Med. 2013;34(4):883–901.
Underwood CJ, Henderson IR, Martienssen RA. Genetic and epigenetic variation of transposable elements in *Arabidopsis*. Curr Opin Plant Biol. 2017;36:135–41.
Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. A distinct small RNA pathway silences selfish genetic elements in the germline. Science. 2006;313(5785):320–4.
Valente S, Mai A. Small-molecule inhibitors of histone deacetylase for the treatment of cancer and non-cancer diseases: a patent review (2011–2013). Expert Opin Ther Pat. 2014;24(4):401–15.
Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods. 2008;5 (9):829–34.
Van der Graaf A, Wardenaar R, Neumann DA, Taudt A, Shaw RG, Jansen RC, et al. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. Proc Natl Acad Sci U S A. 2015;112 (21):6676–81.
Van Dooren T, Silveira A, Gilbaut E, Jimenez-Gomez JM, Martin A, Bach L, et al. Mild drought induces phenotypic and DNA methylation plasticity but no transgenerational effects in *Arabidopsis*. BioRxiv. 2018. https://doi.org/10.1101/370320.
Van Oppen MJH, Gates RD, Blackall LL, Cantin N, Chakravarti LJ, Chan WY, et al. Shifting paradigms in restoration of the world's coral reefs. Glob Chang Biol. 2017;23(9):3437–48.
Van Rooij E, Sutherland LB, Liu N, Williams AH, McAnally J, Gerard RD, Richardson JA, Olson EN. A signature pattern of stress-responsive microRNAs that can evoke cardiac hypertrophy and heart failure. Proc Natl Acad Sci U S A. 2006;103(48):18255–60.
Vanyushin BF, Belozersky AN, Kokurina NA, Kadirova DX. 5-Methylcytosine and 6-methylamino-purine in bacterial DNA. Nature. 1968;218(5146):1066–7.
Vaughn MW, Tanurdžić M, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD, et al. Epigenetic natural variation in Arabidopsis thaliana. PLoS Biol. 2007;5(7):1617–29.
Vergeer P, Ouborg NJ. Evidence for an epigenetic role in inbreeding depression. Biol Lett. 2012;8 (5):798–801. https://doi.org/10.1098/rsbl.2012.0494.
Verhoeven KJF, Jansen JJ, Van Dijk PJ, Biere A. Stress-induced DNA methylation changes and their heritability in asexual dandelions. New Phytol. 2010;185(4):1108–18.
Verhoeven KJF, von Holdt BM, Sork VL. Epigenetics in ecology and evolution: what we know and what we need to know. Mol Ecol. 2016;25(8):1631–8.
Vogt G. Facilitation of environmental adaptation and evolution by epigenetic phenotype variation: insights from clonal, invasive, polyploid, and domesticated animals. Environ Epigenet. 2017;3 (1):1–17.
Volpe TA, Kidner C, Hall IM, Teng G, Grewal SIS, Martienssen RA. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science. 2002;297 (5588):1833–7.
Waalwijk C, Flavell RA. MspI, an isoschizomer of hpaII which cleaves both unmethylated and methylated hpaII sites. Nucleic Acids Res. 1978;5(9):3231–6.

Waddington CH. The epigenotype. Int J Epidemiol. 2012;41(1):10–3.
Wajed SA, Laird PW, DeMeester TR. DNA methylation: an alternative pathway to cancer. Ann Surg. 2001;234(1):10–20.
Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol Cell. 2011;43 (6):904–14.
Wang X, Song S, Wu Y-S, Li Y-L, Chen T, Huang Z, et al. Genome-wide mapping of 5-hydroxymethylcytosine in three rice cultivars reveals its preferential localization in transcriptionally silent transposable element genes. J Exp Bot. 2015;66(21):6651–63.
Wang Y, Sheng Y, Liu Y, Pan B, Huang J, Warren A, et al. N6-methyladenine DNA modification in the unicellular eukaryotic organism *Tetrahymena thermophila*. Eur J Protistol. 2017;58:94–102.
Watanabe A, Yamada Y, Yamanaka S. Epigenetic regulation in pluripotent stem cells: a key to breaking the epigenetic barrier. Philos Trans R Soc Lond B Biol Sci. 2013;368 (1609):20120292.
Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet. 2005;37(8):853–62.
Wedd L, Maleszka R. DNA methylation and gene regulation in honeybees: from genome-wide analyses to obligatory epialleles. Adv Exp Med Biol. 2016;945:193–211. https://doi.org/10.1007/978-3-319-43624-1_9. In: Jeltsch A, Jurkowska R, editors. DNA methyltransferases – role and function
Weinmann A, Farnham P. Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. Methods. 2002;26(1):37–47.
Weksberg R, Smith AC, Squire J, Sadowski P. Beckwith-Wiedemann syndrome demonstrates a role for epigenetic control of normal development. Hum Mol Genet. 2003;12(Spec No 1): R61–8.
Weng MK, Natarajan K, Scholz D, Ivanova VN, Sachinidis A, Hengstler JG, et al. Lineage-specific regulation of epigenetic modifier genes in human liver and brain. PLoS One. 2014;9(7): e102035.
West-Eberhard M. Phenotypic accommodation: adaptive innovation due to developmental plasticity. J Exp Zool Mol Dev Evol. 2005;304(6):610–8.
Wheldon LL, Abakir A, Ferjentsik Z. Transient accumulation of 5-carboxylcytosine indicates involvement of active demethylation in lineage specification of neural stem cells. Cell Rep. 2014;7(5):1353–61.
Whipple AV, Holeski LM. Epigenetic inheritance across the landscape. Front Genet. 2016;7:189.
Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153 (2):307–19.
Wibowo A, Becker C, Marconi G, Durr J, Price J, Hagmann J, et al. Hyperosmotic stress memory in Arabidopsis is mediated by distinct epigenetically labile sites in the genome and is restricted in the male germline by DNA glycosylase activity. Elife. 2016;5 https://elifesciences.org/articles/13546
Wijetunga NA, Delahaye F, Zhao YM, Golden A, Mar JC, Einstein FH, et al. The meta-epigenomic structure of purified human stem cell populations is defined at cis-regulatory sequences. Nat Commun. 2014;5:5195.
Wilson ME, Sengoku T. Developmental regulation of neuronal genes by DNA methylation: environmental influences. Int J Dev Neurosci. 2013;31(6):448–51.
Wion D, Casadesús J. N6-methyl-adenine: an epigenetic signal for DNA-protein interactions. Nat Rev Microbiol. 2006;4(3):183–92.
Wong H-L, Byun H-M, Kwan JM, Campan M, Ingles SA, Laird PW, et al. Rapid and quantitative method of allele-specific DNA methylation analysis. Biotechniques. 2006;41(6):734–9.
Woodward C, Hansen L, Beckwith F, Redman R, Rodriguez R. Symbiogenics: an epigenetic approach to mitigating impacts of climate change on plants. HortScience. 2012;47(6):699–703.
Wu H, Zhang Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. Cell. 2014;156(1–2):45–68.

Wyatt GR, Cohen SS. A new pyrimidine base from bacteriophage nucleic acids. Nature. 1952;170 (4338):1072–3.
Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10:232. http://link.springer.com/article/10.1186/1471-2105-10-232
Xia J, Joyce CE, Bowcock AM, Zhang W. Noncanonical microRNAs and endogenous siRNAs in normal and psoriatic human skin. Hum Mol Genet. 2013;22(4):737–48.
Xiao S, Cao X, Zhong S. Comparative epigenomics: defining and utilizing epigenomic variations across species, time-course, and individuals. Wiley interdisciplinary reviews. Syst Biol Med. 2014;6(5):345–52.
Xiao C-L, Zhu S, He M-H, Chen Y, Yu G-L, Chen D, et al. N6-methyladenine DNA modification in human genome. BioRxiv. 2017;176958.
Xiao CL, Zhu S, He M, Chen, Zhang Q, Chen Y, Yu G, Liu J, Xie SQ, Luo F, Liang Z, Wang DP, Bo XC, Gu XF, Wang K, Yan GR. N(6)-methyladenine DNA modification in the human genome. Mol Cell. 2018;71:306–18 e7.
Xie HJ, Li H, Liu D, Dai WM, He JY, Lin S, et al. ICE1 demethylation drives the range expansion of a plant invader through cold tolerance divergence. Mol Ecol. 2015;24(4):835–50.
Xing X, Cai W, Luo L, Liu L, Shi H. The prognostic value of p16 hypermethylation in cancer: a meta-analysis. Plos One. 2013;8(6):e54970. http://pubmedcentralcanada.ca/pmcc/articles/PMC3689792/
Xu Z, Bolick SC, DeRoo LA, Weinberg CR, Sandler DP, Taylor JA. Epigenome-wide association study of breast cancer using prospectively collected sister study samples. J Natl Cancer Inst. 2013;105(10):694–700.
Xu S, Grullon S, Ge K, Peng W. Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. In: Kidder BL, editor. Stem cell transcriptional networks: methods and protocols. New York: Springer; 2014. p. 97–111. https://doi.org/10.1007/978-1-4939-0512-6_5.
Xue K, Gu JJ, Zhang Q, Mavis C, Hernandez-Ilizaliturri FJ, Czuczman MS, et al. Vorinostat, a histone deacetylase (HDAC) inhibitor, promotes cell cycle arrest and re-sensitizes rituximab- and chemo-resistant lymphoma cells to chemotherapy agents. J Cancer Res Clin Oncol. 2016;142(2):379–87.
Yaish MW, Peng M, Rothstein SJ. Global DNA methylation analysis using methyl-sensitive amplification polymorphism (MSAP). Methods Mol Biol. 2014;1062:285–98.
Yakovlev IA, Fossdal CG. *In silico* analysis of small RNAs suggest roles for novel and conserved miRNAs in the formation of epigenetic memory in somatic embryos of Norway spruce. Front Plant Physiol. 2017;8:674.
Yakovlev I, Fossdal CG, Skrøppa T, Olsen JE, Jahren AH, Johnsen Ø. An adaptive epigenetic memory in conifers with important implications for seed production. Seed Sci Res. 2012;22:63–6.
Yakovlev IA, Carneros E, Lee Y, Olsen JE, Fossdal CG. Transcriptional profiling of epigenetic regulators in somatic embryos during temperature induced formation of an epigenetic memory in Norway spruce. Planta. 2016;243(5):1237–49.
Yan H, Simola DF, Bonasio R, Liebig J, Berger SL, Reinberg D. Eusocial insects as emerging models for behavioural epigenetics. Nat Rev Genet. 2014;15(10):677–88.
Yan H, Bonasio R, Simola DF, Liebig J, Berger SL, Reinberg D. DNA methylation in social insects: how epigenetics can control behavior and longevity. Annu Rev Entomol. 2015;60:435–52.
Yang IV, Richards A, Davidson EJ, Stevens AD, Kolakowski CA, Martin RJ, et al. The nasal methylome: a key to understanding allergic asthma. Am J Respir Crit Care Med. 2017;195 (6):829–31.
Yoo CB, Jones PA. Epigenetic therapy of cancer: past, present and future. Nat Rev Drug Discov. 2006;5(1):37.
Yu M, Hon G, Szulwach K, Song C, Jin P. Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. Nat Protoc. 2012;7(12):2159–70. https://doi.org/10.1038/nprot.2012.137.

Yung PYK, Elsässer SJ. Evolution of epigenetic chromatin states. Curr Opin Chem Biol. 2017;41:36–42.
Zas R, Cendán C, Sampedro L. Mediation of seed provisioning in the transmission of environmental maternal effects in Maritime pine (*Pinus pinaster* Aiton). Heredity. 2013;111(3):248–55.
Zemach A, Zilberman D. Evolution of eukaryotic DNA methylation and the pursuit of safer sex. Curr Biol. 2010;20(17):R780–5.
Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science. 2010;328(5980):916–9.
Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, Chen H, et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell. 2006;126 (6):1189–201.
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.
Zhang Y, et al. APP processing in Alzheimer's disease. Mol Brain. 2011;4(1):3. https://doi.org/10.1186/1756-6606-4-3.
Zhang J, Stevens MF, Bradshaw TD. Temozolomide: mechanisms of action, repair and resistance. Curr Mol Pharmacol. 2012a;5(1):102–14.
Zhang L, Lu X, Lu J, Liang H, Dai Q, Xu G-L, et al. Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. Nat Chem Biol. 2012b;8(4):328–30.
Zhang Y, Zhang X-O, Chen T, Xiang J-F, Yin Q-F, Xing Y-H, et al. Circular intronic long noncoding RNAs. Mol Cell. 2013a;51(6):792–806.
Zhang Y-Y, Fischer M, Colot V, Bossdorf O. Epigenetic variation creates potential for evolution of plant phenotypic plasticity. New Phytol. 2013b;197(1):314–22.
Zhang L, Chen W, Iyer LM, Hu J, Wang G, Fu Y, et al. A TET homologue protein from Coprinopsis cinerea (CcTET) that biochemically converts 5-methylcytosine to 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxylcytosine. J Am Chem Soc. 2014;136(13):4801–4.
Zhang G, Huang H, Liu D, Cheng Y, Liu X, Zhang W, et al. N6-methyladenine DNA modification in *Drosophila*. Cell. 2015a;161(4):893–906.
Zhang J, Liu Y, Xia E-H, Yao Q-Y, Liu X-D, Gao L-Z. Autotetraploid rice methylome analysis reveals methylation variation of transposable elements and their effects on gene expression. Proc Natl Acad Sci U S A. 2015b;112(50):E7022–9.
Zhang H, Lang Z, Zhu J-K. Dynamics and function of DNA methylation in plants. Nat Rev Mol Cell Biol. 2018;19(8):489–506.
Zhao J, Goldberg J, Bremner JD, Vaccarino V. Global DNA methylation is associated with insulin resistance: a monozygotic twin study. Diabetes. 2011. https://doi.org/10.2337/db11-1048.
Zheng X, Chen L, Xia H, Wei H, Lou Q, Li M, et al. Transgenerational epimutations induced by multi-generation drought imposition mediate rice plant's adaptation to drought condition. Sci Rep. 2017;7(1):39843. http://www.nature.com/articles/srep39843
Zhong X. Comparative epigenomics: a powerful tool to understand the evolution of DNA methylation. New Phytol. 2016;210(1):76–80.
Zhu J-K. Active DNA demethylation mediated by DNA glycosylases. Annu Rev Genet. 2009;43:143–66.
Zhu X, Shan L, Wang F, Wang J, Shen G, Liu X, et al. Hypermethylation of BRCA1 gene: implication for prognostic biomarker and therapeutic target in sporadic primary triple-negative breast cancer. Breast Cancer Res Treat. 2015;150(3):479–86.
Zilberman D, Henikoff S. Genome-wide analysis of DNA methylation patterns. Development. 2007;134:3959–65. https://doi.org/10.1242/dev.001131.
Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500(7463):477–81.
Zimmerman KCK, Levitis DA, Pringle A. Beyond animals and plants: dynamic maternal effects in the fungus *Neurospora crassa*. J Evol Biol. 2016;29(7):1379–93.
Zoghbi H, Beaudet A. Epigenetics and human disease. Cold Spring Harb Perspect Biol. 2016;8 (2):479–510. https://doi.org/10.1101/cshperspect.a019497.

Landscape Genomics: Understanding Relationships Between Environmental Heterogeneity and Genomic Characteristics of Populations

Niko Balkenhol, Rachael Y. Dudaniec, Konstantin V. Krutovsky, Jeremy S. Johnson, David M. Cairns, Gernot Segelbacher, Kimberly A. Selkoe, Sophie von der Heyden, Ian J. Wang, Oliver Selmoni, and Stéphane Joost

Abstract Landscape genomics is a rapidly advancing research field that combines population genomics, landscape ecology, and spatial analytical techniques to explicitly quantify the effects of environmental heterogeneity on neutral and adaptive genetic variation and underlying processes. Landscape genomics has tremendous potential for addressing fundamental and applied research questions in various

N. Balkenhol (✉)
Wildlife Sciences, University of Goettingen, Büsgenweg 3, Göttingen 37077, Germany
e-mail: nbalken@gwdg.de

R.Y. Dudaniec
Department of Biological Sciences, Macquarie University, E8B Eastern Road, North Ryde, Sydney, NSW 2109, Australia
e-mail: rachael.dudaniec@mq.edu.au

K.V. Krutovsky
Department of Forest Genetics and Forest Tree Breeding, University of Goettingen, Büsgenweg 2, Göttingen 37077, Germany

Department of Ecosystem Science and Management, Texas A&M University, 2138 TAMU, College Station, TX 77843, USA

N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences, 3 Gubkina Str., Moscow 119333, Russia

Genome Research and Education Center, Siberian Federal University, 50a/2 Akademgorodok, Krasnoyarsk 660036, Russia
e-mail: konstantin.krutovsky@forst.uni-goettingen.de

J.S. Johnson
Department of Geography, Texas A&M University, MS 3147, College Station, TX 77843-3147, USA

School of Forestry, Northern Arizona University, Flagstaff, AZ 86011, USA

Dorena Genetic Resource Center, 34963 Shoreview Dr, Cottage Grove, OR 97424, USA
e-mail: jsjohnson@tamu.edu; jeremy.johnson@nau.edu; jeremysjohnson@fs.fed.us

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)], https://doi.org/10.1007/13836_2017_2,

research fields, including ecology, evolution, and conservation biology. However, the unique combination of different scientific disciplines and analytical approaches also constitute a challenge to most researchers wishing to apply landscape genomics. Here, we present an introductory overview of important concepts and methods used in current landscape genomics. For this, we first define the field and explain basic concepts and methods to capture different hypotheses of landscape influences on neutral genetic variation. Next, we highlight established and emerging genomic tools for quantifying adaptive genetic variation in landscape genomic studies. To illustrate the covered topics and to demonstrate the potential of landscape genomics, we provide empirical examples addressing a variety of research question, i.e., the investigation of evolutionary processes driving population differentiation, the landscape genomics of range expanding species, and landscape genomic patterns in organisms of special interest, including species inhabiting aquatic and terrestrial environments. We conclude by outlining remaining challenges and future research avenues in landscape genomics.

Keywords Adaptive landscape genetics • Environmental association analysis (EAA) • Functional connectivity • Genome-wide association studies (GWAS) • Genotype-environment association (GEA) • Landscape resistance • Local adaptation • Outlier loci • Seascape genomics

D.M. Cairns
Department of Geography, Texas A&M University, MS 3147, College Station, TX 77843-3147, USA
e-mail: cairns@tamu.edu

G. Segelbacher
Wildlife Ecology and Management, University of Freiburg, Tennenbacher Str. 4, Freiburg 79106, Germany
e-mail: gernot.segelbacher@wildlife.uni-freiburg.de

K.A. Selkoe
Bren School of Environmental Science & Management, National Center for Ecological Analysis and Synthesis, University of California Santa Barbara, 735 State Street, Ste. 300, Santa Barbara, CA 93101, USA
e-mail: Selkoe@nceas.ucsb.edu

S. von der Heyden
Evolutionary Genomics Group, Department of Botany and Zoology, University of Stellenbosch, Private Bag X1, Matieland, Stellenbosch 7602, South Africa
e-mail: svdh@sun.ac.za

I.J. Wang
Department of Environmental Science, Policy, and Management, University of California, Berkeley, 130 Mulford Hall #3114, Berkeley, CA 94720-3114, USA
e-mail: ianwang@berkeley.edu

O. Selmoni and S. Joost
Laboratory of Geographic Information Systems (LASIG), School of Civil and Environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Bâtiment GC – Station 18, Lausanne, Vaud 1015, Switzerland
e-mail: oliver.selmoni@epfl.ch; stephane.joost@epfl.ch

1 Introduction

Any geneticist will agree that environmental conditions can substantially affect the genetic variation of natural populations, with important consequences for ecological and evolutionary processes and phenomena. For instance, selection pressures induced by the heterogeneous environment can promote local adaptation by favoring different alleles in different spatial localities (e.g., Hedrick et al. 1976; Richardson et al. 2014), just as historical or contemporary environmental barriers to gene flow (e.g., ice-covered areas during glacial periods, landscape features such as roads) can increase genetic differentiation among formerly connected populations (e.g., Hitchings et al. 1997) due to genetic drift. In some cases, local adaptation due to natural selection and reduced gene flow can ultimately lead to speciation, making genetic variation a fundamental level of the biodiversity hierarchy (Via 2002; Primack 2014). Nevertheless, explicitly accounting for spatial environmental heterogeneity in genetic studies – especially those dealing with recent and contemporary population genetics at fine spatial scales – has seen an unprecedented growth only in the last ca. 15 years (Storfer et al. 2010; Dyer 2015a). In part, this is due to the emergence of the field of landscape genetics, which was formally introduced in a seminal paper by Manel et al. (2003), and later extended towards landscape genomics (Luikart et al. 2003; Joost et al. 2007). In essence, landscape genetics can be defined as ... *research that combines population genetics, landscape ecology, and spatial analytical techniques to explicitly quantify the effects of landscape composition, configuration, and matrix quality on micro-evolutionary processes, such as gene flow, drift, and selection, using neutral and adaptive genetic data*. (Balkenhol et al. 2016a, p. 3). Basically, one can simply replace "population genetics" with "population genomics" to define landscape genomics (see Sect. 2.1 for details). While several earlier landscape genetic studies exist (e.g., Pamilo 1988; Merriam et al. 1989; Barbujani and Sokal 1990; Manicacci et al. 1992; Gaines et al. 1997; see also Manel et al. 2003), the recent growth of landscape genomic studies can be attributed to two main factors. First, landscape genomic studies are nowadays facilitated by novel technologies that make it possible to gather and analyze genetic and environmental data at large quantities and at high qualities. For instance, next-generation sequencing (NGS) allows us to quantify genetic variation in many individuals at dense genomic coverage at decreasing costs (e.g., Luikart et al. 2003; Andrews et al. 2016). Similarly, environmental data can be obtained at high spatial and temporal resolutions for large study areas from remote sensing devices, such as satellites or drones (e.g., Pettorelli et al. 2005; Anderson and Gaston 2013). Increased computational power also enables us to actually handle and analyze these large, spatially explicit data sets in a reasonable amount of time (e.g., Kidd and Ritchie 2006; Paul and Song 2012).

Second, the swift rise of landscape genomics can also be attributed to the increased interest in the ecological and evolutionary consequences of contemporary environmental change, such as habitat loss and fragmentation or human-expedited climate change. Specifically, understanding and predicting the consequences of ongoing environmental changes can be regarded as a major research need in the current Anthropocene, where humans are causing substantial environmental

changes and associated biodiversity losses (e.g., Haddad et al. 2015). Due to the technological advances mentioned above, landscape genomics has tremendous potential for contributing to such research, so it is not surprising that the number of landscape genomic studies has been rising exponentially since 2003 (Storfer et al. 2010; Dyer 2015a). However, getting started in landscape genomics can still be a daunting task, because the field amalgamates concepts, data, and methods from seemingly disparate disciplines that relatively few scientists are familiar with. Furthermore, many of the analytical approaches employed in landscape genomics develop very rapidly, making it challenging to keep up-to-date with all methods and concepts that are relevant for a landscape genomic study.

Here, our goal is to provide a general introduction to the field of landscape genomics. The chapter complements other recent work that provided a more general overview of landscape genetics (Balkenhol et al. 2016b), because we here focus more strongly on genomic approaches for landscape genetics and specifically address geneticists interested in applying landscape genomic approaches. Because of this, we assume that readers are familiar with basic population genetic concepts, which are discussed elsewhere (e.g., Waits and Storfer 2016) and in two other chapters of this book.

The chapter starts with a description of basic concepts and definitions that are a prerequisite for understanding the remaining sections. Next, we briefly summarize concepts and approaches for neutral landscape genomics, before focusing on more novel approaches that are particularly suitable for adaptive landscape genomics (see Sect. 2.1 for definition of the different terms). To illustrate the covered concepts and methods, we provide several empirical examples of landscape genomic applications, and finally outline several remaining challenges and future opportunities in landscape genomics.

2 Basic Concepts and Definitions

We can distinguish two components of genetic variation, namely the amount of genetic variation (genetic diversity, sometimes also called genetic variability) and the spatial distribution of genetic variation (genetic structure, e.g., via measures of genetic differentiation or genetic distances). Both components of genetic variation can be quantified using loci or genomic regions that are affected by selection (adaptive genetic variation) and those loci or regions that are not affected by it (neutral genetic variation). However, it is important to note that genetic data showing signs of selection may not actually be under selection, because selection may actually be acting on other loci or regions that the analyzed genetic data are linked to. Similarly, genetic data that appear to be selectively neutral could still be under selection, for example when selection acts upon highly polygenic traits so that the influence of selection on individual loci or small genomic regions is too small to be detected. In general, it is hard to truly discriminate selectively neutral from selectively adaptive genetic variation, except in a few cases, such as microsatellite or SNP variation controlled by loci in the noncoding regions. Nevertheless, we here follow the vast majority of publications

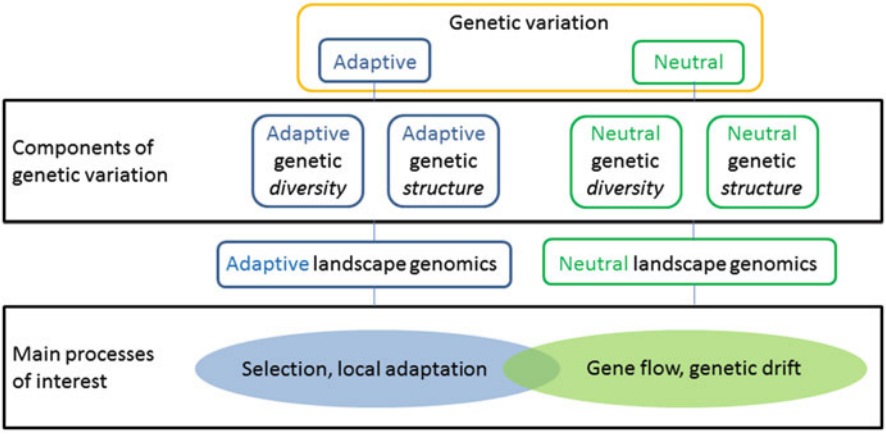

Fig. 1 Conceptual chart illustrating adaptive and neutral landscape genomics. Note that the processes of interest overlap between the two types of landscape genomics, and the greatest insights can be accomplished by combining neutral and adaptive data

that use the term adaptive genetic data to refer to those loci or regions that show signs of selection, and neutral genetic data for those loci or regions that do not show any evidence of selection. Depending on whether we use neutral or adaptive genetic data, we can focus on different underlying processes and research questions (Fig. 1).

Furthermore, landscape genomic studies can focus on analyzing environmental effects on neutral or adaptive genetic variation per se, which is particularly interesting for geneticists investigating the emergence and maintenance of genetic variation in nature. Alternatively, studies can use landscape genomic approaches to investigate how the environment impacts the processes that underlie observed patterns of genetic variation, which is often most interesting for ecologists who study effective dispersal (i.e., migration) and gene flow, and evolutionary biologists studying natural selection, adaptation, etc.

2.1 Landscape Genetics vs. Genomics

Given the different types of genetic variation, processes, and questions that landscape genetic studies can focus on, we can distinguish between "neutral" and "adaptive" landscape genetics (Holderegger et al. 2006). While the former focuses on putatively neutral processes, such as gene flow and genetic drift, the latter focuses on adaptive processes, such as selection and local adaptation. Since genomic approaches greatly facilitate the detection of loci or genomic regions under selection (e.g., Andrews et al. 2016), the term landscape genomics is now often used for studies seeking to identify environmental influences on adaptive genetic variation (Schwartz et al. 2009; Hand et al. 2015). In contrast, the term landscape genetics usually applies to studies dealing with selectively neutral genetic markers

and underlying processes (Manel and Holderegger 2013). However, it is possible to conduct an adaptive landscape genetic study without applying population genomics, for example through the use of established quantitative trait loci (QTLs) that are known to be under selection (Holderegger et al. 2006; Manel et al. 2010). Similarly, it is possible to conduct a neutral landscape genomic study, for example when an NGS approach is used to develop thousands of single-nucleotide polymorphisms (SNPs) and then choosing only those that are likely not under selection (i.e., by excluding outlier loci, e.g., Whitlock and Lotterhos 2015). This subset of tentative selectively neutral SNPs can then be used to evaluate landscape influences on neutral gene flow (e.g., Rasic et al. 2014).

While we acknowledge that not all of the examples we discuss in this chapter have been derived from genome-wide sequencing approaches, genomic data will undoubtedly become the standard for all landscape genetic studies in the near future, and the difference in nomenclature between landscape genetics and genomics will further diminish and eventually disappear. Hence, we here use the term landscape genomics to encompass all studies that explicitly test for environmental impacts on genome-wide genetic variation, even if they do not (yet) rely fully on the whole genome sequencing approaches. However, we distinguish between neutral landscape genomics and adaptive landscape genomics, depending on the type of genetic data and the processes of interest (Fig. 1). Ideally, neutral and adaptive landscape genomics should be combined to fully elucidate the ecological and evolutionary processes affecting different components of genetic variation (Hand et al. 2015; Balkenhol et al. 2016c). However, since underlying assumptions, data types, and methods differ for these two types of landscape genomics, we discuss them separately in Sects. 3 and 4, respectively.

2.2 *Influences of Spatial Environmental Heterogeneity on Genetic Variation*

From conceptual and analytical standpoints, genetic variation can be influenced by *local* environmental conditions, as well as the environment occurring *in-between* sampling locations (Fig. 2).

Local environmental conditions can include, for example, factors such as local climate, or habitat characteristics such as vegetation type and quality. Local environmental conditions can be measured *at* sampling locations, *within* areas, or *around* sampling locations or areas, for example within a certain radius around sampling points or habitat patches (Wagner and Fortin 2013; Pflüger and Balkenhol 2014). Local environmental conditions can induce spatially varying selective pressures, which directly affect adaptive genetic variation (Schoville et al. 2012). The local environment can favor certain alleles over others, or select against migrant individuals or their offspring via natural or sexual selection (e.g., Nosil et al. 2005). Local environmental conditions can also bias dispersal, for example

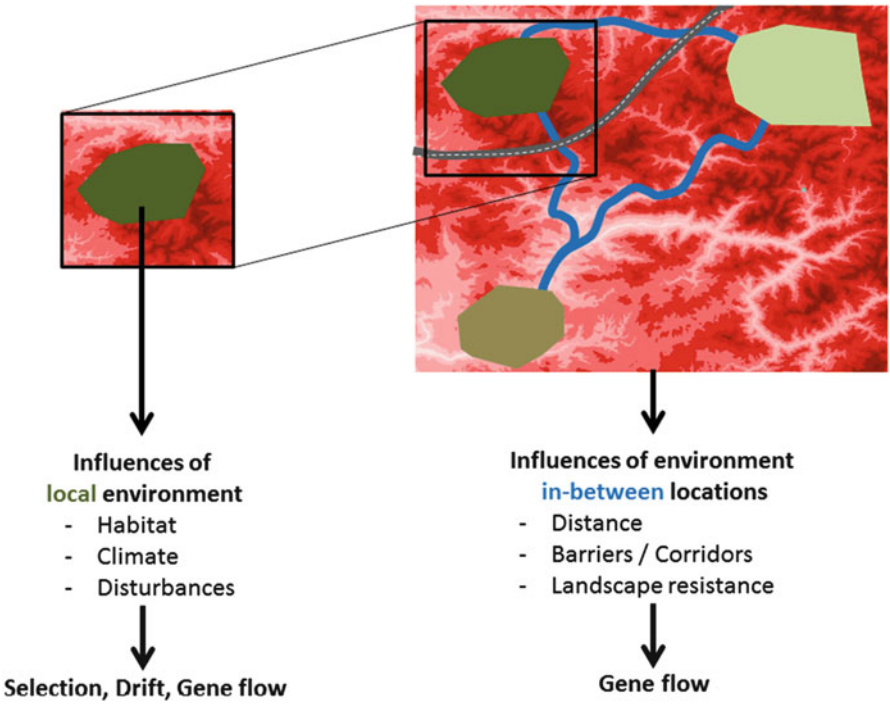

Fig. 2 Potential environmental influences on genetic variation. The green areas show sampling locations or patches, with local environmental conditions indicated by different green color shades. The red gradient represents the heterogeneous quality of the landscape matrix. Left: Local environmental conditions, such as habitat type and quality, can directly impact genetic variation via natural selection, but can also impact drift and effective dispersal, for example when the local carrying capacity affects population size or density-dependent emi- and immigration. Right: Environmental conditions in-between locations can impact effective dispersal and resulting gene flow. In nature, both local and in-between influences will be important for shaping both neutral and adaptive genetic variation

when dispersing individuals prefer to settle into specific environments, or into environments that are similar to their natal environment (Wang and Bradburd 2014). Additionally, local environmental conditions can impact local population sizes and carrying capacities, which impact drift and also density-dependent dispersal and resulting patterns of gene flow (Pflüger and Balkenhol 2014).

The environment in-between locations can include barriers that impede gene flow or corridors that facilitate it. More generally, the environment in-between locations can be characterized in terms of landscape resistance, which essentially reflects the probability that organisms will successfully cross a particular environment (Zeller et al. 2012). The resistance of a landscape is largely determined by the intervening "matrix," which is the term used to describe the landscape found in-between sampling locations that is not primary habitat of the study species. The matrix can alter levels of gene flow among locations, because matrix quality

impacts effective dispersal (i.e., dispersal leading to successful reproduction), either because individuals respond to environmental heterogeneity during dispersal movements, or because the environment experienced during dispersal alters their survival (Zeller et al. 2012). Many studies have shown that the quality and heterogeneity of the matrix (i.e., its resistance) can have profound influences on effective dispersal, gene flow, and resulting patterns of genetic variation (e.g., reviewed in Waits et al. 2016). Analyzing such in-between environmental characteristics is probably the most novel contribution that landscape genomics has made to population genetic studies (Holderegger and Wagner 2008; Manel and Holderegger 2013). While local environmental characteristics, such as patch size or climate, have been considered in many genetic studies even before the terms landscape genetics and landscape genomics were coined (see, e.g., Keyghobadi 2007), explicitly including the effects of landscape structure is relatively new. Landscape structure basically consists of two components, called *composition* and *configuration*. While composition refers to the amount of certain elements within a landscape (e.g., percentage of area covered by different vegetation types), landscape composition refers to the spatial arrangement of these elements (Fig. 3).

The spatial arrangement of the landscape is particularly important for analyzing the effects of the landscape matrix on functional connectivity, or the degree to which a landscape facilitates the dispersal of individual organisms (or their propagules) and resulting gene flow (Taylor et al. 2006).

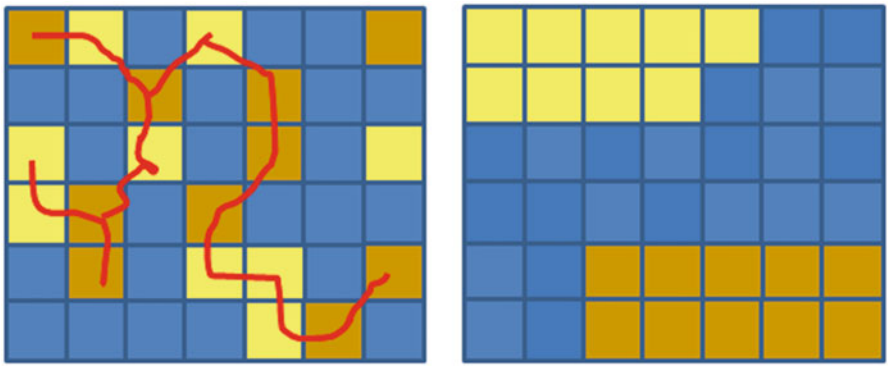

Fig. 3 Landscape composition and configuration. Two hypothetical landscapes with the same composition (i.e., number of blue, yellow, and brown cells), but different configuration (i.e., spatial arrangement of differently colored cells). Assuming that a study species cannot move through the blue cells, the left landscape can actually still be crossed (i.e., the cells in the landscape are still functionally connected, as demonstrated by the red path), even though the left landscape looks more fragmented than the right one

2.3 Analytical Steps

To reflect the different potential genetic influences of the local environment and the matrix, Wagner and Fortin (2013) presented a conceptual framework for the spatial analysis of landscape genetic data. They distinguish among analytical approaches that are based on nodes, neighborhoods, links, or boundaries. While the first two types of methods are most appropriate for analyzing effects of the local environment on genetic variation, the latter two methods are most suitable for analyzing effects of the environment in-between the sampled locations. Importantly, the different environmental influences are not mutually exclusive and probably often interact in nature. Thus, to fully understand the effects of environmental heterogeneity on genetic variation, landscape genomic studies should ideally test and contrast multiple hypotheses relating to different influences.

To assess these various effects, landscape genomic studies essentially have to conduct three analytical steps (Balkenhol et al. 2016a). First, they have to quantify genetic variation so that the genetic data captures the patterns and microevolutionary processes of interest. Second, they have to quantify environmental heterogeneity so that the landscape data reflects different hypotheses to be tested. Finally, they have to statistically link genetic and environmental data so that landscape-genetic hypotheses can be tested explicitly. In reality, these different steps are not always separated, for example because some methods simultaneously accomplish steps 2 and 3. However, to get started with landscape genomics, it often helps to envisage the analysis along these three steps. Hence, we encourage readers to keep the three analytical steps in mind when reading the next sections, where we describe different analytical approaches for neutral and adaptive landscape genomics.

3 Neutral Landscape Genomics

Landscape genomic studies based on selectively neutral genetic data often focus on environmental effects on genetic *structure*. Specifically, most neutral landscape genomic studies analyze environmental effects on functional connectivity, as defined above. To test for environmental effects on functional connectivity, most studies statistically compare measures of genetic similarity with measures of landscape connectivity among sampling units, which may be either individuals or populations (i.e., groups of individual). Individual-based analyses have been shown to have higher power for detecting landscape-genetic relationships (Landguth et al. 2012; Prunier et al. 2013), and are especially suitable for continuously distributed species. However, population-based analyses remain meaningful whenever genetic populations can be defined and delineated with no or little uncertainty, or when analyses are conducted between distinct spatial areas, such as management units.

3.1 Distance-Based Analysis Framework

Regardless of how sampling units are defined, most studies use a distance-based analysis for their inferences (Storfer et al. 2010; DiLeo and Wagner 2016). For this, genetic dissimilarity is usually estimated in the form of genetic distances or indices of genetic differentiation, such as F_{ST} values. Higher values of these pairwise estimates are interpreted as indicating lower levels of gene flow and underlying functional connectivity. These values are then compared to estimates of functional landscape connectivity, which are often calculated in the form of effective distances. These effective distances account for the hypothesized heterogeneity of the landscape and can reflect different landscape-genetic hypotheses (Fig. 4).

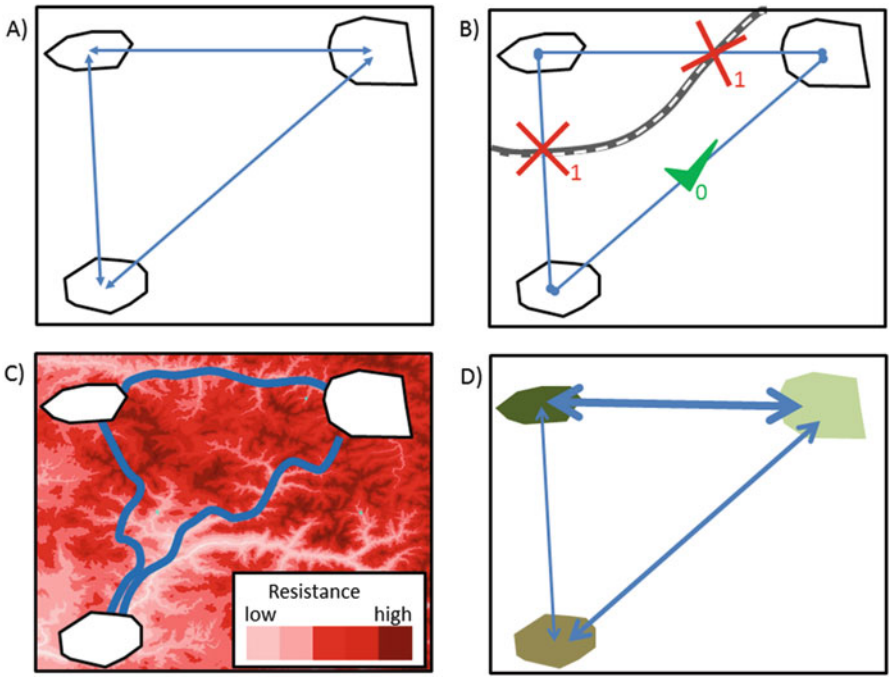

Fig. 4 Hypotheses typically tested in neutral landscape genomics, with blue arrows depicting the effective distances estimated to reflect each hypothesis. (**a**) Isolation-by-distance (IBD) is tested by correlating genetic with geographic distances among locations (straight-line). (**b**) To test isolation-by-barrier (IBB), the occurrence or number of linear barriers found in-between locations is calculated. (**c**) Testing isolation-by-resistance (IBR) requires the estimation of effective distances that account for the heterogeneous resistance of the landscape. (**d**) For isolation-by-environment (IBE), environmental (dis-)similarities are estimated that quantify the differences in local environmental conditions among sampling locations

3.1.1 Isolation-by-Distance (IBD)

An often-tested hypothesis is based on the classical IBD model originally developed by Wright (1943). In this model, gene flow among locations is affected by the geographic (i.e., straight-line) distance separating them (Fig. 4a). The strength of IBD depends on the scale of the study area in relation to the dispersal distance of the species, and the model assumes a homogeneous environment. IBD has been confirmed in many study systems, and it often serves as a null model in both neutral and adaptive landscape genomics.

3.1.2 Isolation-by-Barrier (IBB)

The first alternative to IBD is the hypothesis of IBB (Fig. 4b). Here, effective dispersal and resulting gene flow are assumed to be influenced by linear barriers that cross the study area either partially or completely. Barriers can include mountain ranges, rivers, roads, habitat edges, etc. and can impede gene flow either completely or partially. Analyzing IBB within the distance-based framework can, for example, be accomplished via a dummy matrix, where those sampling location pairs that are separated by a barrier receive a value of 1, whereas locations not separated by a barrier receive a pairwise value of 0. Alternatively, the number of times that a straight line between two sampling locations crosses potential barriers can be used to estimate a pairwise effective distance reflecting IBB.

3.1.3 Isolation-by-Resistance (IBR)

The third and perhaps most prominent landscape-genetic hypothesis is termed IBR (Fig. 4c). This hypothesis considers the heterogeneous resistance of the landscape matrix to movement and gene flow (McRae 2006). To test IBR, various types of effective distances can be estimated that account for the resistance of the environment by calculating the least-costly route among pairs of sampling locations. The two most commonly used effective distances accounting for landscape resistance are based on (a) least-cost and (b) circuit-theoretic algorithms. Least-cost algorithms attempt to find the single, least-costly path, and essentially assume that a dispersing individual has perfect knowledge of the entire landscape and moves through it in an optimal fashion (Adriaensen et al. 2003). In contrast, circuit-theory considers all possible pathways among locations, assumes that individuals move across the landscape randomly, and that their probability of crossing a certain area of the landscape depends on the resistance of that area (McRae et al. 2008). Another option for capturing the landscape resistance among sampling locations is to use transects or corridors of a certain width connecting sampling locations, and to quantify the resistance within these transects (e.g., van Strien et al. 2012).

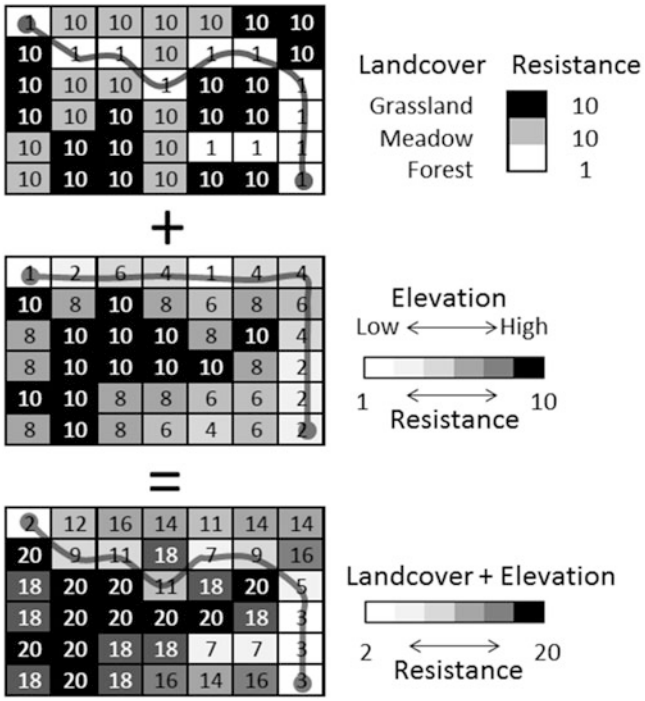

Fig. 5 Illustration of landscape resistance surfaces. For both variables "landcover" and "elevation" resistance values range from 1 (lowest resistance) to 10 (highest resistance). Based on each surface, effective distances can be calculated, for example using least-cost paths (*gray* lines). Note that in this example, the route of the least-cost path is the same for the "landcover" and the combined surface, but the cumulative costs distance is different between all three surfaces

Landscape Resistance Surfaces All of these approaches are based on resistance surfaces, which are spatial data layers where each raster cell of the study landscape receives a value that represents the hypothesized resistance of the landscape in that cell. For example, in Fig. 5, two variables are hypothesized to influence landscape resistance in a hypothetical species that preferentially moves through forest, but tends to avoid high-elevation areas. For the variable "landcover," the resistance value is set to 1 (no resistance), if a cell is covered by forest, and 10 otherwise. Similarly, for the variable "elevation," resistance values range from 1 (lowest elevation) to 10 (highest elevation). The combined resistance surface reflecting both variables ranges from 2 to 20. Based on each surface, effective distances based on least-cost or circuit-theoretic algorithms can be calculated. Many other approaches for parameterizing and optimizing resistance surfaces exist, and they can reflect various hypotheses of linear or nonlinear relationships between the landscape data and resistance to gene flow (e.g., Dudaniec et al. 2013, 2016; Mateo-Sanchéz et al. 2015). The resistance surface leading to effective distances that show the strongest statistical association with the genetic distances is chosen as

the surface that best captures the impacts of landscape resistance on gene flow and spatial genetic structure.

Since resistance values are only hypothesized (i.e., the true impact of the environment on effective dispersal in the study species is not known with certainty), many different combinations of variables and different resistance values of each variable have to be evaluated in empirical studies (Cushman et al. 2006). Due to the complexity of this task, many options exist for parameterizing and evaluating resistance surfaces, and we refer readers to the comprehensive reviews of Spear et al. (2010, 2016) for further details. In landscape genomics, the main result of resistance surface modeling is one or several pairwise matrices of distances (in cost units or electric current values) that measure the effective separation distance between any two sampling locations. Each matrix reflects a different hypothesis of how the study species might be influenced by landscape resistance. The best representation of landscape resistance can then be identified by statistically comparing these different distances to the measures of genetic distances among sampling units, essentially identifying which resistance hypothesis best explains the observed genetic distance.

3.1.4 Isolation-by-Environment (IBE)

The fourth major hypothesis in neutral landscape genomics is IBE (Fig. 4d). Under this hypothesis, the degree of genetic differentiation among sampling units should increase with increasing environmental dissimilarity (Wang and Bradburd 2014). Put differently, higher levels of gene flow should occur among locations that are more similar with respect to local environmental conditions, such as habitat type, temperature, or precipitation. IBE can arise from many different processes, both adaptive and neutral, as discussed in Sect. 5.1.

To test for IBE, effective distances among sampling locations are usually calculated in the form of environmental distances (also called environmental resemblances, similarities, or dissimilarities). These distances essentially estimate a pairwise measure of how close sampling units are in the single or multivariate variable space, so that location pairs with more similar environmental conditions will have lower pairwise environmental distance (i.e., low dissimilarity, high similarity or resemblance). A variety of such environmental distance metrics exist and their advantages and limitations are, for example, discussed in Legendre and Legendre (2012).

Importantly, within the distance-based analytical framework commonly applied in neutral landscape genomics, IBE is the only hypothesis that considers local environmental conditions, rather than influences of the environment in-between sampling locations (i.e., distance, barriers, or resistance).

As stated before, the different hypotheses are not mutually exclusive, and all of them can and should be considered in landscape genomic studies. Also, it should be noted that there are many other approaches for quantifying neutral genetic variation and environmental heterogeneity that we have not discussed here. For example,

there is a suite of methods for detecting (local) linear barriers to gene flow, such as Monmonier's algorithm (Monmonier 1973) or the Wombling method (Womble 1951; Crida and Manel 2007). Results of these methods can be compared to the spatial occurrence of environmental boundaries, such as habitat edges or roads (Blair et al. 2012). Similarly, genetic structure is now often quantified using a variety of genetic clustering methods, for example via the well-known software STRUCTURE (Pritchard et al. 2000; see also François and Waits 2016 for a review of genetic clustering methods). The outcome of such clustering methods can either be used directly for landscape genetic inferences, or they can be used to estimate genetic distances among individuals, which then can be compared to effective distances for enhanced ecological insight (e.g., Balkenhol et al. 2014).

3.2 Statistically Linking Neutral Genetic and Environmental Data

A large variety of statistical approaches exists for the final analytical step, where genetic variation and environmental variables have to be linked statistically (Wagner and Fortin 2013, 2016). Analytical methods commonly used in other fields (e.g., standard regression) can generally not be used in neutral landscape genomics, because we are dealing with pairwise data (genetic and effective distances among all pairs of sampling locations), and because data often show significant positive spatial autocorrelation, basically meaning that data from locations that are close in space tend to be more similar than data from locations far apart. In essence, both of these challenges lead to non-independent data values that violate fundamental assumptions of most standard statistical methods. Hence, the final step often requires the use of special analytical techniques that can deal with the data structure typically encountered in neutral landscape genomics. A detailed review of existing methods for this is beyond the scope of this chapter, and readers are referred to Wagner and Fortin (2016) for a thorough overview of the various available approaches.

Within the distance-based framework outlined above, assessing the relative support for the different landscape genetic hypotheses is often particularly challenging because the various effective distances can be strongly correlated with each other. This challenge is especially severe when effective distances are compared that have been calculated from different parameterizations of the same resistance surface (Zeller et al. 2016). The most commonly used methods are still the Mantel test and the partial Mantel test (Storfer et al. 2010; DiLeo and Wagner 2016), which are basically correlation approaches with significance values estimated via a specific permutation approach. Mantel tests have been severely criticized for various reasons, and alternative approaches should be used whenever possible (Balkenhol et al. 2009; Legendre and Fortin 2010; Guillot and Rousset 2013; Legendre et al. 2015). However, there is currently no consensus on the most appropriate method for

statistically testing associations between environmental and neutral genetic data. More studies are required to assess the relative utility of different methods for specific research questions and data sets.

In sum, a large variety of methods are available for neutral landscape genomics, and their application to different terrestrial and aquatic systems has already led to important findings in ecology, evolution, and conservation biology (see, e.g., Wang et al. 2013; Selkoe et al. 2016a; Waits et al. 2016). This large analytical variety is also one of the main challenges for current studies, because it hinders the comparability of results, and choosing among methods is not trivial (see Sect. 6). However, we highlight that in addition to methodological considerations, another, perhaps even more important aspect for neutral landscape genomics is the precise, a priori definition of multiple, testable hypothesis. This is also important for guiding general study design and sampling in landscape genomics (Balkenhol and Fortin 2016). Thus, rather than getting lost in analytical details, we encourage researchers to first and foremost focus on defining good research questions and design landscape genomic studies that lead to strong scientific inferences.

4 Adaptive Landscape Genomics

Most studies cited above have actually not used genome-wide approaches to quantify genetic variation, but instead have largely relied on microsatellites or relatively short sequences of mitochondrial DNA. Such data are appropriate for addressing questions related to neutral genetic data and underlying processes, but are not quite suitable for addressing landscape genomic research questions dealing with selection and adaptation. For these kinds of questions, we usually have to identify loci or regions under selection, and then statistically relate this adaptive genetic data to environmental heterogeneity. Storfer et al. (2016) distinguish four general frameworks for accomplishing this: (a) a *correlative* framework that is based on outlier detection and/or environmental association analysis, (b) a *phenotypic* framework which relies on quantitative trait loci (QTL) or genome-wide association studies (GWAS), (c) a framework based on *candidate genes*, and (d) a framework based on *exomes and transcriptomes*. We discuss all of these frameworks below, but largely focus on correlative approaches, and especially environmental association analysis (EAA). This is arguably the most widely used framework in landscape genomics right now, it seems to outperform other approaches for identifying adaptive genetic variation in heterogeneous environments (Jones et al. 2013), and it is the only of the four frameworks that directly incorporates environmental data into the detection of selection.

4.1 Correlative Approaches

Correlative landscape genomic approaches often rely on identifying loci or regions under selection via different types of "outlier detection" methods. As pointed out by Luikart et al. (2003), this population genomics approach aims at finding the genes whose diversity patterns do not follow the ones of the rest of the genome. This is greatly facilitated by high-throughput sequencing techniques, which enable genome-wide genotyping and sampling of genetic markers that may be situated in, or linked to, functional genes that are under selection. Popular "reduced representation" sequencing approaches to obtain genome-wide markers such as Single Nucleotide Polymorphisms (SNPs) include Restriction-site Associated DNA Sequencing (RADseq, Miller et al. 2007) or Genotype by Sequencing (GBS, Elshire et al. 2011; Narum et al. 2013) and target enrichment (Dasgupta et al. 2015; Lu et al. 2016, 2017; Suren et al. 2016). Details on these approaches can be found in another chapter of this book specifically focusing on genotyping and sequencing technologies in population genetics and genomics.

These techniques can recover numerous (100s to 1,000s) genetic markers that can then be partitioned into neutral and selective loci using a wide variety of statistical approaches (e.g., Günther and Coop 2013; Whitlock and Lotterhos 2015). The idea beneath is that a gene under selection will not obey to the neutral forces rather than by selection that shape genetic variation and hence show genotype frequencies that cannot be explained by the neutral theory of molecular evolution. The fixation index (F_{ST}) and the deviation from Hardy-Weinberg proportions (F_{IS}) are popular measures of the genetic differentiation among populations for a specific gene. For example, SNPs under putative selection may be identified using F_{ST} outlier tests, which identify loci with higher or lower F_{ST} than expected from the F_{ST} distribution expected under neutrality. Numerous statistical tests have been designed to detect genes significantly differing in these proxy variables in comparison to the rest of the genome. The earliest types of significance tests developed for outlier loci detection are listed by Luikart et al. (2003). More recently, elaborated statistical analysis of F_{ST} has been developed following Bayesian approaches (e.g., BAYESCAN, Foll and Gaggiotti 2008) or analyses based on principal components (e.g., PCadapt, Duforet-Frebourg et al. 2016). Hoban et al. (2016) provide an in-depth review of the advantages and limitations of such "genome-scan" approaches for finding signals of local adaptation. Once adaptive genetic data have been identified, it can be statistically compared among different landscapes or among different environmental categories (e.g., Turner et al. 2010).

4.1.1 Environmental Association Analysis (EAA)

Alternatively or in addition to outlier detection, EAA can be used to detect signatures of local adaptation to environmental heterogeneity. EAA is at the

interface of bioinformatics, genomics, spatial statistics, and landscape ecology and uses correlation studies between the genomic data and the environment to identify genes either potentially linked to candidate genes or the genes themselves under selection. Note that various other terms are used to describe EAA, including that the terms "genetic-environment correlation" or "genotype-environment association" (GEA, e.g., Whitlock and Lotterhos 2015).

Luikart et al. (2003) were the first to realize the potential of combining landscape genetic analyses with population genomic data. The first implementation of such an approach was published by Joost et al. (2007) in a study dedicated to the detection of candidate loci for selection in insect and livestock species.

Environmental association studies relate environmental variation to genetic polymorphisms, searching for correlative indication of evolutionary responses to spatial heterogeneity (Holderegger et al. 2010). Such associations depend on precisely describing environmental conditions, which require elaborated engineering tools for high-resolution, area-wide coverage of microsite characteristics. In parallel, a whole-genome perspective should enable one to identify potentially adaptive loci or genomic elements, which can then be tested for how they correlate with variation in site conditions (Parisod and Holderegger 2012).

Numerous tools have been developed to perform EAA analysis, each differing mainly by the type of model employed, the statistical procedure used to test for the association, and the way population structure is dealt with (see below; Rellstab et al. 2015). EAA can be performed with various statistical approaches, including logistic regressions (Stucki et al. 2016; Joost et al. 2007; Carl and Kühn 2007), matrix correlations (Hancock et al. 2011; Fischer et al. 2013), general linear models (Zulliger et al. 2013; Manel et al. 2012; Bradbury et al. 2013a; Legendre et al. 2012), and mixed effect models (Frichot et al. 2013; Coop et al. 2010). Excellent reviews and comparisons of the different analytical approaches for EAA can be found in Rellstab et al. (2015) and Forester et al. (2016).

4.1.2 Accounting for Population Structure

Regardless of the statistical approach chosen to conduct EAA, a specific issue is the incorporation of population structure as a confounding factor. Individuals close in space tend to be genetically similar, producing a gradient of neutral genetic differentiation that might overlap with environmental gradients and result in false signals of adaptation (Rellstab et al. 2015; Joost et al. 2013).

The earliest EAA methods did not incorporate the neutral genetic structure and simply aimed at testing the association between genotype frequencies and environmental gradients (Joost et al. 2007; Carl and Kühn 2007). These approaches tend to increase the false discovery rate but are less demanding in terms of calculation (Rellstab et al. 2015). More recently, several methods have been developed to take into account population structure (Rellstab et al. 2015). Among them some are employing mixed effects models (Rellstab et al. 2015), such as BayEnv (Coop et al. 2010; Günther and Coop 2013), LFMM (Frichot et al. 2013), or BayPass (Gautier

2015), while BayEscEnv is based on an alternative model (de Villemereuil and Gaggiotti 2015). The inclusion of population structure in the models allows these methods to get a relatively low false positive rate (De Mita et al. 2013; Frichot et al. 2013; Forester et al. 2016). These approaches interpret the overall variance within the genotype matrix as neutral genetic structure (Rellstab et al. 2015). This can be a major drawback, if the considered population is not genetically structured and can, therefore, result in a loss of statistical power. Moreover, these methods use Markov Chain Monte Carlo (MCMC) as stochastic algorithm, which requires multiple runs in order to obtain representative results (Rellstab et al. 2015; Coop et al. 2010; Frichot et al. 2013). The computation time requested is therefore substantial (Rellstab et al. 2015; Stucki et al. 2016). For this reason, high performance computation methods like Samβada (Stucki et al. 2016) can be a valuable alternative. The population structure can be previously investigated using specific tools like ADMIXTURE (Alexander et al. 2009), STRUCTURE (Pritchard et al. 2000), Localdiff (Duforet-Frebourg and Blum 2014), or a principal component analysis (PCA, Patterson et al. 2006). If the population structure is meaningful, the coefficients of membership to the subpopulations can be included in a bivariate model; otherwise a univariate model is employed considering the environmental variables by themselves (Stucki et al. 2016).

4.1.3 Global and Local Spatial Autocorrelation

Beyond detection of selection signatures, it is possible to quantify the level of spatial dependence in the distribution of genotypes analyzed. This measure of spatial autocorrelation refers to similarities or differences among neighboring individuals that cannot be explained by chance. Assessing whether the geographic location has an effect on allele frequency is especially important in landscape genomics since statistical models assume independence between events. Thus, if individuals with similar genotypes tend to concentrate in space, spurious correlations may co-occur with specific values of environmental variables. On the other hand, spatial independence of data strengthens the confidence in the detections.

Samβada software (Stucki et al. 2016) measures the global spatial autocorrelation in the whole dataset with Moran's *I*, as well as the spatial dependency of each point with Local Indicators of Spatial Association (LISA, Moran 1950; Anselin 1995). In practice, LISAs are computed by comparing the value of each point with the mean value of its neighbors as defined by a specific weighting scheme based on a kernel function. Both a spatially fixed kernel type relying on distance only and a varying kernel type considering point density can be used. There are three fixed kernels (moving window, Gaussian, and bi-square) and a varying one (nearest neighbors). The sum of LISAs on the whole dataset is proportional to the Moran's *I* (Anselin 1995). Significance assessment relies on an empirical distribution of the indices. For Moran's *I*, genotype occurrences are permutated among the locations of individuals of the whole dataset and a pseudo *p*-value is computed as the proportion of permutations, for which *I* is equal to – or more extreme (higher for

a positive Moran's I or lower for a negative Moran's I) – than the observed I. For LISA, the pseudo p-value is separately computed for each point (individual), by keeping the value of the individuals of interest fixed and permuting its neighboring points with the rest of the dataset.

Once a diagnostic of spatial dependence in loci of interest has been carried out, a clever approach is to develop spatially explicit models that directly include autocorrelation. SGLMM (Guillot et al. 2014) provides such a model; however, the current *R*-based implementation does not fit the computational requirements of whole-genome analysis. Alternatively, Geographically Weighted Regressions (GWR) measure the spatial stationarity of regression coefficients by fitting a distinct model for each sampling location. The number of neighboring points considered for each sampling location is given by the weighting scheme. These models allow some local coefficients to differ between sampling points while some "global" coefficients are common to all points (Fotheringham et al. 2002; Joost et al. 2013). Thus GWR enable building a null model where the constant term may vary in space and then refining it by adding a global environmental effect for all locations. Comparing these two models would enable assessing whether the global environmental effect is needed to describe the distribution of the genotype. The key advantage of allowing the constant term to vary in space is to take spatial autocorrelation into account in the models. This way, GWR allow investigating the spatial behavior of loci showing selection signatures with standard logistic regressions and may help to distinguish between local adaptation and population structure in landscape genomics. However, GWR models require a fine-tuning of the weighting scheme from the user, which restrains their application to very large datasets. Another method borrows from techniques that examine changes in species community composition through space, but instead assesses the effect of environmental gradients on changing allele frequencies using Generalized Dissimilarity Modeling (GDM) or Gradient Forest (GF) analysis (Fitzpatrick and Keller 2015). The GDM/GF approach may be applied to any system but is particularly useful for range expanding species (see Sect. 5.2), as it allows the effects of geography to be filtered out (e.g., by integrating latitude and longitude into the model) as well as neutral genetic processes (Fitzpatrick and Keller 2015).

4.1.4 Combining EAA Approaches

In the sections above, we listed strengths and drawbacks of several EAA approaches. It is important to point out that all of these approaches have implicit common assumptions concerning the functional relationship between allele distribution and environmental variables (Joost et al. 2013). In particular, such a relationship needs to be constant and requires time to be established after environmental change arises (Joost et al. 2013). A good way to deal with the uncertainty in the results produced by EAA analysis is the combination with other adaptive landscape genomics approaches (see below), and also the combination of multiple EAA approaches. Loci that are detected to be under selection by different methods can alleviate the weaknesses of

each approach, thus leading to greater reliability of inferences (Rellstab et al. 2015). Similarly, combining population genomics approaches (i.e., outlier detection) with EAA can cope with the intrinsic limitations of each paradigm (Rellstab et al. 2015; Joost et al. 2013).

4.2 Phenotypic Approaches

Another approach for identifying loci under selection involves finding associations between genetic variation and fitness-relevant phenotypes. Quantitative Trait Loci (QTL) mapping has been employed way before the advent of genome-wide scanning technologies and represents nonetheless a powerful tool to detect genes responsible for adaptation (Ehrenreich and Purugganan 2006; Stinchcombe and Hoekstra 2008). In QTL mapping, the genetic contribution to a measurable phenotype is investigated by crossing two parental individual differing in this phenotype and by analyzing how the genetic markers segregate with the phenotype in the successive generations (Stinchcombe and Hoekstra 2008). When studying adaptation, it is therefore necessary to focus on a phenotype on which the selection force acts (Ehrenreich and Purugganan 2006). The onset of genomics allowed to increase the accuracy of mapping the QTLs, but didn't help overcoming the major limitations of this approach: the need for an experimental breeding and the certainty about the adaptive phenotype (Borevitz and Chory 2004; Stinchcombe and Hoekstra 2008).

GWAS solve many of the abovementioned shortcomings of the QTL mapping. GWAS accompanied the appearance of high-throughput technologies for genetic marker identification and investigates the association between quantifiable phenotypes and genome-wide genetic markers (McCarthy et al. 2008). This method requires a large sample size, but overcomes the need for experimental breeding (McCarthy et al. 2008). It also has a higher resolution for marker-trait associations, and some of these associations could be rather causative and not only due to the close linkage between markers and causative genes. By using adaptive, trait-related phenotypes, GWAS can facilitate the search for the genetic variants responsible for the adaptation (Morris et al. 2013). After identifying adaptive genetic variation through QTLs or GWAS, this variation can be statistically linked to environmental data. Some EAA studies have even directly employed GWAS-inspired methods considering environmental variables as phenotypes (Bradbury et al. 2013a; Eckert et al. 2009; Porth et al. 2015), stressing therefore the symmetry between these two paradigms (Rellstab et al. 2015).

4.3 Candidate Genes

Rather than scanning the whole genome for loci potentially under selection (as in GWAS), the candidate gene approach makes use of information available from model species (i.e., those with an annotated reference genome), where adaptive genes or gene regions have already been identified. Focusing on these regions in phylogenetically related non-model species can increase the chances of finding signals of selection. The candidate gene approach applied to adaptive traits can lead to the detection of gene variants implied in an adaptive response (Pel et al. 2009). The candidate loci can be chosen because of the homology with genes of known function in other species, proximity to the genomic region associated with a phenotype, because of the function predicted from its sequence, or from studies of mutants (Pflieger et al. 2001; Rellstab et al. 2015). The candidate gene approach represents an appealing source of genetic information in EAA analysis when costs or other technical reasons do not allow for a complete coverage of the genome (Rellstab et al. 2015). In a landscape genomics context, candidate genes have been statistically linked to habitat types (Hoekstra et al. 2006) and climate (Sork et al. 2016), and the framework will likely be used more often in future studies (see also Bragg et al. 2015).

4.4 Exomes and Transcriptomes

The fourth and final framework for adaptive landscape genomics identified by Storfer et al. (2016) also relies heavily on high-throughput sequencing data. The large amounts of high-quality data make it possible to analyze exome and transcriptome variation, i.e., the types and amounts of RNAs that could be associated with differences in gene regulation among different environments. Approaches for exome and transcriptome analysis are discussed by Storfer et al. (2016), who highlight the potential of these approaches for creating novel data, and hence new insights on how environmental change alters gene expression. However, while exome and transcriptome analyses indeed have tremendous potential for elucidating adaptive landscape genomic processes, they have not yet been applied in an actual landscape genomic study.

5 Examples of Landscape Genomics Applications

Both neutral and adaptive landscape genomic approaches have been applied to a large variety of organisms and to address a substantial diversity of research questions. Indeed, the scope of landscape genomic applications is ever-increasing, and is now stretching well beyond the focus on plants and animals inhabiting terrestrial

ecosystems (e.g., Manel et al. 2003). For example, metagenomic sequencing of microorganisms enables understanding of microbial community structure and the landscape or climatic characteristics that determine their diversity and persistence, and landscape genomic analyses are beginning to be explored in this area (reviewed in Dudaniec and Tesson 2016). In addition to community structure, metagenomics may also allow for identification of functional, adaptive genes among microorganisms, which may be used in a landscape genomics framework as done for macroorganisms (Dudaniec and Tesson 2016). Similarly, landscape genomic approaches are now used to understand the spread and dynamics of pathogens in heterogeneous environments (Biek and Real 2010; Alamouti et al. 2014; Schwabl et al. 2017).

Here, we provide several examples of landscape genomic applications that have either not been covered elsewhere, or because they nicely illustrate the high flexibility of landscape genomics approaches in nontypical "landscapes" (e.g., seascapes, see Sect. 5.4). We refer readers to other recent publications that provide more detailed reviews on specific systems or taxa, e.g., Montgelard et al. (2014) for terrestrial mammals, Selkoe et al. (2016a, b) for aquatic systems, and Dyer (2016) for plants in general.

5.1 Landscape Genomics of Terrestrial Organisms

In this section, we provide a critical review of how landscape genomics has been used to assess both neutral and adaptive genomic variation in terrestrial plant and animal organisms. Though our review of the literature is not exhaustive, the articles chosen should serve as a barometer for the state of research in this niche area of the field. We feel that the papers reviewed here provide a sufficient overview of where the field has been, and where it is heading.

5.1.1 Landscape Genomics of Forest Trees

Forest trees are ideal model organisms illustrating the use of landscape genomics within terrestrial systems. Forests are charismatic components of many landscapes and keystone species in many ecosystems. They provide habitat for a wide range of species and are a central component of the landscape matrix. Yet beyond their ecological function, forest trees are economically important and provide a wide range of ecosystem services to society (Costanza et al. 1997). Because of their economic importance many tree species are exploited, and the remaining landscapes can become disconnected and fragmented, leading to concerns about the preservation of genetic diversity and adaptive capacity (Krutovsky et al. 2012; Ratnam et al. 2014). In conjunction with man-made landscape fragmentation, climate change (Allen et al. 2010) and forest disturbance (Dale et al. 2001) threaten large tracts of these long-lived species. As patches of forest become less functionally connected cascading effects may disrupt the flow of energy and nutrients

destabilizing the entire system. In order to better understand how changes in climate and landscape structure will affect the long-term stability and resilience of terrestrial systems, like forests, researchers must find ways of linking the processes of gene flow, dispersal, and adaptation to landscape change.

Because of their unique life history characteristics and high levels of genomic and phenotypic variation, forest trees are good models for understanding evolution and population processes (González-Martínez et al. 2006; Sork et al. 2013). Like other terrestrial plants, trees are fixed in space and cannot directly move away from changing environments. Thus, their ability to respond to change is limited to dispersal, local adaptation, and phenotypic plasticity (Aitken et al. 2008). While assessing genetic versus plastic response in forest trees remains a challenge (but see Benomar et al. 2016), landscape genomics allows us to investigate both migration and adaptation across many spatial scales and across a wide range of environmental heterogeneity (Johnson et al. 2016, 2017a, b).

The Beginnings of Forest Tree Landscape Genomics

Forest ecologists and geneticists have long realized that seed and pollen dispersal must be understood within a spatial and temporal framework (Loiselle et al. 1995; Schupp and Fuentes 1995; Sork et al. 1999). Understanding landscape connectivity and gene flow, with an emphasis on barrier detection and the effects of forest fragmentation, and its impacts on the distribution of genetic variation and diversity has been a common focus in terrestrial systems. Until recently, neutral genetic variation assessed using microsatellite markers (SSRs) was the primary genetic approach. Most of the work assessed functional connectivity. While the use of selectively neutral genetic markers such as SSRs was very important for understanding demographic and neutral processes in forest systems, they did not allow for the direct assessment of adaptation and selection in the face of change. Fortunately, recent progress in genomics and nucleotide sequencing provides researchers with practically unlimited numbers of markers including both selectively neutral (such as SSRs and SNPs in noncoding regions) and potentially affected by selection (for instance, non-synonymous SNPs). Arguments were made for understanding forest fragmentation on seed and pollen gene flow (Smouse and Sork 2004; Sork and Smouse 2006), evolutionary adaptation in forest trees (González-Martínez et al. 2006; Holderegger et al. 2008; Kremer et al. 2012; Sork et al. 2013; Lepais and Bacles 2014), and a broad incorporation of landscape genomics into plant and tree research (Holderegger et al. 2010). Many of these papers highlighted the benefits of landscape genomics in investigating how forests will respond to global ecological change, yet the generality and transferability of single species studies have recently been questioned (Calic et al. 2016), and a shift from descriptive to predictive studies still has not been fully achieved (Manel and Holderegger 2013).

Neutral Landscape Genomics of Forest Trees

An important avenue of landscape genomics using neutral genetic variation has been to identify barriers impacting gene movement. For example, estuarine barriers and ocean currents were found to restrict gene flow in the mangrove species *Avicennia germinans*, *Rhizophora mangle*, and *Rhizophora mucronata* (Ceron-Souza et al. 2012; Wee et al. 2014), and ocean barriers limited gene flow in sandalwood (*Santalum insular*) (Lhuillier et al. 2006). Likewise, barriers have been detected in terrestrial landscapes. Mountain ranges have been found to restrict gene flow in ash (*Fraxinus mandshurica*) (Hu et al. 2010), birch (*Betula maximowicziana*) (Tsuda et al. 2010), and oak (*Quercus lobata*) (Ashley et al. 2015), and decreasing river size was found to restrict gene flow in cottonwood (*Populus fremonti*) (Cushman et al. 2014). In their study of valley oak (*Q. labata*), Ashley et al. (2015) found that not only did mountain ranges restrict gene flow, but wide open expanses also limited the movement of pollen. Though mountain ranges and landscape features can limit gene flow, the finding is not universal. A range wide study of sweet chestnut (*Castanea sativa*) identified population genetic structure, but did not identify barriers to dispersal using Monmonier's algorithm, concluding that the large population differentiation was due to divergent selection and not barriers to gene movement (Martin et al. 2012).

Another major focus of forest landscape genomics using neutral genetic variation is the assessment of the effects of forest fragmentation on the distribution of genetic diversity. From conservation and management perspectives, understanding how the spatial configuration of harvest tracts or the pattern of disturbance impacts landscape connectivity and gene flow is of critical importance to preserving genetic diversity (Krutovsky et al. 2012; Ratnam et al. 2014). In the case of forest fragmentation, many studies have shown trees to be resilient to fragmentation due to long distance dispersal and high levels of pollen gene flow among forest fragments (Savolainen et al. 2007; O'Connell et al. 2006, 2007). Fragmented landscapes were found to be functionally connected in oak species, *Quercus macrocarpa* and *Q. sclerophylla* (Craft and Ashley 2007; Wang et al. 2011, 2012), mountain hemlock, *Tsuga mertensiana* (Ally and Ritland 2007; Johnson et al. 2017a, b), service tree, *Sorbus domestica* (Kamm et al. 2009, 2010), mountain birch, *Betula pubescens* (Truong et al. 2007), white spruce, *Picea glauca* (Fageria and Rajora 2013; O'Connell et al. 2006, 2007), and papaya, *Carica papaya* (Chavez-Pesqueira et al. 2014). In contrast, an analysis of cottonwood, *P. fremonti* found increased fragmentation resulted in decreased genetic diversity with important conservation implications for other terrestrial species (Cushman et al. 2014).

Abiotic factors can also impact gene flow. Temporal patterns of wind speed and direction can shape spatial genetic structure in forest species. The seasonal differences in wind direction explained the spatial genetic structure of *Engelhardia roxburghiana* due to timing of pollen and seed release (Wang et al. 2016). Few studies have incorporated wind timing, or other temporally variable factors into landscape genetic analysis in terrestrial systems.

Within a range wide context, genetic diversity is often structured according to the center-periphery (central-peripheral or central-marginal) model. In this model

higher gene flow from large central populations into small peripheral ones maintains genetic diversity (Kremer et al. 2012). It is often associated with a reduction in the adaptive potential of individuals at the edge, because the influx of individuals adapted to the center of the range counters the impact of selection for traits suitable to the surrounding environment (e.g., gene swamping) (Kirkpatrick and Barton 1997; Lenormand 2002; Gaston 2009; Kubisch et al. 2014). On balance, there are cases where curtailment of gene flow to marginal populations at range edges can reduce genetic diversity within the marginal populations. Populations of eastern white pine (*Pinus strobus*) in northern Ontario had significantly lower allelic diversity and effective population size at the margins than did central populations (Chhatre and Rajora 2014).

Studies assessing postglacial colonization of northern landscapes have shown forest trees to be capable of long distance dispersal and gene flow, despite the negative effects of diversity loss due to bottleneck effects (Roberts and Hamann 2015). This phenomenon has been demonstrated in pequi, *Caryocar brasiliense* (Diniz-Filho et al. 2009), Sitka spruce, *Picea sitchensis* (Holliday et al. 2012), eastern white pine (Zinck and Rajora 2016), and mountain hemlock, *Tsuga mertensiana* (Ally et al. 2000; Johnson et al. 2017a, b).

Adaptive Landscape Genomics of Trees

A handful of papers helped initialize the landscape genomic approach in forest trees using genome-wide SNP and AFLP markers to identify putative loci under selection for different climatic variable using EAA (Table 1). The bioclimatic factors of temperature and precipitation were associated with outlier loci in white spruce, *Picea glauca* (Namroud et al. 2008), black spruce, *Picea mariana* (Prunier et al. 2011), black alder, *Alnus glutinosa* (Cox et al. 2011), and European beech, *Fagus sylvatica* L. (Cuervo-Alarcon 2017). An EAA identified SNP loci associated with aridity, precipitation, and temperature in loblolly pine, *Pinus taeda* (Eckert et al. 2010a, b; Chhatre et al. 2013). A nice example of the EAA approach in forest trees is the association of outlier loci with serotiny in lodgepole pine, *Pinus contorta* (Parchman et al. 2012). It was found that 50% of phenotypic variation was associated with just 11 loci across three different populations.

Recently, landscape genomics has been combined with common garden approaches to associate genomic variation with phenotypic variation across environmental gradients (Table 1). Sork et al. (2010) used SSRs to correlated geographic patterns of genetic variation to climate using ecological niche modeling in California valley oak (*Q. lobata*), illustrating that, historically, the species was connected across its range through dispersal, and that observed genetic structure was thus related to climatic adaptation and not dispersal limitation. By combing common gardens with landscape genomic analysis we have the potential to separate out genotype by phenotype by environment interactions (Lepais and Bacles 2014). The approach has been used in a variety of system to link specific phenotypic traits to geographic gradients of environmental and climate variables. Leaf size was

Table 1 Selected landscape genomic studies of forest trees using adaptive genetic markers

Species	Markers	Analysis and approach	Inference	Reference
Picea glauca	534 SNPs	F_{ST} outlier	Candidate genes were identified and associated with temperature and precipitation	Namroud et al. (2008)
Pinus taeda	3,059 SNPs	Association	8 outlier SNPs associated with aridity and genome-wide population structure	Eckert et al. (2010b)
Pinus taeda	1,730 SNPs	Association/generalized linear mixed model (GLMM)	48 SNPs were correlated to PCs describing temperature, precipitation, and winter aridity	Eckert et al. (2010a)
Alnus glutinosa	163/154 AFLPs	Association	Identified 4 outlier loci associated with climate, mainly temperature	Cox et al. (2011)
Picea sitchensis	339 SNPs	Outlier	14 SNP outliers, asymmetrical gene flow from center to edge effects adaptive capacity	Holliday et al. (2012)
Picea mariana	583 SNPs	Outlier	26 SNPs identified as outliers associated with temperature and precipitation	Prunier et al. (2011)
Populus balsamifera	335 cand. SNPs 412 ref. SNPs	F_{ST} outlier	46 outlier SNPs identified and associated with *Arabidopsis* flower-time network	Keller et al. (2012)
Abies alba *Larix decidua* *Pinus cembra* *Pinus mugo*	249 SNPs 267 SNPs 459 SNPs 693 SNPs	Association	Seasonal minimum temperature was the most important climate variable for all species; genetic data were correlated with geography	Mosca et al. (2012)

(continued)

Table 1 (continued)

Species	Markers	Analysis and approach	Inference	Reference
Pinus contorta	97,616 SNPs	Association	11 loci were associated with serotiny, explaining 50% of genetic variation	Parchman et al. (2012)
Eucalyptus gomphocephala	7 SSRs 11 EST-SSRs	F_{ST} outlier/association	2 EST-SSRs were identified as undergoing diversifying selection; these loci were associated with climate and gradient variables; the study identified adaptive genetic markers	Bradbury et al. (2013a)
Picea rubens	61 SNPs in 36 candidate genes	Outlier/association	This study assessed how climate and pollution led to local adaptation; 7 SNP loci were associated with climate in older trees (cohort), while 3 SNP loci were associated with pollution in younger trees (cohort)	Bashalkhanov et al. (2013)
Pinus taeda	2,665 unigene based SNPs	Outlier/association/principal component analysis/logistic regression/Bayesian mixed linear model implemented in BAYENV	Multiple associations with latitudinal, elevational, and climatic variables were identified	Chhatre (2013); K. V. Krutovsky and V. E. Chhatre, (unpublished)
Alnus glutinosa	1,990 SNPs	Common garden + outlier and association	Phenotypic variation in leaf size is linked with outlier analysis and associated with variation in temperature	De Kort et al. (2014)

(continued)

Table 1 (continued)

Species	Markers	Analysis and approach	Inference	Reference
Populus trichocarpa	29,354 SNPs	Common garden + outlier and association	Heritability was strong for phenology traits compared to biomass/growth traits; variability was associated with latitude, max day length, and temperature of tree origin	McKown et al. (2014)
Abies alba *Larix decidua*	231 SNPs 233 SNPs	IBD vs. IBA/outlier	Both isolation and adaptation occur at landscape scales; 2–7 outlier SNPs were associated with temperature and soil	Mosca et al. (2014)
Pinus massoniana *Pinus hwangshanensis*	25 candidate genes	Isolation with migration and adaptation	Ecological divergence of two species associated with climate	Zhou et al. (2014)
Pinus lambertiana	475 SNPs	Common garden + outlier and association	SNPs were associated with 5 phenotypic traits and 11 environmental variables; 6 SNPs were associated with phenotype and 31 with environmental variables; 2 SNPs associated with both phenotype and environment and 1 of those was associated with carbon isotope and soil/climate factors	Eckert et al. (2015)
Pinus lambertiana	186 SNP candidates	Neutral population genetic structure/outlier	The candidate approach identified 2 population clusters and 9 candidate SNPs associated with drought	Vangestel et al. (2016)

(continued)

Table 1 (continued)

Species	Markers	Analysis and approach	Inference	Reference
Quercus lobata	195 SNPs 40 candidate genes	Candidate gene/ association	3 SNPs were associated with bud burst and flowering; 2 SNPs were associated with temperature and precipitation; these associations varied with climate and provide support for spatially divergent selection	Sork et al. (2016)
Abies alba	267 SNPs in 175 candidate genes	Candidate gene/ association	16 SNPs showed divergent selection; all outlier SNPs were associated with winter drought and one of them showed selection in relation to elevation; 2 F_{ST} outliers suggested adaptive divergence for date of bud flush and growth rate	Roschanski et al. (2016)
Pinus cembra *Pinus mugo*	768 SNPs 1,152 SNPs	Outlier/ association	Outlier and association analysis tested to what degree elevation effected genomic diversity; low genomic differentiation was found; outliers were associated with temperature and precipitation; 5 SNPs were in common between the species and associated with abiotic stress response; temperature was shown to be an important component of adaptive potential	Mosca et al. (2016)

(continued)

Table 1 (continued)

Species	Markers	Analysis and approach	Inference	Reference
Pinus strobus	44 SNPs	F_{ST} outlier	SNPs in 25 candidate genes were identified and associated with 19 bioclimatic variables	Rajora et al. (2016)
Fagus sylvatica L.	13 microsatellite markers and 70 SNPs in 24 climate adaptation related candidate genes	Outlier/association/principal component analysis/logistic regression/Bayesian mixed linear model implemented in BAYENV	Association with environmental variables was detected for 24 (34.3%) SNPs, and 5 (7.14%) of them were identified also as FST outliers	Cuervo-Alarcon (2017)

linked with range wide temperature variation in *A. glutinosa* (De Kort et al. 2014). Variability in phenology was associated with latitude, maximum day length, and temperature of tree origin in black cottonwood, *Populus trichocarpa* (McKown et al. 2014). In a broad analysis of sugar pine (*Pinus lambertiana*) Eckert et al. (2015) associated five different phenotypic traits, including height, bud flush, carbon isotope, and nitrogen concentration, and 11 environmental variables with hundreds of genome-wide SNPs. Their analysis revealed an association with both phenotype (carbon isotope) and environment (soil/climate factors associated with water availability) at a single SNP locus.

The candidate gene framework discussed in Sect. 4.3 has also been increasingly used in forest systems (Table 1). Phenotype (bud burst/flowering) and environment (temperature and precipitation) were found to vary with climate in *Q. lobata* (Sork et al. 2016), and *P. lambertiana* candidate loci were associated with drought (Vangestel et al. 2016). A novel assessment of tree age structure in red spruce (*Picea rubens*) found that candidate loci in older trees were associated with climate while candidate loci in young trees were associated with air pollution (Bashalkhanov et al. 2013). Candidate loci associated with photoperiod were found to vary with latitude, temperature, and precipitation in balsam poplar, *Populus balsamifera* (Keller et al. 2012). Additionally, the approach has been used to assess ecological divergence between two closely related species *Pinus massoniana* and *Pinus hwangshanensis* (Zhou et al. 2014). Though many association studies have focused on single gene effects, multilocus effects may provide a better correlation with environmental predictors. Single gene outlier analysis approaches were compared to single and multilocus environmental association analyses in eastern white pine where among-population multilocus genetic covariance had a much higher correlation with climate

factors than did single gene effects (Rajora et al. 2016). The aforementioned study used a population graph approach (Dyer and Nason 2004; Dyer 2015b) to construct among-population multilocus covariance genetic distances (cGD) separately for the SSR and SNP loci based on the topology of the constructed population graphs. This approach is an emerging method in landscape genomics that can account for patterns of population genetic structure where F_{ST} outlier approaches that rely on unreasonable assumptions fail (Dyer 2015b; Murphy et al. 2016).

Comparative Landscape Genomics in Forest Trees

To date, most landscape genomic studies in forest trees, and terrestrial systems more broadly, have focused on single species. Comparative studies are now being conducted to identify if similar landscape and climate processes affect different species within a landscape in the same way. Assessing neutral genomic variation, a multispecies study found that different landscape processes best explained population genomic structure in *Bursera simarubra*, *Ficus insipida*, and *Brosimum alicastrum* (Poelchau and Hamrick 2012). Differences in pairwise F_{ST} between populations of the three species were each associated with different environmental variables using partial Mantel's tests: environmental niche distance in *B. simarubra*, geographic distance in *F. insipida*, and historic barriers in *B. alicastrum*. A multispecies study in *Abies alba*, *Larix decidua*, *Pinus cembra*, and *Pinus mugo* used SNPs developed from a common gene pool to identify common associations with climate, environment, and population genomic structure (F_{ST}). A multivariate approach found that SNPs in all four species were strongly correlated with principle components corresponding to seasonal minimum temperature. However, individually each species was also correlated with a wide range of environmental variables (Mosca et al. 2012). This study was further expanded upon to test if either isolation by distance (IBD) or isolation by adaptation (IBA) was responsible for the differences among the populations of each of the two species: *Abies alba* and *Larix decidua* (Mosca et al. 2014). As it turns out, both IBD and IBA were present, and even though IBD was stronger than IBA, after accounting for geographic distance, 2–7 outlier loci were associated with temperature and soil (Mosca et al. 2014) (Table 1).

5.1.2 Landscape Genomics of Wildlife

The vast majority of landscape genomic studies in terrestrial wildlife have focused on neutral processes, particularly concerning dispersal movements and resulting patterns of gene flow (Storfer et al. 2010). These landscape genomic studies focusing on wildlife connectivity have recently been reviewed by Waits et al. (2016) and will therefore not be discussed here. Instead, we highlight three recent studies on adaptive landscape genomics in non-model terrestrial vertebrates that nicely illustrate the diversity of landscape genomic approaches for wildlife research.

Relating Candidate Genes to Parasite Load in Red Grouse

Wenzel and Piertney (2015) selected 12 candidate genes previously developed for red grouse (*Trichostrongylus tenuis*) using genomic and transcriptomic data. These genes were chosen based on their association with various physiological functions, including regulation of immune responses. Wenzel and Piertney (2015) then used both population- and individual-based statistical approaches to relate variation in these genes to gastrointestinal nematode burden of the sampled birds. The various population genomic approaches identified only few associations of genetic variation with nematode burden, and these associations varied strongly among different statistical methods. In contrast, the individual-based analysis clearly identified signatures of natural selection, with nine of the 12 tested loci showing significant associations with parasite load. While this study focused on a phenotypic trait (i.e., nematode burden) and did not relate variation in candidate genes to environmental heterogeneity, it accounted for landscape-level management actions and spatial sampling design in the statistical analysis. Importantly, the study demonstrates that landscape genomic approaches where allelic variation of individuals is statistically linked to environmental data have higher power to detect loci under selection compared to classical population genomic approaches.

Detecting Selection-Driven Loci and Environmental Associations in Dall's Sheep

The study by Wenzel and Piertney (2015) took advantage of previously developed candidate genes in a species that is relatively well researched under both natural and experimental settings. When such candidate genes are not available for the study species, information from other closely related species can be helpful. For example, Roffler et al. (2016a) used targeted exon capture to discover SNPs in Dall's sheep (*Ovis dalli dalli*) using available sequences from candidate genes in a closely related wild species (bighorn sheep, *Ovis canadensis*) and the domestic sheep (*Ovis aries*) genome. They used the discovered SNPs to genotype 476 Dall's sheep from across their range and applied two outlier tests and one EAA approach to detect signatures of selection. Across the three statistical methods, nine genes were identified as selection-driven, and they all were significantly correlated with precipitation, temperature, latitude, longitude, and elevation. These results indicate adaptation to local environmental conditions, especially because five of the selection-driven candidate genes are associated with immune responses and respiratory health, respectively, in the species in which they were originally discovered. The study by Roffler et al. (2016a) is an excellent example of how to maximize the interpretability and eco-evolutionary meaningfulness of adaptive landscape genomic studies in non-model wildlife species, for which whole genome data is not yet available.

Disentangling Processes Causing Genomic Differentiation in Islands Foxes

When genomic information for closely related (model) species is not available, high-throughput sequencing technology can be used to sample genomic variation across a large portion of the genome. For example, Funk et al. (2016) used RADseq to develop 5,293 SNP loci for the island fox (*Urocyon littoralis*). They then used this genomic data to test whether differentiation among six island fox populations off the coast of southern California was mainly explained by drift or local adaptation. Using a combination of statistical methods, Funk et al. (2016) concluded that overall genomic differentiation among the six islands was largely explained by strong drift. However, different outlier tests suggested between 3.3 and 6.6% of the loci to be under selection. While none of the utilized EAA approaches could detect any significant associations with these loci and environmental data, genomic differentiation estimated from the outlier loci matched patterns of morphological similarity among the sampled populations. These results suggest that despite strong drift, divergence due to local adaptation explains at least some of the selection-driven genomic variation, and that the tested environmental data do not reflect heterogeneity in underlying selection pressures.

These examples show that adaptive landscape genomics can help to advance our understanding of selection and local adaptation in non-model terrestrial wildlife species. Two particularly interesting research topics that can be addressed using landscape genomics relate to the evolutionary processes underlying IBE, and to range-expanding species in changing environments. We discuss these two topics in the next sections.

5.2 *Using Landscape Genomics to Study IBE and Underlying Evolutionary Processes*

Patterns of IBE can result from many different processes, both selective and neutral (Wang and Bradburd 2014). Divergent selection is frequently invoked and may be the primary mechanism driving IBE (Kawecki and Ebert 2004; Nosil et al. 2005). When populations inhabiting different environments are locally adapted, natural or sexual selection can act against dispersers with phenotypes adapted to a different environment, limiting the reproductive success of dispersers moving between habitats (Servedio 2004; Nosil 2004; Nosil et al. 2005; Safran et al. 2013). We can expect the strength of selection to be proportional to the difference between the environment to which an individual is adapted and the environment to which it has dispersed. Thus, nonnative individuals will experience an increasing reduction in fitness, relative to native individuals, when dispersing into increasingly different environments. This, naturally, reduces gene flow between divergent environments and leads to a pattern in which genetic differentiation increases with the environmental differences between populations (Sexton et al. 2014; Wang and Bradburd

2014). Moreover, when selection is generally weak or incomplete, dispersers that are reproductively successful in a new environment may produce offspring with traits, or combinations of traits, that are not ideally suited to the local environment. This can result in these offspring having reduced fitness compared to offspring from native parents and will lead to a further reduction in long-term gene flow between divergent environments (Servedio 2004; Nosil et al. 2005). The agents of selection that act upon adult dispersers and offspring of native and nonnative parents may be the same, but could also be very different – for instance, if different agents of selection act at different times of year or on different life history stages (Wang and Bradburd 2014). Finally, even when selection is absent, biased dispersal, in which different individuals or populations have a preference for dispersal to different habitats, can also produce a pattern of IBE (Edelaar et al. 2008; Bolnick et al. 2009; Edelaar and Bolnick 2012). Though biased dispersal can be linked to divergent selection – for instance, if a trait associated with biased dispersal provides a fitness advantage in a particular environment – but it need not – for instance, when dispersers avoid novel habitat or have a preference for their natal habitat (Davis and Stamps 2004; Feder and Forbes 2007; Rosenblum and Harmon 2011; Bolnick and Otto 2013).

Under any of these scenarios, studying IBE can be a gateway to investigating how evolutionary processes play out over a landscape. The various selective processes that can generate IBE form natural links to examining divergent natural and sexual selection, and how they drive microevolutionary responses (Lee and Mitchell-Olds 2011; Sexton et al. 2014). What the agents of selection acting on spatial genetic variation are and how selective agents and the strength of selection vary across space are sure to feature prominently in future landscape genomics work (Wang and Bradburd 2014). Even when the mechanisms underlying IBE are not selective in nature, interesting questions about what processes lead to the evolution of biased dispersal or that lead to divergent habitat preferences in different populations can still be asked (Davis and Stamps 2004; Bolnick et al. 2009). There are now a number of rigorous empirical studies that have investigated IBD and IBE to examine how evolutionary processes play out over a landscape and influence the evolution of genetic and phenotypic diversity (Sexton et al. 2014). These studies have been performed in diverse taxa, and there are no particular study organisms in which these studies are more valuable than others. However, several of the earliest studies that explicitly considered IBE were performed in lizards (*Anolis* spp.), and a nice set of empirical studies that were conducted on lizard species in different parts of the world exists and presents excellent examples of how landscape genomics can be used to investigate evolutionary processes.

To examine whether the ecological and evolutionary processes across landscapes will generate similar genetic patterns in closely related species, Wang et al. (2013) performed a comparative landscape genetic analysis of 17 species of *Anolis* lizards from the Greater Antilles using structural equation modeling (SEM). As a form of latent variable modeling, SEM allowed them to infer the contributions of individual environmental variables to IBE without any a priori knowledge or expectation for how they should be weighted. *Anolis* lizards on the Greater Antilles

evolved through repeated adaptive radiations, in which species diversified to fill available environmental niche space. There are now species on each island with traits adapted to particular parts of the vertical habitat structure, and together these species, with convergent traits that have evolved for the microhabitat in which they are found, compose what are called ecomorph classes. The adaptive radiations and ecomorph evolution mean that multiple congeneric species are now found in sympatry, presenting an excellent opportunity to examine how the same landscape affects genetic diversity in different species. Landscape heterogeneity on each island is quite diverse, encompassing a range of ecosystems and environmental clines in temperature and precipitation that often form xeric to mesic habitat gradients. Wang et al. (2013) sampled an average of 21 populations per species, spanning a wide variety of the environmental conditions on each of the four Greater Antillean islands.

The results of their SEM analysis revealed a pattern that was fairly consistent across most species in the study, in which IBD explained 36.3%, and IBE explained 17.9% of the variance in genetic distances between populations (Wang et al. 2013). So, overall, the geographic distance between populations had about twice as much of an effect on genetic divergence as the differences in their environments did. This result was quite consistent between the species within islands, suggesting that congeneric species experience the landscape in similar ways, but there were some distinct differences between islands. For instance, the species on Hispaniola showed a stronger pattern of IBE than IBD, while the species on Jamaica showed only minimal signals of IBE (Fig. 6).

The SEM analysis also found that temperature gradients were the primary drivers of IBE, while precipitation gradients also contributed strongly in some species (Wang et al. 2013). Altogether, these results suggest that, in addition to geographic isolation, local adaptation or biased dispersal also played an important

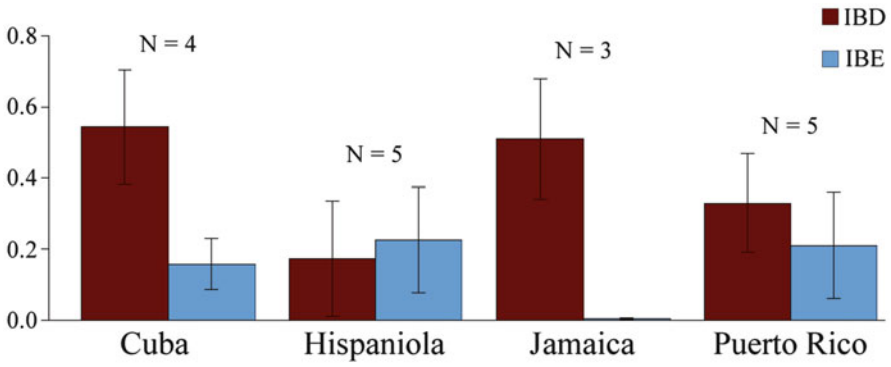

Fig. 6 The proportion of variance in genetic distances explained by isolation-by-distance (IBD) and isolation-by-environment (IBE) among populations of *Anolis* lizards from the four Greater Antillean islands. For each island, the mean estimates of IBD (red) and IBE (blue) for each study species inhabiting the island are presented, error bars represent one standard deviation from the mean (modified from Wang et al. 2013)

role in population genetic divergence in the Greater Antillean *Anolis* lizards, and that ecological and evolutionary processes on landscape (and species responses to landscape composition and configuration) can affect related species in similar ways, resulting in similar evolutionary outcomes.

To test evolutionary hypotheses about the factors that generate population divergence and biodiversity in the tropics, Freedman et al. (2010) investigated divergence and selection along ecological gradients in a species of rainforest skink (*Trachylepis affinis*) in Cameroon using a set of genome-wide amplified fragment length polymorphisms (AFLPs) and generalized dissimilarity modeling, a form of nonlinear regression modeling. Their study landscape included multiple environmental gradients, including habitat clines from lowland to montane forest and a forest-savanna ecotone, and it also encompassed several regions that formed refuges for skink populations during the last glacial maximum. The goal of the study was to test whether diversification occurred across ecological gradients (the ecological speciation/diversification hypothesis, Smith et al. 1997) or out of glacial refugia (the Pleistocene forest refuge hypothesis, Mayr and O'Hara 1986; Moritz et al. 2000).

The results of their GDM analysis identified that neutral genetic differentiation occurred primarily along the forest-savanna ecotone and found evidence of both IBD and IBE, which was primarily associated with variation in precipitation (Freedman et al. 2010). They also found evidence for a set of loci under divergent selection along the forest-savanna ecotone, which matched a pattern of morphological divergence in fitness-related traits along the same ecotone. Two other sets of loci, also bearing signatures of divergent selection, were significantly differentiated across the lowland-montane forest gradient and between glacial refugia. Thus, although divergent selection was detected between lowland and montane forests and between glacial refugia, it does not appear to inhibit gene flow between these different environments. Whether selection and a reduction in gene flow are sufficient to initiate reproductive isolation remains unclear, but the strong evidence for divergent selection on adaptive loci across the forest-savanna ecotone associated with greater genome-wide neutral genetic divergence provides stronger support for the ecological diversification hypothesis than for the Pleistocene forest refuge hypothesis (Freedman et al. 2010).

To investigate the microevolutionary processes that govern population genetic and morphological divergence, Barley et al. (2015) studied IBD and IBE in sun skinks (*Eutropis multifasciata*) in Southeastern Asia using a panel of genome-wide SNPs and Bayesian geostatistical modeling (Bradburd et al. 2013). Skinks, as a clade, express a great deal of morphological conservatism across different species, with considerable cryptic diversity and many species showing only a few distinguishable differences. So, Barley et al. (2015) also measured a suite to morphological traits for all of the individuals they sampled to see if this general pattern was found within species as well. They collected specimens from 20 populations spread out on a landscape encompassing highly heterogeneous habitat on mainland Southeast Asia and the nearby Philippine Islands Archipelago. Thus, some regions of the study area contained relatively contiguous stretches of suitable habitat, while others

were highly fragmented. In addition, both the islands and the mainland housed several broad climate gradients often associated with local topography.

The results of the above skink study showed that, as expected, the more isolated populations in fragmented areas showed higher levels of genetic differentiation overall. The skink populations showed a clear pattern of IBD but no discernible signal of IBE, and thus it appears that divergent selection or biased dispersal plays little, if any, role in generating overall genomic divergence in this species (Barley et al. 2015). Comparing the phenotypic divergence between populations (P_{ST}) to their genetic divergence (F_{ST}) revealed that populations were less morphologically differentiated than expected based on their overall levels of genetic isolation ($P_{ST} < F_{ST}$). Thus, some selective mechanism is needed to explain why the populations have not diverged morphologically to the extent expected purely from neutral population divergence. One such mechanism, as presented by the authors, is stabilizing selection on the measured morphological traits, which effectively constrains phenotypic divergence in this system (Barley et al. 2015). Hence, this study illustrates that even when no pattern of IBE is detected, landscape genomic studies that explicitly account for environmental heterogeneity can still reveal the evolutionary processes driving the distribution of genetic and phenotypic variation on a landscape. In this case, it's actually stabilizing selection, rather than divergent selection, that governs morphological evolution in sun skinks, which appears broadly consistent with the evolutionary processes that control morphological evolution between species of skinks as well (Barley et al. 2015).

Future work investigating IBE is poised to take advantage of new genomic resources and the declining cost of genome sequencing technologies. An exciting area of development will be characterizing how spatial patterns, including IBD and IBE, vary among different sites across the genome. We already know that genomic divergence can be highly heterogeneous (Nosil et al. 2008, 2009), because some evolutionary forces act on the entire genome, while others are highly localized (Nosil et al. 2009; Turner and Hahn 2010; Flaxman et al. 2013). New studies can investigate why some regions of the genome show greater IBE than others, and which environmental factors contribute to these patterns in different loci (Wang and Bradburd 2014). We generally expect that loci associated with adaptive traits will show elevated IBE when populations are locally adapted, but adaptive loci could also show less IBE than the genomic background when they are advantageous across different environments. For instance, Fountain et al. (2016) found evidence for positive selection on loci underlying traits associated with dispersal ability in populations of Glanville fritillary butterflies inhabiting fragmented landscapes, and, intuitively, loci linked to traits like dispersal ability are among those expected to show different patterns of IBE. Hence, many opportunities now exist for increasing our understanding of the diverse evolutionary processes, often linked to spatial environmental variation, that act heterogeneously across the genome and for better understanding the processes driving genome evolution in general.

5.3 *Landscape Genomics of Range-Expanding Species Under Changing Climate Conditions*

As climate change proceeds, the evolutionary processes that govern species range shifts and expansions are becoming increasingly under focus. Landscape genomics approaches, in combination with modeling, offer unmatched tools for examining these processes in changing environments. However, the challenges of detecting patterns of adaptive variation along environmental gradients occupied by range shifting or invasive species are not trivial. During a range expansion, allelic richness and heterozygosity may decline along the axis of the expansion due to a series of founder events and stochastic allele loss (White et al. 2013). Since this can generate genetic drift in the same direction of species' colonization, allele frequencies can be driven to fixation, making it problematic to distinguish neutral from adaptive genetic signatures (Klopfstein et al. 2006; Frichot et al. 2015). This phenomenon of "allele surfing," whereby rare alleles become more frequent at range expansion fronts according to the strength of genetic drift rather than selection, can increase population genetic differentiation and confound signatures of local adaptation (Klopfstein et al. 2006). Notably allele surfing may also promote adaptation as well when beneficial alleles are "surfed" on the wave of expansion (Gralka et al. 2016), but deleterious alleles may also be "surfed" at the range expansion front (Travis et al. 2007).

Considering the effects of genetic drift in EAA analysis is particularly important given that environmental variation typically corresponds with latitudinal or altitudinal gradients from which genetic samples are obtained. Therefore, both environmental gradients and expansion axes are often aligned when species are tracking their environmental niches during expansion (Frichot et al. 2015; Lancaster 2016). Not accounting for genetic drift may increase the likelihood of erroneous EAAs because neutral allele frequencies can behave similarly to those under environmental selection. As mentioned above, spatial analysis of allele frequencies across environmental gradients can be used to tease apart genetic drift from selection processes. Local Indicators of Spatial Association (LISA) analysis is mentioned above and addresses the effect of spatial autocorrelation on allele frequencies (e.g., Stucki et al. 2016). Another approach for applying landscape genomics to range expanding species is the use of GDM/GF analysis mentioned earlier. The GDM/GF approach may be applied to any system but it is particularly useful for range expanding species as it allows the effects of geography and neutral processes to be filtered out (Fitzpatrick and Keller 2015). The latter can be accounted for by integrating a pre-identified set of neutral genes into the model and evaluating their contribution relative to putative candidate genes under selection via their allele frequency response curve (termed "allelic turnover") across environmental gradients (Fitzpatrick and Keller 2015). For example, a putatively adaptive gene in a temperature-sensitive species (e.g., a butterfly) may show a twofold allelic turnover at 18°C along a temperature gradient, but if allelic turnover of the neutral "reference" genes shows an identically shaped fourfold response at 18°C, our confidence

in this gene being adaptive becomes diminished. Alternatively, if the neutral allelic turnover shows a 0.5-fold change with a response curve that differed in shape to the adaptive candidate genes, this would offer greater support that the relationship we see in the candidate gene is adaptive to the environmental gradient. This method offers a useful approach for identifying thresholds of adaptation in species affected by climate change and diverse environments, such as those undergoing range shifts and expansions. The ability to identify thresholds of adaptation can aid in making predictions about species distributions, climate sensitivity, and persistence.

In addition to environmental adaptation, range expanding and invasive species often exhibit differential morphological or developmental traits along their range (e.g., Ducatez et al. 2016), but the evolutionary bases of these traits are rarely teased apart from environmental effects (but see Buckley et al. 2012; Swaegers et al. 2015). Many traits under selection may be highly polygenic (i.e., many small-effect loci under weak selection) and detecting loci under selection may be difficult in these cases. Notably, EAA analyses are likely to be more sensitive at detecting multiple loci under weak selection compared with F_{ST} outlier tests (Frichot et al. 2013, 2015), which are best used for detecting large-effect loci under strong selection (Whitlock and Lotterhos 2015). Given the multiple selection pressures exerted upon species from both abiotic and biotic sources during range expansion, teasing apart which loci are associated with environmental or phenotypic variables is likely to result in many overlapping and correlated loci of both small and large effects.

In the absence of a reference genome or candidate genes, EAA enables identification of loci that may be involved in local adaptation along environmental gradients occupied by range expanding species. There are a few examples to date that document increased signatures of selection for adaptive loci at range expanding edges. For example, a study on European damselflies (*Coenagrion scitulum*, Swaegers et al. 2015) documents parallel, non-neutral evolutionary changes in allele frequencies within independent expanding edge populations with respect to flight performance and thermal regime. In addition, evidence for genetically determined phenotypic differences was obtained along the range expansion (Swaegers et al. 2015), which is an important step for teasing apart changes resulting from heritable genetic variation versus trait plasticity (Merilä and Hendry 2014). An earlier study investigated a butterfly range expansion with AFLP (Amplified Fragment Length Polymorphisms) genetic markers using an "Isolation by Adaptation" (IBA) approach based on partial Mantel tests (Nosil et al. 2008), and found significant associations among allele frequencies with habitat type, independent of colonization history (Buckley et al. 2012). Furthermore, a study examining adaptive evolution in range expanding bank voles (*Myodes glareolus*) found a loss of genetic diversity (with SNPs) towards the range margin due to genetic drift, no increase in deleterious alleles, but an increase in outlier loci that coded for functional genes, suggesting enhanced selection (White et al. 2013). Lower genetic diversity and signatures of natural selection were also detected in marginal populations of eastern white pine (Chhatre and Rajora 2014).

Although more complex models can now be used to examine range expansion effects on local adaptation (e.g., Schumaker 2013; Landguth et al. 2016) empirical studies using EAA methods are still relatively few, but offer a promising approach for validating simulation findings and ultimately understanding how species are spatially responding to climate change. By identifying the spatial distribution of adaptive variation in species that are shifting their ranges, we may be better able to manage for species' current and potential habitat, novel species' interactions, the spread of invasive species, or the diseases that expanding species carry into new areas.

5.4 *Seascape Genomics: Applying Landscape Genomics in Aquatic Environments*

Seascape genomics is a natural extension of landscape genetic approaches, with great relevance to conservation and management of marine species (Gagnaire et al. 2015; Selkoe et al. 2016a, b; Kelly and Phillips 2016; Riginos et al. 2016). However, there are distinct differences between marine and terrestrial settings that affect the spatiotemporal distribution of species and genes. Solid barriers to movement are rare in the sea. While the fluid environment might seem like a vast shapeless surface promoting homogeneous, diffusive spread of migrants, in fact, strong currents create asymmetrical and circuitous pathways that channel drifting particles and counteract diffusion. Mapping dispersal corridors and quantifying dispersal resistance differ dramatically between landscapes and seascapes. Application of resistance modeling (e.g., McRae et al. 2008; Spear et al. 2016) has been stymied by the unidirectional nature of marine dispersal by ocean currents. Most marine species disperse during a tiny larval stage that can last from minutes to months, potentially taking them hundreds of kilometers in the currents before metamorphosis into a sedentary or mobile adult. Successfully reaching suitable adult habitat may require careful timing of reproduction for spawned larvae to exploit countercurrents, upwelling cycles, and gyres that may return offspring to natal habitat after forays into open ocean currents, and also active behaviors to counteract passive drifting (Paris et al. 2007; Morgan 2014). Thus, understanding functional connectivity and not just structural connectivity associated with currents is critical for seascape genetics (Selkoe et al. 2016a). There is strong interest in testing the link between ocean currents and gene flow by comparing outputs of ocean circulation models and genetic data, and in fact, disagreement of outputs is not uncommon (e.g., White et al. 2010; Selkoe and Toonen 2011).

Aside from current flows, sharp gradients in the ocean's temperature, salinity, and oxygen also act as obstacles to successful dispersal and gene flow, even in larger bodied species such as fishes (Caldwell and Gergel 2013). For the American lobster, *Homarus americanus*, two disparate current systems contribute to the neutral genetic divergence of lobster populations, which is further cemented by

local adaptation to temperature (Benestan et al. 2016). Temperature is especially critical to marine ecosystems (Bowen et al. 2016), acting at macro scales to define distributions of marine biodiversity (Belanger et al. 2012), as well as micro scales to, e.g., set timing of spawning (Afán et al. 2015). Due to the combination of large species ranges and exothermic physiology, spatial temperature differences – a.k.a., the "therma-scape," – appear to very commonly shape marine population genetic structure at both neutral and adaptive markers (Conover et al. 2006; Selkoe et al. 2016a).

Marine study systems have an important role to play in uncovering the dynamics of local adaptation in the face of high dispersal (Hauser and Carvalho 2008). Local adaptation may be especially apparent in marine species due to their large effective population sizes which lead to low rates of drift and higher rates of weak selection, selective sweeps, and diversifying selection (Nielsen et al. 2009). Large effective population size also leads to long-lasting genome-wide impacts of founder effects following colonization (Orsini et al. 2013). Seascape genomics studies hold promise to uncover drivers of fine-scale genetic divergence, and are rapidly overturning long-standing beliefs that marine populations are genetically homogeneous over large scales, even in cases of highly mobile species (Gaither et al. 2016). These advances collectively provide powerful insights for conservation and management strategies (Bradbury et al. 2013b; von der Heyden et al. 2014; Gagnaire et al. 2015; Riginos et al. 2016).

Most marine genomic studies utilize SNP discovery in conjunction with outlier tests to detect islands of genomic differentiation and the environmental drivers leading to ecological differentiation rather than allopatric speciation. The low genetic differentiation common to marine populations increases power to detect "true" outliers if the population is at drift-migration equilibrium. However, it can lead to high false positive rates if the population is not at equilibrium. These nonequilibrium situations may be common in nearshore environments due to widespread impacts of past sea level fluctuation, glaciation, or recolonization following stochastic events such as storm surges (Marko and Hart 2011). Increasing the "q-value threshold" in outlier detection with BayScan or similar methods can reduce false positives. For example, BayScEnv incorporates a locus-specific term to account for nonequilibrium effects (de Villemereuil and Gaggiotti 2015), which is needed to minimize false inference of selection, which can ultimately impact management applications such as delineation of fishery stocks. Gagnaire et al. (2015) detail methods well suited to detecting low rates of spatial genetic differentiation and revealing the role of clines in marine connectivity, such as metrics of haplotype sharing and focusing on rare alleles, or "migrant tracts" of DNA segments that resist recombination after admixture.

The handful of seascape genomics studies published to date focused primarily on species with high-value fisheries, and overall there is a strong bias toward temperate systems compared to polar and tropical zones (Selkoe et al. 2016a). Nevertheless, the first crop of seascape genomic studies provides fascinating insights into the interaction between the genome and environment (Riginos et al. 2016). Here, we

highlight a few empirical examples that speak to the complexity of how environment shapes population genetics in a dynamic ocean setting.

The three-dimensionality of the seascape requires careful measurement of an organism's environmental influences. For example, a study on Atlantic cod, *Gadus morhua*, revealed that salinity and oxygen at spawning depth, rather than at the sea surface, best explained outlier loci (Berg et al. 2015). The complex life histories of marine species often factor into which seascape factors drive population structure. In the cod example, some outlier loci were associated with genes involved in egg buoyancy (which impacts a developing egg's exposure to high salinity surface waters), highlighting the critical role of early life stages in shaping local adaptation. Focusing on single life history stages will miss some of the more subtle processes affecting overall structure. Von der Heyden et al. (2007) show differentially structured populations of adult and juvenile Cape hakes (*Merluccius paradoxus*), suggesting different environmental factors influence the spatial genetics of each stage. Marine populations show not only temporal genetic shifts across age classes, but also responses to dramatic inter-annual and decadal-scale changes in the ocean environment. Sampling at single time points can miss these ephemeral dynamics that can produce lasting signatures of selection or isolation. For example, Henriques et al. (2016) used three successive years of sampling of the shelf-associated hake, *Merluccius capensis*, that in years with increased upwelling and associated low oxygen water events show a distinct movement of fish from the northern range southwards associated with physiological tolerance to hypoxia. The ephemeral pattern in these anomalous years disrupts an otherwise stable barrier limiting gene flow between regions.

The future success of seascape genetics will be shaped, in part, by overcoming obstacles to de novo genome assembly. For example, high levels of heterozygosity must often be bred out of organisms in captivity, which is often not possible for marine species. As of 2015, only 18 genomes of marine species had been assembled, compared to ~70 for terrestrial species (Kelley et al. 2016). The power of whole genome assembly and annotation promises to provide deeper insights into the function and interactions of gene regions in local adaptation (Hemmer-Hansen et al. 2014). The stickleback fish was an early target of whole genome assembly due to interest in understanding recurrent marine–freshwater evolution by the lineage (Jones et al. 2012). Strong signatures of directional selection were found on every chromosome for the three-spined stickleback (*Gasterosteus aculeatus*), with many loci linked to strong salinity and temperature gradients over the sampling domain (Guo et al. 2015). In contrast, the Atlantic cod genome shows SNP outliers to be highly clustered within three chromosomes, where chromosomal inversions led to "islands of divergence" within the genome (Bradbury et al. 2013b). Despite very high neutral gene flow, the repression of recombination in the inverted regions has enabled oceanic and coastal cod population to adapt to local oxygen, temperature, and salinity regimes throughout its species range (Sodeland et al. 2016).

Seascape genomics have thus far highlighted an array of factors influencing the spatial distribution of species and genes in marine systems. However, the majority of factors tested are abiotic, thereby limiting our understanding of the ecological

context of species distributions. A major aim of future studies should be the incorporation of ecological factors, such as species and community-level interactions, to begin to elucidate this poorly characterized component shaping marine biodiversity and future responses to climate change.

6 Remaining Challenges and Future Research Avenues in Landscape Genomics

As shown throughout this chapter, the potential of landscape genomics for fundamental and applied research is substantial. The increasing interest in the field has led to the rapid development of a large variety of analytical approaches for assessing landscape-genomics influences. This variety is a challenge in several ways. First, it makes it difficult to obtain an overview of landscape genomics as a beginner, to choose among the available methods for analyzing empirical data sets, and to keep up-to-date as a more experienced landscape geneticist. There will never be a single analytical approach that is optimal for addressing all types of landscape genomics research questions, and interdisciplinary collaboration in research and teaching will continue to be a cornerstone for progress in landscape genomics. Nevertheless, more studies are needed that identify those methods that work particularly well or particularly poorly for different systems and questions, and provide practical advice on how to conduct both neutral and adaptive landscape genomics studies for specific research questions. Simulation studies are an important way of testing methods (Hoban et al. 2012), and they have already provided valuable assessments of different methods and sampling designs in landscape genomics (see Landguth et al. 2016).

Second, the variety of approaches, with their different assumptions, advantages, and limitations, also makes it challenging to synthesize results obtained from landscape genomics studies. The choice of analytical methods, and the choice of landscape-genomic hypotheses tested with them, can strongly influence conclusions of a study (Balkenhol et al. 2009; Jaquiery et al. 2011). Again, simulation studies are an excellent way for testing the reliability of drawn conclusions (e.g., Gauffre et al. 2008), and to assess in how far results of a specific study can be extrapolated to other study systems (e.g., other landscapes or species). In addition to simulations, several studies have suggested that using more than one method for final inferences can increase reliability and certainty in landscape genomics (e.g., Balkenhol et al. 2009; Rellstab et al. 2015; Rajora et al. 2016).

Third, due to the remaining methodological issues in landscape genomics, too little progress has been made on the theoretical and conceptual development of landscape genomics. Currently, landscape genomics is often viewed merely as a set of tools for statistically linking environmental and genomic data, and one can rightfully question whether this justifies the definition of a distinct scientific field. For instance, Dyer (2015a) showed that most studies using the term "landscape genetics" can actually be defined as a highly nonuniform set of population genetic

studies, and that the scope of the so-called landscape genetic studies is still very limited. Dyer (2015a) argued that to develop landscape genomics into a distinct field, much more emphasis should be placed on theory development that links individual- and population-based patterns and processes, and that amalgamates ecology and evolution in a truly interdisciplinary way. This view is supported by Balkenhol et al. (2016b, c), who called for a shift in landscape genomics from the current, statistical, and pattern-oriented framework towards an eco-evolutionary and process-oriented framework. Developing the theory underlying such a framework is a major research task, but will be vital for understanding how and why genetic variation is influenced by environmental heterogeneity across spatial and temporal scales.

Another challenge of current landscape genomic studies is adequate study design and sampling. Too often, genetic and genomic data are gathered for other research purposes, and landscape genomic questions are only considered after sampling is finished. This can impede our ability to draw strong scientific inferences about landscape-genomic relationships, and targeted sampling should instead be preferred (e.g., Storfer et al. 2007; Manel et al. 2010). Specifically, deriving hypotheses or expectation about how the environment potentially influences genetic variation (Fig. 2) *before* sampling can greatly enhance our power to detect these influences, if they indeed exist (Balkenhol and Fortin 2016).

Apart from spatial considerations in sampling design, one must also consider the genomic sequencing strategy employed and the genomic resources available for the study organism. Reduced-representation sequencing methods like Genotyping by Sequencing (GBS), Restriction-Site Associated DNA sequencing (RADseq: Miller et al. 2007), and RNA sequencing are popular methods for obtaining many thousands of single nucleotide polymorphism markers (SNPs) that may be neutral or adaptive (Narum et al. 2013). However, the number of markers obtained and hence the ability to detect genes under selection from such sequencing methods may be influenced by library preparation method, density of SNPs according to genome size, the bioinformatics parameters applied to SNP filtering, and for identifying gene function, the existence of an annotated reference genome or transcriptome (discussed in Lowry et al. 2017). However, the resources available for most projects are well below those required for whole genome sequencing of every sample, making reduced-representation sequencing an appropriate and informative choice for the objectives of most landscape genomics projects.

6.1 Future Research in Landscape Genomics

What will the future of landscape genomics likely hold? First, we envisage that in the next years, the variety and complexity of analytical approaches will increase even further. After all, the increasing availability of genomic data will also lead to new approaches and statistical methods for analyzing them. As stated above, comparing and evaluating existing as well as novel methods are in high demand

for the applicability of landscape genomics. We are hopeful that after a phase of rapid development, landscape genomics will eventually identify a set of methods that work particularly well for specific questions, while other, problematic methods will disappear from the field. For example, several very helpful studies already exist that compare different approaches for identifying adaptive genetic variation in landscape genomics (e.g., Rellstab et al. 2015; Forester et al. 2016; Hoban et al. 2016; Rajora et al. 2016) and that highlight issues with some of the more prominent laboratory techniques used in these approaches (e.g., RADseq, Lowry et al. 2017).

In addition to reducing methodological issues, the future of landscape genomics will likely be characterized by an expanded research scope that will include additional concepts, data types, and processes. Specifically, we see great potential for future research avenues in landscape genomics along the following topics.

6.2 Landscape Genomics and Nongenetic Data

In the future, increasing effort will likely be given to amalgamating landscape genomics with other research approaches producing nongenetic data. For instance, habitat models derived from presence-absence or occurrence data are often used to parameterize resistance surfaces in landscape genomics (e.g., Wang et al. 2008; Engler et al. 2014). However, recent landscape genomic studies have shown that habitat models do not always adequately capture landscape influences on effective dispersal and resulting genetic structures (Mateo-Sanchéz et al. 2015; Roffler et al. 2016b), because habitats are used differently during dispersal compared to other behaviors, such as foraging (e.g., Benz et al. 2016; Ziólkowska et al. 2016; Abrahms et al. 2017). A better option for parameterizing resistance surfaces in animals might hence be actual movement data, which can be gathered at increasingly fine spatial and temporal resolutions using satellite-telemetry. A variety of methods exist for distinguishing different behaviors within individual movement paths (reviewed in Edelhoff et al. 2016) and for quantifying how dispersal behavior is influenced by environmental heterogeneity (e.g., Cushman and Lewis 2010). Similarly, landscape genomic data can be combined with demographic estimates of population size, survival, or fecundity to understand the interplay between local demography and population genomics in heterogeneous environments. Addressing the complex questions involving demography and genomics can already be accomplished via simulations, for example in software CDMetaPOP (Landguth et al. 2017).

6.3 Landscape Genomics and Eco-Evolutionary Dynamics

Collecting demographic data in combination with genomic data is also important for investigating eco-evolutionary dynamics, which refers to tight feedback mechanisms between ecological and evolutionary processes (Pelletier et al. 2009;

Legrand et al. 2016). An increasing number of studies shows that these feedbacks can occur across ecological timescales, meaning that evolution acts fast enough to influence the ecology of organisms within a few generations (e.g., Fronhofer and Altermatt 2015; DeLong et al. 2016). Since a major goal of landscape genomics is to understand and eventually predict the consequences of ongoing contemporary environmental change for genetic variation, future studies should consider eco-evolutionary dynamics and assess how the potential for adaptive evolution impacts the genetic response of populations of species in changing environments. For this, it will be particularly vital to consider the polygenic architecture of many traits that influence the survival and reproductive success of individuals. Understanding environmental selection pressures on these traits and their underlying genes will require landscape genomic approaches that statistically link multi-locus variation to environmental heterogeneity (e.g., Rajora et al. 2016). Most EAA approaches in landscape genomics use only a single locus at a time, but several of the methods discussed in Sect. 4.1.1 can also test for multi-locus signatures of selection. Forester et al. (2017) compared the reliability of several EAA approaches for detecting selection acting on multiple loci and demonstrated substantial differences among methods. These differences likely also depend on underlying demographic history and sampling design, and future studies are needed to clarify how environmental effects on polygenic selection can best be detected.

6.4 Landscape Community Genomics

Recently, Hand et al. (2015) suggested "landscape community genomics (LCG)" as a framework for assessing the eco-evolutionary responses of multiple species in complex and dynamic environments. The sampling design for LCG requires landscape genomic studies to be conducted for at least two interacting species (i.e., a community) in multiple, diverse study landscapes. LCG also requires the combination of adaptive and neutral landscape genomics that we have advocated throughout the chapter, but with several additional benefits. First, analyzing multiple species within the same landscapes makes it possible to assess the generality of findings and hence can help to synthesize results across species and landscapes (e.g., Dudaniec et al. 2016). Second, LCG can account for biotic interactions among species, such as competition or coevolution. These biotic interactions can influence spatial genetic variation, but are seldom considered in current landscape genomic studies. Third, LCG explicitly considers the interaction between biotic and abiotic (i.e., environmental) factors shaping genetic variation, thus potentially resulting in the most thorough picture of how genetic variation is shaped in nature. Finally, LCG can ultimately also help to evaluate how environmental impacts on community genetics will alter ecosystem properties, because important factors shaping ecosystems are indirectly impacted by genetic variation (e.g., the distribution, abundance, structure, demography, and interaction of coexisting populations;

Whitham et al. 2006). In sum, LCG can be seen as the most advanced type of landscape genomic study, and the framework will likely lead to important insights on eco-evolutionary dynamics in heterogeneous and changing environments.

6.5 Application of Landscape Genomics in Conservation Management

Clearly, landscape genomics has great potential for conservation management. However, while the gap between conservation practitioners and population geneticists is already an issue for any genetic study, the uptake by conservation managers from even more complex genomic studies is even more challenging (Hoffmann et al. 2015; Shafer et al. 2015). There is indeed still a large gap between scientists working within the field of genetics and those dealing with conservation problems on the ground. Several studies identified this science policy gap (e.g., Hoban et al. 2013; Taylor and Soanes 2016) which is getting even more prominent when it comes to the field of genomics (Shafer et al. 2015). However, for many of the pressing conservation topics, genomic tools may be able to get us further than using classic neutral genetic approaches. For the field of landscape genomics, this is specifically true when adaptive genetic variation could be used for inferring the potential of local populations to changing environments. However, until now in conservation contexts this has been rarely used so far and then mostly focusing on population correlations or single candidate gene approaches (Shafer et al. 2015). While there is a strong need from the conservation community to better evaluate if species or populations may be able to adapt to certain environmental conditions there is still quite some uncertainty involved in how the results of landscape genomic analyses should actually be interpreted. While more and more adaptive loci can be identified, methods are still developing fast in screening such loci and in analyzing large amounts of data. This calls for validation studies, multispecies approaches, and also common garden experiments to actually demonstrate that our results are of practical relevance.

7 Conclusion

To conclude, landscape genomics provides a complex but powerful framework for addressing fundamental and applied research questions in many different fields. As discussed throughout this chapter, concepts and methods in the field advance rapidly. On the one hand, this makes it challenging to establish and maintain a thorough overview of newest developments and to discern subtle analytical nuances from crucial improvements. On the other hand, the large diversity of neutral and

adaptive landscape genomic approaches associated with the swift progress in the field makes it particularly vibrant and exciting.

Soon we will begin to see whole genome comparisons across heterogeneous landscapes allowing researchers to identify a broad array of ecological and climatic factors influencing neutral and adaptive processes. In the future it will be important to move beyond an assessment of population genetic structure in single species to infer potential future responses to climate and landscape change and to begin to predict how communities of species will respond based on our knowledge of adaptive capacity (Holderegger et al. 2010; Storfer et al. 2010; Neale and Kremer 2011; Manel and Holderegger 2013; Sork et al. 2013). Finding general responses across multiple species and further assessing multilocus effects will continue to be important goals for future landscape genomic studies (Calic et al. 2016; Rajora et al. 2016).

We are convinced that we have only just begun to realize the potential of landscape genomics, but as highlighted above, there is much room – and need – not only for methodological, but also for conceptual and theoretical improvement in landscape genomics (see also Dyer 2015a; Balkenhol et al. 2016c). Hence, we are curious to see how the field will develop from here, and hope that this chapter will help to further motivate population geneticists to apply and enhance landscape genomics.

References

Abrahms B, Sawyer SC, Jordan NR, McNutt JW, Wilson AM, Brashares JS. Does wildlife resource selection accurately inform corridor conservation? J Appl Ecol. 2017;54(2):412–22. https://doi.org/10.1111/1365-2664.12714.

Adriaensen F, Chardon JP, De Blust G, Swinnen E, Villalba S, Gulinck H, Matthysen E. The application of 'least-cost' modelling as a functional landscape model. Landsc Urban Plan. 2003;64(4):233–47.

Afán I, Chiaradia A, Forero MG, Dann P, Ramírez F. A novel spatio-temporal scale based on ocean currents unravels environmental drivers of reproductive timing in a marine predator. Proc R Soc B. 2015;282(1810):20150721.

Aitken SN, Yeaman S, Holliday JA, Wang T, Curtis-McLane S. Adaptation, migration or extirpation: climate change outcomes for tree populations. Evol Appl. 2008;1:95–111.

Alamouti SM, Haridas S, Feau N, Robertson G, Bohlmann J, Breuil C. Comparative genomics of the pine pathogens and beetle symbionts in the genus *Grosmannia*. Mol Biol Evol. 2014;31(6):1454–74. https://doi.org/10.1093/molbev/msu102.

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.

Allen CD, Macalady AK, Chenchouni H, Bachelet D, McDowell N, Vennetier M, Kitzberger T, Rigling A, Breshears DD, Hogg EH, Gonzalez P, Fensham R, Zhang Z, Castro J, Demidova N, Lim J-H, Allard G, Running SW, Semerci A, Cobb N. A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. For Ecol Manage. 2010;259:660–84.

Ally D, Ritland K. A case study: looking at the effects of fragmentation on genetic structure in different life history stages of old-growth Mountain Hemlock (*Tsuga mertensiana*). J Hered. 2007;98:73–8.

Ally D, El-Kassaby Y, Ritland K. Genetic diversity, differentiation and mating system in Mountain Hemlock (*Tsuga mertensiana*) across British Columbia. For Genet. 2000;7:97–108.
Anderson K, Gaston KJ. Lightweight unmanned aerial vehicles will revolutionize spatial ecology. Front Ecol Environ. 2013;11:138–146.
Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016;17:81–92.
Anselin L. Local Indicators of Spatial Association-LISA. Geogr Anal. 1995;27:93–115.
Ashley MV, Abraham ST, Backs JR, Koenig WD. Landscape genetics and population structure in Valley Oak (*Quercus lobata* Nee). Am J Bot. 2015;102:2124–31.
Balkenhol N, Fortin M-J. Basics of study design: sampling landscape heterogeneity and genetic variation for landscape genetic studies. In: Balkenhol N, Cushman S, Storfer A, Waits L, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016. p. 58–75.
Balkenhol N, Waits LP, Dezzani RJ. Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. Ecography. 2009;32:818–30.
Balkenhol N, Holbrook J, Zager P, Rachael J, Onorate D, DeSimone R, White C, Waits LP. A multi-method approach for analyzing hierarchical genetic structures: a case study with cougars (*Puma concolor*). Ecography. 2014;37:552–63.
Balkenhol N, Cushman S, Storfer A, Waits LP. Introduction to landscape genetics: defining, learning and applying an interdisciplinary field. In: Balkenhol N, Cushman S, Storfer A, Waits L, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016a. p. 1–17.
Balkenhol N, Cushman S, Storfer A, Waits LP. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016b.
Balkenhol N, Cushman S, Storfer A, Waits LP. Current status, future opportunities and remaining challenges in landscape genetics. In: Balkenhol N, Cushman S, Storfer A, Waits L, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016c. p. 247–55.
Barbujani G, Sokal RR. Zones of sharp genetic change in Europe are also linguistic boundaries. Proc Natl Acad Sci U S A. 1990;87:1816–9.
Barley AJ, Monnahan PJ, Thomson RC, Grismer LL, Brown RM. Sun skink landscape genomics: assessing the roles of microevolutionary processes in shaping genetic and phenotypic diversity across a heterogeneous and fragmented landscape. Mol Ecol. 2015;24:1696–712.
Bashalkhanov S, Eckert AJ, Rajora OP. Genetic signatures of natural selection in response to air pollution in red spruce (*Picea rubens*, Pinaceae). Mol Ecol. 2013;22:5877–89.
Belanger CL, Jablonski D, Roy K, Berke SK, Krug AZ, Valentine JW. Global environmental predictors of benthic marine biogeographic structure. Proc Natl Acad Sci U S A. 2012;109:14046–51.
Benestan L, Quinn BK, Maaroufi H, Laporte M, Clark FK, Greenwood SJ, Rochette R, Bernatchez L. Seascape genomics provides evidence for thermal adaptation and current-mediated population structure in American lobster (*Homarus americanus*). Mol Ecol. 2016;25:5073–92.
Benomar L, Lamhamedi MS, Rainville A, Beaulieu J, Bousquet J, Margolis HA. Genetic adaptation vs. ecophysiological plasticity of photosynthetic-related traits in young *Picea glauca* trees along a regional climatic gradient. Front Plant Sci. 2016;7:48.
Benz RA, Boyce MS, Thurfjell H, Paton DG, Musiani M, Dormann CF, Ciuti S. Dispersal ecology informs design of large-scale wildlife corridors. PLoS One. 2016;11:e0162989. https://doi.org/10.1371/journal.pone.0162989.
Berg PR, Jentoft S, Star B, Ring KH, Knutsen H, Lien S, Jakobsen KS, Andre C. Adaptation to low salinity promotes genomic divergence in Atlantic cod (*Gadus morhua* L.). Genome Biol Evol. 2015;7:1644–63.
Biek R, Real LA. The landscape genetics of infectious disease emergence and spread. Mol Ecol. 2010;19:3515–31.

Blair C, Weigel DE, Balazik M, Keeley ATH, Walker FM, Landguth E, Cushman S, Murphy M, Waits L, Balkenhol N. A simulation-based evaluation of methods for inferring linear barriers to gene flow. Mol Ecol Resour. 2012;12:822–33.
Bolnick DI, Otto SP. The magnitude of local adaptation under genotype-dependent dispersal. Ecol Evol. 2013;3:4722–35.
Bolnick DI, Snowberg LK, Patenia C, et al. Phenotype-dependent native habitat preference facilitates divergence between parapatric lake and stream stickleback. Evolution. 2009;63:2004–16.
Borevitz JO, Chory J. Genomics tools for QTL analysis and gene discovery. Curr Opin Plant Biol. 2004;7:132–6.
Bowen BW, Gaither MR, DiBattista JD, Iacchei M, Andrews KR, Grant WS, et al. Comparative phylogeography of the ocean planet. Proc Natl Acad Sci U S A. 2016;113:7962–9.
Bradburd GS, Ralph PL, Coop GM. Disentangling the effects of geographic and ecological isolation on genetic differentiation. Evolution. 2013;67:3258–73.
Bradbury D, Smithson A, Krauss SL. Signatures of diversifying selection at EST-SSR loci and association with climate in natural Eucalyptus populations. Mol Ecol. 2013a;22:5112–29.
Bradbury IR, Hubert S, Higgins B, Bowman S, Borza T, Paterson IG, et al. Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. Evol Appl. 2013b;6:450–61.
Bragg JG, Supple RL, Andrew RL, Borevitz JO. Genomic variation across landscapes: insights and applications. New Phytol. 2015;207:953–67.
Buckley J, Butlin RK, Bridle JR. Evidence for evolutionary change associated with the recent range expansion of the British butterfly, *Aricia agestis*, in response to climate change. Mol Ecol. 2012;21:267–80.
Caldwell IR, Gergel SE. Thresholds in seascape connectivity: influence of mobility, habitat distribution, and current strength on fish movement. Landsc Ecol. 2013;28:1937–48.
Calic I, Bussotti F, Martinez-Garcia PJ, Neale DB. Recent landscape genomics studies in forest trees-what can we believe? Tree Genet Genomes. 2016;12:3.
Carl G, Kühn I. Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. Ecol Model. 2007;207:159–70.
Ceron-Souza I, Bermingham E, McMillan WO, Jones FA. Comparative genetic structure of two mangrove species in Caribbean and Pacific estuaries of Panama. BMC Evol Biol. 2012;12:205.
Chavez-Pesqueira M, Suarez-Montes P, Castillo G, Nunez-Farfan J. Habitat fragmentation threatens wild populations of *Carica papaya* (Caricaceae) in a lowland rainforest. Am J Bot. 2014;101:1092–101.
Chhatre VE. Population structure, association mapping of economic traits and landscape genomics of east Texas loblolly pine (*Pinus taeda* L.). PhD thesis, Texas A&M University; 2013. 157 pp.
Chhatre VE, Rajora OP. Genetic divergence and signatures of natural selection in marginal populations of a keystone, long-lived conifer, eastern white pine (*Pinus strobus*) from northern Ontario. PLoS One. 2014;9(5):e97291. https://doi.org/10.1371/journal.pone.0097291.
Chhatre VE, Byram TD, Neale DB, Wegrzyn JL, Krutovsky KV. Genetic structure and association mapping of adaptive and selective traits in the east Texas loblolly pine (*Pinus taeda* L.) breeding populations. Tree Genet Genomes. 2013;9:1161–78. https://doi.org/10.1007/s11295-013-0624-x.
Conover DO, Clarke LM, Munch SB, Wagner GN. Spatial and temporal scales of adaptive divergence in marine fishes and the implications for conservation. J Fish Biol. 2006;69:21–47.
Coop G, Witonsky D, Di Rienzo A, Pritchard JK. Using environmental correlations to identify loci underlying local adaptation. Genetics. 2010;185:1411–23.
Costanza R, D'Arge R, De Groot R, Farber S, Grasso M, Hannon B, et al. The value of the world's ecosystem services and natural capital. Nature. 1997;387:253–60.
Cox K, Broeck AV, Van Calster H, Mergeay J. Temperature-related natural selection in a wind-pollinated tree across regional and continental scales. Mol Ecol. 2011;20:2724–38.
Craft KJ, Ashley MV. Landscape genetic structure of bur oak (*Quercus macrocarpa*) savannas in Illinois. For Ecol Manage. 2007;239:13–20.

Crida A, Manel S. WOMBSOFT: a R package that implements the wombling method to identify genetic boundary. Mol Ecol Notes. 2007;7:588–91.
Cuervo-Alarcon LC. Genetic analysis of European beech populations across precipitation gradients: understanding the adaptive potential to climate change. PhD thesis, Georg-August University of Göttingen, Göttingen; 2017. 153 pp.
Cushman SA, Lewis JS. Movement behavior explains genetic differentiation in American black bears. Landsc Ecol. 2010;25:1613–25.
Cushman SA, McKelvey KS, Hayden J, Schwartz MK. Gene flow in complex landscapes: testing multiple hypotheses with causal modeling. Am Nat. 2006;168:486–99.
Cushman SA, Max T, Meneses N, Evans LM, Ferrier S, Honchak B, et al. Landscape genetic connectivity in a riparian foundation tree is jointly driven by climatic gradients and river networks. Ecol Appl. 2014;24:1000–14.
Dale VH, Joyce LA, McNulty S, Neilson RP, Ayres MP, Flannigan MD, et al. Climate change and forest disturbances. Bioscience. 2001;51:723–34.
Dasgupta MG, Dharanishanthi V, Agarwal I, Krutovsky KV. Development of genetic markers in Eucalyptus species by target enrichment and exome sequencing. PLoS One. 2015;10: e0116528.
Davis JM, Stamps JA. The effect of natal experience on habitat preferences. Trends Ecol Evol. 2004;19:411–6.
De Kort H, Vandepitte K, Bruun HH, Closset-Kopp D, Honnay O, Mergeay J. Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species *Alnus glutinosa*. Mol Ecol. 2014;23:4709–21.
De Mita S, Thuillet AC, Ahmadi N, Manel S, Ronfort J, Vigouroux Y. Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. Mol Ecol. 2013;22:1383–99.
de Villemereuil P, Gaggiotti OE. A new FST-based method to uncover local adaptation using environmental variables. Methods Ecol Evol. 2015;6:1248–58.
DeLong JP, Forbe VE, Galic N, Gibert JP, Laport RG, Phillips JS, Vavra JM. How fast is fast? Eco-evolutionary dynamics and rates of change in populations and phenotypes. Ecol Evol. 2016;6:573–81.
DiLeo MF, Wagner HH. A landscape ecologist's agenda for landscape genetics. Curr Landsc Ecol Rep. 2016;1:115–26.
Diniz-Filho JAF, Nabout JC, Bini LM, Soares TN, de Campos Telles MP, de Marco P Jr, Collevatti RG. Niche modelling and landscape genetics of *Caryocar brasiliense* ("Pequi" tree: Caryocaraceae) in Brazilian Cerrado: an integrative approach for evaluating central-peripheral population patterns. Tree Genet Genomes. 2009;5:617–27.
Ducatez S, Crossland M, Shine R. Differences in developmental strategies between long-settled and invasion-front populations of the cane toad in Australia. J Evol Biol. 2016;29:335–43.
Dudaniec RY, Tesson SVM. Applying landscape genetics to the microbial world. Mol Ecol. 2016;25:3266–75. https://doi.org/10.1111/mec.13691.
Dudaniec RY, Rhodes JR, Worthington-Wilmer J, Lyons M, Lee K, McAlpine CA, Carrick FN. Using multi-level models to identify drivers of landscape genetic structure among management areas. Mol Ecol. 2013;22:3752–65.
Dudaniec RY, Worthington-Wilmer J, Hanson J, Warren M, Bell S, Rhodes JR. Dealing with uncertainty in landscape genetic resistance models: a case of three co-occurring marsupials. Mol Ecol. 2016;25:470–86. https://doi.org/10.1111/mec.13482.
Duforet-Frebourg N, Blum MGB. Nonstationary patterns of isolation-by-distance: inferring measures of local genetic differentiation with Bayesian kriging. Evolution. 2014;68:1110–23.
Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 Genomes data. Mol Biol Evol. 2016;33:1082. https://doi.org/10.1093/molbev/msv334.
Dyer RJ. Is there such a thing as landscape genetics? Mol Ecol. 2015a;24:3518–28.
Dyer RJ. Population graphs and landscape genetics. Annu Rev Ecol Evol Syst. 2015b;46:327–42.

Dyer RJ. Landscapes and plant populations genetics. In: Balkenhol N, Cushman S, Storfer A, Waits L, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016. p. 183–98.

Dyer RJ, Nason JD. Population graphs: the graph theoretic shape of genetic structure. Mol Ecol. 2004;13:1713–27.

Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV, et al. Association genetics of coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. Cold-hardiness related traits. Genetics. 2009;182:1289–302.

Eckert AJ, Bower AD, González-Martínez SC, Wegrzyn JL, Coop G, Neale DB. Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). Mol Ecol. 2010a;19:3789–805.

Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, González-Martínez SC, Neale DB. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics. 2010b;185:969–82.

Eckert AJ, Maloney PE, Vogler DR, Jensen CE, Mix AD, Neale DB. Local adaptation at fine spatial scales: an example from sugar pine (*Pinus lambertiana*, Pinaceae). Tree Genet Genomes. 2015;11:1–17.

Edelaar P, Bolnick DI. Non-random gene flow: an underappreciated force in evolution and ecology. Trends Ecol Evol. 2012;27:659–65.

Edelaar P, Siepielski AM, Clobert J. Matching habitat choice causes directed gene flow: a neglected dimension in evolution and ecology. Evolution. 2008;62:2462–72.

Edelhoff H, Signer J, Balkenhol N. Path segmentation for beginners: an overview of current methods for detecting changes in animal movement patterns. Mov Ecol. 2016;4:21.

Ehrenreich IM, Purugganan MD. The molecular genetic basis of plant adaptation. Am J Bot. 2006;93:953–62.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6(5):e19379.

Engler J, Rödder D, Filz K, Habel J, Balkenhol N. Comparative landscape genetics in three closely related sympatric Hesperid butterflies with diverging ecological traits. PLoS One. 2014;9(9): e106526.

Fageria MS, Rajora OP. Effects of harvesting of increasing intensities on genetic diversity and population structure of white spruce. Evol Appl. 2013;6:778–94.

Feder JL, Forbes AA. Habitat avoidance and speciation for phytophagous insect specialists. Funct Ecol. 2007;21:585–97.

Fischer MC, Rellstab C, Tedder A, Zoller S, Gugerli F, Shimizu KK, et al. Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. Mol Ecol. 2013;22:5594–607.

Fitzpatrick MC, Keller SR. Ecological genomics meets community-level modeling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. Ecol Lett. 2015;18:1–16.

Flaxman SM, Feder JL, Nosil P. Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow. Evolution. 2013;67:2577–91.

Foll M, Gaggiotti OA. Genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics. 2008;180:977.

Forester BR, Jones MR, Joost S, Landguth EL, Lasky JR. Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. Mol Ecol. 2016;25:104–20.

Forester BR, Lasky JR, Wagner HH, Urban DL. Using genotype-environment associations to identify multilocus local adaptation. bioRxiv. 2017; 129460. https://doi.org/10.1101/129460.

Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted regression: the analysis of spatially varying relationships. West Sussex: Wiley; 2002.

Fountain T, Nieminen M, Sirén J, et al. Predictable allele frequency changes due to habitat fragmentation in the *Glanville fritillary* butterfly. Proc Natl Acad Sci U S A. 2016;113:2678–83.

François O, Waits LP. Clustering and assignment methods in landscape genetics. In: Balkenhol N, Cushman S, Storfer A, Waits L, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016. p. 247–55.

Freedman AH, Thomassen HA, Buermann W, Smith TB. Genomic signals of diversification along ecological gradients in a tropical lizard. Mol Ecol. 2010;19:3773–88.

Frichot E, Schoville SD, Bouchard G, Francois O. Testing for associations between loci and environmental gradients using latent factor mixed models. Mol Biol Evol. 2013;30:1687–99.

Frichot E, Schoville SD, de Villemereuil P, Gaggiotti OE, François O. Detecting adaptive evolution based on association with ecological gradients: orientation matters! Heredity. 2015;115:22–8.

Fronhofer EA, Altermatt F. Eco-evolutionary feedbacks during experimental range expansions. Nat Commun. 2015;6:6844.

Funk WC, Lovich RE, Hohenlohe PA, Hofman CA, Morrison SA, Sillett TS, Ghalambor CK, Maldonado JE, Rick TC, Day MD, Polato NR, Fitzpatrick SW, Coonan TJ, Crooks KR, Dillon A, Garcelon DK, King JL, Boser CL, Gould N, Andelt WF. Adaptive divergence despite strong genetic drift: genomic analysis of the evolutionary mechanisms causing genetic differentiation in the island fox (*Urocyon littoralis*). Mol Ecol. 2016;25:2176–94.

Gagnaire PA, Broquet T, Aurelle D, Viard F, Souissi A, Bonhomme F, Arnaud-Haond S, Bierne N. Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. Evol Appl. 2015;8:769–86.

Gaines MS, Diffendorfer JE, Tamarin RH, Whittam TS. The effects of habitat fragmentation on the genetic structure of small mammal population. J Hered. 1997;88:294–304.

Gaither MR, Bowen BW, Rocha LA, Briggs JC. Fishes that rule the world: circumtropical distributions revisited. Fish Fish. 2016;17:664–79.

Gaston KJ. Geographic range limits: achieving synthesis. Proc R Soc Lond B Bio Sci. 2009;276:1395–406.

Gauffre B, Estoup A, Bretangnolle V, Cosson JF. Spatial genetic structure of a small rodent in a heterogeneous landscape. Mol Ecol. 2008;17:4619–29.

Gautier M. Genome-wide scan for adaptive divergence and association with population-specific covariates. Genetics. 2015;201:1555–79.

González-Martínez SC, Krutovsky KV, Neale DB. Forest-tree population genomics and adaptive evolution. New Phytol. 2006;170:227–38.

Gralka M, Stiewe F, Farrell F, Möbius W, Waclaw B, Hallatschek O. Allele surfing promotes microbial adaptation from standing variation. Ecol Lett. 2016;19:889–98.

Guillot G, Rousset F. Dismantling the Mantel tests. Methods Ecol Evol. 2013;4:336–44.

Guillot G, Vitalis R, le Rouzic A, Gautier M. Detecting correlation between allele frequencies and environmental variables as a signature of selection. A fast computational approach for genome-wide studies. Spat Stat. 2014;8:145–55.

Günther T, Coop G. Robust identification of local adaptation from allele frequencies. Genetics. 2013;195:205–20.

Guo B, DeFaveri J, Sotelo G, Nair A, Merilä J. Population genomic evidence for adaptive differentiation in Baltic Sea three-spined sticklebacks. BMC Biol. 2015;13:19.

Haddad NM, Brudvig LA, Clobert J, Davies KF, Gonzalez A, Holt RD, et al. Habitat fragmentation and its lasting impact on Earth's ecosystems. Sci Adv. 2015;1:e1500052.

Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperona G, Toomajian C, et al. Adaptation to climate across the *Arabidopsis thaliana* genome. Science. 2011;334(6052):83–6.

Hand BK, Lowe WH, Kovach RP, Muhlfeld CC, Luikart G. Landscape community genomics: understanding eco-evolutionary processes in complex environments. Trends Ecol Evol. 2015;30:161–8.

Hauser L, Carvalho GR. Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. Fish Fish. 2008;9:333–62.

Hedrick PW, Ginevan ME, Ewing EP. Genetic polymorphism in heterogeneous environments. Annu Rev Ecol Syst. 1976;7:1–32.

Hemmer-Hansen J, Therkildsen NO, Pujolar JM. Population genomics of marine fishes: next-generation prospects and challenges. Biol Bull. 2014;227:117–32.
Henriques R, von der Heyden S, Lipinski MR, du Toit N, Kainge P, Bloomer P, Matthee CA. Spatio-temporal genetic structure and the effects of long-term fishing in two partially sympatric offshore demersal fishes. Mol Ecol. 2016;25:5843–61.
Hitchings SP, Beebee, Trevor JCT. Genetic substructuring as a result of barriers to gene flow in urban *Rana temporaria* (common frog) populations: implications for biodiversity conservation. Heredity. 1997;79:117–27.
Hoban S, Bertorelle G, Gaggiotti OE. Computer simulations: tools for population and evolutionary genetics. Nat Rev Genet. 2012;13:110–22.
Hoban SM, Hauffe H, Pérez-Espona S, Arntzen J, Bertorelle G, Bryja J, et al. Bringing genetic diversity to the forefront of conservation policy and management. Conserv Genet Resour. 2013;5:593–8.
Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, Poss ML, Reed LK, Storfer A, Whitlock MC. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. Am Nat. 2016;188(4):379–97.
Hoekstra HE, Hirschmann RJ, Bundey RJ, Insel P, Crossland JP. A single amino acid mutation contributes to adaptive color pattern in beach mice. Science. 2006;313:101–4.
Hoffmann A, Griffin P, Dillon S, Catullo R, Rane R, Byrne M, Jordan R, Oakeshott J, Weeks A, Joseph L, Lockhart P, Borevitz J, Sgrò C. A framework for incorporating evolutionary genomics into biodiversity conservation and management. Clim Change Resp. 2015;2:1. https://doi.org/10.1186/s40665-014-0009-x.
Holderegger R, Wagner HH. Landscape genetics. BioScience. 2008;58:199–207.
Holderegger R, Kamm U, Gugerli F. Adaptive vs. neutral genetic diversity: implications for landscape genetics. Landsc Ecol. 2006;21:797–807.
Holderegger R, Herrmann D, Poncet B, Gugerli F, Thuiller W, Taberlet P, Gielly L, Rioux D, Brodbeck S, Aubert S, Manel S. Land ahead: using genome scans to identify molecular markers of adaptive relevance. Plant Ecol Divers. 2008;1:273–83.
Holderegger R, Buehler D, Gugerli F, Manel S. Landscape genetics of plants. Trends Plant Sci. 2010;15:675–83.
Holliday JA, Suren H, Aitken SN. Divergent selection and heterogeneous migration rates across the range of Sitka spruce (*Picea sitchensis*). Proc R Soc Lond B Biol Sci. 2012;279:1675–83.
Hu L-J, Uchiyama K, Shen H-L, Ide Y. Multiple-scaled spatial genetic structures of Fraxinus mandshurica over a riparian-mountain landscape in Northeast China. Conserv Genet. 2010;11:77–87.
Jaquiery J, Broquet T, Hirzel AH, Yearsley J, Perrin N. Inferring landscape effects on dispersal from genetic distances: how far can we go? Mol Ecol. 2011;20:692–705.
Johnson JS, Gaddis KD, Cairns DM, Lafon CW, Krutovsky KV. Plant responses to global change: next generation biogeography. Phys Geogr. 2016;37:93–119. https://doi.org/10.1080/02723646.2016.1162597.
Johnson JS, Gaddis KD, Cairns DM, Konganti K, Krutovsky KV. Landscape genomic insights into the historic migration of mountain hemlock in response to Holocene climate change. Am J Bot. 2017a;104(3):439–50. https://doi.org/10.3732/ajb.1600262.
Johnson JS, Gaddis KD, Cairns DM, Krutovsky KV. Seed dispersal at alpine treeline: an assessment of seed movement within the alpine treeline ecotone. Ecosphere. 2017b;8(1):e01649. https://doi.org/10.1002/ecs2.1649.
Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, Birney E. The genomic basis of adaptive evolution in three-spine sticklebacks. Nature. 2012;484:55–61.
Jones MR, Forester BR, Teufel AI, Adams RV, Anstett DN, Goodrich BA, Landguth EL, Joost S, Manel S. Integrating spatially explicit approaches to detect adaptive loci in a landscape genomics context. Evolution. 2013;67:3455–68.

Joost S, Bonin A, Bruford MW, Després L, Conord C, Erhardt G, et al. A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. Mol Ecol. 2007;16:3955–69.
Joost S, Vuilleumier S, Denson JD, Schoville S, Leempoel K, Stucki S, et al. Uncovering the genetic basis of adaptive change: on the intersection of landscape genomics and theoretical population genetics. Mol Ecol. 2013;22:3659–65.
Kamm U, Rotach P, Gugerli F, Siroky M, Edwards P, Holderegger R. Frequent long-distance gene flow in a rare temperate forest tree (*Sorbus domestica*) at the landscape scale. Heredity. 2009;103:476–82.
Kamm U, Gugerli F, Rotach P, Edwards P, Holderegger R. Open areas in a landscape enhance pollen-mediated gene flow of a tree species: evidence from northern Switzerland. Landsc Ecol. 2010;25:903–11.
Kawecki TJ, Ebert D. Conceptual issues in local adaptation. Ecol Lett. 2004;7:1225–41.
Keller SR, Levsen N, Olson MS, Tiffin P. Local adaptation in the flowering-time gene network of balsam poplar, *Populus balsamifera* L. Mol Biol Evol. 2012;29:3143–52.
Kelley JL, Brown AP, Therkildsen NO, Foote AD. The life aquatic: advances in marine vertebrate genomics. Nat Rev Genet. 2016;17:523–34.
Kelly E, Phillips BL. Targeted gene flow for conservation. Conserv Biol. 2016;30:259–67.
Keyghobadi N. The genetic implications of habitat fragmentation for animals. Can J Zool. 2007;85:1049–64.
Kidd MK, Ritchie MG. Phylogeographic information systems: putting the geography into phylogeography. J Biogeogr. 2006;33:1851–65.
Kirkpatrick M, Barton NH. Evolution of a species' range. Am Nat. 1997;150:1–23.
Klopfstein S, Currat M, Excoffier L. The fate of mutations surfing on the wave of a range expansion. Mol Biol Evol. 2006;23:482–90.
Kremer A, Ronce O, Robledo-Arnuncio JJ, Guillaume F, Bohrer G, Nathan R, Bridle JR, Gomulkiewicz R, Klein EK, Ritland K, Kuparinen A, Gerber S, Schueler S. Long-distance gene flow and adaptation of forest trees to rapid climate change. Ecol Lett. 2012;15:378–92.
Krutovsky KV, Burczyk J, Chybicki I, Finkeldey R, Pyhäjärvi T, Robledo-Arnuncio JJ. Gene flow, spatial structure, local adaptation and assisted migration in trees. In: Schnell RJ, Priyadarshan PM, editors. Genomics of tree crops. New York: Springer; 2012. p. 71–116. https://doi.org/10.1007/978-1-4614-0920-5_4.
Kubisch A, Holt RD, Poethke H-J, Fronhofer EA. Where am I and why? Synthesizing range biology and the eco-evolutionary dynamics of dispersal. Oikos. 2014;123:5–22.
Lancaster LT. Widespread range expansions shape latitudinal variation in insect thermal limits. Nat Clim Chang. 2016;6:618. https://doi.org/10.1038/nclimate2945.
Landguth EL, Fedy B, Garey A, Mumma M, Emel S, Oyler-McCance S, et al. Effects of sample size, number of markers, and allelic richness on the detection of spatial genetic pattern. Mol Ecol Resour. 2012;12:276–84.
Landguth E, Cushman S, Balkenhol N. Simulation modeling in landscape genetics. In: Balkenhol N, Cushman S, Storfer A, Waits L, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016. p. 101–13.
Landguth EL, Bearlin A, Day CC, Dunham J. CDMetaPOP: an individual-based, eco-evolutionary model for spatially explicit simulation of landscape demogenetics. Methods Ecol Evol. 2017;8:4–11.
Lee C-R, Mitchell-Olds T. Quantifying effects of environmental and geographical factors on patterns of genetic differentiation. Mol Ecol. 2011;20:4631–42.
Legendre P, Fortin MJ. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Mol Ecol Resour. 2010;10:831–44.
Legendre P, Legendre LP. Numerical ecology. London: Elsevier; 2012.
Legendre P, Borcard D, Roberts DW. Variation partitioning involving orthogonal spatial eigenfunction submodels. Ecology. 2012;93:1234–40.

Legendre P, Fortin MJ, Borcard D. Should the Mantel test be used in spatial analysis? Methods Ecol Evol. 2015;6:1239–47.
Legrand D, Cote J, Fronhofer EA, Holt RD, Ronce O, Schtickzelle N, Travis JMJ, Clobert J. Eco-evolutionary dynamics in fragmented landscapes. Ecography. 2016. https://doi.org/10.1111/ecog.0253.
Lenormand T. Gene flow and the limits to natural selection. Trends Ecol Evol. 2002;17:183–9.
Lepais O, Bacles CF. Two are better than one: combining landscape genomics and common gardens for detecting local adaptation in forest trees. Mol Ecol. 2014;23:4671–3.
Lhuillier E, Butaud J-F, Bouvet J-M. Extensive clonality and strong differentiation in the insular pacific tree *Santalum insulare*: implications for its conservation. Ann Bot. 2006;98:1061–72.
Loiselle BA, Sork VL, Nason J, Graham C. Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). Am J Bot. 1995;82:1420–5.
Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, Storfer A. Breaking RAD: an evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. Mol Ecol Resour. 2017;17:142–52.
Lu M, Krutovsky KV, Nelson CD, Koralewski TE, Byram TD, Loopstra CA. Exome genotyping, linkage disequilibrium and population structure in loblolly pine (*Pinus taeda* L.). BMC Genomics. 2016;17:730. https://doi.org/10.1186/s12864-016-3081-8.
Lu M, Krutovsky KV, Nelson CD, West JB, Reilly NA, Loopstra CA. Association genetics of growth and adaptive traits in loblolly pine (*Pinus taeda* L.) using whole-exome-discovered polymorphisms. Tree Genet Genomes. 2017;13:57. https://doi.org/10.1007/s11295-017-1140-1.
Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet. 2003;4:981–94.
Manel S, Holderegger R. Ten years of landscape genetics. Trends Ecol Evol. 2013;28:614–21.
Manel S, Schwartz MK, Luikart G, Taberlet P. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol Evol. 2003;18:189–97.
Manel S, Gugerli F, Thuiller W, Alvarez N, Legendre P, Holderegger R, et al. Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. Mol Ecol. 2010;19:3760–72.
Manel S, et al. Broad-scale adaptive genetic variation in alpine plants is driven by temperature and precipitation. Mol Ecol. 2012;21:3729–38.
Manicacci D, Olivieri I, Perrot V, Atlan A, Gouyon PH, Prosperi JM, Couvet D. Landscape ecology: population genetics at the metapopulation level. Landsc Ecol. 1992;6:147–59.
Marko PB, Hart MW. The complex analytical landscape of gene flow inference. Trends Ecol Evol. 2011;26:448–56.
Martin MA, Mattioni C, Molina JR, Alvarez JB, Cherubini M, Herrera MA, Villani F, Martin LM. Landscape genetic structure of chestnut (*Castanea sativa* Mill.) in Spain. Tree Genet Genomes. 2012;8:127–36.
Mateo-Sanchéz M, Balkenhol N, Cushman S, Pérez T, Domínguez P, Saura S. A comparative framework to infer landscape effects on population genetic structure: are habitat suitability models effective in explaining gene flow? Landsc Ecol. 2015;8:1405–20.
Mayr E, O'Hara RJ. The biogeographic evidence supporting the Pleistocene forest refuge hypothesis. Evolution. 1986;40:55–67.
McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hitschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9:356–69.
McKown AD, Guy RD, Klapste J, Geraldes A, Friedmann M, Cronk QCB, El-Kassaby YA, Mansfield SD, Douglas CJ. Geographical and environmental gradients shape phenotypic trait variation and genetic structure in *Populus trichocarpa*. New Phytol. 2014;201:1263–76.
McRae BH. Isolation by resistance. Evolution. 2006;60:1551–61.
McRae BH, Dickson BG, Keitt TH, Shah VB. Using circuit theory to model connectivity in ecology and conservation. Ecology. 2008;10:2712–24.

Merilä J, Hendry AP. Climate change, adaptation, and phenotypic plasticity: the problem and the evidence. Evol Appl. 2014;7:1–14.
Merriam G, Kozakiewicz M, Tsuchiya E, Hawley K. Barriers as boundaries for metapopulations and demes of *Peromyscus leucopus* in farm landscapes. Landsc Ecol. 1989;2:227–35.
Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 2007;17:240–8.
Monmonier M. Maximum–difference barriers: an alternative numerical regionalization method. Geogr Anal. 1973;3:245–61.
Montgelard C, Zenboudji S, Ferchaud A, Arnal V, van Vuuren BJ. Landscape genetics in mammals. Mammalia. 2014;78:139–57.
Moran PAP. Notes on continuous stochastic phenomena. Biometrika. 1950;37:17–23.
Morgan SG. Behaviorally mediated larval transport in upwelling systems. Adv Oceanogr. 2014;2014:364214.
Moritz C, Patton JL, Schneider CJ, Smith TB. Diversification of rainforest faunas: an integrated molecular approach. Annu Rev Ecol Syst. 2000;31:533–63.
Morris GP, et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci. 2013;110:453–8.
Mosca E, Eckert AJ, Di Pierro EA, Rocchini D, La Porta N, Belletti P, Neale DB. The geographical and environmental determinants of genetic diversity for four alpine conifers of the European Alps. Mol Ecol. 2012;21:5530–45.
Mosca E, González-Martínez SC, Neale DB. Environmental versus geographical determinants of genetic structure in two subalpine conifers. New Phytol. 2014;201:180–92.
Mosca E, Gugerli F, Eckert AJ, Neale DB. Signatures of natural selection on *Pinus cembra* and *P. mugo* along elevational gradients in the Alps. Tree Genet Genomes. 2016;12:9. https://doi.org/10.1007/s11295-015-0964-9.
Murphy M, Dyer R, Cushman SA. Graph theory and network models in landscape genetics. In: Balkenhol N, Cushman S, Storfer A, Waits L, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016. p. 165–80.
Namroud M-C, Beaulieu J, Juge N, Laroche J, Bousquet J. Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. Mol Ecol. 2008;17:3599–613.
Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA. Genotyping-by-sequencing in ecological and conservation genomics. Mol Ecol. 2013;22:2841–7.
Neale DB, Kremer A. Forest tree genomics: growing resources and applications. Nat Rev Genet. 2011;12:111–22.
Nielsen EE, Hemmer-Hansen JA, Larsen PF, Bekkevold D. Population genomics of marine fishes: identifying adaptive variation in space and time. Mol Ecol. 2009;18:3128–50.
Nosil P. Reproductive isolation caused by visual predation on migrants between divergent environments. Proc R Soc Lond B. 2004;271:1521–8.
Nosil P, Vines TH, Funk DJ. Reproductive isolation caused by natural selection against immigrants from divergent habitats. Evolution. 2005;59:705–19.
Nosil P, Egan SP, Funk DJ. Heterogeneous genomic differentiation between walking-stick ecotypes: "isolation by adaptation" and multiple roles for divergent selection. Evolution. 2008;62:316–36.
Nosil P, Funk DJ, Ortiz-Barrientos D. Divergent selection and heterogeneous genomic divergence. Mol Ecol. 2009;18:375–402.
O'Connell LM, Mosseler A, Rajora OP. Impacts of forest fragmentation on the matting system and genetic diversity of white spruce (*Picea glauca*) at the landscape level. Heredity. 2006;97:418–26.
O'Connell LM, Mosseler A, Rajora OP. Extensive long-distance pollen dispersal in a fragmented landscape maintains genetic diversity in white spruce. J Hered. 2007;98:640–5.

Orsini L, Vanoverbeke J, Swillen I, Mergeay J, Meester L. Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization. Mol Ecol. 2013;22:5983–99.

Pamilo P. Genetic variation in heterogeneous environments. Ann Zool Fenn. 1988;25:99–106.

Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA. Genome-wide association genetics of an adaptive trait in lodgepole pine. Mol Ecol. 2012;21:2991–3005.

Paris CB, Cherubin LM, Cowen RK. Surfing, spinning, or diving from reef to reef: effects on population connectivity. Mar Ecol Prog Ser. 2007;347:285–300.

Parisod C, Holderegger R. Adaptive landscape genetics: pitfalls and benefits. Mol Ecol. 2012;21:3644–6.

Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2:e190.

Paul J, Song YS. Blockwise HMM computation for large-scale population genomic inference. Bioinformatics. 2012;28:2008–15.

Pel MA, Foster SJ, Rietman H, van Arkel G, Jones JD, Van Eck HJ, et al. Mapping and cloning of late blight resistance genes from *Solanum venturii* using an interspecific candidate gene approach. Mol Plant-Microbe Interact. 2009;22:601–15.

Pelletier F, Garant D, Hendry AP. Eco-evolutionary dynamics. Philos Trans R Soc B Biol Sci. 2009;364:1483–9.

Pettorelli N, Vik JO, Mysterud A, Gaillard J-M, Tucker CJ, Stenseth NC. Using the satellite-derived NDVI to assess ecological responses to environmental change. Trends Ecol Evol. 2005;20:503–10.

Pflieger S, Lefebvre V, Causse M. The candidate gene approach in plant genetics: a review. Mol Breed. 2001;7:275–91.

Pflüger F, Balkenhol N. A plea for simultaneously considering matrix quality and local environmental conditions when analyzing landscape impacts on effective dispersal. Mol Ecol. 2014;23:2146–56.

Poelchau MF, Hamrick JL. Differential effects of landscape-level environmental features on genetic structure in three codistributed tree species in Central America. Mol Ecol. 2012;21 (20):4970–82.

Porth I, Klápště J, McKown AD, La Manita J, Guy RD, Ingvrasson PK, et al. Evolutionary quantitative genomics of *Populus trichocarpa*. PLoS One. 2015;10:e0142864.

Primack RB. Essentials of conservation biology. 6th ed. Sunderland: Sinaur Associates; 2014.

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.

Prunier J, Laroche J, Beaulieu J, Bousquet J. Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. Mol Ecol. 2011;20:1702–16.

Prunier JG, Kaufmann B, Fenet S, Picard D, Pompanon F, Joly P, Lena JP. Optimizing the trade-off between spatial and genetic sampling efforts in patchy populations: towards a better assessment of functional connectivity using an individual-based sampling scheme. Mol Ecol. 2013;22:5516–30.

Rajora OP, Eckert AJ, Zinck JWR. Single-locus versus multilocus patterns of local adaptation to climate in eastern white pine (*Pinus strobus, Pinaceae*). PLoS One. 2016;11(7):e0158691.

Rasic G, Filipovic I, Weeks AR, Hoffmann AA. Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, *Aedes aegypti*. BMC Genomics. 2014;15:275.

Ratnam W, Rajora OP, Finkeldey R, Aravanopoulos F, Bouvet J-M, Vallancourt RE, Kanashiro M. Genetic effects of forest management practices: global synthesis and perspectives. For Ecol Manage. 2014;333:52–65.

Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R. A practical guide to environmental association analysis in landscape genomics. Mol Ecol. 2015;24:4348–70.

Richardson JL, Urban MC, Bolnick DI, Skelly DK. Microgeographic adaptation and the spatial scale of evolution. Trends Ecol Evol. 2014;29:165–76.

Riginos C, Crandall ED, Liggins L, Bongaerts P, Treml E. Navigating the currents of seascape genomics: how spatial analyses can augment population genomic studies. Curr Zool. 2016;62:581–601. https://doi.org/10.1093/cz/zow067.

Roberts DR, Hamann A. Glacial refugia and modern genetic diversity of 22 western North American tree species. Proc R Soc B. 2015;282(1804):20142903.

Roffler GH, Amish SJ, Smith S, Cosart T, Kardos M, Schwartz MK, Luikart G. SNP discovery in candidate adaptive genes using exon capture in a free-ranging alpine ungulate. Mol Ecol Resour. 2016a;16:1147–64.

Roffler GH, Schwartz MK, Pilgrim MK, Talbot SL, Sage GK, Adams LG, Luikart G. Identification of landscape features influencing gene flow: how useful are habitat selection models? Evol Appl. 2016b;9:805–17.

Roschanski AM, Csillery K, Liepelt S, Oddou-Muratorio S, Ziegenhagen B, Huard F, Ulrich KK, Postolache D, Vendramin GG, Fady B. Evidence of divergent selection at landscape and local scales in Abies alba Mill. in the French Mediterranean Alps. Mol Ecol. 2016;25:776–94.

Rosenblum EB, Harmon LJ. "Same same but different": replicated ecological speciation at white sands. Evolution. 2011;65:946–60.

Safran RJ, Scordato ESC, Symes LB, Rodríguez RL, Mendelson TC. Contributions of natural and sexual selection to the evolution of premating reproductive isolation: a research agenda. Trends Ecol Evol. 2013;28:643–50.

Savolainen O, Pyhäjärvi T, Knürr T. Gene flow and local adaptation in trees. Annu Rev Ecol Evol Syst. 2007;38:595–619.

Schoville SD, Bonin A, François O, Lobreaux S, MeloDelima C, Manel S. Adaptive genetic variation on the landscape: methods and cases. Annu Rev Ecol Evol Syst. 2012;43:23–43.

Schumaker NH. HexSim Version 2.5.7. Corvallis: U.S. Environmental Protection Agency, Environmental Research Laboratory; 2013. http://hexsim.net

Schupp EW, Fuentes M. Spatial patterns of seed dispersal and the unification of plant population ecology. Ecoscience. 1995;2:267–75.

Schwabl P, Llewellyn MS, Landguth EL, Andersson B, Kitron U, Costales JA, et al. Prediction and prevention of parasitic diseases using a landscape genomics framework. Trends Parasitol. 2017;33(4):264–75.

Schwartz MK, Luikart G, McKelvey KS, Cushman SA. Landscape genomics: a brief perspective. In: Cushman SA, Huettmann F, editors. Spatial complexity, informatics, and wildlife conservation. Tokyo: Springer; 2009. p. 165–74.

Selkoe KA, Toonen RJ. Marine connectivity: a new look at pelagic larval duration and genetic metrics of dispersal. Mar Ecol Prog Ser. 2011;436:291–305.

Selkoe KA, D'Aloia CC, Crandall ED, Iacchei M, Liggins L, Puritz JB, von der Heyden S, Toonen RJ. A decade of seascape genetics: contributions to basic and applied marine connectivity. Mar Ecol Prog Ser. 2016a;554:1–19.

Selkoe KA, Scribner KT, Galindo HM. Waterscape genetics – applications of landscape genetics to rivers, lakes, and seas. In: Balkenhol N, Cushman S, Storfer A, Waits L, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016b. p. 220–46.

Servedio MR. The evolution of premating isolation: local adaptation and natural and sexual selection against hybrids. Evolution. 2004;58:913–24.

Sexton JP, Hangartner SB, Hoffmann AA. Genetic isolation by environment or distance: which pattern of gene flow is most common? Evolution. 2014;68:1–15.

Shafer ABA, Wolf JBW, Alves PC, Bergström L, Bruford M, Brännström I, et al. Genomics and the challenging translation into conservation practice. Trends Ecol Evol. 2015;30:78–87.

Smith TB, Wayne RK, Girman DJ, Bruford MW. A role for ecotones in generating rainforest biodiversity. Science. 1997;276:1855–7.

Smouse PE, Sork VL. Measuring pollen flow in forest trees: an exposition of alternative approaches. For Ecol Manage. 2004;197:21–38.

Sodeland M, Jorde PE, Lien S, Jentoft S, Berg PR, Grove H, et al. 'Islands of divergence' in the Atlantic cod represent polymorphic chromosomal rearrangements. Genome Biol Evol. 2016;8:1012–22. https://doi.org/10.1093/gbe/evw057.

Sork V, Smouse P. Genetic analysis of landscape connectivity in tree populations. Landsc Ecol. 2006;21:821–36.

Sork VL, Nason J, Campbell DR, Fernandez JF. Landscape approaches to historical and contemporary gene flow in plants. Trends Ecol Evol. 1999;14:219–24.

Sork VL, Davis FW, Westfall R, Flint A, Ikegami M, Wang H, Grivet D. Gene movement and genetic association with regional climate gradients in California valley oak (*Quercus lobata* Née) in the face of climate change. Mol Ecol. 2010;19:3806–23.

Sork VL, Aitken SN, Dyer RJ, Eckert AJ, Legendre P, Neale DB. Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. Tree Genet Genomes. 2013;9:901–11.

Sork VL, Squire K, Gugger PF, Steele SE, Levy ED, Eckert AJ. Landscape genomic analysis of candidate genes for climate adaptation in a California endemic oak, *Quercus lobata*. Am J Bot. 2016;103:33–46.

Spear SF, Balkenhol N, McRae B, Scribner K, Fortin M-J. Modeling resistance surfaces for landscape genetics: considerations for parameterization and analysis. Mol Ecol. 2010;19:3576–91.

Spear SF, Cushman SA, McRae BM. Resistance surface modeling in landscape genetics. In: Balkenhol N, Cushman S, Storfer A, Waits L, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016. p. 129–48.

Stinchcombe JR, Hoekstra HE. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity. 2008;100:158–70.

Storfer A, Murphy MA, Evans JS, Goldberg CS, Robinson S, Spear SF, et al. Putting the 'landscape' in landscape genetics. Heredity. 2007;98:128–42.

Storfer A, Murphy MA, Spear SF, Holderegger R, Waits LP. Landscape genetics: where are we now? Mol Ecol. 2010;19:3496–514.

Storfer A, Antolin MF, Manel S, et al. Genomic approaches in landscape genetics. In: Balkenhol N, Cushman S, Storfer A, Waits L, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016. p. 249–164.

Stucki S, Orozco-terWengel P, Bruford MW, Colli L, Masembe C, Negrini R, Taberlet P, Joost S, The NEXTGEN Consortium. High performance computation of landscape genomic models including local indicators of spatial association. Mol Ecol Resour. 2016; arXiv:1405.7658. https://arxiv.org/abs/1405.7658

Suren H, Hodgins KA, Yeaman S, Nurkowski KA, Smets P, Rieseberg LH, Aitken SN, Holliday JA. Exome capture from the spruce and pine giga-genomes. Mol Ecol Resour. 2016;16:1136–46.

Swaegers J, Mergeay J, Van Geystelen A, Therry L, Larmuseau MHD, Stoks R. Neutral and adaptive genomic signatures of rapid poleward range expansion. Mol Ecol. 2015;24:6163–76. https://doi.org/10.1111/mec.13462.

Taylor HR, Soanes K. Breaking out of the echo chamber: missed opportunities for genetics at conservation conferences. Biodivers Conserv. 2016;25:1987–93.

Taylor P, Fahrig L, With K. Landscape connectivity: a return to basics. In: Crooks KR, Sanjayan M, editors. Connectivity conservation. Cambridge: Cambridge University Press; 2006. p. 29–43.

Travis JMJ, Munkemuller T, Burton OJ, et al. Deleterious mutations can surf to high densities on the wave front of an expanding population. Mol Biol Evol. 2007;24:2334–43.

Truong C, Palmé AE, Felber F. Recent invasion of the mountain birch *Betula pubescens* ssp. *tortuosa* above the treeline due to climate change: genetic and ecological study in northern Sweden. J Evol Biol. 2007;20:369–80.

Tsuda Y, Sawada H, Ohsawa T, Nakao K, Nishikawa H, Ide Y. Landscape genetic structure of *Betula maximowicziana* in the Chichibu mountain range, central Japan. Tree Genet Genomes. 2010;6:377–87.

Turner TL, Hahn MW. Genomic islands of speciation or genomic islands and speciation? Mol Ecol. 2010;19:848–50.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nat Genet. 2010;42:260–3.

van Strien MJ, Keller D, Holderegger R. A new analytical approach to landscape genetic modelling: least-cost transect analysis and linear mixed models. Mol Ecol. 2012;21:4010–23.

Vangestel C, Vazquez-Lobo A, Martinez-Garcia PJ, Calic I, Wegrzyn JL, Neale DB. Patterns of neutral and adaptive genetic diversity across the natural range of sugar pine (*Pinus lambertiana* Dougl.). Tree Genet Genomes. 2016;12:51.

Via S. The ecological genetics of speciation. Am Nat. 2002;159(S3):S1–7.

von der Heyden S, Lipinski MR, Matthee CA. Mitochondrial DNA analyses of the Cape hakes reveal an expanding, panmictic population for *Merluccius capensis* and population structuring for mature fish in *Merluccius paradoxus*. Mol Phylogenet Evol. 2007;42:517–27.

von der Heyden S, Beger M, Toonen RJ, van Herwerden L, Juinio-Meñez MA, Ravago-Gotanco R, Fauvelot C, Bernardi G. The application of genetics to marine management and conservation: examples from the Indo-Pacific. Bull Mar Sci. 2014;90:123–58.

Wagner H, Fortin M-J. A conceptual framework for the spatial analysis of landscape genetic data. Conserv Genet. 2013;14:253–61.

Wagner HH, Fortin MJ. Basics of spatial data analysis: linking landscape and genetic data for landscape genetic studies. In: Balkenhol N, Cushman S, Storfer A, Waits L, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016. p. 77–98.

Waits LP, Storfer A. Basics of population genetics: quantifying neutral and adaptive genetic variation for landscape genetic studies. In: Balkenhol N, Cushman S, Storfer A, Waits L, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016. p. 35–57.

Waits LP, Cushman SA, Spear SF. Applications of landscape genetics to connectivity research in terrestrial animals. In: Balkenhol N, Cushman S, Storfer A, Waits L, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016. p. 199–219.

Wang IJ, Bradburd GS. Isolation by environment. Mol Ecol. 2014;23:5649–62.

Wang Y-H, Yang KC, Bridgman CL, Lin LK. Habitat suitability modelling to correlate gene flow with landscape connectivity. Landsc Ecol. 2008;23:989–1000.

Wang R, Compton SG, Chen X-Y. Fragmentation can increase spatial genetic structure without decreasing pollen-mediated gene flow in a wind-pollinated tree. Mol Ecol. 2011;20:4421–32.

Wang R, Compton SG, Shi Y-S, Chen X-Y. Fragmentation reduces regional-scale spatial genetic structure in a wind-pollinated tree because genetic barriers are removed. Ecol Evol. 2012;2:2250–61.

Wang IJ, Glor RE, Losos JB. Quantifying the roles of ecology and geography in spatial genetic divergence. Ecol Lett. 2013;16:175–82.

Wang Z-F, Lian J-Y, Ye W-H, Cao H-L, Zhang Q-M, Wang Z-M. Pollen and seed flow under different predominant winds in wind-pollinated and wind-dispersed species *Engelhardia roxburghiana*. Tree Genet Genomes. 2016;12:19.

Wee AKS, Takayama K, Asakawa T, Thompson B, Onrizal, Sungkaew S, Tung NX, Nazre M, Soe KK, Tan HTW, Watano Y, Baba S, Kajita T, Webb EL. Oceanic currents, not land masses, maintain the genetic structure of the mangrove *Rhizophora mucronata* Lam. (Rhizophoraceae) in Southeast Asia. J Biogeogr. 2014;41:954–64.

Wenzel MA, Piertney SB. Digging for gold nuggets: uncovering novel candidate genes for variation in gastrointestinal nematode burden in a wild bird species. J Evol Biol. 2015;28:807–25.

White C, Selkoe KA, Watson J, Siegel DA, Zacherl DC, Toonen RJ. Ocean currents help explain population genetic structure. Proc R Soc B. 2010;277:1685–94.

White TA, Perkins SE, Heckel G, Searle JB. Adaptive evolution during an ongoing range expansion: the invasive bank vole (*Myodes glareolus*) in Ireland. Mol Ecol. 2013;22:2971–85.

Whitham TG, Bailey JK, Schweitzer JA, Shuster SM, Bangert RK, LeRoy VJ, et al. A framework for community and ecosystem genetics: from genes to ecosystems. Nat Rev Genet. 2006;7:510–23.

Whitlock MC, Lotterhos KE. Reliable detection of loci responsible for local adaptation: inference of a neutral model through trimming the distribution of FST. Am Nat. 2015;186(S1):S24–36.

Womble W. Differential systematics. Science. 1951;28:315–22.

Wright S. Isolation by distance. Genetics. 1943;28:114–38.

Zeller KA, McGarigal K, Whiteley AR. Estimating landscape resistance to movement: a review. Landsc Ecol. 2012;27:777–97.

Zeller KA, Creech TG, Millette KL, Crowhurst RS, Long RA, Wagner HH, Balkenhol N, Landguth EL. Using simulations to evaluate Mantel-based methods for assessing landscape resistance to gene flow. Ecol Evol. 2016;6:4115–28.

Zhou Y, Zhang L, Liu J, Wu G, Savolainen O. Climatic adaptation and ecological divergence between two closely related pine species in Southeast China. Mol Ecol. 2014;23:3504–22.

Zinck JWR, Rajora OP. Post-glacial phylogeography and evolution of a wide-ranging highly-exploited keystone forest tree, eastern white pine (*Pinus strobus*) in North America: single refugium, multiple routes. BMC Evol Biol. 2016;16:56. https://doi.org/10.1186/s12862-016-0624-1.

Ziólkowska E, Ostapowicz K, Radeloff VC, et al. Assessing differences in connectivity based on habitat versus movement models for brown bears in the Carpathians. Landsc Ecol. 2016;31:1863. https://doi.org/10.1007/s10980-016-0368-8.

Zulliger D, Schnyder E, Gugerli F. Are adaptive loci transferable across genomes of related species? Outlier and environmental association analyses in Alpine Brassicaceae species. Mol Ecol. 2013;23:1626–39.

Paleogenomics: Genome-Scale Analysis of Ancient DNA and Population and Evolutionary Genomic Inferences

Tianying Lan and Charlotte Lindqvist

Abstract Ancient DNA analysis has in the last 30 years grown into a compelling research tool that has radically transformed many scientific fields. In particular, methods of extracting ancient DNA that is often highly degraded and advances in genome sequencing technologies within the last decade have revolutionized genetic research of extinct and ancient lineages. Insights into ancient genomes, and their links to modern ones, hold unparalleled promise to capture the numerous processes of organismal evolution and their responses to a changing world. Hence, genomic-scale sequencing of up to several-thousand-year-old remains has contributed substantially to our understanding of the impacts of Pleistocene glaciations in shaping the Earth's biodiversity and organismal distributions, the process of domestication, the history of diseases, and our own history as humans. In this chapter, we review some of the advances in ancient DNA sequencing and give examples of recent case studies in paleogenomic research.

Keywords Ancient DNA · DNA degradation · Domestication · Genomics · Metagenomics · Next-generation sequencing · Targeted enrichment

1 Introduction

The study of ancient DNA (aDNA) refers to the analysis of historical or ancient biological samples that have not been archived with the intent of subsequent molecular analyses, e.g., specimens stored in museum collections or excavated from caves, permafrost, ice cores, or archeological or paleontological sites. After the death of an organism, its DNA is rapidly degraded by endogenous nucleases and other chemical processes, as well as exogenous microorganisms that feed on and

T. Lan · C. Lindqvist (✉)
Department of Biological Sciences, University at Buffalo (SUNY), Buffalo, NY, USA

School of Biological Sciences, Nanyang Technological University, Singapore, Singapore
e-mail: CL243@buffalo.edu

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*,
Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2017_7,

degrade macromolecules, resulting in DNA fragmentation, base modifications, and a reduction of overall DNA amounts (Pääbo et al. 2004). Although some environmental circumstances, such as rapid desiccation, low temperatures, and high salt concentrations, can slow down the DNA degradation process (Lindahl 1993a, b), eventually the accumulative effects of postmortem DNA decay will become so extensive that no useful molecules are left (Hofreiter et al. 2001). The methods of DNA amplification developed in the 1980s brought incredible reports of DNA in specimens many millions of years old, but it was subsequently predicted that no measurable DNA would survive for more than 100,000 years (Lindahl 1993a, b). However, with the advent of genome sequencing technologies and continuous improvements in aDNA methods, this date has continually been pushed back. Because DNA from extinct species and populations can provide a direct window into the past, aDNA facilitates observations of changes in genetic diversity over time. Over the last 30 years, aDNA research based mostly on high copy number mitochondrial DNA (mtDNA) sequence fragments has been informative in answering questions about organismal relationships and timings of divergence, as well as hypotheses of geological and environmental events and their impact on genetic changes and biogeographic patterns. However, it is in the last decade with advances in high-throughput massively parallel DNA sequencing – also termed next-generation sequencing – that we have experienced the most transformative period in aDNA research. Such scaled-up analyses of genomic-level aDNA, or paleogenomics, are delivering insights into the impacts of the Pleistocene glaciations at the organismal level and even complex evolutionary histories of ancient and extinct organisms, including our own history, and are transforming our knowledge of the history of plant and animal domestications and sweeping historical pandemics. This chapter gives a review of the history of aDNA research and the status of the applications and promises of paleogenomics. We present some case studies of ancient and extinct organisms, from humans to microbes. While recognizing the immense value of decade-old historical samples, e.g., specimens archived in museum collections that can give insight into historical responses to anthropogenic perturbations, and despite that they exhibit patterns of DNA degradation and must often be treated similarly to ancient specimens, here we mainly focus on studies of specimens and remains hundreds to thousands of years (kyr) old.

2 From Paleogenetics to Paleogenomics

Over three decades ago, aDNA sequencing was initiated by using bacterial cloning of short fragments of DNA retrieved from a museum specimen of the quagga, an extinct member of the horse family, as well as of Egyptian mummies (Higuchi et al. 1984; Pääbo 1985). Although these early results were irreproducible (Pääbo et al. 2004), these studies suggested that endogenous DNA, which is generally limited to very low levels of short and damaged fragments, was retrievable from ancient specimens.

The development of the polymerase chain reaction (PCR) in the 1980s (Saiki et al. 1985) greatly facilitated targeting and amplifying the low level of surviving DNA molecules in ancient specimens. Consequently, the number and range of aDNA studies expanded rapidly (Fig. 1), and remarkable reports of DNA recovered from up to many million-year-old amber inclusions were published (Cano et al. 1992; DeSalle et al. 1992; Cano and Borucki 1995). Many of these claims, also referred to as "antediluvian" DNA (Lindahl 1993a, b), were later revealed to be impossible to replicate or clear the results of contamination. However, with increasing knowledge about postmortem DNA damage patterns, improvements of DNA extraction methodology, and rigorous laboratory standards of contamination control and authentication criteria, results based on the PCR-based method became more reliable, and they continue to be used routinely in aDNA studies to obtain mostly short, overlapping DNA amplicons. As a result, a large number of comprehensive phylogenetic and population genetic studies on both extinct and extant species have been carried out, achieving significant discoveries. Nevertheless, due to the requirement for prior knowledge about the target sequences, the limited inferential possibilities from analyses of short sequence fragments, the focus on high copy number mitochondrial and chloroplast DNA, and the nonrecombining nature of such organellar DNAs, PCR-based approaches are infeasible in large-scale studies and for interpretations of complex evolutionary histories.

2.1 Next-Generation Sequencing and Related Technological Advancements in Paleogenomic Studies

Sanger sequencing, as a first-generation sequencing technology, is still an effective way of sequencing smaller sets of amplified fragments from degraded DNA templates. However, its limited throughput and relatively high cost generate major barriers to large-scale genomic studies, as demonstrated by the costly and long process of sequencing the first human genome (Lander et al. 2001). In the mid-2000s, the advent of next-generation sequencing (NGS) technologies, which allow billions of molecules to be sequenced simultaneously, opened the gate for massively parallel genome sequencing and to the "paleogenomic" era.

Importantly, aDNA is characterized by extremely short fragments, which is actually ideal for NGS, since most NGS platforms as a first step require that the input DNA template is fragmented into smaller pieces. Along with the continually increasing accuracy and throughput of NGS and the decreasing price per nucleotide base, the number of aDNA studies has increased dramatically in the last two decades (Fig. 1). Just within the last few years, many ancient genomes, ranging from pathogenic bacteria to vertebrate genomes and numerous ancient human genomes, have been sequenced (see Table 1 for some references). In addition to enabling high-coverage nuclear and mitochondrial genome sequencing, NGS made it possible to obtain ultrashort aDNA fragments (~30–50 bp) that were abundant in a diverse range of ancient specimens but were too short to be routinely amplified by PCR. Although

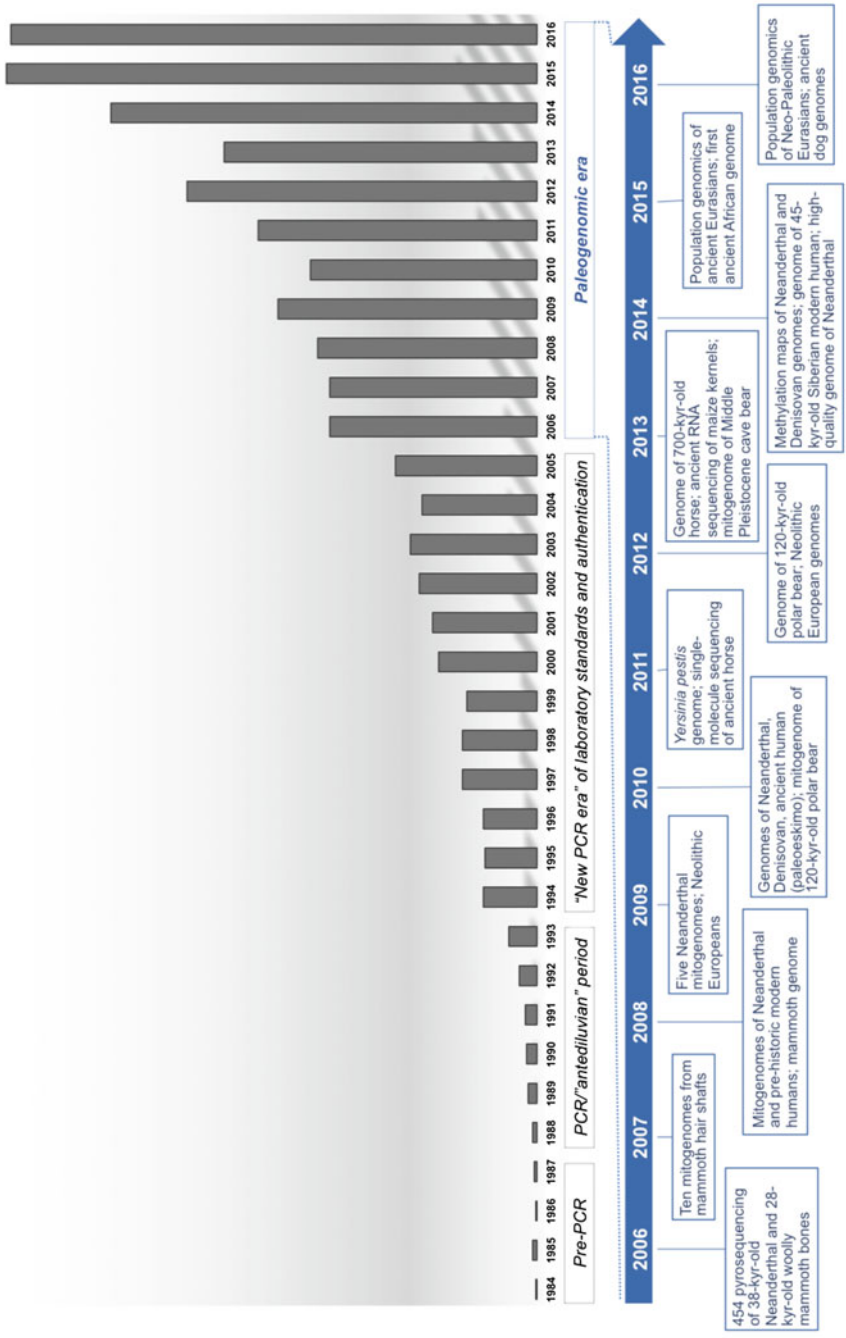

Fig. 1 Three decades of ancient DNA research. Histogram of aDNA literature showing a gradually expanding number of research papers from one report of the cloning of quagga DNA in 1984 (Higuchi et al. 1984) to over 350 papers in 2016. These three decades of aDNA research are split into four major periods: (1) The "pre-PCR" period

NGS technologies provide a much more feasible and efficient strategy for aDNA sequencing and clearly have transformed aDNA research, they are not free of problems. For example, sequencing error rates (~0.1–15%) are higher than that of traditional Sanger sequencing, and the damaged nucleotides in aDNA further inflate these error rates. Moreover, the substantial loss of endogenous DNA during library construction and the inflation of exogenous DNA content during amplification remain major challenges. In order to take full advantage of NGS technologies in aDNA research, comprehensive studies seeking to improve the efficiency of the key technological steps, such as DNA extraction, library construction, and hybridization enrichment, have been performed and are discussed in the following sections.

2.2 *Ancient DNA Extraction*

Despite the massive technological improvements and decreasing costs of NGS sequencing, the efficiency of the NGS approach for sequencing ancient genomes is largely determined by the amount of endogenous DNA recovered from archeological and paleontological samples (hereafter referred to as ancient samples), which contain both endogenous and exogenous DNA. It has been demonstrated that in most ancient samples, endogenous DNA contents are very low, often less than 1% of the total DNA content (Orlando et al. 2011). Only in very rare cases, e.g., particular environmental conditions (Miller et al. 2008) or sample types (Gilbert et al. 2007), will endogenous DNA content ratios exceed 50% (Rasmussen et al. 2010; Reich et al. 2010; Prufer et al. 2014; Palkopoulou et al. 2015; Botigue et al. 2016). This is due to the spontaneous damage taking place after organismal death, such as oxidation and hydrolysis, which results in fragmentation and chemical modification of the DNA molecules (Thomas and Gilbert 2006). In contrast, exogenous DNA, which is usually from environmental contaminants such as bacteria and fungi, is often abundant. Unlike modern genome studies, it is impractical to obtain more DNA by increasing starting material in ancient genome studies, since the material available for destructive sampling is usually limited. Therefore, multiple DNA extraction methods, proven replicable and robust and maximizing the yield of endogenous DNA, have been developed.

Fig. 1 (continued) where aDNA research constituted sequencing of bacterial clones (Pääbo 1985); (2) The PCR period where astonishing reports of DNA from up to many million-year-old specimens were published (Golenberg et al. 1990; Cano et al. 1992; DeSalle et al. 1992; Soltis et al. 1992; Cano et al. 1993; Woodward and Bunnell 1994), also referred to as "antediluvian" DNA (Lindahl 1993a, b); (3) A more somber period in aDNA research where new laboratory standards and authentication of aDNA were introduced; and (4) The paleogenomic era where NGS and other new technologies have been applied to ancient DNA. The chart comprises a total of 3,166 publications, including original research, technical, and review papers from public database searches (as of December 2016) (SciFinder, Scopus, Web of Science, PubMed, Wiley), followed by manual curation in EndNote X7. Timeline of selected milestones in the first decade of paleogenomic research is shown

Table 1 A selection of paleogenomic case studies within the last 6 years (as of December 2016), including information about estimated endogenous content from mapping to the nearest modern reference genome, average genome coverage, and methods

Species	Age (kyr)	Endogenous content (%)	Coverage (fold)	DNA extraction method[a]	Library construction method[b]	Sequencing platform	Reference
Homo							
Neanderthal	38, 44, undated	<5	1.3	SS	454 converted into Illumina	Illumina	Green et al. (2010)
	~50 and 60–70	70	52 and 0.5	SS	DS and SS	Illumina	Prufer et al. (2014)
Denisovan	~50	70	1.9	SS	DS	Illumina	Reich et al. (2010)
	~50	70	30	SS	SS	Illumina	(Meyer et al. 2012)
Modern human	~4	93.17	20	SS	DS	Illumina	Rasmussen et al. (2010)
	5.3	37.9	7.6	PC	SOLiD	SOLiD	Keller et al. (2012)
	~45	1.8–10	42	SS	SS	Illumina	Fu et al. (2014)
	~24	17	1	SC	DS	Illumina	Raghavan et al. (2014a, b)
	~12.6	0.5–28.2	14.4	SS	DS	Illumina	Rasmussen et al. (2014)
	~7–8	n/a	Up to 22	SC	DS and SS	Illumina	Lazaridis et al. (2014)
	~37	n/a	2.42	SS and SC	DS	Illumina	Seguin-Orlando et al. (2014)
	~4.5	n/a	12.5	SC	DS	Illumina	Llorente et al. (2015)
	~0.2–6	n/a	0.003–1.7	SS	DS	Illumina	Raghavan et al. (2015)
	~8.5	0.4 and 1.4	~1	SS	DS	Illumina	Rasmussen et al. (2015a, b)
	~4–7.6	n/a	2–7	PC	DS	Illumina	Hofmanova et al. (2016)
	1.25–3.15	18.9–40	0.004–7.25	SC	DS	Illumina	Jeong et al. (2016)
Vertebrates							
Woolly mammoth	18.5	58–90	<1	PC	454	454	Miller et al. (2008)
	4.3 and 44.8	~80	17.1 and 11.2	SC	DS	Illumina	Palkopoulou et al. (2015)

Polar bear	~120	4.7 (Illumina), 59 (Ion Torrent)	1.83	SS and SC	DS and Ion Torrent	Illumina and Ion Torrent	Miller et al. (2012); Lan et al. (2016)
Wild auroch	6.75	28.10	6.23	SC	DS	Illumina	Park et al. (2015)
Horse	~43 and 700	0.47 (Illumina), 4.21 (Helicos)	1.78 and 1.12	SS	DS	Illumina and Helicos	Orlando et al. (2013)
	~42.7 and 16	0.6 and 0.03	7.4 and 24.3	SS	DS	Illumina	Schubert et al. (2014)
Dog	7 and 4.7	>67	9	SC and PC	DS	Illumina	Botigue et al. (2016)
	4.8	85.14	28	SC	SS	Illumina	Frantz et al. (2016)
Plants							
Maize	5.31	70	1.73	SC and PC	SS	Illumina	Ramos-Madrigal et al. (2016)
Barley	6	0.4–96.4	0.19 to 20	PTB	DS	Illumina	Mascher et al. (2016)
Microorganisms							
P. infestans	~0.17 and 0.13	n/a	16 and 22	CTAB	DS	Illumina	Martin et al. (2013)
	~0.17	n/a	>20	SC and PTB	DS	Illumina	Yoshida et al. (2013)
Y. pestis	~0.67	n/a	30	PC	DS	Illumina	Bos et al. (2011)
M. leprae	~1	n/a	>100	SS	DS	Illumina	Schuenemann et al. (2013)
M. tuberculosis	~1	n/a	>20	SS	DS	Illumina	Bos et al. (2014)
V. cholerae	~0.17	n/a	15	FFPE	DS	Illumina	Devault et al. (2014a, b)
Variola virus	~0.37	n/a	18	FFPE	DS	Illumina	Duggan Ana et al. (2016)

[a]DNA extraction methods were mainly based on silica-in-solution (SS) (Rohland and Hofreiter 2007), silica columns (SC) (Dabney et al. 2013), traditional phenol-chloroform (PC), PTB or CTAB (Ristaino et al. 2001; Kistler 2012), and FFPE (formalin-fixed paraffin-embedded) tissue extraction (Okello et al. 2010)
[b]Two Illumina library construction methods were mostly applied: the double-stranded protocol (DS) (Meyer and Kircher 2010) and the single-stranded (SS) protocol (Gansauge and Meyer 2013)

Since the successful recovery of DNA material from dried muscle of the extinct quagga in 1984 (Higuchi et al. 1984), a range of aDNA extraction methods has been extensively tested and developed for different types of ancient samples, including bones, dental remains, feces, hairs, soft tissues, and sediments (Gamba et al. 2016). One of the most commonly adopted aDNA extraction methods is the in-solution, silica-based DNA extraction method developed by Rohland and Hofreiter (2007). This method was optimized for degraded bones and dental remains and later optimized for increasing sample throughput by using columns for washing and elution steps (Rohland et al. 2010). The in-solution silica-based method has been successful in various aDNA studies, including retrieval of DNA from Neanderthal, Denisovan (Meyer et al. 2012), and, the oldest successfully ancient genome to date, an ~700-kyr-old permafrost horse sample (Orlando et al. 2013). More recently, Dabney and colleagues (2013) developed a silica column-based method that significantly outperformed the in-solution method in the recovery of ultrashort fragments (<50 bp). By using this method, mitochondrial genomes were reconstructed from an ~400-kyr-old cave bear and an ~400-kyr-old hominin, the oldest non-permafrost samples successfully sequenced to date. DNA extraction methods based on silica columns increase not only the total yield of DNA but also the relative abundance of endogenous DNA and its molecular diversity (Gamba et al. 2016). Optimizations focusing on reducing the exogenous fraction have also been tested, such as eliminating exogenous DNA prior to sample digestion (Korlevic et al. 2015; Cruz-Dávalos et al. 2016), adding a bleach wash or additional digestion steps (Ginolhac et al. 2012; Der Sarkissian et al. 2014; Damgaard et al. 2015; Boessenkool et al. 2016), and using enrichment for methylation marks depleted in bacterial genomes (Seguin-Orlando et al. 2015). The bleach wash step, for example, can have a negative impact on DNA yield in some cases (Lan et al. unpublished data). Furthermore, it has been shown that aDNA yields differ in different parts of the bones or teeth. For example, significantly higher yield was observed in the denser part of the ancient human petrous bone (Pinhasi et al. 2015). Thus, sampling strategy, such as using the densest part of an ancient bone or the dentine of a tooth (Allentoft et al. 2015; Damgaard et al. 2015), should also be taken into account for optimizing the aDNA extraction procedure.

Despite successes achieved by improvement in aDNA extraction methods, studies have shown that only a small fraction (or even none) of the endogenous DNA can be recovered by most DNA extraction methods (Gilbert et al. 2007; Barta et al. 2014). By comparing the fragment length distributions from the ~700-kyr-old horse and ~400-kyr-old cave bear sequences, Hofreiter et al. (2015) speculated that substantially more DNA may be recovered from the horse sample if using a more efficient extraction method. Therefore, further optimizations maximizing the amount and representation of endogenous DNA are required.

2.3 NGS Library Preparation of aDNA

As one of the key steps in the NGS sequencing procedure, several standard library preparation protocols were developed for different NGS platforms (Margulies et al.

2005; Bentley et al. 2008). The first aDNA studies applying NGS approaches (Green et al. 2006; Poinar et al. 2006) achieved important breakthroughs in the aDNA research field. However, the standard library preparation protocols that were used were soon demonstrated to be poorly suitable for the extremely low quantities of damaged aDNA and resulted in substantial loss of endogenous DNA (Noonan et al. 2005; Green et al. 2006). A number of research groups have worked on optimizing library preparation protocols and improving the conversion efficiency of aDNA into NGS libraries for both double- and single-stranded-based methods (Fig. 2) (Maricic and Paabo 2009; Meyer and Kircher 2010; Briggs and Heyn 2012; Gansauge and Meyer 2013; Bennett et al. 2014).

2.3.1 Double-Stranded Library Preparation

The common strategy for NGS library preparation requires ligation of sequencing adapters to fragmented double-stranded DNA followed by multiple purification steps. Two methods have been used for the ligation of adapters. The blunt-end method, which was developed by 454 Life Sciences, ligates partially double-stranded adapter pairs to the blunt-end-repaired double-stranded DNA template. The Y-shaped adapter method, which was introduced by Illumina, ligates the Y-shaped adapter with a T-overhang to both ends of DNA templates that carry A-overhangs created by an A-tailing reaction. Improvements to specifically increase the conversion efficiency in aDNA library preparations include using heat treatment rather than NaOH to release streptavidin-coated beads, using quantitative PCR to determine the amounts of libraries for emulsion PCR, removing uracil miscoding lesions, which are abundant in damaged DNA, and minimizing purification steps (Fortes and Paijmans 2015). It is noteworthy that the application of enzymatic treatment (e.g., USER enzyme mix), which removes abasic cytosines deaminated into uracils, followed by repairing of the DNA fragments prior to adapter ligation, has shown great success in increasing the accuracy of DNA sequences determined while maintaining DNA sequence yield (Briggs et al. 2010; Reich et al. 2010; Meyer et al. 2012; Fu et al. 2014; Seguin-Orlando et al. 2014; Orlando et al. 2015; Rasmussen et al. 2015a, b; Rohland et al. 2015; Bos et al. 2016; Hofmanova et al. 2016). However, drawbacks due to the nature of aDNA cannot be avoided. For example, if damaged DNA molecules carry single-stranded breaks on both strands, they cannot be utilized, and there is significant loss of DNA molecules caused by the washing steps when silica spin columns are used (20–80%, Lan et al. unpublished data) or when carboxylated beads are employed. Also, when using the blunt-end method, 50% of the DNA templates that receive nondistinct adapters by chance are lost from the library (Bennett et al. 2014).

2.3.2 Single-Stranded Library Preparation

Recently, a single-stranded library preparation method was specifically developed for the sequencing of aDNA on the Illumina platform (Gansauge and Meyer 2013)

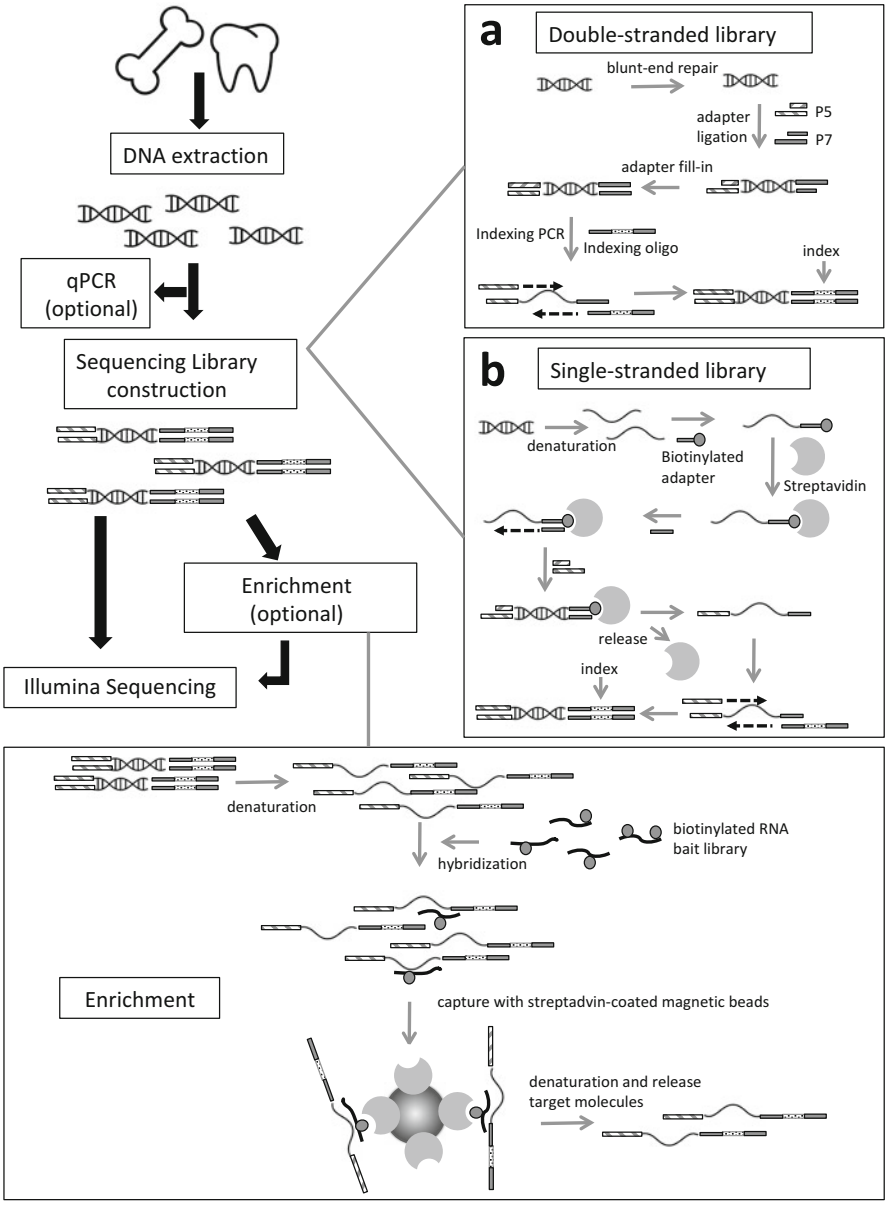

Fig. 2 A pipeline for performing Illumina sequencing from ancient specimen. Following DNA extraction, qPCR can be performed to examine endogenous DNA level for estimating input for library preparation and targeted enrichment (Enk et al. 2013). Illumina sequencing libraries are usually constructed through either (**a**) a double-stranded protocol (Meyer and Kircher 2010) or (**b**) a single-stranded protocol (Gansauge and Meyer 2013). A bead capture enrichment protocol (Carpenter et al. 2013; Enk et al. 2014) can be performed to enrich target sequences prior to sequencing

and then expanded to the Ion Torrent platform (Bennett et al. 2014). This method initially uses a single-stranded DNA ligase and a biotinylated adapter to capture and immobilize single-stranded DNA molecules to streptavidin-coated beads. A PCR reaction using the bound-to-bead DNA molecules as template is then performed to generate double-stranded DNA, followed by a second adapter blunt-end ligation. The single strand is then released from the beads by heated denaturation and used to complete the adapter sequence through an amplification reaction. All reaction steps are carried out while the DNA is tightly bound to the beads, and this largely reduces the loss of molecules during washing steps. In addition, DNA molecules with single-stranded breaks on both strands are disassembled into multiple fragments in the single-stranded method, and each fragment has an independent chance of being recovered in the library. Hence, this method has been highly efficient in building NGS libraries from Denisovan and Neanderthal samples (Meyer et al. 2012; Prufer et al. 2014), as well as 30–50 bp DNA fragments from a ~400-kyr-old cave bear (Dabney et al. 2013) and a ~400-kyr-old hominin (Meyer et al. 2014). The conversion efficiency of this method, i.e., the efficiency of conversion of input DNA fragments to DNA molecules that can be sequenced, was calculated to be about 30–70%, and the sequence yield is at least sixfold higher compared with the double-stranded method when using the same DNA extract (Gansauge and Meyer 2013). It was also suggested that the single-stranded method converts a higher proportion of endogenous DNA relative to exogenous DNA in most samples studied (Bennett et al. 2014). Although this method is currently believed to be the most efficient, improvements to increase the conversion efficiency and reduce the cost and preparation time are still needed.

2.4 Targeted Enrichment

For ancient samples with low endogenous DNA contents, substantial shotgun sequencing is often needed to generate sufficient coverage for the genomic regions of interest. The direct shotgun sequencing approach is not only costly but also impossible in some cases due to limited quantities of ancient material. To address this issue, hybridization enrichment (also referred to as targeted capture) has been used to enrich sequencing libraries for selected genomic regions, e.g., a single chromosome, exome, a subset of SNPs, or organelle genomes and even whole nuclear genomes (Briggs et al. 2009; Burbano et al. 2010; Maricic et al. 2010; Carpenter et al. 2013; Fu et al. 2013, 2016; Castellano et al. 2014; Enk et al. 2014). This approach uses synthesized baits that have high-sequence similarity to the target genomic regions to capture and immobilize the desired DNA fragments through hybridization, either in solution (Fig. 2) (Bos et al. 2014; Enk et al. 2014; Haak et al. 2015; Duggan Ana et al. 2016; Fu et al. 2016) or on microarrays (Burbano et al. 2010; Fortes and Paijmans 2015; Bos et al. 2016; Paijmans et al. 2016; Spyrou et al. 2016), whereas the non-hybridized fragments are washed away via several washing steps. The end product should then contain a much higher ratio of targeted endogenous DNA versus exogenous DNA, making the subsequent sequencing more precise and yielding higher coverage at a

lower cost. In addition, it has been suggested that shorter DNA fragments have higher hybridization efficiency than longer fragments (Hodges et al. 2007), naturally introducing a bias against exogenous contaminating DNA that is generally expected to be longer than endogenous aDNA. One study showed that endogenous DNA had been significantly enriched from less than 5% in the pre-enrichment library to over 70% in the after-enrichment library following in-solution hybridization enrichment (Enk et al. 2014). Furthermore, the flexibility of hybridization enrichment has been shown to allow for cross-species capture for species with no prior sequence information available, using baits designed from a close relative – an effective approach to investigate aDNA from non-model organisms or extinct lineages.

To date, many hybridization enrichment protocols have been applied to aDNA research from a variety of species, including hominids (Briggs et al. 2009; Burbano et al. 2010; Carpenter et al. 2013; Fu et al. 2013, 2016; Castellano et al. 2014), mammals (Dabney et al. 2013; Zhang et al. 2013; Enk et al. 2014; Mohandesan et al. 2016), plants (Avila-Arcos et al. 2011), and microbes (Bos et al. 2011; Devault et al. 2014a, b; Bos et al. 2016; Spyrou et al. 2016), demonstrating the tremendous potential of this approach. Commercial hybridization enrichment assays are available for studying human and model organisms and their close relatives (Bodi et al. 2013; Elhaik et al. 2013), while custom-designed assays also have been developed and optimized for non-model organisms (Maricic et al. 2010; Bi et al. 2012; Enk et al. 2014). Small-scale mitogenome enrichment, which has been broadly applied in many aDNA studies, is usually feasible for most ancient samples due to the high abundance of mtDNA, while whole-genome enrichment is still not very cost-effective for large-scale sequencing of ancient genomes, especially for population-level studies on samples with very low levels of endogenous DNA content. Whole-genome enrichment has, however, been successfully carried out in Neanderthal and mammoth genome studies (Carpenter et al. 2013; Enk et al. 2014). Furthermore, for extinct species with no closely related reference genome available, it is infeasible to design highly efficient whole-genome custom baits. Whole exome capture (Castellano et al. 2014) may be an alternative approach, since the exome is generally more conserved than other regions of the genome, making cross-species capture over medium evolutionary distances more feasible (Cosart et al. 2011; Bi et al. 2012).

For designing a custom enrichment protocol, it is important to be aware that aDNA hybridization enrichment is very sensitive. It has been found that even within the same experiment, the enrichment rate (endogenous DNA proportion after enrichment to the proportion before enrichment) for different samples often varies significantly (Carpenter et al. 2013; Enk et al. 2013, 2014). The reason for different enrichment rates lies in many factors that can impact the efficiency of hybridization, e.g., sequence similarity between bait and target, hybridization temperature, bait type, bait tiling, bait concentration, and post-hybridization washing temperatures (Avila-Arcos et al. 2011; Bodi et al. 2013; Li et al. 2013). In addition, intrinsic factors, such as low levels of endogenous DNA, highly fragmented DNA molecules, and complexity of aDNA handling, are added to factors inherent to the capture protocol. Several studies have been carried out to investigate the impact of different parameters for aDNA capture and to optimize experimental conditions, such as bait-tilling densities, bait molecular features, and amounts of starting DNA (Enk et al.

2013; Cruz-Dávalos et al. 2016; Paijmans et al. 2016). However, due to the complexity of the enrichment approach and the characteristics of aDNA, capture parameters need to be carefully addressed during the design stage of the capture assay, and more investigations of the effects of different capture parameters are warranted.

3 Paleogenomics Applications: Paleogenomic Case Studies from Humans to Pathogens

Advances in genome sequencing technologies have dramatically transformed the use of aDNA. In a first attempt at paleogenomic sequencing, shotgun sequencing of amplification-independent metagenomic libraries was performed and yielded ~27 kb of cave bear (*Ursus spelaeus*) genome sequence (Noonan et al. 2005). In the following year, using a combination of Sanger and massively parallel pyrosequencing, ~65 kb of nuclear genome sequence data was recovered from a 38-kyr-old Neanderthal specimen (Noonan et al. 2006). These first experiments demonstrated two major problems of the cloning-based, high-throughput sequencing approach: most of the sequences produced came from exogenous contaminant DNA, so only tiny fractions of the genome could be reconstructed, and bacterial cloning posed a heavy experimental load. Additionally, the cost for sequencing was high. Nevertheless, these first studies demonstrated the prospects of genome-scale sequencing from ancient specimens containing very low levels of endogenous DNA, and advances in NGS technologies and library preparation protocols have largely resolved some of these problems. Consequently, new paleogenomes are being reported at an increasing rate today. To date, many complete or partial paleogenomes have been recovered, particularly from hominins but also other animals, plants, and microorganisms (Table 1), addressing many key questions in unraveling and reconstructing the evolution history of extinct and extant species. In the following sections, we will present major applications of paleogenomes and review some recent findings for a selection of species, ranging from hominins to pathogens.

3.1 Human Paleogenomics

3.1.1 Archaic Hominins

Representing a groundbreaking milestone in paleogenomics, the sequencing and analysis of genomes from two extinct archaic hominins, Neanderthals and Denisovans, which are the closest relatives of anatomically modern humans, not only offered a complementary approach for understanding hominin evolutionary history but also provided remarkable technical advancements for generating and analyzing aDNA data. Since their first discovery in 1856, remains left by Neanderthals have been found across Eurasia, from Western Europe to Central and Northern Asia. In 2010, the first draft Neanderthal genome was assembled at ~1.3-fold coverage from a combined sequencing dataset derived from three female bone specimens found in the Vindija Cave in Croatia and

dated to 38–44 kyr ago (Green et al. 2010). Four years later, as a consequence of improvements in aDNA extraction and sequencing techniques, as well as owing to the exceptionally high endogenous DNA content and minimal contamination levels, a 52-fold high-quality genome was obtained from a female Neanderthal fossil (the Altai Neanderthal) recovered in a ~50-kyr-old layer in the Denisova Cave in Southern Siberia Altai Mountains, and a low-coverage draft genome (0.5-fold) was generated in the same study from an approximately 60–70-kyr-old Neanderthal infant from the Mezmaiskaya Cave in the Caucasus (Prufer et al. 2014).

Unlike the abundant fossil records of Neanderthals, Denisovans are known only from a finger bone and two molars excavated at the Denisova Cave. Initially, the morphologies of the finger bones and the molar were not informative enough to ascertain which hominin group they belonged to. Although mtDNA sequence indicated that the mitogenome diverged from the Neanderthal-modern human clade approximately 1 million years ago (Krause et al. 2010), the identity of the finger bone remained questionable until researchers published a ~1.9-fold coverage nuclear genome in 2010 (Reich et al. 2010), suggesting that this bone belongs to a group distinct from both Neanderthals and modern humans. Later in 2012, by using a new single-stranded DNA library preparation method, a draft Denisovan genome was obtained at 30-fold coverage from the same finger bone with less than 1% modern human contamination, showing a quality comparable to that of modern genomes sequenced at a similar depth (Meyer et al. 2012).

Analyses of these genomes have suggested that Neanderthal and Denisovan populations shared a common origin and split approximately 381–473 kyr ago and that their common ancestor diverged from the ancestors of modern humans around 550–765 kyr ago (Prufer et al. 2014). Demographic histories showed that a decline in population size occurred in both archaic populations sometime before 1 myr ago, whereas an expansion took place in the ancestral population to modern humans. The population size of Neanderthals was estimated to have been only about a tenth that of modern humans, despite a broad distribution across Eurasia. Similar population size was estimated for the Denisovan genome (Meyer et al. 2012; Prufer et al. 2014). Coinciding with the reduced effective population size, extremely low heterozygosity was found in both the Neanderthal and Denisovan genomes, indicating that the genetic diversity of the populations to which the ancient specimens belonged was very low compared with that of modern humans.

Admixture analyses based on the vast genomic data have revealed extensive information regarding gene flow among archaic hominins and modern humans. All Neanderthal genomes sequenced to date showed significantly more derived alleles shared with non-Africans than with sub-Saharan Africans and that Neanderthal ancestry contributed 1.5–2.1% of modern Eurasian genomes (Prufer et al. 2014), suggesting that gene flow occurred between the ancestors of non-Africans and Neanderthals outside Africa. Further studies indicated this gene flow was from Neanderthal to the common ancestor of modern Eurasians and that this most likely occurred at an early stage of the out-of-Africa expansion around 50–65 kyr ago, before the divergence of Europeans (Sankararaman et al. 2012; Fu et al. 2014). Consistent with living at a time closer to the time of gene flow, two genomes from 37- to 45-kyr-old ancient Eurasians show longer segments of

Neanderthal ancestry than those in present-day humans (Fu et al. 2014; Seguin-Orlando et al. 2014). Interestingly, a significantly higher level of Neanderthal ancestry was found in East Asians than in Europeans, suggesting either that there was an additional pulse of Neanderthal gene flow into the ancestor of Eastern Asians after they diverged from Europeans or that the proportion of Neanderthal ancestry in Europeans was reduced by interbreeding with a modern human group that did not admix with Neanderthals (Wall et al. 2013; Kim and Lohmueller 2015; Vernot and Akey 2015). Similarly, the Denisovan genome was found to contribute to the gene pool of modern humans, however, through a different history of admixture. Denisovan ancestry was estimated to contribute 4–6% in genomes of Melanesians, aboriginal Australians and other Southeast Asian islanders, and to a lower level (~0.2%) in other mainland Asian and Native American populations (Reich et al. 2011; Meyer et al. 2012; Prufer et al. 2014). It was suggested that gene flow from Denisovans to the ancestor of modern populations in Oceania possibly occurred in Southeast Asia approximately 44–54 kyr ago (Reich et al. 2011; Sankararaman et al. 2016). Additionally, Denisovans themselves appear to have received admixture from other archaic hominin groups. It has been proposed that gene flow from Neanderthals contributes at least 0.5% of the Denisovan genome, and an additional gene flow that contribute 0.5–8% was from an unknown archaic hominin group, possibly *Homo erectus*, which diverged from other hominins more than 1 myr ago (Prufer et al. 2014). Overall, it is evident that the evolutionary history of archaic hominins and modern humans is complex, and more admixture events will likely be revealed through future finding on new fossils. Unfortunately, DNA is not likely to be preserved in the many interesting fossil remains from equatorial or otherwise warm regions, including that of the enigmatic *Homo floresiensis* from Malesia, extraction and sequencing efforts on which have so far been unsuccessful (Stringer 2014).

Genome sequences of archaic hominin genomes have also helped pinpoint genetic changes that may set modern humans apart from their extinct relatives by identifying DNA features carried by modern humans that differ from archaic hominins and great ape genomes. A genome-wide catalog of these genetic changes was established based on 1,094 modern human genomes and the high-quality archaic hominin genomes; it consists of ~30,000 single nucleotide substitutions and ~4,000 indels, among which only 96 amino acid substitutions were identified (Prufer et al. 2014). Among the 87 genes carrying fixed amino acid changes, several genes related to brain development were found to express more often in the ventricular zone of the developing neocortex than genes carrying silent substitutions (Prufer et al. 2014). However, our limited understanding of how those genetic changes relate to phenotypes makes it difficult to accurately predict their functional consequences, and further functional investigations will be necessary to clarify whether or how these changes affect phenotypes in modern humans.

3.1.2 Ancient Anatomically Modern Humans

Genome sequences from ancient remains of anatomically modern humans have facilitated a number of breakthroughs in inferring past population histories by providing access to genetic variation that existed in past populations but possibly lost in the modern-day

genetic pool. The first ancient human genome was sequenced from permafrost-preserved hair of a ~4-kyr-old Paleo-Eskimo at an average depth of 20× (Rasmussen et al. 2010). The ancestry analysis suggested the population the Paleo-Eskimo individual belonged to had migrated from Siberia to North American Arctic approximately 5.5 kyr ago. This migration is proposed to be independent of the ones giving rise to the Inuit and other Native Americans (Rasmussen et al. 2010). The genetic continuity displayed in the Paleo-Eskimo population indicated that extensive gene flow occurred among local groups, which likely resulted in a large overall effective population size. The Paleo-Eskimos survived in the Arctic for more than 4 kyr until they were eventually replaced by the Neo-Eskimos, ancestors of present-day Inuit, less than 700 years ago (Raghavan et al. 2014a, b).

Since the sequencing of the Paleo-Eskimo genome, numerous ancient human genomes have been analyzed, shedding light on major questions in the population history of anatomically modern humans. One such question is who were the first inhabitants of the Americas (Fig. 3). The hypothesis of a Siberian origin for contemporary Native Americans is supported by the genome sequencing of a 24-kyr-old Siberian individual from Mal'ta in South Central Siberia (Raghavan et al. 2014a, b). The Mal'ta individual was found to share genetic affinities to both West Eurasian and modern-day Native Americans and was estimated to contribute to 14–38% of current Native American ancestry. Considering the strong Eastern Asian genetic component found among Native Americans (Schurr and Sherry 2004), it is likely that Native Americans are admixed between populations that were related to the Mal'ta lineage and one or more unknown East Asian lineages (Raghavan et al. 2014a, b). The first and the oldest ancient Native American genome was sequenced from remains of a child excavated at the Anzick site, Montana, USA (Rasmussen et al. 2014). The Anzick child, buried approximately 12.6 kyr ago, is believed to belong to the Clovis culture – the earliest archeological complex in the Americas. The analyses suggested that the Clovis metapopulation, from which the Anzick child genome originates, is closely related to all indigenous American populations and was directly ancestral to many contemporary Native Americans, supporting a pre-Clovis occupation of the Americas. Interestingly, similar amounts of the Mal'ta genetic signal were found in the Anzick and contemporary Native American genomes, indicating the early admixture event involving the Mal'ta-like lineage happened at least 12.6 kyr ago. Furthermore, the Anzick population is more closely related to Central and South Americans than to northern North Americans, whereas the genome sequences obtained from the 8,500-year-old Kennewick Man found in Washington showed great affinity with several contemporary Native North American populations (Rasmussen et al. 2015a, b). This evidence suggests that the divergence of two Native American lineages likely occurred more than 12.6 kyr ago. Consistent with this finding, a large-scale genomic study involving 31 modern and 23 ancient human genomes also inferred that northern North Americans diverged from southern North Americans and Central and South Americans ~13 kyr ago, while all Native Americans diverged from their ancestors ~20 to 23 kyr ago (Raghavan et al. 2015).

The peopling of Europe is another major controversial topic that has been enlightened by paleogenomic studies. Decades of debates have been focused on understanding the origin of agriculture in Europe, including whether the spread of agriculture was from cultural diffusion within indigenous hunter-gatherers or from demic

Fig. 3 Population history of Native Americans. Genomic evidence from a 24-kyr-old Siberian individual Mal'ta suggests that ancestors of contemporary Native Americans migrated from Siberia in a single wave no earlier than 23 kyr ago, separate from the Inuit (*smaller arrow*), and that they were likely admixed with one or more East Asian lineages (*dashed line*). Evidence obtained from a 12.6-kyr-old Anzick-1 and an 8.5-kyr-old Kennewick Man suggested a split between northern North American lineages and Central and South American lineages likely occurred more than 12.6 kyr ago. Adapted from Raghavan et al. (2015)

diffusion of Near-Eastern farmers during the Neolithic transition (Bramanti et al., 2009). In recent years, numerous genomes have been obtained from ancient humans in Eurasia that have provided invaluable clues into the complex genetic structure of

Europeans, providing compelling evidence to support the demic diffusion model (Keller et al. 2012; Skoglund et al. 2012; Lazaridis et al. 2014; Raghavan et al. 2014a, b; Seguin-Orlando et al. 2014; Allentoft et al. 2015; Haak et al. 2015; Mathieson et al. 2015; Fu et al. 2016; Hofmanova et al. 2016; Lazaridis et al. 2016). The first ancient European genome was sequenced from a 5.3-kyr-old Copper Age Tyrolean Iceman discovered in the Ötztal Alps (Keller et al. 2012). Surprisingly, the Iceman genome showed a greater genetic affinity to extant Sardinians than to the present-day population inhabiting the Alps. Similarly, a 5-kyr-old Scandinavian farmer was found to have closer genetic ties to extant Southern Europeans than to extant Northern Europeans (Skoglund et al. 2012). These findings suggest that admixture likely occurred between ancient Southern European farmers and Northern European hunter-gatherers due to the demic diffusion of Neolithic farmers 8–6 kyr ago (Keller et al. 2012; Skoglund et al. 2012). Substantial evidence of demic diffusion was obtained from genomes of several Anatolian and Aegean individuals dated to 4,000–7,600 BC; these data similarly suggested that Near-Easterners migrated into Europe via two independent routes and that admixture between migrating farmers and local hunter-gatherers occurred throughout the Neolithic era (Hofmanova et al. 2016). It is becoming clear that the peopling of Europe has been a very complex process and that multiple migrations from distinct populations might have influenced the present-day genetic make-up of Europeans. First, a revised three-way mixture demic diffusion model was introduced by studies of genomes from several 7–8-kyr-old individuals from Western and Central Europe (Lazaridis et al. 2014), the 24,000-year-old Mal'ta Siberian genome (Raghavan et al. 2014a, b), and the 37-kyr-old Eastern Eurasian (Seguin-Orlando et al. 2014). According to this model, at least three different ancient populations contributed to the ancestry of present-day Europeans: (1) indigenous Paleolithic West European hunter-gatherers, (2) ancient North Eurasians related to Upper Paleolithic Siberians, and (3) early Near-Eastern Neolithic farmers. A genome-wide SNP analysis of 69 ancient Europeans later suggested a massive westward migration from the Pontic steppe ~4.5 kyr ago that involved the Yamnaya – an Early Bronze Age population whose culture appeared to have replaced the Neolithic farming cultures in temperate Eastern Europe by 3,000 BC, in what is also known as the Bronze Age transition. It was found that the Yamnaya share ancestry with ancient North Eurasians (Haak et al. 2015). Together with the findings from over 100 Bronze Age ancient Eurasian genomes (Allentoft et al. 2015), the demic diffusion model was further refined by identifying the Yamnaya population as one of the sources of ancient North Eurasian ancestry that contributed to present-day Europeans.

Because aDNA survival is negatively correlated with temperature, nuclear genomic sequences from fossils at lower latitudes, such as Africa, South America, and Australia, are much less abundant. To date, only one paleogenomic study successfully reconstructed an ancient African genome, which was sequenced at 12.5× coverage from a 4.5-kyr-old individual found in Mota Cave in southeastern Ethiopia (Llorente et al. 2015). This individual was found to have a genetic tie to the contemporary Ethiopian populations that live in this region, indicating population continuity until present. Given that the Mota population predates the West Eurasian backflow that occurred ~3 kyr ago, there was no West Eurasian introgression found in the Mota genome, and it was suggested that this

genome can act as an ideal unadmixed African reference for future studies. In Eastern Asia, paleogenomic studies are also scarce. Only a partial genome was obtained from a 40-kyr-old modern human discovered in the Tianyuan Cave near Beijing. It was inferred that this individual belonged to the common ancestral population of present-day East Asians and Native Americans and that it had already diverged from the ancestors of present-day Europeans (Fu et al. 2016). Recently, genomes were recovered from several individuals of prehistoric Himalayan populations dated to 3.15–1.25 kyr ago that showed a strong genetic similarity to contemporary high-altitude East Asians (Jeong et al. 2016). These genomes were found to share high-altitude-adaptive allelic signatures with present-day Tibetans, suggesting that the high-altitude populations maintained a long-term stability in genetic structure since East Asians colonized this region. Thus, acculturation or cultural diffusion rather than large-scale introgression from non-East Asians might have been responsible for the temporal changes in material culture and mortuary behavior in the region (Jeong et al. 2016).

3.2 Non-hominin Vertebrate Paleogenomics

Although most paleogenomic studies have so far focused on uncovering ancient human history, another major branch is based on non-hominin vertebrates. Successful retrieval of genomes from fossils dating back beyond 100 kyr (Miller et al. 2012; Orlando et al. 2013; Lan et al. 2016) broke the Middle Pleistocene time barrier of paleogenomics and gave promise for rebuilding genomes from Early Pleistocene fossils. Although recovery of such aDNA is still largely from remains preserved in permafrost and at high latitudes, the potential exists for retrieving genome-scale data from many vertebrate and other animal groups, also from lower-latitude environments. In the following, we review some recent advances in mammalian paleogenomics, focusing on nearly complete or partial ancient genomes reconstructed from the extinct woolly mammoth (*Mammuthus primigenius*), polar bear (*Ursus maritimus*), dog (*Canis lupus familiaris*), and horse (*Equus ferus*) and the impact these studies have had from genetic changes associated with extinction and domestication to ancient demographic trajectories and past interspecific admixture events.

3.2.1 Woolly Mammoth

The woolly mammoth was among the most abundant megafaunal species during the Pleistocene and early Holocene of the Northern Hemisphere, but it became extinct in its mainland range about 10 kyr ago (Stuart et al. 2004). A few isolated island populations persisted into the Holocene and finally went extinct roughly 3.7 kyr ago (Vartanyan et al. 2008). The woolly mammoth is one of the most studied prehistoric animals due to ample specimens of soft tissue and hair mostly recovered from permafrost, which is considered to be an excellent environment for preservation of DNA (Poinar and Stankiewicz 1999; Smith et al. 2001; Willerslev et al. 2004). A number of partial or full mitogenomes of mammoths were retrieved early on, and they were used to infer phylogenetic relationships to extant

elephants (Krause et al. 2006; Rogaev et al. 2006; Debruyne et al. 2008; Gilbert et al. 2008). The first efforts to sequence the woolly mammoth nuclear genome also occurred during the first stages of the paleogenomic era and not surprisingly resulted in low-genome coverage (Poinar et al. 2006; Miller et al. 2008). More recently, high-quality genomes were sequenced to investigate mammoth demographic history and genetic changes leading up to its extinction (Palkopoulou et al. 2015), as well as pinpointing genes possibly associated with adaptations to the Arctic (Lynch et al. 2015). One specimen, which yielded 11-fold genome coverage, was derived from a mainland population 45 kyr old, when the populations were flourishing, while another that yielded 17-fold genome coverage was from the last-surviving island population. The demographic trajectories of the two specimens appeared to be nearly identical, revealing an ancient bottleneck and a steep decline in effective population size in the ancestral island populations at the start of the Holocene. Reduced genome-wide heterozygosity and accumulation of detrimental mutations were found in the island specimen, delivering the first direct evidence of genetic stochasticity due to small population size contributing to population extinction (Palkopoulou et al. 2015; Rogers and Slatkin 2016).

3.2.2 Polar Bear

The polar bear, whose main habitat is largely shaped by the extent of Arctic sea ice, has become a symbolic species for understanding how climate change and glacial oscillations impact species evolution and biodiversity. Complete genomes from ancient polar bear remains might provide invaluable clues regarding the adaption of the species to the extreme conditions of the Arctic, after splitting from its lower-latitude sister species, the brown bear. Such fossil genomes could also illuminate the intertwined evolutionary histories of the two species. However, the polar bear fossil record is extremely poor because their remains most likely disappear into the ocean after they die on the sea ice. In 2012, a draft genome was sequenced from the oldest known polar bear fossil, a stratigraphically validated ~120-kyr-old jawbone from the Norwegian archipelago of Svalbard (Miller et al. 2012), for the first time successfully pushing the limits of a vertebrate genome toward the Middle Pleistocene. This specimen was recently re-investigated to obtain a new draft assembly at ~1.8-fold coverage. Genome-wide, SNP-based admixture analyses of this ancient genome together with multiple modern genomes suggest that the ancient polar bear shares less gene flow with brown bears compared to extant polar bears (Lan et al. 2016). These findings imply introgression from ancestral brown bears into the ancestor of the modern polar bear lineage. A more complete genome from the ancient polar bear remains may help clarify whether past interspecies introgression correlated with climate oscillations and how it impacted the adaptation of polar bear to the Arctic environment.

3.2.3 Dog Domestication

Through a complex domestication process, the dog emerged as the first domestic animal before the advent of settled agriculture. Although many archeological fossils and numerous genome sequences from modern dogs are available (Freedman et al. 2014; Pilot et al. 2015; Shannon et al. 2015), claims on their origin, the domestication process, and the early evolutionary history still remain controversial, largely because of highly insufficient genetic studies on ancient dog genomes. In 2016, two studies almost simultaneously published the first high-quality ancient dog genomes and conducted comprehensive analyses to reconstruct their intricate evolutionary histories (Botigue et al. 2016; Frantz et al. 2016). One study recovered a 28-fold coverage genome from a Late Neolithic dog dated back to ~4.8 kyr ago and mitochondrial sequences from 59 ancient dogs that lived between ~3 and 14 kyr ago in Ireland (Frantz et al. 2016). The other study obtained two ninefold coverage genomes, one each from an Early Neolithic ~7-kyr-old dog and a Late Neolithic ~4.7-kyr-old dog from central Europe (Botigue et al. 2016). Consistently, both studies discovered that all ancient dog genomes carry a low copy number of the *AMY2B* gene, which is associated with starch digestion and normally is present at high copy number in modern dog genomes, indicating the adaptation to starch-rich diets occurred after the advent of agriculture during the Neolithic period. However, the two studies proposed contradictory hypotheses on the domestication process. The former study suggested that two independent domestication events from two genetically distinct wolf populations might have occurred in Eastern and Western Eurasia, respectively. During the Late Neolithic, dogs from Eastern Eurasia dispersed westward together with human migrations and partially replaced the indigenous Paleolithic ancient domesticates (Frantz et al. 2016). Opposing this hypothesis, the latter study observed genomic continuity and shared ancestry from the Early Neolithic to modern dog genomes in Europe, although substantial gene flow from Indian dogs was found in the Late Neolithic genome (Botigue et al. 2016). To clarify the long-lasting debate on dog domestication, sequencing more genomes from ancient dogs across Eurasia is critically needed.

3.2.4 Horse Domestication

The earliest archeological evidence of horse domestication dates back to ~5.5 kyr ago. However, the genetic processes underlying horse domestication remain enigmatic, largely because the near-extinct wild relatives are not available for comparative genomic studies. Several studies in recent years have investigated ancient horse genomes at a series of time points to reconstruct horse domestication history (Orlando et al. 2013; Schubert et al. 2014; Der Sarkissian et al. 2015; Librado et al. 2015). The first ancient horse genome was sequenced at approximately 1.12-fold coverage from a ~700-kyr-old permafrost-preserved fossil found at the Thistle Creek site, Canada (Orlando et al. 2013). The success of this study raised the upper limits of DNA survival in vertebrate remains by almost six times, from ~120 kyr BP, the oldest polar bear genome, to ~735 kyr BP, setting the current record for the oldest eukaryote genome ever sequenced. A

subsequent study implemented genome-wide scans of positive selection to uncover the genetic changes associated with domestication by comparing two high-coverage (7.4- and 24.3-fold) ancient genomes obtained from pre-domestication horses with modern genomes from a wide collection of domestic breeds (Schubert et al. 2014). This study identified a set of 125 loci that possibly have undergone domestication-related positive selection, including genes likely associated with physiological adaptations to human utilization, including cognitive changes associated with taming horses (Schubert et al. 2014). In addition, the study discovered that the modern domestic horse genome carries significantly higher deleterious mutation loads than pre-domestication horse genomes, reflecting the side effect of selective breeding or "cost of domestication," which is proposed to be caused by repetitive bottlenecks during the domestication process (Schubert et al. 2014). Similarly, another study demonstrated that a bottleneck associated with ~110 years of captivity has largely impacted the genetic diversity of Przewalski's horse (Der Sarkissian et al. 2015). Genomes of horse remains excavated from Yakutia were observed to have significant genetic discontinuity between a 5.2-kyr-old ancient genome and contemporary genomes, indicating that present-day Yakutian horses are genetically distinct from the now-extinct population that inhabited Yakutia in the mid-Holocene (Librado et al. 2015). The introduction of modern Yakutian horses possibly accompanied the migration of the Yakut people around the thirteenth to fifteenth centuries.

3.3 Plant Paleogenomics

During the last three decades of aDNA research, ancient plant remains in general have attracted much less attention than remains from ancient humans or vertebrates. Theoretically, ancient plant remains, most of which are seeds, pollen, and wood, should preserve aDNA well. There have been successful extractions of aDNA from various types of plant tissue preserved in charred, desiccated, or waterlogged conditions (Schlumbaum et al. 2008). Small-scale genetic analyses of ancient plant DNA have been performed on several domesticates, such as maize (*Zea mays*) (Jaenicke-Despres et al. 2003), barley (*Hordeum vulgare*) (Palmer et al. 2009), cotton (*Gossypium*) (Palmer et al. 2012), and wheat (*Triticum*) (Li et al. 2011; Bilgic et al. 2016), and have greatly informed research into the origin, domestication process, patterns of adaptation, and subsequent diffusion of crops. However, despite major advances in NGS technologies and aDNA protocols, progress toward reconstructing complete genomes from ancient plants has been lagging. The main challenge owes to a lack of modern reference genomes, a usually large genome size, and complex genome organization with large portions of repetitive elements and varying ploidy levels. Recently, the first attempts to obtain high-quality genomes of ancient barley and maize remains were finally successful.

3.3.1 Barley

As one of the founder crops of the Early Neolithic agricultural societies, domesticated barley (*Hordeum vulgare* L.) remains from ~10,000 BC were found at archeological sites in the Fertile Crescent (Zohary et al. 2012). A recent study reconstructed paleogenomes from five 6-kyr-old barley grains excavated at a cave in the Judean Desert in Israel (Mascher et al. 2016). Preserved under hot and arid conditions, the ancient desiccated barley grains contained sufficient amounts of endogenous DNA. The highest depth of coverage among the obtained paleogenomes was 20-fold, which is remarkable considering the barley reference genome (barley cv. Morex) is ~5.1Gb. Combining both ancient and modern sequence data, the results showed that the genomes of ancient domesticated barley appear remarkably similar to those of proximate extant landraces, although gene flow was observed between domesticated and sympatric wild populations. Thus, the major domestication events likely occurred more than 6 kyr ago.

3.3.2 Maize

The complex evolutionary history of maize (*Zea mays* L. ssp. *mays*) and its domestication, which occurred between 10 and 6.25 kyr ago in Mexico, has long been studied using genomic data from modern samples. Recent large-scale paleogenomic studies profoundly complemented the previous findings to clarify the process of domestication and early adaptations (da Fonseca et al. 2015; Ramos-Madrigal et al. 2016). One of the studies used a capture-based approach to acquire ~10× coverage of exons from 348 genes in 32 ancient Mexican maize samples dating back 0.75–6 kyr ago. Population genetic analysis identified several genes showing evidence of adaptation to arid Southwest American conditions, including genes relevant to drought tolerance and sugar content (da Fonseca et al. 2015). Another paleogenome at 1.7× coverage was successfully rebuilt from a 5.31-kyr-old maize cob excavated in the Tehuacán Valley of Mexico. Compared against modern landraces and the wild teosinte grasses, this ancient sample was demonstrated to represent the basal lineage to all modern varieties. The observation of a mix of both ancestral and derived states in genes related to domestication suggested that the ancient sample is an intermediate step between maize and teosinte, highlighting the gradual process of maize domestication.

3.4 Paleogenomics of Pathogenic Microorganisms

To date, high-quality paleogenomes of pathogenic microorganisms have been obtained for multiple infectious agents, including *Yersinia pestis* (Bos et al. 2011, 2016; Wagner et al. 2014; Rasmussen et al. 2015a, b; Feldman et al. 2016; Spyrou et al. 2016); *Mycobacterium leprae* (Schuenemann et al. 2013), a member of the *Mycobacterium tuberculosis* complex (Bos et al. 2014); a fungus-like eukaryote pathogen *Phytophthora*

infestans (Martin et al. 2013; Yoshida et al. 2013); *Vibrio cholera* (Devault et al. 2014a, b); a periodontal pathogen *Tannerella forsythia* (Warinner et al. 2014); and the smallpox *Variola* virus. Genomes of ancient pathogens have provided invaluable insights into the origin, phylogeography, pathogenicity, evolution, and adaptation of the targeted pathogen in emerging and reemerging infections. We here review some of these studies.

3.4.1 *Yersinia pestis*

In 2011, the first microbial paleogenome was reconstructed at 30-fold coverage for the bacterium *Yersinia pestis* from European Black Death victims dating to the fourteenth century. As one of the most virulent pathogens, *Y. pestis* is known to have been responsible for three major human plague pandemics throughout history, namely, the Plague of Justinian (sixth and eighth centuries AD), the second waves of pandemics (mid-fourteenth-century Black Death until the mid-eighteenth century AD), and the third pandemic (mid-nineteenth century until mid-twentieth century AD). However, deciphering the evolutionary history of *Y. pestis* by using molecular clocks has been greatly compromised due to extensive variation in nucleotide substitution rates among lineages (Cui et al. 2013; Wagner et al. 2014) and resulted in considerable uncertainty over the origin of this bacterium, how long it caused epidemic disease in human populations, and whether the lineages causing the three pandemics shared common etiologic agents. To answer these questions, several research groups reconstructed multiple *Y. pestis* paleogenomes from human skeletal remains dating back to the historical pandemics, as well as long before any recorded history of pandemics (Bos et al. 2011; Wagner et al. 2014; Rasmussen et al. 2015a, b; Bos et al. 2016; Feldman et al. 2016; Spyrou et al. 2016), and it was confirmed that *Y. pestis* was the causative agent for the historical pandemics. The lineages associated with the first pandemic were found to be different from the ones associated with the other pandemics and the modern lineage likely derived from the lineage associated with the second pandemic wave (Wagner et al. 2014; Feldman et al. 2016). It was also inferred that the most recent common ancestor (MRCA) of all *Y. pestis* lineages existed more than 5 kyr ago, and an ancient less-pathogenic lineage acquired the genetic changes making it highly virulent by 3 kyr ago (Rasmussen et al. 2015a, b).

3.4.2 *Mycobacterium tuberculosis*

Mycobacterium tuberculosis infections, also known as TB, were widespread in the past and still remain an emerging global threat. Although TB has had a long history with humans, the origin, the earliest hosts, and the evolutionary history of this disease still remain unclear. Ancient *M. tuberculosis* DNA fragments have been retrieved from samples from ancient Egypt, ancient Rome, and pre-Columbian America (Spigelman et al. 2015), allowing researchers to track genetic changes in the ancestral strains, as well as how they established and adapted in human populations. The significance of these aDNA studies was well illustrated by a paleogenomic analysis focusing on elucidating the transmission path of TB in the New World (Bos et al. 2014). Previous phylogeographic evidence based on modern strains of *M. tuberculosis* suggested that TB was introduced

post contact in the New World; however, abundant archeological evidence indicates the presence of tuberculosis in the New World before European contact (Roberts and Buikstra 2003; Bos et al. 2014). To resolve the conflict, three mycobacterial genomes were recovered from 1-kyr-old pre-Columbian Peruvian human skeletons, directly revealing that a member of the *M. tuberculosis* complex infected humans before contact (Bos et al. 2014). Meanwhile, dating approaches, which showed vastly different coalescence estimates in previous studies due to lack of ancient data, indicated a MRCA of the *M. tuberculosis* complex less than 6 kyr ago. Indeed, the history of TB has been proven to be far more complex than previously expected, and more ancient genomes recovered for *M. tuberculosis* in the near future will likely help clarify the phylogeographic history of TB.

3.4.3 *Mycobacterium leprae*

Leprosy, which results from infection with the pathogen *Mycobacterium leprae*, is another disease that afflicted humankind throughout history. The pathogen was widespread in Europe until it suddenly declined and essentially disappeared between the fourteenth and sixteenth centuries. To investigate the disappearance of leprosy from Europe and the relationship between ancient and modern strains, multiple genomes of *M. leprae* were obtained from tenth- to fourteenth-century human skeletal remains (Schuenemann et al. 2013; Mendum et al. 2014). Comparative genomics revealed significant genomic conservation during the past 1,000 years, suggesting the sudden decline of leprosy in the sixteenth century was not likely due to the loss of virulence in the ancient strain. Additionally, phylogenetic studies suggest a European origin for leprosy in the Americas and that a MRCA of all *M. leprae* strains existed ~3 kyr ago.

3.4.4 *Phytophthora infestans*

Phytophthora infestans, an oomycete that is the causative agent of potato late blight, was responsible for the catastrophic Irish potato famine and severe crop losses in the rest of Europe during the nineteenth century. Genetic information on the *P. infestans* strain that caused the Great Famine and its relationship to modern strains remained unknown until two research groups simultaneously reconstructed several paleogenomes of *P. infestans* from nineteenth-century herbarium collections (Martin et al. 2013; Yoshida et al. 2013). High endogenous DNA content retained in the herbarium collections allowed high-coverage assemblies of the relatively large genome (~240 Mb) of *P. infestans*. Interestingly, the historical strains are closely related to a modern genotype, but compared to modern strains, the historical strains that caused the Great Famine possessed a number of different infection-related genes. Presumably, these genes were gradually replaced by a

new set of infection-related genes due to selection pressure caused by introduction of resistance genes from wild potato relatives into the potato genome to fight against potato blight. Divergence time estimation indicated that the historical *P. infestans* strains only persisted for about 50 years until being replaced by the closely related modern genotype.

3.4.5 Variola

Compared to genome sequencing of ancient bacterial pathogens, paleogenomics of viruses is highly underdeveloped, largely due to the fragility of most viruses and the relatively low number of virions relative to host or environmental DNA in infected materials (Harkins and Stone 2015). Among the most devastating human diseases over the past hundreds of years, smallpox is the only disease eradicated by vaccination. This infectious disease is caused by the variola virus (VARA), and hypotheses on the origin of VARA and how it evolved following immunization date to more than two centuries ago but remain a matter of debate (Li et al. 2007; Babkin and Babkina 2015). Studies of ancient smallpox viral strains could be an exceptional opportunity to elucidate virus biology and evolution, but very few studies dedicated to ancient VARA have been successful. A ~180 kb draft genome of ancient VARA from a Lithuanian child mummy dating back to the seventeenth century showed strong conservation in gene content and arrangement between genomes from seventeenth- and twentieth-century strains, while molecular dating indicated that the VARA lineages eradicated during the twentieth century had only been in existence for 200 years (Duggan Ana et al. 2016). The impact of widespread vaccination on the selection pressure acting on virulence evolution remains uncertain but may be addressed with recovery of additional paleogenomes reconstructed from ancient strains.

3.5 Paleometagenomics

Since the initial analyses based on PCR and cloning of coprolites from the extinct Shasta ground sloth (*Nothrotheriops shastensis*) (Poinar et al. 1998), next-generation sequencing of DNA from environmental samples (eDNA), including the human microbiome (Adler et al. 2013; Warinner et al. 2014; Warinner et al. 2015), has taken aDNA research to the ecosystem level. Usually based on a metabarcoding approach, or high-throughput sequencing of amplicons targeting microbial, fungal, plant, and animal DNA, analysis of ancient DNA from sediments (sedaDNA) is a relatively new tool that can provide ecological insights into past environments and habitats, including faunal and floral changes over time (Pedersen et al. 2015; Thomsen and Willerslev 2015; Birks and Birks 2016). Particularly in very cold conditions, DNA can be preserved in sediments

that are hundreds of thousands of years old (Willerslev et al. 2003, 2007), but various other environments and time frames have also been analyzed (e.g., Haile et al. 2007; Parducci et al. 2012; Boessenkool et al. 2014; Willerslev et al. 2014; Epp et al. 2015; Smith et al. 2015; Alsos et al. 2016). Analysis of sedaDNA offers an important complement to classical morphological analysis to obtain a more complete picture of past biodiversity in paleoecological surveys (Parducci et al. 2015). For example, in a study of lake sediments from Svalbard going back over 8,000 years, all but two genera of vascular plants identified as macrofossils were identified with sedaDNA, which also identified additional taxa, including algae and bryophytes, and more species per sample (Alsos et al. 2016). The moderate changes in terrestrial vegetation over the Holocene suggested a resilience of the tundra flora in the face of climatic changes. Another study explored both nematode and plant diversity in 242 sediment samples from across the Arctic, spanning the last 50 kyr (Willerslev et al. 2014). This study showed that the Arctic vegetation was dominated by non-graminoid herbaceous vascular plants (forbs) until around the last glacial maximum where diversity declined markedly and after which the vegetation became dominated by woody plants and grasses. The authors suggested that this turnover may have been associated with the massive decline of megafaunal populations after the LGM (Willerslev et al. 2014). However, there are caveats using a metabarcoding approach based on universal primers and PCR, including potential bias toward preferential amplification of certain taxa and failure to detect particularly rare taxa (Pedersen et al. 2015). The extent of future promises of sedaDNA will likely depend on moving from metabarcoding approaches to true metagenomic shotgun sequencing.

4 Beyond the Genome

In addition to sequencing of paleogenomes, it has been demonstrated that proteins, epigenetic patterns in DNA, and RNA preserved in ancient specimens can also be used to infer evolutionary relationships and adaptation associated with molecular changes. Proteins, which were shown to be able to survive in fossils for periods of time orders of magnitude greater than for DNA (Collins et al. 2002; Torres et al. 2002; Palmqvist et al. 2003; Asara et al. 2007; Schweitzer et al. 2007; Organ et al. 2008; Schweitzer et al. 2009; Buckley and Collins 2011; Wadsworth and Buckley 2014), may provide a source of genetic information that is otherwise not accessible via aDNA analysis, especially in fossils beyond the limits of aDNA survival. Protein-derived information was initially retrieved in fossils using immunological approaches (Lowenstein 1980; Shoshani et al. 1985; Avci et al. 2005; Schweitzer et al. 2007). However, these approaches were considered unreliable owing to possible non-specific reactions with contaminants (Brandt et al. 2002). Recently, advances in protein sequencing and mass spectrometry techniques have enabled the complex mixture of proteins – the proteome – to be analyzed in ancient specimens.

Protein sequences have been obtained from many ancient samples, including dinosaurs (Asara et al. 2007; Organ et al. 2008; Schweitzer et al. 2009, 2013), Neanderthals (Nielsen-Marsh et al. 2005), South American ungulates (Buckley 2015; Welker et al. 2015), giant ground sloths (Buckley et al. 2015), and mammoth (Cappellini et al. 2012), as well as extant horses (Ostrom et al. 2006) and bovids (Wadsworth and Buckley 2014). Currently, ancient proteomic analysis mainly focuses on inferring evolutionary relationships from long-extinct organisms (Schweitzer et al. 2009; Buckley 2015; Buckley et al. 2015; Welker et al. 2015). For example, the close relationship between birds and dinosaurs proposed by morphological studies was confirmed by phylogenetic analysis of collagen sequences obtained from dinosaurs (Schweitzer et al. 2009).

Paleoepigenetics, which refers to the investigation of epigenetic modifications in ancient genomes, offers an opportunity to examine the evolutionary process of phenotypic plasticity and to gain insight into evolutionary processes that cannot be inferred from genetics alone. It has been demonstrated that DNA methylation patterns can survive in ancient specimens preserved under various environmental conditions and over a large temporal span, with examples including Neanderthals, Denisovans, ancient modern humans, bison, woolly mammoths, polar bears, equids, and barley (Briggs et al. 2010; Llamas et al. 2012; Gokhman et al. 2014; Pedersen et al. 2014; Smith et al. 2014; Seguin-Orlando et al. 2015; Hanghoj et al. 2016). Postmortem deamination converts methylated cytosine into thymine, while it converts non-methylated cytosine into uracil. Thus, an ancient epigenome can be reconstructed by discriminating between deaminated cytosine bases that become thymine versus those that become uracil. This strategy was successfully applied to generate a nucleosome map from a Saqqaq individual (Pedersen et al. 2014) and to reconstruct methylation maps from a Neanderthal and a Denisovan (Gokhman et al. 2014). Overall, the DNA methylation patterns of archaic humans and the Saqqaq individual do not differ significantly from modern humans, with the exception of the limb development-related HOXD gene clusters, of which the epigenetic changes may have played a key role in the recent evolution of human limbs (Gokhman et al. 2014). A genomic methylation profile was also investigated in archeological barley, showing an elevated degree of genomic methylation in one of the samples, which potentially was caused by environmental stress, such as viral infections (Smith et al. 2014). More recently, a bioinformatic pipeline (epiPALEOMIX) was developed to automate the characterization of ancient epigenomes. This pipeline was successfully applied to obtain methylation signatures from shotgun sequence data of 35 ancient genomes ranging from ancient humans to mammals (Hanghoj et al. 2016).

Ancient RNA represents a new dimension in the study of ancient biomolecules. Several studies, mostly from plant seeds, show that RNA can be preserved for decades to hundreds of years (Fordyce et al. 2013; Guy 2013; Ng et al. 2014; Smith et al. 2014; Guy and Gerard 2016). A partial transcriptome obtained from ancient maize kernels has offered an opportunity to directly test evolutionary changes in gene expression during the domestication process (Fordyce et al. 2013). The characterization of viral RNA genomes in ancient seeds may also shed light on species interactions (Guy 2013; Ng et al. 2014; Smith et al. 2014; Guy and Gerard 2016).

5 Conclusions and Future Perspectives

The bulk of paleogenomic research so far has largely focused on human evolution, from leading-edge studies of extinct archaic hominin genomes and their implications for our knowledge of anatomically modern human evolution to population genomic analyses of ancient remains of modern humans from the Americas and Eurasia. However, at an accelerating rate over the last decade, and paralleling advances in genome-scale sequencing technologies and increasing understanding of DNA damage patterns and ancient DNA recovery, ancient DNA research has also contributed to discoveries among microbes, plants, and other animals and has provided increasing insight into paleodiets and paleoenvironments. As methods continue to improve, and more complete and annotated genomes become available from living organisms that can function as references against which paleogenomes can be mapped, we will undoubtedly see increasing numbers of genomes from extinct species and populations as well. DNA degradation and exogenous contamination will continue to pose limitations to the extent of paleogenomic sequencing and assembly, but high-coverage archaic hominin genomes (Meyer et al. 2012; Prufer et al. 2014) that have brought the standard of paleogenomes to that of modern genomes give promise to the future of paleogenomic research. Future steps in technology, including transcriptomic, epigenetic, and proteomic analyses coupled with functional studies, may further bring paleogenomics toward the level of modern genomics. Nevertheless, studies of ancient DNA will continue to provide an unparalleled complement to modern genomics and multiple other disciplines. For example, recovery of ancient DNA unearthed an entire new lineage of extinct hominins, the Denisovans, based on genome sequencing alone (Meyer et al. 2012). Future genome sequencing of additional Denisovan and Neanderthal individuals coupled with epigenetic, transcriptional, and functional studies could offer remarkable insights into ancient human phenotypes and how they may have impacted the evolution of modern humans (Simonti et al. 2016). Similarly, future studies of ancient genomes, epigenomes, and transcriptomes from a multitude of other species, populations, and ecosystems can be expected to continue to expand our understanding of past environmental impacts in shaping the Earth's biodiversity and organismal demography, distributions, and adaptive responses and help inform current and future trends.

Finally, the resurrection of extinct species, also referred to as "de-extinction," is a new area of interest that has gained attention in recent years. For example, it has been suggested that resurrecting the mammoth has become a realistic prospect because of CRISPR/Cas9 (Cong et al. 2013; Mali et al. 2013), a novel gene-editing technology that could help create a hybrid embryo, in which mammoth traits such as small ears, subcutaneous fat, long shaggy hair, and cold-adapted blood would be programmed into an Asian elephant. However, multiple ethical issues associated with such efforts have been raised (Sherkow and Greely 2013; Sandler 2014). Future survival of such hybrid animals that resurrected species would likely represent would depend on available habitats, interactions with other species and their environment, and close monitoring and management, and most such animals would likely have to live a life in captivity. Knowing that elephants often do not fare well in captivity and that Asian elephants today are on the brink of extinction raises further questions regarding the ethics and potential success of woolly mammoth de-extinction experiments.

References

Adler CJ, Dobney K, Weyrich LS, Kaidonis J, Walker AW, Haak W, et al. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat Genet. 2013;45:450–5.

Allentoft ME, Sikora M, Sjogren KG, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of Bronze Age Eurasia. Nature. 2015;522:167–72.

Alsos IG, Sjogren P, Edwards ME, Landvik JY, Gielly L, Forwick M, et al. Sedimentary ancient DNA from Lake Skartjorna, Svalbard: assessing the resilience of arctic flora to Holocene climate change. Holocene. 2016;26:627–42.

Asara JM, Schweitzer MH, Freimark LM, Phillips M, Cantley LC. Protein sequences from mastodon and Tyrannosaurus rex revealed by mass spectrometry. Science. 2007;316:280–5.

Avci R, Schweitzer M, Boyd R, Wittmeyer J, Terán Arce F, Calvo J. Preservation of bone collagen from the late Cretaceous period studied by immunological techniques and atomic force microscopy. Langmuir. 2005;21:3584–90.

Avila-Arcos MC, Cappellini E, Romero-Navarro JA, Wales N, Moreno-Mayar JV, Rasmussen M, et al. Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. Sci Rep. 2011;1(74)

Babkin IV, Babkina IN. The origin of the variola virus. Virus. 2015;7:1100–12.

Barta JL, Monroe C, Teisberg JE, Winters M, Flanigan K, Kemp BM. One of the key characteristics of ancient DNA, low copy number, may be a product of its extraction. J Archaeol Sci. 2014;46:281–9.

Bennett EA, Massilani D, Lizzo G, Daligault J, Geigl EM, Grange T. Library construction for ancient genomics: single strand or double strand? Biotechniques. 2014;56:289–300.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53–9.

Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good JM. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics. 2012;13:403.

Bilgic H, Hakki EE, Pandey A, Khan MK, Akkaya MS. Ancient DNA from 8400 Year-Old Çatalhöyük wheat: implications for the origin of neolithic agriculture. PLoS One. 2016;11:e0151974.

Birks HJB, Birks HH. How have studies of ancient DNA from sediments contributed to the reconstruction of Quaternary floras? New Phytol. 2016;209:499–506.

Bodi K, Perera A, Adams P, Bintzler D, Dewar K, Grove D, et al. Comparison of commercially available target enrichment methods for next-generation sequencing. J Biomol Tech. 2013;24:73–86.

Boessenkool S, Mcglynn G, Epp LS, Taylor D, Pimentel M, Gizaw A, et al. Use of ancient sedimentary DNA as a novel conservation tool for high-altitude tropical biodiversity. Conserv Biol. 2014;28:446–55.

Boessenkool S, Hanghoj K, Nistelberger HM, Der Sarkissian C, Gondek A, Orlando L, et al. Combining bleach and mild pre-digestion improves ancient DNA recovery from bones. Mol Ecol Resour. 2016;17(4):742–51.

Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, et al. A draft genome of Yersinia pestis from victims of the Black Death. Nature. 2011;478:506–10.

Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature. 2014;514:494–7.

Bos KI, Herbig A, Sahl J, Waglechner N, Fourment M, Forrest SA, et al. Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an historical plague focus. Elife. 2016;5:e12994.

Botigue L, Song S, Scheu A, Gopalan S, Pendleton A, Oetjens M, et al. Ancient European dog genomes reveal continuity since the early Neolithic. BioRxiv. 2016:68189/68181–68189/68130

Brandt E, Wiechmann I, Grupe G. How reliable are immunological tools for the detection of ancient proteins in fossil bones? Int J Osteoarchaeol. 2002;12:307–16.

Bramanti B, Thomas MG, Haak W, Unterländer M, Jores P, Tambets K, et al. Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. Science. 2009;326 (5949):137–40.

Briggs AW, Heyn P. Preparation of next-generation sequencing libraries from damaged DNA. *Methods Mol Biol*. 2012;840:143–54.

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, et al. Targeted retrieval and analysis of five neandertal mtDNA genomes. Science. 2009;325:318–21.

Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Paabo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 2010;38:E87.

Buckley M. Ancient collagen reveals evolutionary history of the endemic South American "ungulates". Proc Biol Sci. 2015;282(1806):20142671.

Buckley M, Collins MJ. Collagen survival and its use for species identification in Holocene-lower Pleistocene bone fragments from British archaeological and paleontological sites. Antiqua. 2011;1:1.

Buckley M, Farina RA, Lawless C, Tambusso PS, Varela L, Carlini AA, et al. Collagen sequence analysis of the extinct giant ground sloths Lestodon and Megatherium. PLoS One. 2015;10(12): e0144793.

Burbano HA, Hodges E, Green RE, Briggs AW, Krause J, Meyer M, et al. Targeted investigation of the Neandertal genome by array-based sequence capture. Science. 2010;328:723–5.

Cano RJ, Borucki MK. Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. Science. 1995;268:1060–4.

Cano RJ, Poinar HN, Roubik DW, Poinar GO Jr. Enzymatic amplification and nucleotide sequencing of portions of the 18s rRNA gene of the bee Proplebeia dominicana (Apidae: Hymenoptera) isolated from 25-40 million year old Dominican amber. Med Sci Res. 1992;20:619–22.

Cano RJ, Poinar HN, Pieniezak NS, Poinar GO Jr. Enzymatic amplification and nucleotide sequencing of DNA from 120-135 million year old weevil. Nature. 1993;363:536–8.

Cappellini E, Jensen LJ, Szklarczyk D, Ginolhac A, da Fonseca RAR, Stafford TW, et al. Proteomic analysis of a pleistocene mammoth femur reveals more than one hundred ancient bone proteins. J Proteome Res. 2012;11:917–26.

Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, Sikora M, et al. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. Am J Hum Genet. 2013;93:852–64.

Castellano S, Parra G, Sanchez-Quinto FA, Racimo F, Kuhlwilm M, Kircher M, et al. Patterns of coding variation in the complete exomes of three Neandertals. Proc Natl Acad Sci U S A. 2014;111:6666–71.

Collins MJ, Nielsen-Marsh CM, Hiller J, Smith C, Roberts J, Prigodich R, et al. The survival of organic matter in bone: a review. Archaeometry. 2002;44:383–94.

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339:819–23.

Cosart T, Beja-Pereira A, Chen S, Ng SB, Shendure J, Luikart G. Exome-wide DNA capture and next generation sequencing in domestic and wild species. BMC Genomics. 2011;12:347.

Cruz-Dávalos DI, Llamas B, Gaunitz C, Fages A, Gamba C, Soubrier J, et al. Experimental conditions improving in-solution target enrichment for ancient DNA. Mol Ecol Resour. 2017;17(3):508–22.

Cui Y, Yu C, Yan Y, Li D, Li Y, Jombart T, et al. Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. Proc Natl Acad Sci. 2013;110:577–82.

Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci U S A. 2013;110:15758–63.

Damgaard PB, Margaryan A, Schroeder H, Orlando L, Willerslev E, Allentoft ME. Improving access to endogenous DNA in ancient bones and teeth. Sci Rep. 2015;5:11184.

Debruyne R, Chu G, King CE, Bos K, Kuch M, Schwarz C, et al. Out of America: ancient DNA evidence for a new world origin of late quaternary woolly mammoths. Curr Biol. 2008;18:1320–6.

Der Sarkissian C, Ermini L, Jonsson H, Alekseev AN, Crubezy E, Shapiro B, Orlando L. Shotgun microbial profiling of fossil remains. Mol Ecol. 2014;23:1780–98.

Der Sarkissian C, Ermini L, Schubert M, Yang MA, Librado P, Fumagalli M, et al. Evolutionary genomics and conservation of the endangered Przewalski's horse. Curr Biol. 2015;25:2577–83.
DeSalle R, Gatesy J, Wheeler W, Grimaldi D. DNA sequences from a fossil termite in Oligo-Miocene amber and their phylogenetic implications. Science. 1992;257:1933–6.
Devault AM, Golding GB, Waglechner N, Enk JM, Kuch M, Tien JH, et al. Second-pandemic strain of Vibrio cholerae from the Philadelphia cholera outbreak of 1849. N Engl J Med. 2014a;370:334–40.
Devault AM, McLoughlin K, Jaing C, Gardner S, Porter TM, Enk JM, et al. Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array. Sci Rep. 2014b;4:4245/4241–8.
Duggan Ana T, Perdomo Maria F, Piombino-Mascali D, Jankauskas R, Marciniak S, Poinar D, et al. 17 (th) century variola virus reveals the recent history of smallpox. Curr Biol. 2016;26(24):3407–12.
Elhaik E, Greenspan E, Staats S, Krahn T, Tyler-Smith C, Xue Y, et al. The GenoChip: a new tool for genetic anthropology. Genome Biol Evol. 2013;5:1021–31.
Enk J, Rouillard J-M, Poinar H. Quantitative PCR as a predictor of aligned ancient DNA read counts following targeted enrichment. Biotechniques. 2013;55:300–9.
Enk JM, Devault AM, Kuch M, Murgha YE, Rouillard J-M, Poinar HN. Ancient whole genome enrichment using baits built from modern DNA. Mol Biol Evol. 2014;31:1292–4.
Epp LS, Gussarova C, Boessenkool S, Olsen J, Haile J, Schroder-Nielsen A, et al. Lake sediment multi-taxon DNA from North Greenland records early post-glacial appearance of vascular plants and accurately tracks environmental changes. Quat Sci Rev. 2015;117:152–63.
Feldman M, Harbeck M, Keller M, Spyrou MA, Rott A, Trautmann B, et al. A high-coverage Yersinia pestis genome from a sixth-century justinianic plague victim. Mol Biol Evol. 2016;33:2911–23.
da Fonseca RR, Smith BD, Wales N, Cappellini E, Skoglund P, Fumagalli M, et al. The origin and evolution of maize in the Southwestern United States. Nat Plants. 2015;1:14003.
Fordyce SL, Avila-Arcos MC, Rasmussen M, Cappellini E, Romero-Navarro JA, Wales N, et al. Deep sequencing of RNA from ancient maize kernels. PLoS One. 2013;8:e50961.
Fortes GG, Paijmans JLA. Analysis of whole mitogenomes from ancient samples. Methods Mol Biol. 2015;1347:179–95.
Frantz LAF, Mullin VE, Pionnier-Capitan M, Lebrasseur O, Ollivier M, Perri A, et al. Genomic and archaeological evidence suggest a dual origin of domestic dogs. Science. 2016;352:1228–31.
Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, et al. Genome sequencing highlights the dynamic early history of dogs. PLoS Genet. 2014;10:e1004016.
Fu QM, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, Paabo S. DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci U S A. 2013;110:2223–7.
Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014;514:445–9.
Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, et al. The genetic history of Ice Age Europe. Nature. 2016;534:200–5.
Gamba C, Hanghoj K, Gaunitz C, Alfarhan AH, Alquraishi SA, Al-Rasheid KAS, et al. Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. Mol Ecol Resour. 2016;16:459–69.
Gansauge M-T, Meyer M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat Protoc. 2013;8:737–48.
Gilbert MTP, Tomsho LP, Rendulic S, Packard M, Drautz DI, Sher A, et al. Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. Science. 2007;317:1927–30.
Gilbert MTP, Drautz DI, Lesk AM, Ho SYW, Qi J, Ratan A, et al. Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. Proc Natl Acad Sci U S A. 2008;105:8327–32.
Ginolhac A, Vilstrup J, Stenderup J, Rasmussen M, Stiller M, Shapiro B, et al. Improving the performance of true single molecule sequencing for ancient DNA. BMC Genomics. 2012;13:177.
Gokhman D, Lavi E, Prufer K, Fraga MF, Riancho JA, Kelso J, et al. Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. Science. 2014;344:523–7.

Golenberg EM, Giannasi DE, Clegg MT, Smiley CJ, Durbin M, Henderson D, Zurawski G. Chloroplast DNA sequence from a miocene Magnolia species. Nature. 1990;344:656–8.
Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, et al. Analysis of one million base pairs of Neanderthal DNA. Nature. 2006;444:330–6.
Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the neandertal genome. Science. 2010;328:710–22.
Guy PL. Ancient RNA? RT-PCR of 50-year-old RNA identifies peach latent mosaic viroid. Arch Virol. 2013;158:691–4.
Guy P, Gerard P. White clover cryptic virus-1 in New Zealand and eastern Australia. Ann Appl Biol. 2016;168:225–31.
Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522:207–11.
Haile J, Holdaway R, Oliver K, Bunce M, Gilbert MTP, Nielsen R, et al. Ancient DNA chronology within sediment deposits: are paleobiological reconstructions possible and is DNA leaching a factor? Mol Biol Evol. 2007;24:982–9.
Hanghoj K, Seguin-Orlando A, Schubert M, Madsen T, Pedersen JS, Willerslev E, Orlando L. Fast, accurate and automatic ancient nucleosome and methylation maps with epiPALEOMIX. Mol Biol Evol. 2016;33:3284–98.
Harkins KM, Stone AC. Ancient pathogen genomics: insights into timing and adaptation. J Hum Evol. 2015;79:137–49.
Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC. DNA sequences from the quagga, an extinct member of the horse family. Nature. 1984;312:282–4.
Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet. 2007;39:1522–7.
Hofmanova Z, Kreutzer S, Hellenthal G, Sell C, Diekmann Y, Diez-del-Molino D, et al. Early farmers from across Europe directly descended from Neolithic Aegeans. Proc Natl Acad Sci U S A. 2016;113:6886–91.
Hofreiter M, Serre D, Poinar HN, Kuch M, Paabo S. Ancient DNA. Nat Rev Genet. 2001;2:353–9.
Hofreiter M, Paijmans JLA, Goodchild H, Speller CF, Barlow A, Fortes GG, et al. The future of ancient DNA: technical advances and conceptual shifts. Bioessays. 2015;37:284–93.
Jaenicke-Despres V, Buckler ES, Smith BD, Gilbert MTP, Cooper A, Doebley J, Paabo S. Early allelic selection in maize as revealed by ancient DNA. Science. 2003;302:1206–8.
Jeong C, Ozga AT, Witonsky DB, Malmstrom H, Edlund H, Hofman CA, et al. Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc. Proc Natl Acad Sci U S A. 2016;113:7485–90.
Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, et al. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun. 2012;3:698.
Kim BY, Lohmueller KE. Selection and reduced population size cannot explain higher amounts of Neandertal ancestry in East Asian than in European human populations. Am J Hum Genet. 2015;96:454–61.
Kistler L. Ancient DNA extraction from plants. Methods Mol Biol. 2012;840:71–9.
Korlevic P, Gerber T, Gansauge MT, Hajdinjak M, Nagel S, Aximu-Petri A, Meyer M. Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. Biotechniques. 2015;59:87–93.
Krause J, Dear PH, Pollack JL, Slatkin M, Spriggs H, Barnes I, et al. Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. Nature. 2006;439:724–7.
Krause J, Briggs AW, Kircher M, Maricic T, Zwyns N, Derevianko A, Paabo S. A complete mtDNA genome of an early modern human from Kostenki, Russia. Curr Biol. 2010;20:231–6.
Lan T, Cheng J, Ratan A, Miller W, Schuster S, Farley S, et al. Genome-wide evidence for a hybrid origin of modern polar bears. BioRxiv. 2016:047498.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513:409–13.

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016;536:419–24.
Li Y, Carroll DS, Gardner SN, Walsh MC, Vitalis EA, Damon IK. On the origin of smallpox: correlating variola phylogenics with historical smallpox records. Proc Natl Acad Sci. 2007;104:15787–92.
Li CX, Lister DL, Li HJ, Xu Y, Cui YQ, Bower MA, et al. Ancient DNA analysis of desiccated wheat grains excavated from a Bronze Age cemetery in Xinjiang. J Archaeol Sci. 2011;38:115–9.
Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJ. Capturing protein-coding genes across highly divergent species. Biotechniques. 2013;54:321–6.
Librado P, Sarkissian CD, Ermini L, Schubert M, Jonsson H, Albrechtsen A, et al. Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. Proc Natl Acad Sci U S A. 2015;112:E6889–97.
Lindahl T. Instability and decay of the primary structure of DNA. Nature. 1993a;362:709–15.
Lindahl T. Recovery of antediluvian DNA. Nature. 1993b;365:700.
Llamas B, Holland ML, Chen K, Cropley JE, Cooper A, Suter CM. High-resolution analysis of cytosine methylation in ancient DNA. PLoS One. 2012;7:e30226.
Llorente MG, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, et al. Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science. 2015;350:820–2.
Lowenstein JM. Species-specific proteins in fossils. Naturwissenschaften. 1980;67:343–6.
Lynch VJ, Bedoya-Reina OC, Ratan A, Sulak M, Drautz-Moses DI, Perry GH, et al. Elephantid genomes reveal the molecular bases of woolly mammoth adaptations to the arctic. Cell Rep. 2015;12:217–28.
Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science. 2013;339:823–6.
Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376–80.
Maricic T, Paabo S. Optimization of 454 sequencing library preparation from small amounts of DNA permits sequence determination of both DNA strands. Biotechniques. 2009;46:51–52, 54–57.
Maricic T, Whitten M, Paabo S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS One. 2010;5:e14004.
Martin MD, Cappellini E, Samaniego JA, Zepeda ML, Campos PF, Seguin-Orlando A, et al. Reconstructing genome evolution in historic samples of the Irish potato famine pathogen. Nat Commun. 2013;4:2172.
Mascher M, Schuenemann VJ, Davidovich U, Marom N, Himmelbach A, Hubner S, et al. Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. Nat Genet. 2016;48:1089–93.
Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528:499–503.
Mendum TA, Schuenemann VJ, Roffey S, Taylor GM, Wu HH, Singh P, et al. Mycobacterium leprae genomes from a British medieval leprosy hospital: towards understanding an ancient epidemic. BMC Genomics. 2014;15:270.
Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010;2010(6):pdb.prot5448.
Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic denisovan individual. Science. 2012;338:222–6.
Meyer M, Fu Q, Aximu-Petri A, Glocke I, Nickel B, Arsuaga J-L, et al. A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature. 2014;505:403–6.
Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, et al. Sequencing the nuclear genome of the extinct woolly mammoth. Nature. 2008;456:387–90.
Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina OC, Zhao F, et al. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proc Natl Acad Sci. 2012;109:E2382–90.

Mohandesan E, Speller CF, Peters J, Uerpmann HP, Uerpmann M, De Cupere B, et al. Combined hybridization capture and shotgun sequencing for ancient DNA analysis of extinct wild and domestic dromedary camel. Mol Ecol Resour. 2016;17(2):300–13.

Ng TFF, Chen LF, Zhou YC, Shapiro B, Stiller M, Heintzman PD, et al. Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. Proc Natl Acad Sci U S A. 2014;111:16842–7.

Nielsen-Marsh CM, Richards MP, Hauschka PV, Thomas-Oates JE, Trinkaus E, Pettitt PB, et al. Osteocalcin protein sequences of Neanderthals and modern primates. Proc Natl Acad Sci U S A. 2005;102:4409–13.

Noonan JP, Hofreiter M, Smith D, Priest JR, Rohland N, Rabeder G, et al. Genomic sequencing of pleistocene cave bears. Science. 2005;309:597–9.

Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, et al. Sequencing and analysis of Neanderthal genomic DNA. Science. 2006;314:1113–8.

Okello JB, Zurek J, Devault AM, Kuch M, Okwi AL, Sewankambo NK, et al. Comparison of methods in the recovery of nucleic acids from archival formalin-fixed paraffin-embedded autopsy tissues. Anal Biochem. 2010;400:110–7.

Organ CL, Schweitzer MH, Zheng W, Freimark LM, Cantley LC, Asara JM. Molecular phylogenetics of mastodon and Tyrannosaurus rex. Science. 2008;320:499.

Orlando L, Ginolhac A, Raghavan M, Vilstrup J, Rasmussen M, Magnussen K, et al. True single-molecule DNA sequencing of a pleistocene horse bone. Genome Res. 2011;21:1705–19.

Orlando L, Ginolhac A, Zhang GJ, Froese D, Albrechtsen A, Stiller M, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature. 2013;499:74–8.

Orlando L, Gilbert MTP, Willerslev E. Reconstructing ancient genomes and epigenomes. Nat Rev Genet. 2015;16:395–408.

Ostrom PH, Gandhi H, Strahler JR, Walker AK, Andrews PC, Leykam J, et al. Unraveling the sequence and structure of the protein osteocalcin from a 42 ka fossil horse. Geochim Cosmochim Acta. 2006;70:2034–44.

Pääbo S. Molecular cloning of ancient Egyptian mummy DNA. Nature. 1985;314:644–5.

Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, et al. Genetic analyses from ancient DNA. Annu Rev Genet. 2004;38:645–79.

Paijmans JLA, Fickel J, Courtiol A, Hofreiter M, Forster DW. Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. Mol Ecol Resour. 2016;16:42–55.

Palkopoulou E, Mallick S, Skoglund P, Enk J, Rohland N, Li H, et al. Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. Curr Biol. 2015;25:1395–400.

Palmer SA, Moore JD, Clapham AJ, Rose P, Allaby RG. Archaeogenetic evidence of ancient nubian barley evolution from six to two-row indicates local adaptation. PLoS One. 2009;4:e6301.

Palmer SA, Clapham AJ, Rose P, Freitas FO, Owen BD, Beresford-Jones D, et al. Archaeogenomic evidence of punctuated genome evolution in gossypium. Mol Biol Evol. 2012;29:2031–8.

Palmqvist P, Gröcke DR, Arribas A, Fariña RA. Paleoecological reconstruction of a lower Pleistocene large mammal community using biogeochemical (δ13C, δ15N, δ18O, Sr: Zn) and ecomorphological approaches. Paleobiology. 2003;29:205–29.

Parducci L, Jorgensen T, Tollefsrud MM, Elverland E, Alm T, Fontana SL, et al. Glacial survival of boreal trees in northern Scandinavia. Science. 2012;335:1083–6.

Parducci L, Valiranta M, Salonen JS, Ronkainen T, Matetovici I, Fontana SL, et al. Proxy comparison in ancient peat sediments: pollen, macrofossil and plant DNA. Philos Trans R Soc Lond B Biol Sci. 2015;370:20130382.

Park SDE, Magee DA, McGettigan PA, Teasdale MD, Edwards CJ, Lohan AJ, et al. Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. Genome Biol. 2015;16:234.

Pedersen JS, Valen E, Velazquez AMV, Parker BJ, Rasmussen M, Lindgreen S, et al. Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. Genome Res. 2014;24:454–66.

Pedersen MW, Overballe-Petersen S, Ermini L, Sarkissian CD, Haile J, Hellstrom M, et al. Ancient and modern environmental DNA. Philos Trans R Soc Lond B Biol Sci. 2015;370(0130383)

Pilot M, Malewski T, Moura AE, Grzybowski T, Olenski K, Rusc A, et al. On the origin of mongrels: evolutionary history of free-breeding dogs in Eurasia. Proc Biol Sci. 2015;282:20152189.
Pinhasi R, Fernandes D, Sirak K, Novak M, Connell S, Alpaslan-Roodenberg S, et al. Optimal ancient DNA yields from the inner ear part of the human petrous bone. PLoS One. 2015;10(6): e0129102.
Poinar HN, Stankiewicz BA. Protein preservation and DNA retrieval from ancient tissues. Proc Natl Acad Sci U S A. 1999;96:8426–31.
Poinar HN, Hofreiter M, Spaulding WG, Martin PS, Stankiewicz BA, Bland H, et al. Molecular coproscopy: dung and diet of the extinct ground sloth Nothrotheriops shastensis. Science. 1998;281:402–6.
Poinar HN, Schwarz C, Qi J, Shapiro B, MacPhee RDE, Buigues B, et al. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science. 2006;311:392–4.
Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505:43–9.
Raghavan M, DeGiorgio M, Albrechtsen A, Moltke I, Skoglund P, Korneliussen TS, et al. The genetic prehistory of the New World Arctic. Science. 2014a;345(6200):1255832.
Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 2014b;505:87–91.
Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, et al. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science. 2015;349:aab3884.
Ramos-Madrigal J, Smith BD, Moreno-Mayar JV, Gopalakrishnan S, Ross-Ibarra J, Gilbert MTP, Wales N. Genome sequence of a 5,310-year-old maize cob provides insights into the early stages of maize domestication. Curr Biol. 2016;26(23):3195–201.
Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010;463:757–62.
Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW Jr, et al. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature. 2014;506:225–9.
Rasmussen M, Sikora M, Albrechtsen A, Korneliussen TS, Moreno-Mayar JV, Poznik GD, et al. The ancestry and affiliations of Kennewick Man. Nature. 2015a;523(7561):455–8.
Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjogren K-G, et al. Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago. Cell. 2015b;163:571–82.
Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468:1053–60.
Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, et al. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet. 2011;89:516–28.
Ristaino JB, Groves CT, Parra GR. PCR amplification of the Irish potato famine pathogen from historic specimens. Nature. 2001;411:695–7.
Roberts CA. Buikstra JE. The bioarchaeology of tuberculosis: a global view on a reemerging disease. Gainesville: University Press of Florida; 2003
Rogaev EI, Moliaka YK, Malyarchuk BA, Kondrashov FA, Derenko MV, Chumakov I, Grigorenko AP. Complete mitochondrial genome and phylogeny of Pleistocene mammoth Mammuthus primigenius. PLoS Biol. 2006;4:e73.
Rogers RL, Slatkin M. Genomic disintegration in woolly mammoths on Wrangel island. arXiv.org, e-Print Arch., Quant Biol. 2016:1–32
Rohland N, Hofreiter M. Ancient DNA extraction from bones and teeth. Nat Protoc. 2007;2:1756–62.
Rohland N, Siedel H, Hofreiter M. A rapid column-based ancient DNA extraction method for increased sample throughput. Mol Ecol Resour. 2010;10:677–83.
Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. Partial uracil—DNA—glycosylase treatment for screening of ancient DNA. Philos Trans R Soc Lond B Biol Sci. 2015;370(1660):20130624.
Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H, Arnheim N. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. 1985;230:1350–4.
Sandler R. The ethics of reviving long extinct species. Conserv Biol. 2014;28:354–60.

Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. The date of interbreeding between Neandertals and modern humans. PLoS Genet. 2012;8:e1002947.
Sankararaman S, Mallick S, Patterson N, Reich D. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. Curr Biol. 2016;26:1241–7.
Schlumbaum A, Tensen M, Jaenicke-Despres V. Ancient plant DNA in archaeobotany. Veg Hist Archaeobotany. 2008;17:233–44.
Schubert M, Jonsson H, Chang D, Der Sarkissian C, Ermini L, Ginolhac A, et al. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. Proc Natl Acad Sci U S A. 2014;111:E5661–9.
Schuenemann VJ, Singh P, Mendum TA, Krause-Kyora B, Jager G, Bos KI, et al. Genome-wide comparison of medieval and modern Mycobacterium leprae. Science. 2013;341:179–83.
Schurr TG, Sherry ST. Mitochondrial DNA and Y chromosome diversity and the peopling of the Americas: evolutionary and demographic evidence. Am J Hum Biol. 2004;16:420–39.
Schweitzer MH, Suo Z, Avci R, Asara JM, Allen MA, Arce FT, Horner JR. Analyses of soft tissue from Tyrannosaurus rex suggest the presence of protein. Science. 2007;316:277–80.
Schweitzer MH, Zheng W, Organ CL, Avci R, Suo Z, Freimark LM, et al. Biomolecular characterization and protein sequences of the Campanian hadrosaur B. canadensis. Science. 2009;324:626–31.
Schweitzer MH, Zheng WX, Cleland TP, Bern M. Molecular analyses of dinosaur osteocytes support the presence of endogenous molecules. Bone. 2013;52:414–23.
Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas AS, Manica A, Moltke I, et al. Genomic structure in Europeans dating back at least 36,200 years. Science. 2014;346:1113–8.
Seguin-Orlando A, Gamba C, Der Sarkissian C, Ermini L, Louvel G, Boulygina E, et al. Pros and cons of methylation-based enrichment methods for ancient DNA. Sci Rep. 2015;5:11826.
Shannon LM, Boyko RH, Castelhano M, Corey E, Hayward JJ, McLean C, et al. Genetic structure in village dogs reveals a Central Asian domestication origin. Proc Natl Acad Sci. 2015;112:13639–44.
Sherkow JS, Greely HT. What if extinction is not forever? Science. 2013;340:32–3.
Shoshani J, Lowenstein JM, Walz DA, Goodman M. Proboscidean origins of mastodon and woolly mammoth demonstrated immunologically. Paleobiology. 1985;11:429–37.
Simonti CN, Vernot B, Bastarache L, Bottinger E, Carrell DS, Chisholm RL, et al. The phenotypic legacy of admixture between modern humans and Neandertals. Science. 2016;351:737–41.
Skoglund P, Malmstroem H, Raghavan M, Stora J, Hall P, Willerslev E, et al. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science. 2012;336:466–9.
Smith CI, Chamberlain AT, Riley MS, Cooper A, Stringer CB, Collins MJ. Neanderthal DNA: not just old but old and cold? Nature. 2001;410:771–2.
Smith O, Clapham AJ, Rose P, Liu Y, Wang J, Allaby RG. Genomic methylation patterns in archaeological barley show de-methylation as a time-dependent diagenetic process. Sci Rep. 2014;4:5559.
Smith O, Momber G, Bates R, Garwood P, Fitch S, Pallen M, et al. Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago. Science. 2015;347:998–1001.
Soltis PS, Soltis DE, Smiley CJ. An rbcL sequence from a Miocene Taxodium (bald cypress). Proc Natl Acad Sci U S A. 1992;89:449–51.
Spigelman M, Donoghue HD, Abdeen Z, Ereqat S, Sarie I, Greenblatt CL, et al. Evolutionary changes in the genome of Mycobacterium tuberculosis and the human genome from 9000 years BP until modern times. Tuberculosis. 2015;95:S145–9.
Spyrou MA, Tukhbatova RI, Feldman M, Drath J, Kacki S, Beltran de Heredia J, et al. Historical Y. pestis genomes reveal the European Black Death as the source of ancient and modern plague pandemics. Cell Host Microbe. 2016;19:874–81.
Stringer C. Human evolution: small remains still pose big problems. Nature. 2014;514:427–9.
Stuart AJ, Kosintsev P, Higham T, Lister AM. Pleistocene to Holocene extinction dynamics in giant deer and woolly mammoth. Nature. 2004;431:684–9.
Thomas M, Gilbert P. Postmortem damage of mitochondrial DNA. Nucleic Acids Mol Biol. 2006;18:91–115.
Thomsen PF, Willerslev E. Environmental DNA—an emerging tool in conservation for monitoring past and present biodiversity. Biol Conserv. 2015;183:4–18.

Torres JM, Borja C, Olivares EG. Immunoglobulin G in 1.6 million-year-old fossil bones from Venta Micena (Granada, Spain). J Archaeol Sci. 2002;29:167–75.
Vartanyan SL, Arslanov KA, Karhu JA, Possnert G, Sulerzhitsky LD. Collection of radiocarbon dates on the mammoths (Mammuthus primigenius) and other genera of Wrangel Island, northeast Siberia, Russia. Quatern Res. 2008;70:51–9.
Vernot B, Akey JM. Complex history of admixture between modern humans and Neandertals. Am J Hum Genet. 2015;96:448–53.
Wadsworth C, Buckley M. Proteome degradation in fossils: investigating the longevity of protein survival in ancient bone. Rapid Commun Mass Spectrom. 2014;28:605–15.
Wagner DM, Klunk J, Harbeck M, Devault A, Waglechner N, Sahl JW, et al. Yersinia pestis and the Plague of Justinian 541–543 AD: a genomic analysis. Lancet Infect Dis. 2014;14:319–26.
Wall JD, Yang MA, Jay F, Kim SK, Durand EY, Stevison LS, et al. Higher levels of neanderthal ancestry in East Asians than in Europeans. Genetics. 2013;194(1):199–209.
Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, et al. Pathogens and host immunity in the ancient human oral cavity. Nat Genet. 2014;46:336–44.
Warinner C, Speller C, Collins MJ. A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome. Philos Trans R Soc Lond B Biol Sci. 2015;370(1660):20130376.
Welker F, Collins MJ, Thomas JA, Wadsley M, Brace S, Cappellini E, et al. Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. Nature. 2015;522:81–U192.
Willerslev E, Hansen AJ, Binladen J, Brand TB, Gilbert MTP, Shapiro B, et al. Diverse plant and animal genetic records from Holocene and Pleistocene sediments. Science. 2003;300:791–5.
Willerslev E, Hansen AJ, Poinar HN. Isolation of nucleic acids and cultures from fossil ice and permafrost. Trends Ecol Evol. 2004;19:141–7.
Willerslev E, Cappellini E, Boomsma W, Nielsen R, Hebsgaard MB, Brand TB, et al. Ancient biomolecules from deep ice cores reveal a forested southern Greenland. Science. 2007;317:111–4.
Willerslev E, Davison J, Moora M, Zobel M, Coissac E, Edwards ME, et al. Fifty thousand years of arctic vegetation and megafaunal diet. Nature. 2014;506:47–51.
Woodward NW, Bunnell M. DNA sequence from Cretaceous period bone fragments. Science. 1994;266:1229–32.
Yoshida K, Schuenemann VJ, Cano LM, Pais M, Mishra B, Sharma R, et al. The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. Elife. 2013;2:e00731.
Zhang HC, Paijmans JLA, Chang FQ, Wu XH, Chen GJ, Lei CZ, et al. Morphological and genetic evidence for early Holocene cattle management in northeastern China. Nat Commun. 2013;4:2755.
Zohary D, Hopf M, Weiss E. Domestication of Plants in the Old World. The origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin. Oxford University Press on Demand; 2012

Genome-Wide Association Studies and Heritability Estimation in the Functional Genomics Era

Dunia Pino Del Carpio, Roberto Lozano, Marnin D. Wolfe, and Jean-Luc Jannink

Abstract Genome-wide association studies (GWAS) are designed to detect the statistical association between genomic markers and phenotypic data in order to identify loci that control complex traits and more recently to quantify the relative amount of trait variance that arises from genetic sources. Moreover, many genomic resources have been generated and analytical tools developed to bring together information linking GWAS results to causal variants. This book chapter is an incredible effort to bring together information about current aspects of genome-wide studies and the concept of heritability. In the first section of this book chapter, we discuss the most critical concepts and experimental considerations in order to follow GWAS. In the later sections, we explore how researchers are trying to answer the question of whether using functional genomic data can improve the power of GWAS in complex phenotypes and if so far has led us to important biological insights. We review the core concept of heritability, its practical applications, and the classical (pre-genomics) methods for measurement, which largely remain relevant. Finally, we outline the genomic resources available for GWA studies. Also, based on what is available for humans, we identify what are the most critical

D. Pino Del Carpio (✉)
Agriculture Research Division, Agriculture Victoria, Melbourne, VIC, Australia

R. Lozano
Plant Breeding and Genetics, School for Integrative Plant Science, Cornell University, Ithaca, NY, USA
e-mail: rjl278@cornell.edu

M. D. Wolfe
Section on Plant Breeding and Genetics, School of Integrative Plant Sciences, Cornell University, Ithaca, NY, USA
e-mail: mw489@cornell.edu

J. -L. Jannink
United States Department of Agriculture, Agricultural Research Service, R.W. Holley Center for Agriculture and Health, Ithaca, NY, USA
e-mail: jeanluc.jannink@ars.usda.gov

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_12,

resources that need to be developed for other species by contrasting the human genomic resources with resources being developed in plant and animal models.

Keywords Data mining · Functional genomics · Genomics · Genome-wide association study · Heritability · Meta-analysis · Network · Pathways · Post-GWAS

1 Introduction

Many important traits in humans, animals, and agricultural crops have a complex nature: they are controlled by many genes and by environmental factors. The genotype-phenotype connection has been studied, i.e., through linkage mapping approaches, such as quantitative trait locus (QTL mapping) in biparental crosses and genome-wide association studies (GWAS) in diversity and crossing panels.

To unravel the genetics of complex diseases, genome-wide association studies (GWAS) have become one of the most important genomics and statistical approaches to increase our understanding of the biology of complex traits through the identification of how many loci control a trait and the estimation of effects of polymorphisms on a trait. In human genetics, GWAS is commonly applied to identify genetic risk factors and understand the genetic basis of common diseases, and in plants it is a breeding tool for the identification of markers for marker-assisted selection.

In order to identify marker-trait associations, linkage disequilibrium (LD) has to occur in the chosen mapping population/diversity set. LD is an important concept in population genetics because it summarizes the genetic variation that occurred within a population through its evolutionary history. LD is a statistical measure of genetic distance among genetic variants that is dynamically affected over time. The extent of chromosomal linkage is affected by several factors such as natural selection, genetic drift, population subdivision and bottlenecks, inbreeding, inversions, and gene conversion. LD decay over time reflects the history of recombination, and it can affect our ability to precisely identify disease/trait variants; long-distance linkage disequilibrium lowers mapping resolution, while short-distance linkage disequilibrium increases the resolution to identify QTL regions.

Among the key factors that have increased the power of genome-wide studies are high-throughput phenotyping, availability of genome-wide sequence data, large study samples, and unbiased analytical tools. In many GWAS, marker polymorphisms have been detected through genotyping by sequencing or SNP arrays that represent a subset of the whole-genome polymorphism present in a species. Further, whole-genome sequencing and methodological developments such as genotype imputation, which relies on nearby allelic LD, have dramatically increased the number of variants available for association studies. In these studies, a marker is declared to be significantly associated with a trait, because first the marker passed a stringent level of statistical significance and, second, because the marker can be *directly* causative or *indirectly* associated with a trait due to its high LD with a causal variant.

Although the list of putative SNPs associated with traits has incremented, it is still difficult to unravel the underlying biology behind these associations. After GWAS reveals a number of statistically significant associations, the next obvious step has been to accumulate evidence that will help the researcher to prioritize variants for follow-up studies. The question of how to discern causal from LD-related variants within a haplotype block has deserved special attention in order to avoid false positives.

Prioritization approaches include, among others, the incorporation of diverse sources of omics data information, in silico genome annotation (including noncoding regulatory regions), and pooling samples for meta-analyses.

In brief, the design of a GWA study requires the following common steps: (1) scanning of genetic and phenotypic variation in a sample population; (2) selection of a statistical model with the inclusion of a correction, for population structure or relatedness to remove spurious associations; (3) identification of associated and linked variants; and (4) follow-up post-GWAS sustained by biological evidence and functional annotation (Fig. 1). More recent studies, for validation purposes, specifically target candidate functional SNPs through genomic engineering approaches, i.e., genome editing (Sander and Joung 2014), to compare the effects of modified alleles on a trait.

Due to the steady reduction in genotyping costs, the creation of large research consortia, and database development, massive amounts of genome-wide data are currently available for scientists to uncover the genetic basis of complex traits at a large scale (see later sections). In this book chapter, we cover the basis of GWAS providing a description of emerging advances and considerations of experimental design, statistical methods, functional annotation, post-GWAS and meta-analyses, and genomic resources, and we discuss the use of GWAS models to estimate the amount of heritable genetic variation. We finally highlight the advantages of GWAS for the identification of causal genes, describe its limitations, and provide future directives and perspectives to follow genome-enabled studies.

2 Experimental Design of Genome-Wide Association Studies

To follow a genome wide association study it is important to consider the following: traits more amenable to genome-wide association studies, sample size and allelic diversity, phenotyping, and population stratification and relatedness. Following on these considerations, it seems likely that for most complex traits increasing sample size, to capture smaller effect or rarer genetic variants, improving phenotype quantification, carefully considering genetic diversity panels, and scaling trait capturing methods will increase the power to detect genetic signals. In the following section, we briefly describe the experimental design parameters and statistical considerations to follow genome-wide association studies.

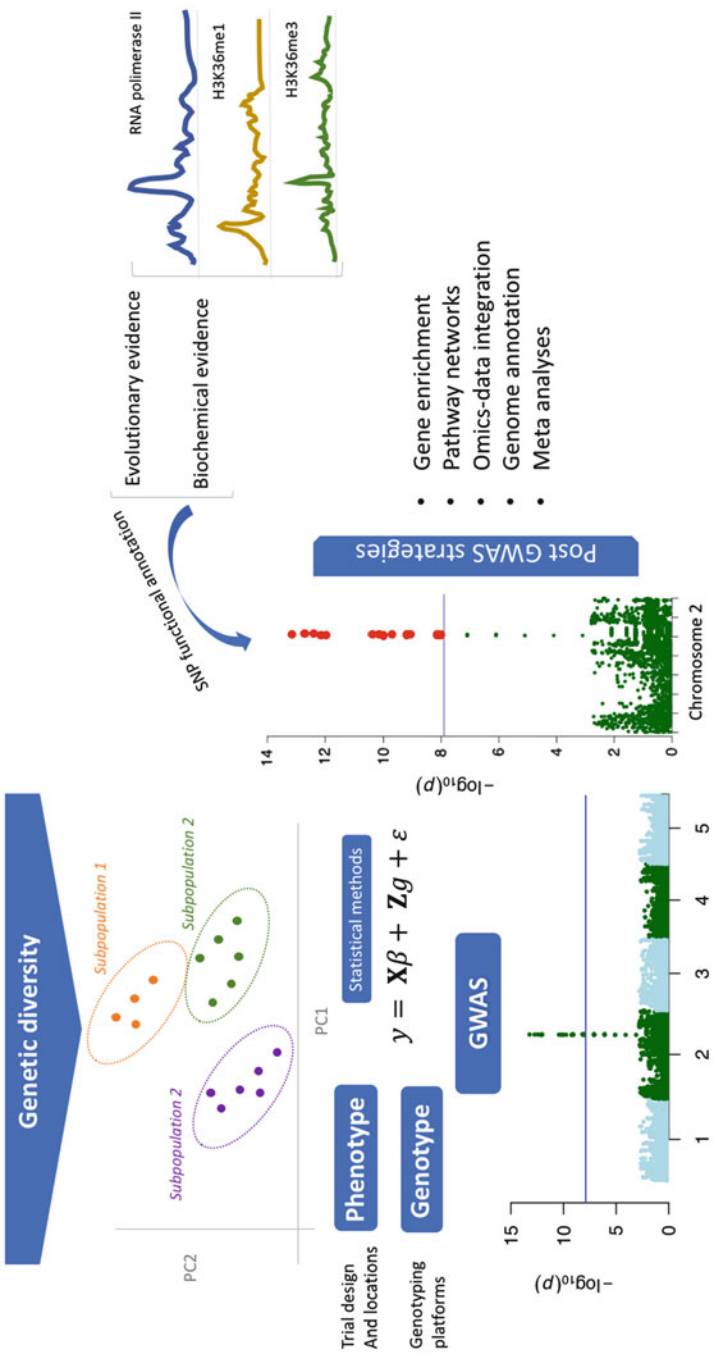

Fig. 1 Schematic representation of a genome-wide association study

2.1 Sample Size and Allelic Diversity

Under the infinitesimal model, quantitative traits are controlled by many loci each with an infinitely small effect (Fisher 1918; Bulmer 1971). Genome-wide association studies (GWAS) have been successful in identifying common variants (by convention those with allele frequency above 5%) associated with common diseases or quantitative traits. The loci detected by GWAS are merely the largest effect sizes drawn from a gamma or similar distribution (Hayes and Goddard 2001) with an underlying additive effect. Traits underpinned by common variants can display favorable results in GWA studies because these variants generate large trait variation in the mapping population; therefore, the GWAS will have an adequate detection power. On the other hand, the scenario where the trait under study is controlled by many genomic locations (polygenic model) with solely small effects or by many rare variants presents a challenge for GWAS. The adequate sample size of a GWA study to detect associations for polygenic traits is a function of the genotyping method, allele frequency, and effect size (Visscher et al. 2017). In plants, sample size is also a function of the mating system: in selfing species, LD is maintained over long genetic or physical distances, while in outcrossing species, LD declines more rapidly (see later section). In general, association studies for polygenic traits, where the proportion of variance explained by individual variants is small, benefit larger experimental sample sizes to detect associated loci.

For many experiments, current sample sizes available for genome-wide association studies are not sufficient to detect the majority of the associated variants. One affordable solution is to have a population sample where we can try to maximize the level of genetic variance to be representative of the allelic diversity present in a species. In plants these sampled populations are also called "core collections" and are usually composed from germplasm seedbank accessions that have been extensively phenotyped. However, by following such approach, we can introduce genetic heterogeneity that can reduce the correlation between phenotype and a specific variant (Korte and Farlow 2013). Although genome-wide studies have increased sample size by densely sampling local populations, this approach can reduce the presence of variants demonstrative of the global diversity in that species (Long et al. 2013; Huber et al. 2014).The issue of sample size has been brought to the forefront by the observation of so-called missing heritability. When a GWAS is not sufficiently powered, it will only uncover variants accounting for a small fraction of the heritability, mostly due to the stringent threshold used to identify significant associations. Joint consideration of all common SNPs has increased the proportion of heritability explained although it does not lead directly to an improved genetic predictor of risk because it does not identify individual loci (Yang et al. 2010; Gibson 2012).

2.2 Phenotyping

Another important aspect for the detection of genetic variants for complex traits is how traits are measured, with measuring problems hindering the detection of association signals (van der Sluis et al. 2010). Using simulation studies van der Sluis et al. (2010) found that phenotypic components, such as complexity, measurement bias, and resolution, could dilute the genetic signal. These issues are important for complex traits, such as psychological, psychiatric, and other (e.g., medical) traits, but do not necessarily apply to phenotypes for which the measurement is simple. Quantitative traits are difficult to dissect because there may be underlying attributes which can be either a group of correlated variables or variables that are closely related at the biological level (Mackay 2014; Sun and Wu 2015). An additional constraint is that many of the important traits in crop plants are the product of dynamic processes that are difficult to assess and measure. With the advent of new technologies for high-throughput phenotyping, in plant science, mostly motivated by the need for high-yielding and stress-tolerant plants, it is currently possible to obtain multidimensional phenotypic data. Phenomics enables the characterization of the "phenome," which refers to the phenotype as a whole (Soulé 1967) through the use of multidisciplinary techniques, such as sensors, image analysis, and robotics (Houle et al. 2010). High-throughput phenotyping improves collection over time, which allows the modeling of the dynamical behavior of phenotypic traits. Most advances in phenotyping have taken place in image analysis and sensor technology development. Various imaging methodologies, such as visible light imaging, infrared imaging, fluorescence imaging, imaging spectroscopy, etc., are being used to collect multilevel phenotype data from macroscopic to molecular scale over a few seconds to weeks (Sozzani et al. 2014).

The analysis of digital images has emerged as a nondestructive method to extract plant trait information in a high-throughput holistic manner (Clark et al. 2013; Fahlgren et al. 2015; Walter et al. 2015). From image processing, a wide number of morphological traits in plant can be extracted using customized analytical tools (Wang et al. 2009; Hartmann et al. 2011; Green et al. 2012; Karaletsos et al. 2012; Bucksch et al. 2014). In a recent *Arabidopsis* study, the observation of growth dynamics by automatic imaging and growth modeling led to the identification of time-specific quantitative trait loci (QTLs) and QTLs related to the whole growth curve (Bac-Molenaar et al. 2015).

2.3 Population Stratification and Relatedness

Confounding effects can cause spurious associations when the GWAS population under study is a mixture of samples with differences in their allele frequencies and in the trait of interest. That is, any marker allele that is at different frequencies in subpopulations with different levels of the trait will be associated with the phenotype (Ewens and Spielman 1995; Pritchard and Rosenberg 1999). These spurious

associations arise between a phenotype and markers that are not linked to any causal loci, which can result in the identification of false-positive associations in the structured populations (Lander and Schork 1994).

In the presence of population structure, statistical models can account for phenotypic covariance due to population structure and familial relatedness with a genomic relationship matrix (Yu et al. 2006; Kang et al. 2008, 2010; Zhang et al. 2010b). The genomic relationship matrix is built with the available single nucleotide polymorphisms (SNPs) and included as a random term in a linear mixed model.

The two most widely used approaches to modeling population structure are model-based estimation of ancestry and principal component analysis (PCA). Genetic ancestry has been introduced as a correction for population stratification. With this approach samples are assigned to subpopulation clusters. Model-based clustering approaches, such as STRUCTURE and ADMIXTURE, have been widely used to infer population structure (Pritchard et al. 2000; Alexander et al. 2009). Principal component analysis is a statistical approach that has also been used to model the ancestry differences between cases and controls in human diseases (Price et al. 2006). Using principal component-based methods by estimating major axes of the pairwise genetic similarity matrix, family relatedness can be captured but not all of the sample structure. Software such as EIGENSTRAT included in the EIGENSOFT package has been used to compute principal components to correct for stratification and increase power to detect true associations (https://www.hsph.harvard.edu/alkes-price/software/).

2.4 Recombination Rate and Linkage Disequilibrium

In any association study, with a number of markers coming from a genotyping platform, markers are distributed over the genome under study and causal mutations are not necessarily assayed directly. In these cases, we rely on the presence of strong linkage disequilibrium (LD) between the identified marker loci and the underlying causal variants (Garner and Slatkin 2003). The extent of LD is relevant in the context of an association study because it will determine the number and density of markers as well as the experimental design (Flint-Garcia et al. 2003). Additionally, after a GWA study is performed, LD decay is generally taken into account to define the extent of a region in base pairs, in which to identify SNPs in high LD with a representative marker. Once the SNPs in a region have been defined, these are generally used for prioritization purposes such as annotation to predict functional SNPs, or they can be annotated into structural genomic categories, i.e., genic or intergenic locations (Chen et al. 2014; Pal et al. 2015). Linkage disequilibrium (LD) can be defined as the covariance of the inheritance of an allele of one SNP and the inheritance of an allele of another SNP within a sampled population. Two variants that differ greatly in allele frequency cannot be in high LD. Thus, if an observed SNP and an unobserved causal variant segregate at different frequencies, they will be in low LD regardless of their physical distance from each other (Nei and Li 1973), and the power to associate them

will be low (Muller-Myhsok and Abel 1997). Among the statistics proposed to measure LD, the descriptive statistic D takes a value that is specific for a set of alleles such that $D_{AB} = 0$ indicates linkage equilibrium and $D_{AB} \neq 0$ linkage disequilibrium or nonrandom association of alleles. In its simplest form, linkage disequilibrium between a pair of loci can be represented as $D_{AB} = p_{AB} - p_A p_B$ which is the difference between the frequency of gametes carrying the pair of alleles A and B at two loci (p_{AB}) and the product of the frequencies of those alleles (p_A and p_B) (Slatkin 2008).

Other popular statistical approaches that look to normalize D are $|D'|$ and r^2. D' uses the theoretical maximum of D to do the normalization, while r^2 uses a correlation coefficient (Lewontin 1964; Hill and Robertson 1968). The maximum value of r^2 is a function of the allele frequencies of two loci, and it decreases with the magnitude of the minor allele frequency difference between the loci (Wray 2005; Amos 2007; Eberle et al. 2007; VanLiere and Rosenberg 2008). Within a population, the degree of LD reflects the influence of factors such as population history, population size, recombination rate, and forces that cause gene frequency evolution (Garner and Slatkin 2003; Wray 2005; Slatkin 2008). LD between a particular pair of loci or in a genomic region depends on local recombination rates. In a population undergoing random mating, recombination events cause chromosomal segments which contain linked loci to be broken apart until they eventually become independent (Zhu et al. 2008). The effective recombination rate is related to the reproduction mode that a species exhibits (Nordborg and Donnelly 1997). In selfing species, recombination is less effective because individuals are more likely to be homozygous at a given locus than in outcrossing species. In rice (*Oryza sativa*), *Arabidopsis* (*Arabidopsis thaliana*), and wheat (*Triticum aestivum*) (Nordborg 2000; Garris et al. 2005; Zhang et al. 2010a), which are self-pollinating species, LD extends much further in comparison with outcrossing species, such as maize (*Zea mays*), grapevine (*Vitis vinifera*), and rye (*Secale cereale*) (Tenaillon et al. 2001; Myles et al. 2009; Li et al. 2011), and conifers (e.g., Pavy et al. 2012). Genome structure and the rate of recombination in different regions across the genome also affect the structure of LD. In human, the relationship between LD decay and "hotspots" in recombination for the major histocompatibility complex (MHC) is well documented as well as the different patterns of LD in the genome (Jeffreys et al. 2001; Teo et al. 2009). In plants, for maize, there is broad evidence for heterogeneity in rates of recombination across the genome, where repetitive regions showed virtually no recombination events (Yao et al. 2002).

3 GWAS Models, Methods, and R Packages and Software Available

3.1 Statistical Models and Methods

Statistical methods used in GWAS have been extensively but briefly reviewed (Zhu et al. 2008; Hayes 2013; Lipka et al. 2015). One of the frequently employed GWAS

model is the linear mixed model which allows the incorporation of fixed effects which account for population structure as a covariate and a random effect correction that accounts for the level of relatedness within a population (Yu et al. 2006). In the implementation of linear mixed models, population stratification (population structure and principal components) is included as a fixed effect in the regression model, while the random effect covariance is included to model the genetic correlation between individuals. Following Henderson's matrix notation (Henderson 1975), the linear mixed-model regression takes the form:

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{\varepsilon}.$$

Here, $\boldsymbol{y}$ is a $n \times 1$ vector of the phenotype. The unknown vector $\boldsymbol{\beta}$ contains estimates of fixed effects, usually including genetic marker and population structure, with design matrix X. The vector $\boldsymbol{u}$ is a random effect, the best linear unbiased prediction (BLUP) which represents the predicted breeding value [given the specification of a covariance matrix (see below)], for each individual. Z is a design matrix pointing observations to genotype identities and $\boldsymbol{\varepsilon}$ is a vector of residuals. Typically, we assume residuals in ε are independent and identically distributed with mean of zero, and variance σ^2_{E} and a covariance matrix often denoted R but equal to the identity matrix I, where the $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are normally distributed with null mean and variance additive genetic variance σ^2_{A}:

$$\mathrm{Var}\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix}$$

In a genetic mixed model, we expect the elements of $\boldsymbol{u}$, in contrast to $\boldsymbol{\varepsilon}$, to be correlated because individuals have common ancestors and thus share alleles that are identical by descent. We model this by specifying a covariance matrix for the levels of $\boldsymbol{u}$, often denoted $G = \sigma^2_{\mathrm{A}}K$, where K is a genetic relatedness kinship matrix derived from markers.

The vector of phenotypes then has the following covariance matrix:

$$\mathrm{var}(\boldsymbol{y}) = V = ZGZ^T + R$$

where $R = I\sigma^2_e$ and σ^2_e is the residual variance.

The variance in the levels of $\boldsymbol{u}$ is the variance in breeding values or, in other words, the additive genetic variance σ^2_{A}, the crucial parameter for heritability estimation (which will be the focus of a latter section).

3.2 Multiple Testing Corrections

The appropriate interpretation of GWAS results for complex traits is dependent on determining the correct P-value threshold for statistical significance. As the number of statistical tests increases with the number of markers included in the analysis, so

does the probability of finding at least one of them to be statistically significant when it is not (i.e., a "type I" error). Statistical procedures accounting for multiple testing have been used in the genome-wide setting and vary depending on how conservative they are. Among conservative methods are the Bonferroni correction (Bonferroni 1935), which can be overly conservative, and the Šidàk-Bonferroni approach (Abdi 2007) which becomes very conservative when the number of comparisons becomes large and when the tests are not independent. A less stringent method is to control for the proportion of false-positive associations typically expressed as false discovery rate (FDR). In the Benjamini and Hochberg correction, *P*-values are ranked from smallest to largest. Each *P*-value is then compared to a Benjamini-Hochberg critical value: (rank/number of tests) × (false discovery rate) (Benjamini and Hochberg 1995).

3.3 GWAS Analysis Software and R Packages

Statistical models for association mapping have been implemented in several software packages (briefly described below). Development of new algorithms has been motivated by the need for an improvement in computational speed. The computation time of each method can be broken down into three steps: (1) building the genomic relationship matrix (GRM/kinship), (2) estimating variance components, and (3) computing association statistics for each SNP (Yang et al. 2014).

3.3.1 EMMAX: Efficient Mixed-Model Association eXpedited

When mixed linear models were first developed for association mapping, the variance parameters were estimated accounting for the fact that the total variance explained by all markers except by the candidate marker may vary across candidate markers in the case of markers of large effect (Kang et al. 2008). Because of the increase in the number of markers for computational efficiency, the repetitive variance component estimation was avoided (Kang et al. 2010). Following this approach, the variance parameters are estimated only once for each dataset and globally applied to each marker. The algorithms "EMMA eXpedited" (called EMMAX) and "population parameters previously determined" (called P3D) (Zhang et al. 2010b) use pre-estimated variance components. These approximate methods implemented in the software programs EMMA eXpedited (EMMAX) and TASSEL (P3D) have since then been widely used in GWAS (genome.sph.umich.edu/wiki/EMMAX).

3.3.2 FaST-LMM: Factored Spectrally Transformed Linear Mixed Models

FaST-LMM is a program written in python for performing both single-SNP and SNP-set genome-wide association studies (GWAS) on extremely large datasets (Lippert et al. 2011). It provides a fast implementation of an exact or approximate model for computation of the test statistics (https://github.com/MicrosoftGenomics/FaST-LMM). In this approach, the estimation of the relationship matrix is produced using a selected set of SNPs, which show the strongest linear correlation with the trait of interest through the FaST-LMM-Select procedure. This algorithm also reduces the computational time by a spectral decomposition of the genetic similarity matrix.

FaST-LMM has the advantage, along with EMMAX and Mendel, of internally imputing missing data at any (genetic or non-genetic) covariate, which can make it convenient for implementing stepwise conditional analyses.

3.3.3 GRAMMAR-Gamma (GenABEL R Package)

The Grammar-Gamma method is a fast variance component-based two-step method implemented in the software GenABEL (Aulchenko et al. 2007b). This method is derived from the original GRAMMAR method in which the residuals from the LMM are first estimated and treated as phenotypes for a genome-wide association study using a standard linear model (Amin et al. 2007; Aulchenko et al. 2007a). Similarly to GRAMMAR, the GRAMMAR-Gamma method produces unbiased SNP effect estimates and test statistics that do not require any deflation but involve the calculation of a GRAMMAR-Gamma correction factor γ (Svishcheva et al. 2012). Developers of the GRAMMAR-Gamma method suggest the use of this method for association testing in whole-genome re-sequencing studies of large human cohorts.

3.3.4 MTMM: Multitrait Mixed Model

MTMM method is an extension of a standard linear mixed model used to perform GWAS of correlated traits (Korte et al. 2012). This multitrait mixed-model method can be applied to phenotypes from different measurements, with correlation due to pleiotropy, and to traits that were measured under different environmental conditions. A GitHub repository of the MTMM R scripts can be found at https://github.com/Gregor-Mendel-Institute/mtmm.

3.3.5 GCTA

This software was developed in the context of addressing the "missing heritability" problem estimating the variance explained by chromosomes or at a whole-genome level (Yang et al. 2011a). GCTA uses files in PLINK format and estimates the genomic relationship matrix from the user input genomic markers. In building the GRM, candidate markers can be excluded via a leave-one-chromosome-out analysis implemented in GCTA software (GCTA-LOCO) (http://cnsgenomics.com/software/gcta/index.html).

3.3.6 TASSEL

Tassel is a stand-alone or command line program which implements several GWAS methods (Bradbury et al. 2007). The "compressed MLM" option decreases the effective sample size of datasets by clustering individuals into groups. The computing complexity function is thus reduced from the cubic of the number of individuals to the cubic of a smaller number of groups. Additional implemented approaches are "population parameters previously determined" (P3D) (Zhang et al. 2010b) that eliminate the need to recompute variance components. Further options in TASSEL 5 are imputation by FILLIN and FSFHap; kinship and PCA can be calculated with user input genomic markers (Swarts et al. 2014).

3.3.7 GAPIT

Genomic Association and Prediction Integrated Tool (GAPIT) is an R package (Lipka et al. 2012). GAPIT was developed to perform genome-wide association study and genomic prediction following the unified mixed model, EMMA, the compressed mixed linear model, and P3D/EMMAX (Kang et al. 2008; Zhang et al. 2010b). After running, GAPIT automatically reports results in a series of tables and graphs (http://zzlab.net/GAPIT/).

4 Making Sense of GWAS with Functional Genomic Data

In a classic GWAS approach, thousands or even millions of SNPs are tested individually for a statistical association with the trait under investigation. In this kind of analysis, each SNP is treated the same way irrespective of its location in the genome or its proximity to regulatory elements or genes known to be important. Over the years, GWA studies have been able to identify thousands of reproducible statistical associations for a wide range of phenotypes (Pickrell 2014). Complex traits are usually influenced by many genes with small effects, and the associated

SNPs have been found to be located in regions outside genes, suggesting an important role for regulatory elements (Glazier et al. 2002; Hindorff et al. 2009; Visscher et al. 2012; Wallace et al. 2014). Moreover, some studies have shown consistent patterns of enrichment of these polygenic effects in specific genome annotation or categories: SNPs tagging regulatory and genic elements highly enriched, introns little enriched, and intergenic regions negatively enriched (Schork et al. 2013).

In the next section, we review different methods to identify functional regions across the genome and how researchers have started to use this information to make a better sense of GWAS results.

4.1 Functional Genomic Data

As soon as the human genome was sequenced in 2001, it was evident that 99% of the more than three billion base pairs that constitute our genome do not code for proteins (Lander et al. 2001; Kellis et al. 2014). Two sources of evidence supported the fact that noncoding regions might have functionally significant elements: evolutionary conserved regions in noncoding elements and GWAS hits for disease-causing variants in these elements. In this context the "Encyclopedia of DNA Elements" (ENCODE) was launched in September 2003 as a public research project with the aim to identify all functional elements in the human genome (ENCODE Project Consortium 2012). Defining "function," however, has been a source of major controversy in the human genomics community; the ENCODE Project, for example, stated that their data allow them to "assign biochemical function for 80% of the genome, in particular outside of the well-studied protein-coding regions" (ENCODE Project Consortium 2012). On the other hand, evidence suggests that less than 10% of the genome is evolutionarily conserved through purifying selection (Graur et al. 2013; Rands et al. 2014), which is in disagreement with previous estimates. In an effort to define functional DNA elements in the human genome and alleviate the controversy, the ENCODE Project presented an article reviewing the three different sources of functional evidence with its advantages and pitfalls (Kellis et al. 2014). The first line of evidence is comprised of the genetic approaches that evaluate the phenotypic consequences of mutations (as in GWAS); the second line of evidence is evolutionary, which quantifies selective constraints; and the final line of evidence is biochemical that measures evidence of molecular activity. An important conclusion is that our understanding of "genomic function" is still limited and that efforts to better define genome elements should focus on integrating these three approaches to gain better insight into the role they play in human biology. In the next section, we explore the evolutionary and biochemical evidence of "function" and the methods developed within each category. Later in the chapter, we will learn how scientists are integrating functional elements of the genome with GWAS information.

4.2 Evolutionary Evidence

This type of evidence is mainly based on comparative genomics and detects sequences that are likely to have undergone selection. The main concept relies on alignment of the genome of divergent species and looks for sequences that have maintained their similarity through evolution. Software like GERP (Cooper et al. 2005), PhastCons (Siepel et al. 2005), and GERP++ (Davydov et al. 2010) are able to identify sites under evolutionary constraint. Briefly, Genomic Evolutionary Rate Profiling (GERP) is a framework that produces position-specific estimates of evolutionary constraint by using maximum likelihood evolutionary rate estimation. GERP identifies regions with nucleotide substitution deficits and quantifies them in terms of the "rejected substitution" (RS) score, which is defined as the number of substitutions under neutrality minus the number of substitutions observed at each position (Cooper et al. 2005). PhastCons is also a software program that identifies evolutionary conserved elements through a multiple alignment given a phylogenetic tree (Siepel et al. 2005). However, this software uses a phylogenetic hidden Markov model (phylo-HMM) rather than the maximum likelihood approach used by GERP. Finally, GERP++ uses the same framework as GERP but with significantly faster and more statistically robust maximum likelihood estimation. Additionally, this algorithm can group constrained positions into constrained elements and assign *P*-values to the predictions (Davydov et al. 2010). High GERP conservation scores can be correlated with GWAS hits, showing an overlap of genetic and evolutionary evidence of function at the significant SNP hit positions (Tulah et al. 2013; Al-Tassan et al. 2015).

4.3 Biochemical Evidence

This category is comprised of any evidence of cellular or enzymatic activity processes acting on DNA. The ENCODE Project applied a wide variety of assays to identify genomic regions falling into five categories: regions expressing long and short RNAs, regions occupied by transcription factors or other regulatory elements, open chromatin regions, regions showing methylation or specific histone modifications, and finally genomic regions that are able to physically interact with each other (Kellis et al. 2014). We will explore some of the more than forty assays that ENCODE has used to discover these functional elements. None of the following assays are restricted to human cells, and most of them are applicable to other animals, plants, or bacteria.

4.3.1 Gene Expression (RNA-Seq)

RNA sequencing (RNA-seq) (Nagalakshmi et al. 2008) comprises a set of experimental procedures that generates cDNA sequences derived from RNA molecules

that are later deep sequenced (Han et al. 2015). RNA-seq allows the characterization of RNAs present in a sample and to quantify their abundance at the same time. Briefly, millions of short segments are sequenced from random positions of the input RNAs, called short reads. These reads are later mapped onto a reference genome. The number of reads aligned to each gene gives a measure of its expression level (Finotello and Di Camillo 2015). RNA-seq allows for analysis of the transcriptome with single base pair resolution and low background noise. Since it offers a greater specificity and sensitivity for both detection of transcripts and estimates of expression, it has replaced any hybridization-based technologies (microarrays). Additionally, this technique is not limited to species or transcript specific probes, so that RNA-seq can detect novel transcripts, indels, and previously unknown changes, even in species lacking a reference genome. Recently, after analyzing a large set of RNA-seq data, it was found that more that 85% of the human genome is transcribed (Hangauer et al. 2013). Most of these sequences belong to a specific class of intergenic transcripts: the long intergenic noncoding RNAs (lincRNAs). Other transcript classes which have been identified using RNA-seq include miRNAs (microRNAs), siRNAs (small interfering RNAs), and other small RNAs such as eRNAs (enhancer RNAs) or snRNAs (small nucleolar RNAs) which are involved in regulation of RNA stability, modulation of chromatin states, or protein translation (Kim et al. 2010; Trapnell et al. 2010; Andersson et al. 2014; Han et al. 2015).

4.3.2 Chromatin Immunoprecipitation Assays

Chromatin immunoprecipitation assays followed by deep sequencing (ChIP-seq) is a technique capable of detecting protein-DNA physical interactions and chemical modifications of histone proteins (Furey 2012). Through the use of this technique, a detailed map of binding sites for transcription factors and core transcriptional machinery can be identified. Together with nucleosome positioning and the dynamic modification of histones, these binding sites are paramount to the understanding of the regulatory networks that govern biological processes (Jiang and Pugh 2009; Farnham 2009). The ChIP process enriches a DNA sample with the transcription factor (TF ChIP-seq) or modified nucleosomes (ChIP-seq) of interest using specific antibodies. In the TF ChIP-seq, the proteins and DNA are cross-linked and then digested with an exonuclease; later, the cross-linked segments are immunoprecipitated, and the DNA is purified and deep sequenced. The process is quite similar in ChIP assays that aim to map histone marks; however, in this case micrococcal nuclease (MNase) digestion is used to fragment the DNA. Protocols can vary depending on the target protein or tissue; basic steps and modifications have been reviewed elsewhere (Park 2009; Ku et al. 2011; Furey 2012). The ENCODE Project hosts the "Factorbook" repository that presents results on ChIP-seq for 167 transcription factors across 837 experiments as of October 2016 (Factorbook, http://www.factorbook.org/). This page also integrates the data with other ENCODE assays such as ChIP-seq of histone marks and nucleosome occupancy. The modENCODE database (modENCODE, http://www.modencode.org/) has available

transcription factor binding sites for *D. melanogaster* and *C. elegans*. ChIP-seq is not restricted to animal models and has been applied in the model plant *A. thaliana* to study the genome-wide positional distribution of the transcription factor binding sites (Yu et al. 2016).

4.3.3 Open Chromatin (DNase-Seq, FAIRE-Seq, MNase-Seq, ATAC-Seq)

The chromatin state in eukaryotic cells depends largely on the stage of the cell cycle. For example, during interphase, the chromatin is structurally loose allowing the transcription machinery to reach DNA. But even in this overall "loose" state, the local structure of chromatin is tightly regulated and plays a central role in gene regulation (Mellor et al. 2005). It is now well understood that transcriptional activation coincides with nucleosome perturbation (Boeger et al. 2003; Lee et al. 2004; Shu et al. 2011; Tsompana and Buck 2014) at promoters, enhancers, silencers, and insulators due to transcription factor binding (Tsompana and Buck 2014). These "perturbed" or "open chromatin" sites are now known as the primary positions for regulatory elements (John et al. 2011). One of the main differences between chromatin accessibility approaches and ChIP-seq is that the former does not use antibodies, thus capturing the genome-wide chromatin landscape and reducing the bias. Below we briefly discuss the most commonly used methodologies for genome-wide chromatin profiling.

DNase-seq

Early observations that active genes were preferentially digested by the nonspecific double-strand endonuclease DNase I (Weintraub and Groudine 1976) provided the first demonstration that active genes exhibited an altered chromatin conformation that made them susceptible to digestion. These genomic regions were named thereafter as "DNase I hypersensitive sites" (DHSs) (Weintraub and Groudine 1976; Keene et al. 1981; Tsompana and Buck 2014). High-throughput DNase assays (DNase-seq) have been developed taking advantage of the drop of sequencing prices and the increased quality of the data (Crawford et al. 2006; Tsompana and Buck 2014). Briefly, chromatin is digested by DNase I endonuclease: fragments generated are size selected to enrich for fragments produced in highly sensitive regions. Afterward, fragments are amplified, sequenced, and mapped back to the reference genome (Meyer and Liu 2014). One drawback of DNase-seq is that it requires a large number of cells as starting material and that the DNase I enzyme carries an intrinsic cleavage bias cutting preferentially in certain sequences independently of its chromatin state (Dingwall et al. 1981; Meyer and Liu 2014). Despite these inconveniences, DNase-seq is being widely used by the ENCODE consortium and is currently the gold standard chromatin accessibility assay in humans.

FAIRE-Seq

FAIRE-seq stands for "Formaldehyde-Assisted Isolation of Regulatory Elements," and while it is considered one of the simplest open chromatin methods, it has the highest background noise (Giresi et al. 2007; Giresi and Lieb 2009). In FAIRE-seq formaldehyde is employed to cross-link chromatin to capture protein-DNA interactions. Thereafter, chromatin is sheared using sonication and phenol-chloroform is used to isolate the fragmented DNA which is sequenced later (Giresi and Lieb 2009).

MNase-Seq

MNase-seq is a technique which employs an enzymatic cleavage approach in which the enzyme used for DNA digestion is the micrococcal nuclease (MNase), an endo-exonuclease that digests DNA not protected within a nucleosome (Meyer and Liu 2014). This technique differs from the ones previously described because it provides an indirect measurement of chromatin accessibility. Rather than reporting open chromatin regions, MNase-seq provides a genome-wide characterization of average nucleosome occupancy and positioning (Barski et al. 2007; Schones et al. 2008).

ATAC-Seq

ATAC-seq that stands for Assay for Transposase-Accessible Chromatin with high-throughput sequencing is a recently developed method for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins, and nucleosome position (Buenrostro et al. 2013). This technique employs a hyperactive Tn5 transposase loaded with adaptors for high-throughput DNA sequencing in a process called "tagmentation" (Adey et al. 2010). This assay can simultaneously fragment and tag the genome with sequencing adaptors. The idea behind this technique is that Tn5 transposase is not able to integrate in DNA regions packed in nucleosomes; thus, integration and posterior sequencing will be enriched for open chromatin regions.

One of the advantages of this technique is that it can detect nucleosome positioning, chromatin accessibility, and TF binding simultaneously. Moreover, the amount of starting material is three to five orders of magnitude lower, while the sensitivity and specificity obtained are similar to DNase-seq (Buenrostro et al. 2013). Moreover, the protocol is less complex and not as time consuming as other assays such as DNase-seq or FAIRE-seq. This technique is envisioned to become the preferred method for the study of open chromatin in the future (Tsompana and Buck 2014).

4.3.4 Chromosome Conformation Capture (Hi-C)

It is now recognized that three-dimensional organization of chromatin can affect gene expression. Hi-C is a technique that captures genome-wide long-distance DNA interactions (Belton et al. 2012; Ay et al. 2015). The basic Hi-C procedures consists of cross-linking cells with formaldehyde, then digesting the DNA with a restriction enzyme, marking the cut-ends with biotin, and ligating them. Finally, DNA is purified, and fragments with biotin are sequenced using paired-end reads (Ay et al. 2015). After using this technique, we can have a clear panorama of interacting DNA regions. This interaction can happen even when the DNA fragments are largely separated in regard to their physical position on a genome. Biologically important processes like the enhancer-promoter interactions can be captured by this technique. Hi-C experiments have been conducted in a wide variety of organisms, including plants (Grob et al. 2014; Liu et al. 2016), bacteria (Le et al. 2013; Marbouty et al. 2015), and fruit fly (Li et al. 2015b), and in numerous human and mouse cell lines (Lieberman-Aiden et al. 2009; Zhang et al. 2012; Rao et al. 2014; Grubert et al. 2015).

5 Measuring Heritability in Relation to GWAS and the Genetic Architecture of Complex Traits

The goal of a GWAS is, in general, to identify key genetic loci underlying phenotypic variability in traits of interest. The relative amount of trait variance that is due to genetics versus environmental and measurement variability is an important indicator of how successful a GWAS can be expected to be effective. That is, when most variability for a trait is not genetic in origin, it will be difficult to isolate genetic signal from nongenetic noise.

The proportion of total variability in a trait that arises from genetic sources is known as the *heritability* (Falconer and MacKay 1996; Visscher et al. 2008; Vinkhuyzen et al. 2013). Indeed, the *heritability* of a phenotypic character is one of the most central genetic parameters in all genetics-related fields. For researchers seeking to understand the genetic basis of complex traits, using GWAS, heritability represents the theoretical upper limit for discovery. Thus, an estimate of the heritability of a trait should be viewed as an essential statistic for interpreting the results of a GWAS.

In the section that follows, we discuss how the completion of the first GWAS led to a very large discrepancy between traditional estimates of heritability and the (tiny) amount of cumulative variance explained by significant hits. This led to the concept that heritability was “missing” and had to be found. Efforts to understand the “missing” heritability have pushed the field forward in a number of ways, notably including the identification of techniques for using genome-wide markers to obtain estimates of genetic variance.

After reviewing the theoretical concept and the classical methods for estimation, we put particular focus on modern methods that leverage GWAS data in order to estimate heritability. Estimation methods that utilize GWAS-type data include the Haseman-Elston regression (HE) and various forms of the linear mixed model (LMM).

Heritability is covered as an essential topic in statistical genetics (Lynch and Walsh 1998, Falconer and MacKay 1996), and more modern developments in theory/methodology have also been reviewed recently (Visscher 2008; Vinkhuyzen et al. 2013). However, the continuing fall in price of genome-wide data and the corresponding explosion in efforts to elucidate the genetic architecture of traits in many organisms have driven rapid developments in methodology for measuring heritability, particularly using genome-wide data (Visscher et al. 2008; Yang et al. 2010; Vinkhuyzen et al. 2013; Speed and Balding 2015).

5.1 Defining Heritability

Heritability is measured by partitioning the phenotypic variability $\left(\sigma_P^2\right)$ in a population into genetic $\left(\sigma_G^2\right)$ and environmental $\left(\sigma_E^2\right)$ variances. The genetic variance, σ_G^2, is frequently partitioned further into variance due to a linear allele substitution or additive effects at each causal locus $\left(\sigma_A^2\right)$, a nonlinear dominance or interaction between alleles at the same locus $\left(\sigma_D^2\right)$, and to epistasis, the interactions of alleles at different loci $\left(\sigma_{Epi}^2\right)$. Epistatic variance can, of course, be defined more specifically in terms of two-locus $\left(\sigma_{AA}^2, \sigma_{AD}^2, \sigma_{DD}^2\right)$, three-locus (e.g., $\sigma_{AAA}^2, \sigma_{AAD}^2, \sigma_{ADD}^2$, etc.), and higher-order interactions; we will use σ_{Epi}^2 for short. Given this partition of variance, heritability is defined as having two types:

Heritability in the broad-sense (H^2):

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}$$

Heritability in the narrow-sense (h^2):

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}$$

5.2 Nonanalytical Factors Influencing Heritability

Before covering the old and new methods for estimating heritability, which can drastically influence the value and meaning of the estimate, it is important to recognize the biological factors that influence the heritability. Heritabilities are not

fixed, knowable quantities; they can only be estimated and estimates for various reasons can vary widely. Genetic variance depends on allele frequencies in the population. Genetic variances can change across generations (e.g., under natural or artificial selection), populations, and datasets as alleles are lost, fixed, or gained. Furthermore, there can be genotype-by-environment (GxE) interaction such that, in effect, genetic architecture of a trait and, thus, heritability can differ between environments (Falconer and MacKay 1996; Lynch and Walsh 1998; Holland et al. 2003; Visscher et al. 2008; Vinkhuyzen et al. 2013).

The denominator of the heritability depends on environmental and other nongenetic sources of variability. Environmental sources of phenotypic variance come in many forms, including but not limited to maternal environments, local-scale or field-level variability, seasonal or climatic, cage or heard effects, age, etc. These factors are important to recognize, particularly when comparing multiple estimates for the same trait.

Finally, we note that the manner in which σ^2_E is handled depends on the goal of the research. For example, in human genetics and plant and animal breeding, environmental factors that can be explicitly controlled are most often removed from σ^2_P before analysis (Visscher et al. 2008), the reasoning for breeding being that a known and controllable factor (e.g., location) is not expected to influence selection accuracy and thus is not relevant to the prediction of selection response (Holland et al. 2003).

Geneticists studying humans are not interested in selection and selection response. Instead, the goal is usually the trait prediction based on information about pedigree or genome (de Los Campos et al. 2013). Thus fixed, controllable parameters are typically removed before estimating heritability (Visscher et al. 2008). Ultimately for breeding and human genetics, the goal is often to know what proportion of unaccounted for variation comes from genetic sources. In contrast to breeders, geneticists studying evolution under natural selection must incorporate all phenotypic variability in the denominator. The reason for this is that evolutionary fitness depends on the full expression of the phenotype regardless of its source.

5.2.1 Heritability and Response to Selection

Although we alluded to this above, we emphasize here the use of heritability for understanding and predicting responses to selection. The univariate (single trait) response to selection can be modeled by the breeder's equation:

$$R = h^2 S$$

Here the cross-generation change in the mean phenotype of the population, R, is equal to the heritability, h^2, times the selection differential, S. The selection differential, S, is the difference between mean of individuals selected as parents of the next generation and the mean of current population as a whole (Falconer and MacKay 1996; Lynch and Walsh 1998). Heritability also represents, by definition, the correlation between true breeding values and the phenotype and thus tells about

maximum possible accuracy for predicting the phenotype (Meuwissen et al. 2001; Heffner et al. 2009; Goddard 2009). As mentioned previously, a number of environmental and other experiment-dependent factors influence heritability estimates. Therefore, estimation of heritability, which involves partitioning these variances, can be a useful exercise for comparing trials, although this should be done with caution because of reasons described above.

Heritability has been measured on many traits, in many different species, and has been reviewed in a number of contexts in several excellent publications (e.g., Falconer and MacKay 1996; Visscher et al. 2008; Vinkhuyzen et al. 2013). Our goal in this section is not to highlight any particular trait or organism, but rather to focus on the methodology and the purpose of heritability estimation, generally.

5.3 Classical Methods of Heritability Estimation

Individuals who share common ancestors resemble to each other to a degree often proportional to the closeness of their relationship (Fisher 1918; Falconer and MacKay 1996). This is a crucial observation in quantitative genetics; that is why all methods of estimating heritability revolve around the measurement of phenotypic similarity and its covariance with the (expected) proportion of the genome-shared identical by descent (IBD) between relatives (Falconer and MacKay 1996; Speed and Balding 2014, 2015).

The choice of the types of relatives to be used for heritability estimation often depends on practical circumstances, and different choices have advantages and limitations. In general, more variation around the expected proportion of the genome-shared IBD is expected for more distant relatives leading to greater sampling variance in heritability estimation (but see below regarding the measurement of genomic realized relationships between distant relatives). In contrast, the closer the relatives used in a study, the more precise the estimate of heritability becomes, but a higher potential bias occurs because of environmental and nonadditive (e.g., dominance) genetic factors (Falconer and MacKay 1996; Visscher et al. 2008; Zuk et al. 2012; Vinkhuyzen et al. 2013; Speed and Balding 2014, 2015).

5.3.1 Parent-Offspring Regression

The use of parent-offspring relationships is relatively straightforward. Phenotypic measurements of each of a sample of offspring are regressed on the mean value of their parents (also called the midparent value), and the slope of regression line measures the heritability. Alternatively, offspring values can be regressed on one of the parents, for example, on the father's phenotypic values. Parent-offspring regression is simple to do and is not as biased by common environmental factors as some other methods are, but requires large sample sizes for precision (Vinkhuyzen et al. 2013).

5.3.2 Sibling Analysis

In nonhuman genetic studies, large full- and half-sib populations are often available. It is then possible, using analysis of variance, to partition variance within and among families and due to male and female parents. An easy example is a population, where each of a series of sires is mated to several dams and, ideally, multiple progenies per dam are measured. It is easy to see how variance can be partitioned into among-dam $\left(\sigma^2_{\text{Dam}}\right)$, among-sire $\left(\sigma^2_{\text{Sire}}\right)$ and within-family $\left(\sigma^2_{\text{Within}}\right)$ components. However, the translation of these components into additive and nonadditive genetic variance components necessary to obtain a heritability estimate can be complicated and is covered in detail in several other texts (Falconer and MacKay 1996; Lynch and Walsh 1998). For example, it can be shown that the variance among sires is the covariance among half-siblings and is equivalent to $\frac{1}{4}\sigma^2_{\text{A}}$. One of the more serious limitations to this approach, aside from the need to create the necessary population in the first place, is for a balanced sample size in terms of numbers of sires, dams, and offspring.

5.3.3 Twin Studies

In humans, family sizes are small and mating designs obviously cannot be implemented. For this reason, one of the most useful methods for estimating heritability involves the comparison of mono- (MZ) to dizygotic (DZ) twins. Differences between MZ twins are expected to be purely due to (common) environmental factors (as they are clones), while differences among DZ twins can include both genetic and environmental sources. Specifically, $\text{cov(MZ)} = \sigma^2_{\text{G}} + \sigma^2_{\text{CE}}$, where σ^2_{CE} is the variance due to common environment. In contrast, $\text{cov(DZ)} = \frac{1}{2}\sigma^2_{\text{A}} + \frac{1}{4}\sigma^2_{\text{D}} + \frac{1}{4}\sigma^2_{\text{AA}} + \frac{1}{8}\sigma^2_{\text{AD}} \cdots \sigma^2_{\text{CE}}$. In a typical design, only the cov (MZ), cov(DZ), and σ^2_{P} are measured. Thus, aside from σ^2_{A}, a trade-off must be made where either nonadditive variance or common environmental effects are partitioned. This leads to an potential upward bias to twin-based estimates of h^2 (Falconer and MacKay 1996; Zuk et al. 2012; Vinkhuyzen et al. 2013). Estimates of heritability done using this approach, using very large samples, have been treated as the "truth" by many researchers and play a major role in the controversy over "missing" heritability (see below).

5.3.4 Haseman-Elston Regression

Haseman-Elston regression is a relatively simple approach, originally developed for estimating variance explained at a marker locus, that is easily extendable to a complex pedigree or, rather, a population with variation in relatedness beyond parent-offspring. The squared difference in phenotype between pairs of individuals

is regressed on a measure of their additive genetic relatedness, and the slope obtained is equal to $-2\sigma^2_A$, and the intercept is $2\sigma^2_P$ (Haseman and Elston 1972; Yang et al. 2010; Chen 2014; Golan et al. 2014). Measures of genetic relatedness can come from pedigree *or* from genomic relationship matrices (GRMs) constructed using SNP data available in typical GWAS datasets. The Haseman-Elston (HE) regression is currently the state-of-the-art method for estimating SNP heritability from ascertained case/control samples (Haseman and Elston 1972).

5.4 Linear Mixed-Model Estimation

The heritability estimation procedures described above are generally applied either using closed-form equations or least-squares methods, which are susceptible to several limitations: imbalance in datasets cannot usually directly handle experimental cofactors with the exception of Haseman-Elston regression and cannot take advantage of complex pedigrees. The linear mixed model (LMM) described in the first section of this book chapter, known commonly as the "animal model," was developed in the 1950s and 1960s (Henderson 1953, 1975, 1976; Vinkhuyzen et al. 2013; Gianola and Rosa 2015) and addresses all of these limitations. The original purpose of this model in animal science was breeding value prediction, for example, for bulls in dairy herds that could not be directly evaluated for milk production (Henderson 1975). For the purpose of heritability estimation, this model is useful because one of its few parameters is the variance in breeding values (i.e., additive genetic variance component, σ^2_A) (Visscher et al. 2008).

In order to estimate heritability using the LMM, the matrix G must be specified based on some kind of information, namely, pedigree, DNA markers, and/or sequence data (Vinkhuyzen et al. 2013; Speed and Balding 2015), and therefore much of our remaining discussion will be concerned with the construction of G. Originally, mixed models were very challenging to apply because, unlike least squares, they require iterative procedures like restricted maximum likelihood (REML) or Markov chain Monte Carlo (MCMC). However, advances in computing power and analytic methods have made LMM estimation tractable even on very large (thousands or even millions) records.

Note that in GWAS, a genome-wide marker-based estimate of kinship is often incorporated in a LMM to control for population structure (Yu et al. 2006; Kang et al. 2008, 2010). In other words, the contrast between heritability estimation and GWAS with LMMs is whether the random genotype effect is the focus (estimating h^2) or a nuisance variable (GWAS). Although our focus is on continuous traits, we note that when applied with appropriate caution, LMMs can be extended to apply to case-control and other kinds of binary and categorical traits (Lee et al. 2011).

5.5 Measurements of Relatedness and Estimates of Heritability with Genetic Markers

The matrix G, like all variance-covariance matrices, is square and symmetrical, with dimensions $N \times N$, where N is the number of individuals to be analyzed. The diagonals contain measures of within-individual variance, which intuitively is related to the degree of homozygosity and thus inbreeding (Vinkhuyzen et al. 2013); specifically the diagonals are defined as $1 + F$, where F is the individual inbreeding coefficient. The classical kinship matrix, sometimes called the numerator relationship matrix, is constructed from pedigrees. Pedigree-based matrices measure the expected (average) relationships calculated as the proportion of the genome-shared IBD. That is, they assume pedigree founders are completed unrelated and noninbred (inbreeding coefficient, $F = 0$), and they assign equal relatedness to all members within a family. Relationship coefficients are $1/2^k$, with k being the number of generations separating a pair of individuals (Falconer and MacKay 1996; Vinkhuyzen et al. 2013), and can be summed if individuals share multiple common ancestors.

Genetic markers, especially dense genome-wide SNP marker data, have made it possible, not just to conduct mapping with GWAS but also to measure genetic relatedness with high precision. The first attempts to use molecular markers for heritability estimation were motived by a lack of pedigree in wild populations of plants and animals, using either a method similar to Haseman-Elston regression (Ritland 1996, 2000; Mousseau et al. 1998) or the animal model (LMM, Kruuk 2004).

Dense genome-wide markers have enabled a major advance in this area. Pedigrees measure only the expected relationship among and within families, with individuals within a nuclear family all assumed to be equally related to each other and to their parents (Falconer and MacKay 1996). However, because parents and founders of the pedigree are not in fact unrelated, and because of stochasticity in recombination and segregation of chromosomes, there is variation around the expectation of relatedness within and among families; this phenomenon is known as Mendelian sampling (Speed and Balding 2015).

With a high density of markers, it is possible to precisely estimate the realized (actual) relationships in any dataset, even among nominally unrelated individuals. One of the first uses of genome-wide markers for heritability estimation was by Visscher et al. (2006), who calculated realized IBD proportions using 1717 markers conditioned on known pedigree for 3375 full-sib pairs. The authors used this information to estimate the heritability of human height and obtained an estimate of 0.80, in agreement with previously published values. Because the estimate was done within families, it was possible for the authors to claim that the measure was free of the assumptions of typical twin studies regarding common environmental and dominance variance. In a follow-up study, with more markers and over 11,000 full-sib pairs, it was possible to partition the genetic variance between chromosomes in order to compare with the extant of GWAS results and set a target for future mapping studies (Visscher et al. 2007).

In a landmark paper, Yang et al. (2010), following work by animal and plant breeders (VanRaden 2008, 2009; Heffner et al. 2009; Hayes et al. 2009), demonstrated the use of genome-wide markers to construct the kinship matrix G and estimate the heritability of human height with LMM. The development was in part motivated by the results of GWAS in which significant SNPs failed to explain more than a few percent of the expected heritability of complex traits (see GWAS and the "missing" heritability below). The genomic relationship between a pair of individuals j and k is denoted A_{jk} and was specified as follows:

$$A_{jk} = \frac{1}{M} \sum_{i}^{N} \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}$$

Here x_i is the number of reference alleles (0, 1 or 2) at the ith SNP for individual j or k, p_i is the frequency of the ith reference allele, and M is the number of SNP markers.

By solving the LMM with the matrix A as above, it was possible, in effect, to fit all markers simultaneously rather than the single-marker regression and P-value threshold approach of GWAS. This work by Yang et al. (2010) expanded on previous sibling studies (e.g., Visscher et al. 2006) by analyzing nominally unrelated individuals. The approach of Yang et al. (2010) enabled the analysis of GWAS datasets, which greatly expanded the available number of samples for heritability estimation. Furthermore, the authors reasoned that using unrelated individuals eliminated confounding of common environments, although it opened the possibility of bias due to population stratification, which they controlled and tested for with various strategies. Yang et al. (2010) analyzed human height variation in Europeans with G constructed with ~300 K SNP markers and obtained an estimate of $h^2 = 0.45$. Although lower than previous twin-based estimates, this result was significant because it explained much more variance than had existing GWAS (see below). The LMM and related methods are implemented in the free and continuously underdevelopment software, GCTA (Yang et al. 2011a).

5.6 GWAS and the "Missing" Heritability

Before discussing further the developments involving the estimation of heritability using genome-wide markers, we cover the topic of "missing" heritability. This topic is at the intersection of GWAS (in humans) and the heritability of complex traits.

Following the completion of some of the first, large-scale GWA studies (e.g., Gudbjartsson et al. 2008; Lettre et al. 2008; Weedon et al. 2008), researchers found that only a very small portion of the heritability for human height that they expected (based largely on twin studies) could be explained by genome-wide significant markers. Human height serves as a model for complex trait genetics, and its study actually led to the development of regression analysis (Galton 1886) and the linking of Mendelian genetics with quantitative variation (Fisher 1918). Human height is

known to have an extremely polygenic genetic architecture and high heritability (~0.8 or greater; Silventoinen et al. 2003; van Dongen et al. 2012), which is in stark contrast to the GWAS explained variance circa 2008 of ~5%. The discrepancy was perhaps dramatically dubbed "the case of the missing heritability" (Maher 2008).

Taking all possible traits, genetic architectures, populations, and other factors into account, it is very unlikely that a single mechanism explains all of the missing heritability, and several key hypotheses can be considered (Maher 2008; Manolio et al. 2009; Gibson 2012). The first hypothesis is that there are very many loci with small effects, which GWAS are not powerful enough to detect. Support for this hypothesis would come from a gradual decrease in the "missing heritability" as sample size and marker density of GWAS increase (Maher 2008; Manolio et al. 2009; Gibson 2010). Initial studies explained less than 5% of the variance for height and had ~100 K people and 500 K variants (Gudbjartsson et al. 2008; Lettre et al. 2008; Weedon et al. 2008). Increasing sample size and density to 180 K people with 2.3 million SNPs explained 10% of the variance (Lango Allen et al. 2010). Increasing further to 250K people with 2.5M markers explained 16% of the variance (Wood et al. 2014), and those authors were able to capture 29% of the variance by lowering the significance threshold and implementing the method of Yang et al. (2010, 2011a, b). These results taken together provide support for the small effects spectrum argument, at least for human height.

The second major possibility is that there are rare variants (frequency $<1\%$) with moderate effect size, which cannot be tagged by marker loci. This is a problem because GWAS and marker-based heritability estimates rely on LD to tag causal variants. If causal variants are rarer than genotyped markers (usually $>1\%$), then their variance will be poorly tagged (Gibson 2012; Speed et al. 2012; de los Campos et al. 2015). Therefore, it is argued that sequencing-based, haplotype, and other forms of association studies will be needed (Maher 2008; Manolio et al. 2009; Gibson 2012; Zuk et al. 2014). Haplotype mapping projects, for example, in humans (1000 Genomes Project Consortium 2010), cattle (Daetwyler et al. 2014), and maize (Chia et al. 2012), involving the whole-genome sequencing of large numbers of genetically representative individuals, have been implemented in part to address this issue. Methods to adjust LMM estimates for LD differences between markers and causals are also an active area of research, which is covered in more details below (Speed et al. 2012; Gusev et al. 2013; Lee et al. 2013; de los Campos et al. 2015; Yang et al. 2015a).

There are assortments of additional factors. One of the strongest is simply the fact that the "missingness" of heritability is based upon comparison to twin studies, which can be upwardly biased estimates because of environmental and/or nonadditive genetic confounding (Falconer and MacKay 1996; Visscher et al. 2008; Zuk et al. 2012). Others include the existence of causal structural or copy number variants that aren't well tagged by existing marker data, the shear amount of noise (i.e., false negatives) in GWAS data, genotype-by-environment interaction, and epigenetic or trans-generational and genetic background effects (Maher 2008; Manolio et al. 2009; Eichler et al. 2010; Gibson 2010). Whatever the cause of "missing" heritability for a particular trait or dataset is, explaining it, however intellectually satisfying, is not necessarily practical or the end goal of GWAS (Manolio et al. 2009).

5.7 Recent Developments in Mixed-Model Estimation of Heritability

Here, we review some of the latest advancements in mixed-model estimation of heritability. Most notably are efforts to adjust estimates for the bias that LD can create. Yang et al. (2010) first acknowledged that differences in allele frequency between markers and causal variants could bias estimates. In an excellent review and simulation study, de los Campos et al. (2015) show that even sequencing-level studies, which include the causal variants in the data, will be biased because many markers will be in linkage equilibrium. They argue that methods to adjust for LD are needed, but will likely never fully solve the problem.

Using extensive simulations, Speed et al. (2012) found that mixed-model estimates of heritability were robust to violations of four key assumptions: (1) the genetic architecture is infinitesimal in nature, (2) all markers have equal variances and (3) are drawn from a Gaussian (normal) distribution, and (4) the same is true of residual or error variances. However, they identified variation in the level of LD across the genome as a major source of bias leading to over- or underestimation of variance in different regions. They proposed scaling marker genotypes in order to down-weight SNPs with lots of local LD. They do this by multiplying SNP genotypes by $\sqrt{w_j}$, where w_j is a weight on the jth marker. The analytical method of deriving w_j is somewhat involved, but it has the effect of producing an unbiased estimate of heritability regardless of the patterns of LD across the genome. This approach is implemented in the free and user-friendly software LDAK (Speed et al. 2012; Speed and Balding 2014). Despite the performance of their method in simulation, the authors obtained the same answer for human height as did Yang et al. (2010), a result which they hypothesized was because over-tagged regions and under-tagged ones balance out, at least in this case.

Lee et al. (2013) subsequently argued that LD weighting was not an ideal approach for very dense genotyping data. Instead, they propose a fairly simple approach: minor allele frequency or MAF stratification. The MAF stratification approach works by partitioning genetic variance with multiple variance components, each with a genomic relationship matrix constructed with SNPs in a different MAF bin (e.g., 0.01 to 0.1, 0.1 to 0.2, etc.). Because high LD is only possible when markers have similar allele frequency, partitioning the variance in this way enables better tagging of causals across the frequency spectrum and leads to an estimate of h^2 that is robust to a range of genetic architectures. In response to this study, Speed et al. (2013) argue that if properly tuned, their LD-weighting method is still valid for ultradense marker data and shows that the combination of LDAK plus MAF stratification can improve upon either approach.

Two additional methods for LD adjustment have been published. The first of these is a weighting method called the "LD residual" in which each marker's genotypes are regressed against the set of other markers within a sliding window (e.g., 100 kb) and the residuals of the regression are then used to construct a genomic relationship matrix (Gusev et al. 2013). The method was effective but was very similar to LDAK in comparisons.

The final and most recent approach, GREML-LDMS, combines MAF stratification with LD stratification and is implemented in the popular GCTA software (Yang et al. 2015b). GREML-LDMS involves first calculating an LD score for each SNP, defined as sum of LD r^2 between that variant and all variants within a 20 megabase region centered on the focal variant. The mean LD score across variants in a sliding window is then calculated. Markers were then stratified into four quartiles of mean LD scores and further grouped into seven MAF bins, leading to the creation of 28 LD and MAF stratified genomic relationship matrices, which were used in a mixed-model analysis. The authors compared GREML-LDMS to both MAF-stratified LDAK (Speed et al. 2012, 2013; Speed and Balding 2014) and the LD-residual approach (Gusev et al. 2013) and found it to be more stable and less biased then either across a range of scenarios.

The study of Yang et al. (2015a, b), using GREML-LDMS, includes an empirical analysis of 44 K individuals with 17.6 M imputed variants. The authors found an excess of variance, which was explained by low frequency and rare variants and that the effects of those variants, based on GWAS, tended to be negative. They suggest an evolutionary interpretation in which new mutations tend to decrease height or increase obesity (BMI) and are thus deleterious to fitness and are thus kept at low frequencies in populations by purifying selection. Finally, GREML-LDMS explained 56% of the variance in height. The authors cite newer twin studies that estimate the heritability at 69%, which they argue may still be overestimated. Combined with a calculation, incorporating even rarer variants into their analysis would increase their estimate further: the posit that the heritability of human height is likely between 0.6 and 0.7 and thus the "missing" heritability is in fact negligible.

Finally, while the majority of our discussion focuses on heritability in the narrow-sense, both parametric (Vitezica et al. 2013; Muñoz et al. 2014; Wolfe et al. 2016) and nonparametric (Gianola and van Kaam 2008; De los Campos et al. 2010), genomic relationship matrices can be used to partition the genetic variance into both additive and nonadditive components and thus to estimate heritability in the broad- and narrow-senses, simultaneously.

5.8 *Partitioning Heritability/Variance Based on Functional Genome Annotations*

For complex traits, much of the heritability is explained by SNPs that do not reach the significance threshold in GWAS (Yang et al. 2010). For most traits, the associated variants cumulatively explain just a small proportion of the total heritability (Manolio et al. 2009). Genomic relationship matrices combined with linear mixed models have also been used to partition genetic variance in several other contexts. These include chromosome-scale heritability (Yang et al. 2011b; Speed and Balding 2014) and different functional categories of SNPs (Gusev et al. 2014). Others have used this methodology for GWAS in an approach known as heritability mapping

(Nagamine et al. 2012; Shirali et al. 2016). In this context, Yang et al. (2011a, b) suggested an alternative to hypothesis testing and QTL identification. Instead, the authors propose to focus on the variance explained by all SNPs together (Yang et al. 2011b; Schork et al. 2013). The authors calculated the variance explained by each chromosome separately. This research opened the path to a broad area of study in quantitative genetics, the partitioning of genetic variance or heritability that is associated with functional categories rather than just chromosome segments. Recently, joint estimation of heritability from functional, category-specific variance components was proposed to assess enrichment (Gusev et al. 2014). In this study GWAS data from over 100,000 samples were analyzed for 11 traits. Using variance component methods, they confirmed that some functional categories contributed disproportionally to the heritability, but unlike previous studies, this approach inferred relevant biological function from all SNPs simultaneously instead of one GWAS hit at a time. One of the constraints for the variance components approach is the need of individual genotypes as input, whereas currently the largest GWAS analyses are conducted through meta-analysis using only summary statistics available from individual studies. In an effort to make this kind of approach more feasible, a new methodology for partitioning heritability was introduced called "stratified LD score regression," which uses summary statistics from GWAS and LD information from an external reference panel (Finucane et al. 2015).

6 Meta-analysis Methods

A myriad of GWAS studies have successfully identified variants associated with a phenotype of interest. However, single GWAS can be underpowered, due to small population size, and in many cases, the associated variants explain little of the disease risk variability.

The NHGRI-EBI Catalog of published genome-wide association studies (http://www.ebi.ac.uk/gwas/home) as of September 2016 contains 2520 studies and 24,218 unique SNP-trait associations (P-value $\leq 5.0 \times 10^{-8}$).

In meta-analyses, GWAS from independent studies are aggregated without the need of phenotypic and genotypic data to be available. As a result of combining information, power is gained for the identification of statistically significant variants that exceed a study-wide threshold. Initially, meta-analyses were retrospective studies carried out by combining summary statistics from previously published studies. However, the International HapMap Project and the 1000 Genomes Project (HapMap 2003; The 1000 Genomes Project Consortium 2015) have provided to the research community resources that are now commonly used for imputation in single GWAS and meta-analysis. Both projects characterize human genetic variation by itemizing common and rare/low-frequency variants and describe the patterns of linkage disequilibrium in the human genome. A large list of profiles for more than

400 consortia can be found at consortiapedia (http://consortiapedia.fastercures.org/consortia/) with information on consortium's mission, structure, data sharing, partners, and more.

Currently, prospective meta-analysis studies go through preliminary stages in which certain criteria have to be met such as compatibility in the design of the study and definition and measurement of the trait under analysis (Evangelou and Ioannidis 2013). To achieve the compatibility in the design of the GWAS, a number of research groups can agree to cooperate prior to the beginning of the study. Individual research groups that are part of a large meta-analysis study conduct quality control checks for each SNP and impute at a whole-genome level usually with a reference panel. Further in the study, association statistics for each SNP are computed, and these summary statistics are provided to the meta-analysis centers. A comprehensive protocol describing state-of-the-art procedures to conduct and perform QC of large-scale genome-wide association meta-analysis GWAMAs has been published (Winkler et al. 2014). Several studies have been successful in leveraging the information from isolated studies by increasing sample size. Meta-analyses of GWAS of blood lipids, BMI, blood pressure, and other disease biomarkers have led to the identification of new loci undetected by earlier, smaller GWAS (Swerdlow et al. 2016).

Meta-analysis studies can combine GWAS results from different phenotypes with a similar disease connotation. For example, in an inflammatory bowel disease (IBD) study by combining autosomal genotype-level data from 15 GWAS of Crohn's disease and/or ulcerative colitis, two common forms of IBD, 71 new associations to IBD were identified (Jostins et al. 2012). After the first stage of the meta-analysis in which significant variants are identified, a set of secondary analyses can be performed such as expression quantitative locus, gene expression, pathway analyses, and protein-protein interaction analyses, among others (Panagiotou et al. 2013). Additionally, to prioritize variants that met genome-wide significance threshold, associated genes can be screened against co-expression networks and tested for enrichment in gene ontology terms.

A meta-analysis including data from 16 studies from the EArly Genetics and Lifecourse Epidemiology (EAGLE) Consortium and the Australian Asthma Genetics Consortium (AAGC) identified ten loci influencing allergic sensitization. In this study, to identify the molecular mechanisms underlying each of the ten loci, they searched for cis-acting expression quantitative trait loci (eQTLs) using gene expression data obtained from six cell types or tissues (Bønnelykke et al. 2013). Functional and enrichment analyses are frequently the follow-up procedure to genome-wide meta-analyses. In a meta-analysis for Dupuytren's disease (DD), based on three datasets comprising in total 1580 cases and 4480 control samples from Germany, Switzerland, and the Netherlands to understand the pathogenesis of DD, GWAS results were integrated with the whole transcriptome data (Becker et al. 2016). In this study functional modules were identified of genes/proteins overrepresented in DD case/control datasets.

The final output of a meta-analysis is an overall synthesis of the results of the series of single GWAS included in the analysis. The main statistical approach to perform a meta-analysis is to combine *P*-values or to estimate the magnitude of the

effects sizes from the combined studies (Li and Ghosh 2014). Possible differences among the GWA studies are detected through heterogeneity tests with the most popular ones being I^2 and Cochran's Q (Zeggini and Ioannidis 2009).

Meta-analysis methods can be classified as *P*-value based or regression coefficient based. Fisher's combined probability test (Fisher 1932) and Stouffer's *Z*-test model the data or the effect sizes from the combined studies. In Fisher's method, $X_F^2 = -2\sum_{i=1}^{k} \ln[P_i]$, where P_i is the *P*-value for the *i*th study and *k* is the number of studies in the meta-analysis. When all of the null hypotheses of the *k* tests are true, the X_F^2 will have a X^2 distribution with $2k$ degrees of freedom.

In the *Z*-transform test, $\sum_{i=1}^{k} Z_i/\sqrt{k}$, the sum of these *Zi*'s, divided by the square root of the number of tests, *k*, has a standard normal distribution if the common null hypothesis is true. An alternative method is the weighted *Z*-test, which has more power and more precision than does Fisher's test. One of the most important limitations of the methods described above is that they do not address between-study heterogeneity.

Based on regression coefficients, meta-analysis can be conducted on the basis of a fixed effect or random effect model. Under the fixed effect model, the true effect size is assumed to be identical for all studies, and the effect size variation between studies is considered as a random error (Ioannidis et al. 2007). Under the random effects model, the goal is not to estimate one true effect, but to estimate the mean of a distribution of effects. Software available for meta-analyses have been extensively described and reviewed elsewhere (Bax et al. 2007; Wallace et al. 2009; Magi and Morris 2010; Viechtbauer 2010; Willer et al. 2010). Commonly used bioinformatics tools and software used for GWAS meta-analysis are METAL, GWAMA, MetABEL, PLINK, and functions within R packages. The CRAN Meta-Analysis *Task View* (https://cran.r-project.org/web/views/MetaAnalysis.html) lists R packages classified according to their functionally within the different stages of a meta-analysis.

7 Post-GWAS Prioritization/Data Mining

A GWAS output is typically a list of selected loci with *P*-values below a significant threshold. However, the interpretability of results is severely impaired by the uncertainty over which exact SNP is the causal variant due to the existence of variants in linkage disequilibrium. Dense genotyping arrays that contain all common SNPs within previously identified risk loci have been used in fine mapping studies (Spain and Barrett 2015). Genotyping platforms used by international consortia in GWAS for diverse diseases and traits comprise the Immunochip (Trynka et al. 2012a), the Metabochip (Voight et al. 2012), the iCOGS array (Michailidou et al. 2013), and the OncoArray (http://epi.grants.cancer/gov/oncoarray/).

With the availability of whole-genome sequence data, SNP variants prioritized through GWAS can be annotated and identified as causal variants based on their

position in coding and noncoding regions of a genome. Challenges arise when these variants are functionally annotated and the identified SNPs are outside of protein-coding genes, which can indicate an underlying regulatory role. Many GWAS signals have been found to be significantly overrepresented in regulatory regions of the genome and directly implicated in complex disease etiology (Hindorff et al. 2009; Nica et al. 2010; Nicolae et al. 2010; Ernst et al. 2011; Maurano et al. 2012; Schaub et al. 2012).

Different criteria and methods can be applied to prioritize the genetic variants identified in GWAS. Most of these methods involve enrichment tests, functional characterization, and integration into biological pathways. A popular approach is to identify the genes carrying significant variants and test for enrichment in gene ontology terms or prioritize genes based on protein-protein interaction networks or pathway databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, Molecular Signatures Database (MSigDB), the network data exchange (NDEx), and ConsensusPathDB, among others (Ogata et al. 1999; Rossin et al. 2011; Pratt et al. 2015; Herwig et al. 2016). Possible shortcoming when following theseannotation-based enrichment methods is a strong bias toward prior available knowledge.

With advances in high-throughput technologies, data from a breadth of large-scale omics profiling can be used to model the interactions that occur from DNA to phenotype through intermediate molecular traits. GWAS results from complex traits have shown the involvement of many loci with small effects possibly interacting with few genes of moderate effect. Within this context, systems genetic approaches have been proposed to reveal the genetic basis of complex traits through a framework that includes the identification of loci through GWAS, genomic analyses, investigation of the effect and variation of multiple intermediate molecular phenotypes and network modeling, and causal inference analysis to construct the molecular circuitry from genotype to phenotype (van der Sijde et al. 2014; Ritchie et al. 2015).

Recently, Claussnitzer et al. (2015) provided a candidate mechanistic basis for the association between the *FTO* obesity-associated locus and obesity. The FTO study is the result of a combination of public resources (epigenomic annotations, chromosome conformation, and regulatory motif conservation), targeted experiments for risk and nonrisk haplotypes (enhancer tiling, gene expression, and cellular profiling), and directed perturbations in human primary cells and mouse models (regulator-target knockdown and overexpression and CRISPR-Cas9 genome editing).

Methods such as Data-driven Expression Prioritized Integration for Complex Traits (DEPICT, www.broadinstitute.org/depict) are not driven by phenotype-specific hypotheses and consider multiple lines of complementary evidence to accomplish gene prioritization, pathway analysis, and tissue/cell type enrichment analysis (Pers et al. 2015). A powerful resource to dissect gene regulatory networks across tissues and higher-order networks across multiple tissues is the Encyclopedia of DNA Elements (ENCODE). Similar databases, like the Genotype-Tissue Expression (GTEx) Project (http://www.gtexportal.org/home/), study the relationship between genetic variation and gene expression and other molecular phenotypes in multiple reference tissues.

7.1 Integrating GWAS Results and Functional Elements in the Genome

As previously reviewed, most of the SNPs associated with human diseases are located in noncoding regions even after accounting for LD and fine mapping (Tak and Farnham 2015). Moreover, most SNPs in LD with these "index" SNPs fall also into these noncoding regions leaving researchers with a large possible list of causative SNPs to test. It is necessary then to prioritize SNPs from this list for further analysis when testing for causality. Besides the methods previously described, SNPs associated with a specific disease or trait can be prioritized using functional annotations discovered with the methods listed in the genome functional annotation section.

In a typical functional enrichment analysis, the most strongly associated SNPs in a GWAS are examined to test whether they fall disproportionately under a certain genomic category (Pickrell 2014). These kinds of studies first identified that GWAS were enriched in protein-coding exons, promoters, and UTRs (Hindorff et al. 2009; Schork et al. 2013), but recently with the availability of resources, such as ENCODE, an increasing number of studies are finding significant enrichments toward functional genomic regions far away from coding sequences. For example, one study systematically investigated the association of multiple types of ENCODE data with disease-associated SNPs across 4724 GWAS for 470 different traits. The study found that 36% of the associated SNPs are in DNase I hypersensitive sites (DHS) and 20% fall within a ChIP-seq peak in at least one cell line (Schaub et al. 2012). Similarly, Ernst et al. (2011) mapped nine chromatin marks across nine human cell types to define 15 chromatin states and found that disease variants are enriched in enhancers identified in the relevant cell types (Ernst et al. 2011). These findings are also consistent with a recent association study in type 1 diabetes (T1D) that found that fine-mapped T1D-associated SNPs are located in active enhancers (Onengut-Gumuscu et al. 2015; Li et al. 2015a). This pattern of enrichment toward functional or regulatory genomic regions is not limited to human cells; Rodgers-Melnick et al. (2016) used a modified version of the classic MNase-seq protocol that tags open chromatin preferentially under light digestion (Vera et al. 2014), to investigate the chromatin landscape in maize. They found that MNase-hypersensitive (MNase HS) regions were associated with gene expression, epigenetic modifications, and patterns of recombination and that, although they map to less than 1% of the maize genome, consistently explain a large portion of the heritable phenotypic variance (~40%) in several complex traits (Rodgers-Melnick et al. 2016).

7.2 Modeling with Functional Annotations

Enrichment test and heritability partitioning assays have shown that several functional regions are important for the regulation of several traits and diseases helping to make sense of GWAS results. However, can the rich source of functional genomic

information that is available nowadays lead to an improvement in GWAS power? We will explore in the next paragraphs some of the efforts that carried tried out to answer this question.

Several studies have found that, in some cases, SNPs associated with a trait are enriched in regulatory regions in specific cell types (Ernst et al. 2011; Trynka et al. 2012b; Gerasimova et al. 2013; Karczewski et al. 2013). Gerasimova et al., for example, were able to predict which cell types contribute to asthma, based on an overlap of the disease-associated SNPs with regulatory regions in a given cell type (Gerasimova et al. 2013). Moreover, they envisioned that it would be possible to rerun the GWAS analysis with higher power by limiting the SNPs to those being in regulatory regions of cell types relevant to the trait or disease. Taking this one step further, Pickrell (2014) built a hierarchical model to jointly analyze GWAS and multiple genomic annotations (Pickrell 2014). Using this statistical framework, the study analyzed data for 18 diseases and 450 genome annotations. The joint model was able to identify a sparse set of biologically interpretable annotations without prior knowledge of the biology of the phenotype. More importantly, the joint model approach was able to exploit the functional annotations to identify high-confidence associations that did not reach genome-wide significance (Pickrell 2014).

Bayesian mixture model (BayesR) is a method that jointly fits all genotypes and allows to map causal variants, study genetic architecture, and provide genomic predictions for genotyped but un-phenotyped individuals (Kemper et al. 2015; Moser et al. 2015). However, this Bayesian framework assumes that all genotypes are equally likely to affect the trait of interest, ignoring any prior biological knowledge. MacLeod et al. (2016) presented an extension of BayesR that can incorporate prior biological information by defining categories of variants likely to be enriched for causal mutations (MacLeod et al. 2016). In the proof-of-concept study, BayesRC increased both the power to detect causal variants and the accuracy of genomic prediction.

7.3 Networks and Pathways

A pathway is composed of a number of genes that are coordinated to accomplish a biological process. In GWA studies, the identified variants can be mapped onto genes, which can in turn be assigned to a biological pathway or to networks of expression data/protein complex (Wang et al. 2010; Leiserson et al. 2013). Pathway-based analysis methods can be broadly defined based on the algorithms used. Examples include overrepresentation analysis, gene set-based scoring, multivariate approaches, and topological-based analysis (Jin et al. 2014). The original pathway-based analyses were performed with microarray data motivated by the fact that functionally related genes which are also found to be co-expressed can help to identify relevant pathways (Subramanian et al. 2005). Pathway Commons stores biological pathway information that is outsourced from public pathways

(http://www.pathwaycommons.org/) and provides links to different apps to visualize and analyze pathways.

For the interpretation of GWAS results, the use of interaction networks can help to prioritize genes that are known to interact with other genes, which are part of a biological pathway of interest. In a network, the levels of interaction among genes are represented by relationships as the ones established in molecular experiments. In general, the vertices of a network represent genes and their encoded protein product, while the edges join these vertices due to a biochemical interaction. Edge weights represent biological evidence coming from sources such as tissue-specific expression, pathway membership, common functional annotations, and similar domain composition (Gilman et al. 2011).

Protein-protein interaction networks (PPIs) describe interactions, such as the relationship of proteins within a protein complex or under certain biological conditions. The subsets of PPI are often obtained in small-scale studies that describe protein binding, and the results are often available in online databases (Sardiu and Washburn 2011). The most frequently used protein-protein interaction (PPI) networks include experimental and prediction databases such as Human Protein Interaction Database HPRD (Peri et al. 2004), Biological General Repository for Interaction Datasets BioGRID (Stark et al. 2006), search tool for the retrieval of interacting genes/proteins STRING (von Mering et al. 2005), database of interacting proteins DIP (Xenarios et al. 2002), Munich Information Center for Protein Sequences (MIPS) (Mewes et al. 1999), and Reactome (Vastrik et al. 2007; Fabregat et al. 2016). Network-assisted search for enriched protein-protein interactions (PPIs) is based on the *P*-values for genes identified through GWAS. Reactome FI (http://www.reactome.org/ used as the PPI reference dataset. This dataset contains 11,879 nodes and 217,249 edges and is by design enriched for true biologically functional relationships (Wu et al. 2010).

Protein interaction network-based pathway analysis (PINBPA) for genome-wide association studies (GWAS) has been developed as a Cytoscape app, to enable analysis of GWAS data in a network fashion (Wang et al. 2015a). PINBPA requires gene-level summary statistics (*P*-values) generated usually by Versatile Gene-based Association Study (VEGAS) program. VEGAS reads in SNP association *P*-values annotate SNPs in genes, produce a gene-based test statistic, and then use simulation to calculate an empirical gene-based *P*-value (Liu et al. 2010). For the network analysis with PINBPA, a set of genes with significant *P*-values is used to search for gene-modules using a greedy search algorithm. Relevant modules are selected based on a z-score and network size.

A popular genomic tool implemented as an R package is dmGWAS which is a dense module searching (DMS) method that applies a greedy algorithm to search for modules in a PPI network (Jia et al. 2011). Recently, the R package dmGWAS 3.0 was used to search for enriched modules in Dupuytren's disease using gene-based *P*-values as node weights and differential co-expression of genes as edge weights (based on the whole transcriptome dataset) (Becker et al. 2016). Additionally, the newly developed EW_dmGWAS algorithm implemented in dmGWAS 3.0 was used to integrate GWAS signals and gene expression profiles to extract dense modules

from a background PPI network. In EW_dmGWAS differential gene co-expression (DGCE) is used to infer edge weights of network modules in the human PPI network (Wang et al. 2015b).

Although pathway membership can help to prioritize genes from sets of significant variants, its effectiveness is limited when biological functional information is incomplete and when genes are solely selected based on a similar functional annotation. Alternatively, to overcome this problem, it has been proposed that SNPs reported by GWAS act as seeds for LD-based query regions. PrixFixe is an R package (http://llama.mshri.on.ca/~mtasan/GranPrixFixe/html/), which uses shared-function or "cofunction" networks (CFNs) to prioritize genes. In brief, SNPs are mapped to a genomic region within an LD range, nearby genes are defined as candidates, and functional connections are identified between genes across distinct candidate sets (Tasan et al. 2014).

8 Epigenome-Wide Association Studies (EWAS)

The comprehensive search for causative variation has expanded from eQTLs, transcription factor binding sites, DNase hypersensitive sites, histone modifications to microRNAs, and DNA methylation studies. MicroRNAs (miRNAs) are a family of small noncoding RNAs that play important regulatory roles in many physiological and disease processes principally at the posttranscriptional level (Ambros 2004; Miska 2005). An integrative analysis of the NHGRI GWAS catalog and the 1000 Genomes databases identified 211,687 trait-/disease-associated single nucleotide polymorphisms (TASs). Of these, 12, 41, and 2041 TASs occur within miRNA precursors, miRNA promoter regions, and 3-UTRs, respectively (Bulik-Sullivan et al. 2013). miRNAs may regulate chromatin structure by regulating key histone modifiers, and taken together, miRNAs can be considered important players in the epigenetic control of gene expression (Chuang and Jones 2007; Wise et al. 2015).

The term epigenetics was first used by Conrad Hal Waddington back in 1939; he defined epigenetics as the study of changes in phenotype without changes in genotype (Waddington 2012; Allis and Jenuwein 2016). Now we know that most of these changes are transduced by epigenetic mechanisms that can modify the expression patterns of an organism cells in a heritable fashion without altering the DNA sequence (Allis and Jenuwein 2016). One classic example of an epigenetic phenomenon is cell differentiation in a multicellular organism; while cells have an identical genotype, the developmental process generates a vast number of cells with differentiated expression profiles and cellular functions (Goldberg et al. 2007).

Over the years, however, epigenetics has adopted different meanings with independent roots; Adrian Bird, for example, proposed a modern and broader definition as "the structural adaptation of chromosomal regions so as to register, signal, or perpetuate altered activity states" (Bird 2007). Despite the troubles in defining epigenetics, the molecular mechanisms underlying it are well described and include DNA

methylation, histone modifications, noncoding RNAs, nucleosome remodeling, and histone variants (Handy et al. 2011; Rakyan et al. 2011; Allis and Jenuwein 2016).

Epigenome-wide association studies (EWAS) have emerged as a different way of applying genome-wide assays to identify regions of the genome that could be affecting individual phenotypes (Birney et al. 2016). Most EWAS use DNA methylation, as this epigenetic mark is easily profiled using current microarray and sequencing technologies (Yong et al. 2016). DNA methylation marks in the genome can be interpreted as a proxy for the GxE interaction, and at specific sites, methylation can affect gene expression (Lin et al. 2016). EWAS looks for differentially methylated regions (DMR), where the pattern observed is discriminatory between cases and controls. EWAS assumes that different levels of methylation at certain locus would be associated or even be causal of an observed phenotype (Rakyan et al. 2011; Birney et al. 2016).

EWAS have been used in large-scale studies in humans including cancer (Michels et al. 2013), autoimmune diseases (Jeffries and Sawalha 2015), mental health problems (Shimada-Sugimoto et al. 2017), and chronic conditions (Ligthart et al. 2016), among others. This novel method presents exciting opportunities to discover association that might be heavily affected by environmental interactions and to generate new insights into disease mechanisms and transcriptional regulation affecting phenotypes.

The epigenetic signatures are variable over time complicating the design and interpretation of EWAS results (Michels et al. 2013). Several studies have stressed the importance of using good practices to minimize spurious associations and noncausal associations (Michels et al. 2013; Birney et al. 2016). These studies agreed that to improve the interpretability of epigenetic studies, a good experimental hypothesis is paramount for success. Other considerations include accounting for cell subtype heterogeneity, controlling for population structure, choosing an appropriate DNA methylation profiling protocol, and whenever possible trying to validate the results with a different molecular technique. Birney et al. (2016) suggested that genotyping and profiling the transcriptome of the same individuals under study would allow for a better interpretation of the epigenetic changes.

9 Genomic Resources for Genome-Wide Association Studies

In this section, we provide some examples of the genomic resources available for a few representative species. Most of the information covered here is related to the development of large consortia for whole-genome sequencing projects and the generation of databases to share tools and genotypic and phenotypic resources.

9.1 Resources for Genome-Wide Association Studies in Humans

In humans, during the past years, genome-wide association studies have identified a large number of associations between genomic regions and complex human disease including type 1 diabetes (Barrett et al. 2009), type 2 diabetes (Saxena et al. 2007; Scott et al. 2007; Zeggini et al. 2007) , bipolar disorder (Scott et al. 2009), Crohn's disease (Franke et al. 2010), and several others (WTCCC 2007; Pickrell 2014).

Genome-wide association studies have been successful in humans due to the design of these studies in which hundreds of thousands or millions of SNPs are genotyped in large cohorts of individuals. A catalog of published GWA studies is kept in the NHGRI-EBI Catalog of published genome-wide association studies (http://www.ebi.ac.uk/gwas). In Table 1, we summarize some of the results found in this repository for the leading causes of death and disability in the United States according to the CDC: heart disease, stroke, cancer, type 2 diabetes, obesity, and arthritis (https://www.cdc.gov/chronicdisease/overview/). Many of these discoveries are the result of the development of human genetic variation catalogs of such as the HapMap, the Wellcome Trust Case Control Consortium (WTCCC), and the *1000 Genomes Project*.

The *International HapMap Project* was launched in October 2002 as a database of human sequence variation at a genome-wide level. It was composed of research groups located worldwide with genotyping centers being responsible for genotyping samples on chromosomal regions previously assigned (HapMap 2003). Single nucleotide polymorphisms (SNPs) identified throughout the genome have been made accessible through dbSNP database in NCBI (http://www.ncbi.nlm.nih.gov) following the data-release principles of a "community resource project." However, the NCBI HapMap webpage on June 16, 2016, announced that NCBI is retiring the HapMap resource due to low number of users accessing the data through this portal.

The *1000 Genomes Project* ran between 2008 and 2015 and set out to provide a comprehensive description of common human genetic variation by the application of whole-genome sequencing to a diverse set of individuals from multiple populations.

The project has finalized with the reconstruction of the genomes of 2504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and high-density microarrays (Consortium et al. 2010; The 1000 Genomes Project Consortium 2015). The primary use of the *1000 Genomes Project* data has been the imputation of SNP variants in lower SNP density datasets. In order to maintain the resources generated by this project, the international genome sample resource (IGSR) has been created (http://www.1000genomes.org).

The *Wellcome Trust Case Control Consortium* (WTCCC) was formed in 2005 to explore the utility, design, and analyses of GWA. Fifty research groups from the United Kingdom active in researching the genetics of common human diseases compose this consortium. In 2007, the WTCCC presented GWA studies of 2000 cases and 3000 shared controls for seven complex human diseases of

Table 1 GWAS catalog genes-regions associated with most deadly chronic diseases and conditions (http://www.ebi.ac.uk/gwas)

Studies	Associations	Trait	Reported genes				
98	526	Breast cancer	ABCC4	GRIK1	LSP1	DKFZp761E198	KIAA1752
			ADAM29	HNF4G	PEX14	DNAJC1	KLF12
			ADHB8	INHBB	PRC1	EBF1	KLF5
			ANKLE1	Intergenic	PTHLH	ECHDC1	LOC643714
			ANKRD16	ITPR1	RAB3C	MAP3K1	MIR1208
			ARHGEF5	KLF12	RAD51L1	MDM4	MIR1972-2
			ARRDC3	KLF4	EMID1	MERIT40	MLLT10
			ATF7IP	KLF5	ERBB4	METAP1D	MYEOV
			BCL2L15	LGR6	ESR1	MKL1	N4BP2L1
			BRCA2	LINC00160	FAM126B	RALY	NDUFB3
			C19orf61	MRPS30	FAM46A	RANBP9	NOBOX
			CCDC88C	MYC	FBN1	ROPN1L	OVOL1
			CDCA7	ORAOV1	FGF3	SIAH2	PIK3C2B
			CDKN2A	NRIP1	FGFR2	SLC45A1	RHBDD3
			CDYL2	NTN4	FOXQ1	SLC4A7	UBE2T
			CHST9	PAX9	FTO	SOGA2	ABHD8
			COL1A1	PDE4D	GALC	SSBP4	CLK1
			COX11	ACTL7A	GLG1	TAB2	DCLRE1B
			CTNNA2	AP4B1	EWSR1	TCF7L2	FGF19
			DIRC3	BARX2	LYPD5	TERT	MUS81
			EIF2S2	C19orf62	MDM4	TET2	PTPN22
			FBXO18	C6orf37	N4BP2L2	TGFBR2	FGF4
			ISYNA1	C6orf97	PTPN7	TMEM45B	TNIP3
			KCNN4	CDKN2B	RAD23B	TNFSF10	TNRC9
			ZNF365	DLX1	SNX32	RNF146	TOX3

(continued)

Table 1 (continued)

Studies	Associations	Trait	Reported genes				
			ZNF577	EGOT	CFL1	SLC25A21	WDR43
			LMO4	LRRN2	CHEK2	ASIP	ZC3H11A
			PPIL3	DLX2	HIPK1	CCND1	ZMIZ1
			ZNF283	ELL			
15	182	Lung cancer	BRCA2	ADH1B	DAXX	CHRNA3	PRSS16
			CHEK2	ENPEP	ZBTB9	CHRNA5	PSMA4
			PSMA4	NAF1	LINC01626	CLPTM1L	PSORS1C1
			TERT	MIR1269A	LINC-PINT	COL11A2	RIF1
			SLC30A8	MIR548I2	LOC401312	COL6A3	RIMS1
			Intergenic	LOC100506688	CROT	COL6A6	RP1
			C6orf72	SLC12A7	PER4	COPS2	RPRM
			CHRNA5	ARAP3	PVT1	CSK	SIX2
			CRP	RUFY1	DLGAP2	CTD-3080P12.3	SLC43A1
			IL1RAP	CD164	TRIM35	DCBLD1	SLK
			BAT3	RFX6	CHRNA2	DRD5	SP3
			CLPTM1L	RNASET2	NRG1	EPHX2	SPRED2
			NPY1R	FGFR1OP	C8orf4	FRY	VGLL2
			CHRNA3	LRRC16A	SOX17	GNL1	XCL2
			TGM5	SLC17A4	NR4A3	GUSBP2	ZMAT4
			TP63	TRIM38	MEGF9	HCG4	ZNF311
			MIPEP	HIST1H4H	LOC101448202	HIST1H1A	ZNF322
			MTMR3	BTN3A1	MIR31HG	HIST1H2BL	ZSCAN16-AS1
			GPC5	ABT1	MTAP	HLA-C	ZSCAN23
			RAD52	ZNF322	CDKN2B-AS1	HLA-DMB	ZSCAN31
			BPTF	LINC00240	SLF2	HLA-DOB	CHRNB4
			TNFAIP6	MIR3143	OBFC1	HLA-DPA1	LOC123688

			VTI1A	ZNF184	AKR1C6P	HLA-DRA	TMEM237
			ROS1	LOC100131289	LINC00900	HLA-E	IKZF2
			HLAclassII	HIST1H2BO	MPZL2	hTERT	COPS8
			DQ141194	ZSCAN12P1	TRIM21	KATNA1	ITSN2
			ACVR1B	PGBD1	RTN4RL2	KIFC1	NCOA1
			SYNGR2	ZSCAN12	PCED1B	LINC00673	SIX3
			ADAR	ZBED9	XRCC6BP1	LINC01012	ACTR2
			LOC101928565	LINC01556	WNK1	LINC01623	LOC102800447
			AK5	OR2J2	FRY	LOC100130451	ETAA1
			ZZZ3	GABBR1	PDS5B	LOC101927237	COL6A5
			FAM73A	IFITM4P	SNRPA1	LOC101927653	CLSTN2
			DNAJB4	ZNRD1	CATSPER2	LOC101927701	DAZL
			GIPC2	TRIM26	WDR76	LOC541472	MSH5-SAPCD1
			MGC27382	TRIM39-RPP21	SEMA6D	LOC646938	SLC44A4
			PTGFR	HLA-E	SECISBP2L	METTL4	TNXB
			LRRC8D	PPP1R18	GALK2	MGC27382	NOTCH4
			ARL6IP6	DDR1	DTWD1	MIR3689A	BTNL2
			ZAK	MUC22	CYP1A2	MIR3939	HLA-DQA1
			CDCA7	C6orf15	CHRNB4	MIR4494	HLA-DQB2
			LOC101927156	HCG27	ADAMTS7	MOG	LOC100294145
			CYP2A6	HCP5	MORF4L1	MSH5	HLA-DOA
			ADH1C	APOM	SOX9	MTAP	HLA-DPB2
			AKR1C1	OR52K2	LINC00470	NCOA1	C6orf10
			ATP8B4	PCDH1	NUMBL	NDUFA4	CCR6
			BAK1	PITX2	OR14J1	NKD2	BUD13
			PLCL2	OR2B2			

(continued)

Table 1 (continued)

Studies	Associations	Trait	Reported genes				
19	102	Coronary heart disease	ACAD10	FN	PCSK9	MORF4L1	WDR12
			ALDH2	FNDC1	PDGFD	PEMT	ZC3HC1
			APOA1	GUCY1A3	PHACTR1	PSRC1	ZNF259
			APOA4	HCG27	PPAP2B	SMCR3	ALDH2
			APOA5	HHIPL1	PPP1R12B	SORT1	ANKS1A
			APOC3	HLA	pseudogene	ATP2B1	COL4A1
			ATP5G1	ILR6	PSRC1	BTNL2	CUX2
			C12orf51	Intergenic	RASD1	C12orf43	CXCL12
			C6orf10	KIAA1462	RPL6-PTPN11	C12orf51	FHL5
			CDKN2B	LDLR	SH2B3	SRR	FLT1
			CNNM2	LIPA	SLC22A3	UBE2Z	HNF1A
			COL4A2	LPA	SMAD3	WDR35	LPA
			CYP17A1	MIA3	SMG6	ABO	LPAL2
			DRB-DQB	MRAS	SNF8	ADAMTS7	MYL2
			GIP	MRPS6	SORT1	CDKN2A	NT5C2
			HLA-C	MTAP	TCF21	CDKN2A/2B	UBE2Q2P1
			VEGFA	MTHFD1L	TTC32	CDKN2A/B	CELSR2
37	135	Stroke	ABCC13	FOXF2	RNU6-36	ADAMTS13	PARK7
			AQP9	FUT2	SH2B3	ADCYAP1R1	PCDH7
			FOXQ1	FUT8	SLC26A11	AGBL1	PITX2
			GAPDHP71	GATA3	SLCO1B1	AIM1	PTPRD
			GLULP3	HDAC9	SPINK2	AKR1D1	PTPRG
			HABP2	HPS4	SPSB4	ALDH1A2	QRICH2
			IL1F10	IL1RN	SUMO2P6	ALDH2	RASIP1
			IL36RN	IMPA2	SUPT3H	ALKBH8	RNF219

			LIPC	Intergenic	TCF7L2	ALPL	CRP
			MIR135A2	JPH3	TCN1	ARID1A	CUBN
			NKX2-5	LIP1	TGFBI	ELTD1	F5
			OBP2B	LOC100507163	TMEM163	ERRFI1	FAF1
			SUPT3H	MICAL2	TRNAK27	ZDHHC22	CHD3
			UBASH3B	MSX2	TSG1	ZFHX3	CLDN17
			WNK1	MYT1L	TSPAN2	C21orf81	DAB1
			ZCCHC14	NBPF3	TTLL5	CACNB2	DPPA3
			ABCC1	NINJ2	VWF	CDC5L	FLJ45455
			ABO	NRCAM	WDFY4	CDKN2C	FLRT2
51	453	Type 2 diabetes	HLA-DQA2	KIF11	MARCKS	ACSL1	GABRA4
			LAMA1	BCL11A	PAX4	SLC35D3	MIR129
			HMG20A	ZBED3	POU5F1	MNX1	GPSM1
			TCF7L2	KLF14	PPARG	PLEKHA1	SLC16A13
			CDKAL1	TP53INP1	R3HDML	HSD17B12	UBE2E2
			SLC30A8	CHCHD9	RBM38	MAP3K11	MAEA
			ADCY5	CENTD2	RBMS1	NRXN3	AP3S2
			FTO	HMGA2	RND3	CMIP	HNF1B
			HHEX	ZFAND6	SACS	ZZEF1	DUSP9
			IG2BP2	PRC1	SATB2	GLP2R	SRBD1
			PPARG	IRS1	SLC16A11	GIP	TYW5
			KCNJ11	WFS1	SNX29	GCKR	LRIG1
			CDKN2B	GRK5	SSR1	ROR2	BCL6
			CDKAL	FAM58A	TCF2	OSBPL1A	SOCS5P5
			CDKN2A	ARF5	TH	LIMS2	MTHFD1L
			KCNQ1	TMEM163	TPT1	DAAM1	ADAMTS20
			IGF2BP2	MAP3K1	TSPAN8	JAG1	GTF2F2
			Intergenic	TGFBR3	XRCC4	JAZF1	LINC00348

(continued)

Table 1 (continued)

Studies	Associations	Trait	Reported genes				
			C2CD4B	DNER	ATP6AP1L	LGR5	C18orf42
			CDC123	TMEM154	HMG1L1	THADA	KLF5
			SPRY2	RREB1	HNF4A	ADAMTS9	CPPED1
			SLC44A3	FAF1	SND1	ST6GAL1	COMMD7
			RBM43	TCF19	PCK1	VPS26A	BCL9
			GALNTL4	LPP	ADAM30	HNF4A	TMEM167A
			TMEM45B	ARL15	AKAP12	GLIS3	MTNR1B
			ANK1	MPHOSPH9	AKAP2	FITM2	RASGRP1
			MGC21675	TMEM75	APOE	ZBTB20	JMJD1C
			PALM2	TMEM18	ARAP1	KIAA1456	VEGFA
			SYK	RNF6	BARX2	SALL4P5	SYN2
			PEX5L	ETV1	C2CD4A	IDE	TMEM175
			PCNXL2	PCBD2	CAMK1D	IGFBP2	CCDC85A
			CR2	LYPLAL1	CAPN13	ANKRD55	FAM60A
			LPIN2	C6orf173	CDKN2B	GCC1	DMRTA1
			C6orf57	C10orf35	CENTD2	FSCN3	ASB3
			ACHE	IL20RA	CHD1L	ARAP1	ATP8B2
			TCERG1L	CRHR2	COX7B2	C2CD4A/B	MIR4686
			PLS1	PTEN	F3	MHC	INAFM2
			FLJ16165	TLE4	GCC1	MACF1	RPL19P16
			HNF1A	DGKB	HHEX	TOMM40	C16orf74
			C14orf70	ST64GAL1	IDE	HLA-DQA1	ADH5P4
			SRR	PROX1	KBTBD8	ABO	WWOX
			PTPRD	GIPR	KLF12	LINGO2	ZPLD1
			ITGB6	KLHDC5	LEP	CTCFL	F3
			TH	C2CD4A	LOC729013	ADAM30	GCC1

			TPT1	TLE1	LPP	AKAP12	HHEX
			TSPAN8	ZMIZ1	MARCKS	AKAP2	IDE
			XRCC4	MC4R	PAX4	APOE	KBTBD8
			ATP6AP1L	SGCG	POU5F1	PAX4	KLF12
			HMG1L1	RHOU	PPARG	PSMD6	LEP
			HNF4A	WISP1	R3HDML	ZFAND3	LOC729013
			SND1	NXN	RBM38	PEPD	LPP
			PCK1	GALNT14	RBMS1	KCNK16	NOTCH2
			ARAP1	HLA-B	RND3	C2CD4A	DCD
			BARX2	HECW1	SACS	CAMK1D	HUNK
			GRB14	CPA6	SATB2	CAPN13	SSR1
			ASCL2	INTS8	SLC16A11	CDKN2B	TCF2
			ANXA2	INS-IGF2	SNX29	CENTD2	CHD1L
			COX7B2				
9	104	Obesity	FTO	TRIM66	SEC16B	LOC105372912	CASC15
			LOC105373769	LINC01122	BDNF	KCNMA1	RPTOR
			FAT1	LOC105371116	TFAP2B	BDNF-AS	ADCY3
			NAT1	LINGO2	BCDIN3D	HS6ST3	CADM2
			LHFPL3	RPS17P5	LOC105378797	ZZZ3	BCDIN3D-AS1
			WWOX	ZNF646P1	GIPR	GNAT2	LINC01122
			TTC28-AS1	ETV5	ETV5	HNF4G	LOC105378803
			NCAM2	LOC441087	SH2B1	LOC105372666	ADCY9
			PCDH9	LOC105371116	FPGT-TNNI3K	POC5	LOC107985132
			INHBB	RABEP2	HOXB-AS3	ALPK1	TMEM18
			ARG1	RNU6-1075P	LOC105378803	LOC107986313	PRKRIRP9
			NRXN3	LOC101928230	TTC28	HOXB5	MAP2K5
			CMYA5	RNU4-17P	TNNI3K	MDFIC	NPC1

(continued)

Table 1 (continued)

Studies	Associations	Trait	Reported genes				
27	321	Arthritis	SALL3	Intergenic	TAGAP	IL2	CD40
			BLK	HLA-DQA2	TEC	SH2B3	PRKCQ
			TRHDE	DPP4	TNFRSF9	BATF	TNFIP3
			ARHGEF3	CDK5RAP2	TPD52	IKZF3	APOM
			HLA-DRB1	MHC	TRAF6	UBASH3A	B3GNT2
			ARL15	CDH6	TXNDC11	STAT4	ANXA3
			GATSL3	SMTNL2	TYK2	REL	CSF2
			PADI4	CD200R1	YDJC	CTLA4	CD83
			SPRED2	RPS12P4	WDFY4	POU3F1	NFKBIE
			AFF3	HLA	ZNF438	KIF3	ARID5B
			CCR6	LDHAL3	COG5	RTKN2	PDE2A
			TRAF1	KCNIP4	MCF2L	IGFBP1	PLD4
			TNFRSF14	ACOXL	DOT1L	GMCL1L	PTPN2
			PTPN22	ANKRD55	GNL3	MED1	ETS1
			RBPJ	ATG5	CDC5L	MTF1	GCH1
			HLA-DRB1	ATM	FTO	P2RY10	PRKCH
			TNFAIP3	C1QBP	DUS4L	PLCL2	ZNF774
			IL2RA	C4orf52	ANKRD55	AHNAK2	PRKCB1
			KIF5A	CCL19	ARAP1	PPIL4	SUPT3H
			TRAF1-C5	CD2	C5	PVT1	TNFRSF14
			MMEL1	CD226	CCL21	RAD51B	FADS3
			CDK6	CD28	CFLAR	RASGRP1	OLIG3
			CCL21	CD5	FADS2	RCAN1	PFKL
			FCRL3	CDK2	FLI1	RUNX1	PIP4K2C
			GATA3	CDK4	GGT6	SFTPD	PLD4

			GRHL2	CEP57	GLT8D1	SYNGR1	PTPN11
			IFNGR2	CASP8	GPR125	IRF8	PXK
			CSF3	CLNK	HLA-DQA1	AIRE	RAG1
			IL20RB	COG6	HLA-DQB1	HLAlocus	RAG2
			IL2RB	CXCR5	IKZF3	IL6ST	IRF4
			C6orf10	ABHD6	IL21	C5orf30	JAZF1
			DNASE1L3	EOMES	IL3	PXK	LBH
			LOC145837	FADS1	INPP5B	IRF5	LOC100506023
			LOC339442	FCGR2A	LOC100506403	CD247	IL6R
			UBE2L3	IRAK1	MHC		

major public health importance: bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D) (WTCCC 2007). Further rounds of GWAS WTCCC1-3 have been carried out since the founding of the consortium (http://www.wtccc.org.uk).

9.2 Resources for Genome-Wide Association Studies in Plants

In plants, the research community has mirrored the progress made in humans with the HapMap Project with the generation of similar resources. Due to its relevance as a crop model, the initiative was led by maize (*Zea mays*) researchers in 2009 with the publication of a first-generation haplotype map of maize (Gore et al. 2009). The project targeted the genetic variation of the maize genome using sequencing technology. HapMap 1 characterized a diverse panel of 27 inbred lines (representative of maize breeding efforts and worldwide diversity)—founders of the maize nested association mapping (NAM) population. HapMap 2 characterized 103 inbred lines representing a wide breadth of the *Z. mays* lineage, comprising 60 improved maize lines, including the parents of the maize nested association mapping (NAM) population, 23 maize landraces, and 19 wild relatives (17 *Z. mays* ssp. *parviglumis* and 2 *Z. mays* ssp. *mexicana*) (Chia et al. 2012). For the maize haplotype version 3 (HapMap3), an international consortium of maize research groups combined resources to build whole-genome sequencing data from 916 maize lines, covering pre-domestication and domesticated *Zea mays* varieties across the world (Bukowski et al. 2017).

Medicago truncatula is a close relative of alfalfa (*M. sativa*) that serves as a model for investigating the genetics and evolution of legume-rhizobia symbiosis (Stacey et al. 2006; De Mita et al. 2007; Heath and Tiffin 2007) in addition to the genetics and evolution of symbiosis with rhizobia and mycorrhizal fungi (Harrison 2005). The Medicago HapMap Project is an international consortium with the task of re-sequencing a diversity panel of 384 inbred lines using Illumina next-generation technology (www.medicagohapmap.org). The resources generated by this project have been used for GWA studies of symbiotic and agronomic traits, linkage disequilibrium, evolution, and drought, among others (Branca et al. 2011; Paape et al. 2013; Stanton-Geddes et al. 2013; Kang et al. 2015).

In wheat (*Triticum*), the first article from the wheat HapMap Project described the global patterns of diversity and selection in the wheat genome (Jordan et al. 2015). The Wheat HapMap variation data of 62 diverse wheat lines was re-sequenced using the whole-exome capture (WEC), and genotyping-by-sequencing (GBS) approaches are freely available for download at the sequence repository hosted by URGI at INRA (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Variations).

The NextGen Cassava Project was launched with the objective—among others—of implementing genomic selection in African breeding programs (http://www.nextgencassava.org/). In addition, phenotypic and genotypic data generated is available in CassavaBase, a centralized, user-friendly, and reliable database (Fernandez-Pozo et al. 2015). Moreover, within the project framework, a cassava haplotype map through deep sequencing 241 diverse accessions identified >28 million segregating variants.

9.3 Resources for Genome-Wide Association Studies in Animals

Due to its important contribution to human nutrition, the generation of genomic resources in animals has been more significant and rapid in the case of the domesticated cattle (*Bos taurus* and *Bos taurus indicus*). The Bovine Genome Consortia composed by both HapMap and Bovine Genome Sequencing and Analysis Consortium had led this effort and represent an extensive collaboration involving more than 300 scientists from 25 different countries (Elsik et al. 2009; Gibbs et al. 2009). In 2009, the Bovine Genome Sequencing and Analysis Consortium published the genome sequence of the taurine cattle, which provided a valuable resource that has since then helped to accelerate livestock genetic improvement for milk and meat production (Elsik et al. 2009).

The Bovine HapMap Consortium characterized the genome structure and annotation with 37,470 single-nucleotide polymorphisms (SNPs) and examined the relatedness in 497 cattle from 19 geographically and biologically diverse breeds (Gibbs et al. 2009). The 1000 Bull Genomes Project on the other hand aimed to build a dataset of sequence variant genotypes of individuals that can provide the bovine research community with genomic resources for imputation of genetic variants into smaller genotypic dataset for genomic prediction and genome-wide association studies in all cattle breeds (http://www.1000bullgenomes.com/). In the first phase of the 1000 Bull Genomes Project, whole genomes of 234 cattle were sequenced, and following a GWA approach, they identified variants associated with milk production and curly coat (Daetwyler et al. 2014).

The 1000 Bull Genomes Project data has been used to test the accuracy of imputation (van Binsbergen et al. 2014) and more recently for targeted imputation in QTL regions based on significance in previous pathway studies, GWAS, BayesR analysis, and signature of selection in the genome (Raven et al. 2016). Similar HapMap initiatives have been followed for other species such as *Ovis aries* (sheep) (http://www.sheephapmap.org/) and *Sus scrofa* (pig) (Groenen et al. 2010).

10 Final Thoughts on Genome-Wide Association Studies

10.1 Advantages

The availability of large datasets has allowed scientists to work toward the identification, in human genetics, of genetic risk factors for common diseases and in plants of markers that can be utilized in marker-assisted selection approaches for breeding. The creation of consortia toward the identification of genes/markers associated with complex traits has increased collaboration among labs. More importantly the scientific community from different fields has benefited from data generated that can be used for functional annotation of associated markers.

Whereas, in the past, most studies of genetic architecture required extensive pedigree or specially designed populations (mating designs), it is now possible to estimate the amount of heritable genetic variation in almost any dataset where genome-wide marker and phenotype information are available. The mixed-model framework for investigating genetic architecture is extremely flexible and can be used to fit almost any model and thus almost any dataset. As we have discussed, genomic mixed models do not only allow estimation of the overall additive genetic variance but can be used in a variety of additional ways. These include partition variance according to chromosome or genome segments, estimating dominance and epistatic variances, testing GWAS results for gene ontology enrichments, and assessing the relative importance of rare versus common alleles.

10.2 Limitations

Perhaps the most important caveat for both GWAS and genomic-mixed model-based is the unavailability of whole-genome sequence datasets; this is the case particularly in plants. Although imputed datasets are available, there is a risk that the genetic variants in low frequency are not detected because of insufficient sample sizes for imputation of rare variants. So far, genome-wide-associated variants can only explain a certain proportion of the genetic variance for a number of phenotypes; the heritability that remains unexplained can be due to variation on the estimates because genetic architecture is not a fixed, completely knowable quantity. Instead, genetic architecture is a property of a definable population at a particular moment in time. We will never discover all causal variants for a trait nor will describe the full range of their interactions both with other causal variants (epistasis) and the environment (GxE).

Genomic mixed models can be computationally intensive to fit. When the number of individuals increases beyond a few thousands, especially if there is replication and/or additional variance components to fit, computers with large amounts of memory and many processors may be required. Most software designed for very

large datasets (tens of thousands to millions) are developed for human genetics and do not handle replication.

As described above, variation in the amount linkage disequilibrium between markers across the genome can generate bias upward or downward of heritability estimates. While, as we discussed, there are approaches to attempt to deal with this, there is currently no way to know how much or what kind of bias occurs. Similarly, unaccounted for or otherwise unknown experimental error and environmental variance can impact variance estimates. For example, in human genetics, datasets unaccounted for shared environments among study individuals could lead to inflation of genetic variance.

10.3 Future Directions and Perspectives

Most of the associated loci identified to date contain hundreds of variants in LD and are located outside protein-coding regions. Current GWAS approaches make it very difficult to identify specific causal variant due to a lack of power and pervasive LD patterns between the individuals analyzed. As a consequence, the biological mechanisms driving these marker-trait associations are poorly understood.

In the future, we foresee GWAS approaches relying on a systems biology approach that will allow to integrate different sources of information, including genetic, biochemical, and evolutionary evidence coupled with appropriate molecular assays that will allow to fully assess the impact of genetic changes onto phenotypes. For complex quantitative traits, the challenge is even bigger as they might be driven by a large number of variants with small effects. In this context, we think that new approaches are necessary; the "omnigenic model" (Boyle et al. 2017), for example, proposed that "disease-risk is largely driven by genes with not direct relevance to disease" and that these genes might be affecting the phenotypes indirectly by interacting with a much smaller number of core genes with direct effect. This kind of approaches together with the development of efficient GWAS frameworks that incorporate functional information (Yang et al. 2017) might supercharge the discovery of causal mutations and the biology underlying each phenotype investigated.

The methodology for estimating heritability with marker data, such as those available in GWAS datasets, remains a rapid area of growth and development. The data available for GWAS and heritability estimation continues to grow and expand at a rapid rate. Not only are marker densities continuing to grow in density and shrink in cost, but also the size of populations and the number of species in which these methods can be applied are still growing quickly.

The number and complexity of phenotype data that are measured for the purpose of studying their genetic architecture is also growing. Increasingly, researchers are collecting transcriptomes, metabolomes, proteomes, and a multitude of highly multivariate, whole-organismal phenotypic measures on hundreds and thousands of genotyped individuals. Highly multivariate phenotype data means that multivariate mixed-model approaches will be a major area for development in the field of

genomic heritability estimation. We will be challenged not just to consider univariate variances but also the genetic covariances among traits. Indeed, the structure of genetic variances and covariances among traits is a key property of a population, which can determine what trait combinations can and cannot evolve.

There is still much computation and software development that is needed to enable fast and easy analysis of unprecedented amounts of data. Mixed models need to be flexible, fast, and easy to apply by the increasingly broad range of researchers with diverse backgrounds and analytic needs. Some examples of theoretical and methodological developments that we expect in the future include improvements to our ability to detect and adjust for linkage disequilibrium in genomic mixed models and improvements in the partitioning of genetic variance components.

11 Conclusions

It is estimated that almost 40,000 associations have been found in humans. However, some people may argue that GWAS have not yet delivered its initial promise as we have not developed cures for all the diseases we have found associations for. We must rethink GWAS not as the magic bullet but just as another layer of information that will allow us to grasp onto the molecular mechanisms that ultimately generate the phenotypes observed.

Genome-wide association analysis and estimation of heritability should now be considered complementary endeavors in the study of the inheritance of complex traits. While GWAS will locate and describe the effects of key QTLs, usually those that have larger effects, genomic mixed-model estimates of heritability and related genetic variances will provide the context for interpreting the importance of those QTL relative to other genetic and environmental factors.

References

Abdi H. Bonferroni and Sidak corrections for multiple comparisons. In: Salkind NJ, editor. Encyclopedia of measurement and statistics. Thousand Oaks: Sage; 2007. p. 103–7.

Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 2010;11:R119.

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.

Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nat Rev Genet. 2016;17:487–500.

Al-Tassan NA, Whiffin N, Hosking FJ, Palles C, Farrington SM, Dobbins SE, et al. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. Sci Rep. 2015;5:10442.

Ambros V. The functions of animal microRNAs. Nature. 2004;431:350–5.

Amin N, van Duijn CM, Aulchenko YS. A genomic background based method for association analysis in related individuals. PLoS One. 2007;2:e1274.

Amos CI. Successful design and conduct of genome-wide association studies. Hum Mol Genet. 2007;16.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.

Aulchenko YS, De Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics. 2007a;177:577–85.

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007b;23:1294–6.

Ay F, Noble WS, Dekker J, Rippe K, Dekker M, Kleckner N, et al. Analysis methods for studying the 3D architecture of the genome. Genome Biol. 2015;16:183.

Bac-Molenaar JA, Vreugdenhil D, Granier C, Keurentjes JJB. Genome-wide association mapping of growth dynamics detects time-specific and general quantitative trait loci. J Exp Bot. 2015;66:5567–80.

Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet. 2009;41:703–7.

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129:823–37.

Bax L, Yu L-M, Ikeda N, Moons KG. A systematic comparison of software dedicated to meta-analysis of causal studies. BMC Med Res Methodol. 2007;7:40.

Becker K, Siegert S, Toliat MR, Du J, Casper R, Dolmans GH, et al. Meta-analysis of genome-wide association studies and network analysis-based integration with gene expression data identify new suggestive loci and unravel a Wnt-centric network associated with Dupuytren's disease. PLoS One. 2016;11:e0158101.

Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y. Hi–C: a comprehensive technique to capture the conformation of genomes. Methods. 2012;58:268–76.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995;57:289–300.

van Binsbergen R, Bink MC, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I, et al. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2014;46:41.

Bird A. Perceptions of epigenetics. Nature. 2007;447:396–8.

Birney E, Smith GD, Greally JM. Epigenome-wide association studies and the interpretation of disease -omics (GS Barsh, Ed). PLoS Genet. 2016;12:e1006105.

Boeger H, Griesenbeck J, Strattan JS, Kornberg RD. Nucleosomes unfold completely at a transcriptionally active promoter. Mol Cell. 2003;11:1587–98.

Bonferroni C. Il calcolo delle assicurazioni su gruppi di teste. In: In Studi in Onore del Professore Salvatore Ortu Carboni, Bardi. 1935. pp 13–60.

Bønnelykke K, Matheson MC, Pers TH, Granell R, Strachan DP, Alves AC, et al. Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. Nat Genet. 2013;45:902–6.

Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169:1177–86.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.

Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, et al. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume Medicago truncatula. Proc Natl Acad Sci U S A. 2011;108:E864–70.

Bucksch A, Burridge J, York LM, Das A, Nord E, Weitz JS, et al. Image-based high-throughput field phenotyping of crop roots. Plant Physiol. 2014;166:470–86.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10:1213–8.
Bukowski R, Guo X, Lu Y, et al. Construction of the third generation Zea mays haplotype map. Gigascience. 2017. https://doi.org/10.1093/gigascience/gix134 [Epub ahead of print].
Bulik-Sullivan B, Selitsky S, Sethupathy P. Prioritization of genetic variants in the microrna regulome as functional candidates in genome-wide association studies. Hum Mutat. 2013;34:1049–56.
Bulmer MG. The effect of selection on genetic variability. Am Nat. 1971;105:201–11.
de los Campos G, Sorensen D, Gianola D. Genomic heritability: what is it? PLoS Genet. 2015;11:e1005048.
Chen GB. Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. Front Genet. 2014;5:1–14.
Chen C-Y, Chang I-S, Hsiung CA, Wasserman WW. On the identification of potential regulatory variants within genome wide association candidate SNP sets. BMC Med Genomics. 2014;7:34.
Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet. 2012;44:803–7.
Chuang JC, Jones PA. Epigenetics and microRNAs. Pediatr Res. 2007;61:24R–9R.
Clark RT, Famoso AN, Zhao K, Shaff JE, Craft EJ, Bustamante CD, et al. High-throughput two-dimensional root system phenotyping platform facilitates genetic analysis of root growth and development. Plant Cell Environ. 2013;36:454–66.
Claussnitzer M, Dankel SN, Kim K-H, et al. FTO obesity variant circuitry and adipocyte browning in humans. N Engl J Med. 2015;373:895–907.
Consortium 1000 Genomes Project, others, Africa W, Consortium T 1000 genomes project, Durbin RM, Altshuler DL, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.
Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15:901–13.
Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. 2006;16:123–31.
Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–65.
Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++ (WW Wasserman, Ed). PLoS Comput Biol. 2010;6:e1001025.
De los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res (Camb). 2010;92:295–308.
De Mita S, Ronfort J, McKhann HI, Poncet C, El Malki R, Bataillon T. Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in nod factor signaling in Medicago truncatula. Genetics. 2007;177:2123–33.
Dingwall C, Lomonossoff GP, Laskey RA. High sequence specificity of micrococcal nuclease. Nucleic Acids Res. 1981;9:2659–74.
van Dongen J, Slagboom PE, Draisma HHM, Martin NG, Boomsma DI. The continuing value of twin studies in the omics era. Nat Publ Gr. 2012;13:640–53.
Eberle MA, Ng PC, Kuhn K, Zhou L, Peiffer DA, Galver L, et al. Power to detect risk alleles using genome-wide tag SNP panels. PLoS Genet. 2007;3:1827–37.
Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11:446–50.
Elsik CG, Tellam RL, Worley KC. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science. 2009;324:522–8.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.
Evangelou E, Ioannidis JPA. Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet. 2013;14:379–89.
Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet. 1995;57:455–64.
Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2016;44:D481–7.
Fahlgren N, Gehan MA, Baxter I. Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. Curr Opin Plant Biol. 2015;24:93–9.
Falconer D, Mackay T. Introduction to quantitative genetics. 4th ed. Harlow, Essex: Longmans Green; 1996.
Farnham PJ. Insights from genomic profiling of transcription factors. Nat Rev Genet. 2009;10:605–16.
Fernandez-Pozo N, Menda N, Edwards JD, Saha S, Tecle IY, Strickler SR, et al. The Sol Genomics Network (SGN)-from genotype to phenotype to breeding. Nucleic Acids Res. 2015;43:D1036–41.
Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. Brief Funct Genomics. 2015;14:130–42.
Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47:1228–35.
Fisher RA. XV.—the correlation between relatives on the supposition of mendelian inheritance. Trans R Soc Edinb. 1918;52:399–433.
Fisher RA. Statistical methods for research workers. Can Med Assoc J. 1932;27:460.
Flint-Garcia SA, Thornsberry JM, Buckler ES 4th. Structure of linkage disequilibrium in plants. Annu Rev Plant Biol. 2003;54:357–74.
Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010;42:1118–25.
Furey TS. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. Nat Rev Genet. 2012;13:840–52.
Galton F. Regression toward mediocrity in hereditary stature. J Anthropol Inst Gt Britain Irel. 1886;15:246–63.
Garner C, Slatkin M. On selecting markers for association studies: patterns of linkage disequilibrium between two and three diallelic loci. Genet Epidemiol. 2003;24:57–67.
Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S. Genetic structure and diversity in Oryza sativa L. Genetics. 2005;169:1631–8.
1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.
Gerasimova A, Chavez L, Li B, Seumois G, Greenbaum J, Rao A, et al. Predicting cell types and genetic variations contributing to disease by combining GWAS and epigenetic data (Y-H Hsu, Ed). PLoS One. 2013;8:e54359.
Gianola D, Rosa GJM. One hundred years of statistical developments in animal breeding. Annu Rev Anim Biosci. 2015;3:19–56.
Gianola D, van Kaam JBCHM. Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics. 2008;178:2289–303.
Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, Gill CA, et al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science. 2009;324:528–32.
Gibson G. Hints of hidden heritability in GWAS. Nat Genet. 2010;42:558–60.
Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2012;13:135–45.

Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron. 2011;70:898–907.
Giresi PG, Lieb JD. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). Methods. 2009;48:233–9.
Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. 2007;17:877–85.
Glazier AM, Nadeau JH, Aitman TJ. Finding genes that underlie complex traits. Science. 2002;298:2345–9.
Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009;136:245–57.
Golan D, Lander ES, Rosset S. Measuring missing heritability: inferring the contribution of common variants. Proc Natl Acad Sci. 2014;111:E5272–81.
Goldberg AD, Allis CD, Bernstein E. Epigenetics: a landscape takes shape. Cell. 2007;128:635–8.
Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, et al. A first-generation haplotype map of maize. Science. 2009;326:1115–7.
Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013;5:578–90.
Green JM, Appel H, Rehrig EM, Harnsomburana J, Chang J-F, Balint-Kurti P, et al. PhenoPhyte: a flexible affordable method to quantify 2D phenotypes from imagery. Plant Methods. 2012;8:45.
Grob S, Schmid MW, Grossniklaus U. Hi-C analysis in Arabidopsis identifies the KNOT, a structure with similarities to the flamenco locus of Drosophila. Mol Cell. 2014;55:678–93.
Groenen MAM, Amaral A, Megens HJ, et al. The porcine HapMap project: genome-wide assessment of nucleotide diversity, haplotype diversity and footprints of selection in the pig. In: Abstract from plant and animal genomes XVIII conference. San Diego; 2010.
Grubert F, Zaugg JB, Kasowski M, Ursu O, Spacek DV, Martin AR, et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell. 2015;162:1051–65.
Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, et al. Many sequence variants affecting diversity of adult human height. Nat Genet. 2008;40:609–15.
Gusev A, Bhatia G, Zaitlen N, Vilhjalmsson BJ, Diogo D, Stahl EA, et al. Quantifying missing heritability at known GWAS loci. PLoS Genet. 2013;9:10–4.
Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. 2014;95:535–52.
Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced applications of RNA sequencing and challenges. Bioinform Biol Insights. 2015;9:29–46.
Handy DE, Castro R, Loscalzo J. Epigenetic modifications: basic mechanisms and role in cardiovascular disease. Circulation. 2011;123:2145–56.
Hangauer MJ, Vaughn IW, McManus MT, Hindorff L, Sethupathy P, Junkins H, et al. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs (JL Rinn, Ed). PLoS Genet. 2013;9:e1003569.
HapMap CTI. The International HapMap Project. Nature. 2003;426:789–96.
Harrison MJ. Signaling in the arbuscular mycorrhizal symbiosis. Annu Rev Microbiol. 2005;59:19–42.
Hartmann A, Czauderna T, Hoffmann R, Stein N, Schreiber F. HTPheno: an image analysis pipeline for high-throughput plant phenotyping. BMC Bioinformatics. 2011;12:148.
Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet. 1972;2:3–19.
Hayes B. Overview of statistical methods for genome-wide association studies (GWAS). Methods Mol Biol. 2013;1019:149–69.
Hayes B, Goddard ME. The distribution of the effects of genes affecting quantitative traits in livestock. Genet Sel Evol. 2001;33:209–29.

Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res (Camb). 2009;91:47–60.
Heath KD, Tiffin P. Context dependence in the coevolution of plant and rhizobial mutualists. Proc Biol Sci. 2007;274:1905–12.
Heffner EL, Sorrells ME, Jannink J-L. Genomic selection for crop improvement. Crop Sci. 2009;49:1.
Henderson CR. Estimation of variance and covariance components. Biometrics. 1953;9:226–52.
Henderson C. Best linear unbiased estimation and prediction under a selection model. Biometrics. 1975;31:423–47.
Henderson CR. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics. 1976;32:69–83.
Herwig R, Hardt C, Lienhard M, Kamburov A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. Nat Protoc. 2016;11:1889–907.
Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theor Appl Genet. 1968;38:226–31.
Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106:9362–7.
Holland JB, Nyquist WE, Cervantes-Martinez CT. Estimating and interpreting heritability for plant breeding: an update. Plant Breed Rev. 2003;22:9–112.
Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. Nat Rev Genet. 2010;11:855–66.
Huber CD, Nordborg M, Hermisson J, Hellmann I. Keeping it local: evidence for positive selection in Swedish Arabidopsis thaliana. Mol Biol Evol. 2014;31:3026–39.
Ioannidis JP, Patsopoulos NA, Evangelou E. Heterogeneity in meta-analyses of genome-wide association investigations. PLoS One. 2007;2:e841.
Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet. 2001;29:217–22.
Jeffries MA, Sawalha AH. Autoimmune disease in the epigenetic era: how has epigenetics changed our understanding of disease and how can we expect the field to evolve? Expert Rev Clin Immunol. 2015;11:45–58.
Jia P, Zheng S, Long J, Zheng W, Zhao Z. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics. 2011;27:95–102.
Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet. 2009;10:161–72.
Jin L, Zuo X-Y, Su W-Y, Zhao X-L, Yuan M-Q, Han L-Z, et al. Pathway-based analysis tools for complex diseases: a review. Genomics Proteomics Bioinformatics. 2014;12:210–20.
John S, Sabo PJ, Thurman RE, Sung M-H, Biddie SC, Johnson TA, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet. 2011;43:264–8.
Jordan KW, Wang S, Lun Y, Gardiner L-J, MacLachlan R, Hucl P, et al. A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biol. 2015;16:48.
Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491:119–24.
Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008;178:1709–23.
Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42:348–54.
Kang Y, Sakiroglu M, Krom N, Stanton-Geddes J, Wang M, Lee YC, et al. Genome-wide association of drought-related and biomass traits with HapMap SNPs in Medicago truncatula. Plant Cell Environ. 2015;38:1997–2011.
Karaletsos T, Stegle O, Dreyer C, Winn J, Borgwardt KM. ShapePheno: unsupervised extraction of shape phenotypes from biological image collections. Bioinformatics. 2012;28:1001–8.
Karczewski KJ, Dudley JT, Kukurba KR, Chen R, Butte AJ, Montgomery SB, et al. Systematic functional regulatory assessment of disease-associated variants. Proc Natl Acad Sci U S A. 2013;110:9607–12.

Keene MA, Corces V, Lowenhaupt K, Elgin SC. DNase I hypersensitive sites in Drosophila chromatin occur at the 5′ ends of regions of transcription. Proc Natl Acad Sci U S A. 1981;78:143–46.
Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A. 2014;111:6131–8.
Kemper KE, Reich CM, Bowman PJ, Vander Jagt CJ, Chamberlain AJ, Mason BA, et al. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genet Sel Evol. 2015;47:29.
Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465:182–7.
Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013;9:29.
Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet. 2012;44:1066–71.
Kruuk LE. Estimating genetic parameters in natural populations using the “animal model”. Philos Trans R Soc Lond B Biol Sci. 2004;359:873–90.
Ku CS, Naidoo N, Wu M, Soong R. Studying the epigenome using next generation sequencing. J Med Genet. 2011;48:721–30.
Lander ES, Schork NJ. Genetic dissection of complex traits. Science. 1994;265:2037–48.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467:832–8.
Le TBK, Imakaev MV, Mirny LA, Laub MT. High-resolution mapping of the spatial organization of a bacterial chromosome. Science. 2013;342:731–4.
Lee C-K, Shibata Y, Rao B, Strahl BD, Lieb JD. Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat Genet. 2004;36:900–5.
Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011;88:294–305.
Lee SH, Yang J, Chen GB, Ripke S, Stahl EA, Hultman CM, et al. Estimation of SNP heritability from dense genotype data. Am J Hum Genet. 2013;93:1151–5.
Leiserson MDM, Eldridge JV, Ramachandran S, Raphael BJ. Network analysis of GWAS data. Curr Opin Genet Dev. 2013;23:602–10.
Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, et al. Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet. 2008;40:584–91.
Lewontin RC. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics. 1964;49:49–67.
Li Y, Ghosh D. Meta-analysis based on weighted ordered P-values for genomic data with heterogeneity. BMC Bioinformatics. 2014;15:226.
Li Y, Haseneyer G, Schön C, Ankerst D, Korzun V, Wilde P, et al. High levels of nucleotide diversity and fast decline of linkage disequilibrium in rye (Secale cereale L.) genes involved in frost response. BMC Plant Biol. 2011;11:1–14.
Li H, Chen H, Liu F, Ren C, Wang S, Bo X, et al. Functional annotation of HOT regions in the human genome: implications for human disease and cancer. Sci Rep. 2015a;5:11633.
Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong C-T, et al. Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. Mol Cell. 2015b;58:216–31.
Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.

Ligthart S, Marzi C, Aslibekyan S, Mendelson MM, Conneely KN, Tanaka T, et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. Genome Biol. 2016;17:255.
Lin X, Barton S, Holbrook JD. How to make DNA methylome wide association studies more powerful. Epigenomics. 2016;8:1117–29.
Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, et al. GAPIT: genome association and prediction integrated tool. Bioinformatics. 2012;28:2397–9.
Lipka AE, Kandianis CB, Hudson ME, Yu J, Drnevich J, Bradbury PJ, et al. From association to prediction: statistical methods for the dissection and selection of complex traits in plants. Curr Opin Plant Biol. 2015;24:110–8.
Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011;8:833–5.
Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A versatile gene-based test for genome-wide association studies. Am J Hum Genet. 2010;87:139–45.
Liu C, Wang C, Wang G, Becker C, Zaidem M, Weigel D. Genome-wide analysis of chromatin packing in Arabidopsis thaliana at single-gene resolution. Genome Res. 2016;26:1057–68.
Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, et al. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. Nat Genet. 2013;45:884–90.
de Los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 2013;9:e1003608.
Lynch M, Walsh B. Genetics and analysis of quantitative traits. Sunderland, MA: Sinauer Associates; 1998.
Mackay TFC. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat Rev Genet. 2014;15:22–33.
MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016;17:144.
Magi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics. 2010;11:288.
Maher B. The case of the missing heritability. Nature. 2008;456:18–21.
Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.
Marbouty M, Le Gall A, Cattoni DI, Cournac A, Koh A, Fiche J-B, et al. Condensin- and replication-mediated bacterial chromosome folding and origin condensation revealed by Hi-C and super-resolution imaging. Mol Cell. 2015;59:588–602.
Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5.
Mellor J, Adam M, Robert F, Larochelle M, Gaudreau L, Adkins MW, et al. The dynamics of chromatin remodeling at promoters. Mol Cell. 2005;19:147–57.
von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005;33(Database issue):D433–7.
Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S, et al. MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 1999;27:44–8.
Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat Rev Genet. 2014;15:709–21.
Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet. 2013;45:353–61.
Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Greally JM, Gut I, et al. Recommendations for the design and analysis of epigenome-wide association studies. Nat Methods. 2013;10:949–55.

Miska EA. How microRNAs control cell division, differentiation and death. Curr Opin Genet Dev. 2005;15:563–8.
Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet. 2015;11(4):e1004969.
Mousseau TA, Ritland K, Heath DD. A novel method for estimating heritability using molecular markers. Heredity (Edinb). 1998;80:218–24.
Muller-Myhsok B, Abel L. Genetic analysis of complex diseases. Science. 1997;275:1328–9.
Muñoz PR, Resende MFR, SA G, Deon M, Resende V. Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics. 2014;198:1759–68.
Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, et al. Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell. 2009;21:2194–202.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320(5881):1344–9.
Nagamine Y, Pong-Wong R, Navarro P, Vitart V, Hayward C, Rudan I, et al. Localising loci underlying complex trait variation using Regional Genomic Relationship Mapping. PLoS One. 2012;7(10):e46501.
Nei M, Li WH. Linkage disequilibrium in subdivided populations. Genetics. 1973;75:213–9.
Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, Dermitzakis ET. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. 2010;6:e1000895.
Nicolae DL, Gamazon E, Zhang W, Duan S, Eileen Dolan M, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6 (4):e1000895.
Nordborg M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics. 2000;154:923–9.
Nordborg M, Donnelly P. The coalescent process with selfing. Genetics. 1997;146:1185–95.
Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 1999;27:29–34.
Onengut-Gumuscu S, Chen W-M, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat Genet. 2015;47:381–6.
Paape T, Bataillon T, Zhou P, Kono TJY, Briskine R, Young ND, et al. Selection, genome-wide fitness effects and evolutionary rates in the model legume Medicago truncatula. Mol Ecol. 2013;22:3525–38.
Pal LR, Yu C-H, Mount SM, Moult J. Insights from GWAS: emerging landscape of mechanisms underlying complex trait disease. BMC Genomics. 2015;16(Suppl 8):S4.
Panagiotou OA, Willer CJ, Hirschhorn JN, Ioannidis JPA. The power of meta-analysis in genome-wide association studies. Annu Rev Genomics Hum Genet. 2013;14:441–65.
Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10:669–80.
Pavy N, Namroud M-C, Gagnon F, Isabel N, Bousquet J. The heterogeneous levels of linkage disequilibrium in white spruce genes and comparative analysis with other conifers. Heredity. 2012;108(3):273–84.
Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, et al. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res. 2004;32: D497–501.
Pers TH, Karjalainen JM, Chan Y, Westra H-J, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun. 2015;6:5890.
Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet. 2014;94:559–73.
Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, et al. NDEx, the network data exchange. Cell Syst. 2015;1:302–5.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–9.
Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 1999;65:220–8.
Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.
Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011;12:529–41.
Rands CM, Meader S, Ponting CP, Lunter G. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. PLoS Genet. 2014;10:e1004525.
Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.
Raven LA, Cocks BG, Kemper KE, Chamberlain AJ, Vander Jagt CJ, Goddard ME, et al. Targeted imputation of sequence variants and gene expression profiling identifies twelve candidate genes associated with lactation volume, composition and calving interval in dairy cattle. Mamm Genome. 2016;27:81–97.
Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet. 2015;16:85–97.
Ritland K. Marker-based method for inferences about quantitative inheritance in natural populations. Evolution. 1996;50:1062–73.
Ritland K. Marker-inferred relatedness as a tool for detecting heritability in nature. Mol Ecol. 2000;9:1195–204.
Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES. Open chromatin reveals the functional maize genome. Proc Natl Acad Sci U S A. 2016;113:E3177–84.
Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet. 2011;7(1):e1001273.
Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol. 2014;32:347–55.
Sardiu ME, Washburn MP. Building protein-protein interaction networks with proteomics and informatics tools. J Biol Chem. 2011;286:23645–51.
Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PIW, Chen H, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science. 2007;316:1331–6.
Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22:1748–59.
Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell. 2008;132:887–98.
Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, et al. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs (G Gibson, Ed). PLoS Genet. 2013;9:e1003449.
Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science. 2007;316:1341–5.
Scott LJ, Muglia P, Kong XQ, Guan W, Flickinger M, Upmanyu R, et al. Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. Proc Natl Acad Sci U S A. 2009;106:7501–6.
Shimada-Sugimoto M, Otowa T, Miyagawa T, Umekage T, Kawamura Y, Bundo M, et al. Epigenome-wide association study of DNA methylation in panic disorder. Clin Epigenetics. 2017;9:6.

Shirali M, Pong-Wong R, Navarro P, Knott S, Hayward C, Vitart V, et al. Regional heritability mapping method helps explain missing heritability of blood lipid traits in isolated populations. Heredity (Edinb). 2016;116:333–8.

Shu W, Chen H, Bo X, Wang S. Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. Nucleic Acids Res. 2011;39:7428–43.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15:1034–50.

van der Sijde MR, Ng A, Fu J. Systems genetics: from GWAS to disease pathways. Biochim Biophys Acta Mol Basis Dis. 2014;1842:1903–9.

Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, et al. Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res. 2003;6:399–408.

Slatkin M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008;9:477–85.

van der Sluis S, Verhage M, Posthuma D, Dolan CV. Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. PLoS One. 2010;5(11):e13929.

Soulé M. Phenetics of natural populations I. Phenetic relationships of insular populations of the side-blotched lizard. Evolution. 1967;21:584–91.

Sozzani R, Busch W, Spalding EP, Benfey PN. Advanced imaging techniques for the study of plant growth and development. Trends Plant Sci. 2014;19:304–10.

Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Hum Mol Genet. 2015;24:R111–9.

Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 2014;24:1550–7.

Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet. 2015;16:33–44.

Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet. 2012;91:1011–21.

Speed D, Hemani G, Johnson MR, Balding DJ. Response to Lee et al.: SNP-based heritability analysis with dense data. Am J Hum Genet. 2013;93:1155–7.

Stacey G, Libault M, Brechenmacher L, Wan J, May GD. Genetics and functional genomics of legume nodulation. Curr Opin Plant Biol. 2006;9:110–21.

Stanton-Geddes J, Paape T, Epstein B, Briskine R, Yoder J, Mudge J, et al. Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in Medicago truncatula. PLoS One. 2013;8(5):e65688.

Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34:D535–9.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.

Sun L, Wu R. Mapping complex traits as a dynamic system. Phys Life Rev. 2015;13:155–85.

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS. Rapid variance components-based method for whole-genome association analysis. Nat Genet. 2012;44:1166–70.

Swarts K, Li HH, Navarro JAR, An D, Romay MC, Hearne S, Acharya C, et al. Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome. 2014;7:1–12.

Swerdlow DI, Kuchenbaecker KB, Shah S, Sofat R, Holmes MV, White J, et al. Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. Int J Epidemiol. 2016;45(5):1600–16.

Tak YG, Farnham PJ. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics Chromatin. 2015;8:57.

Tasan M, Musso G, Hao T, Vidal M, MacRae CA, Roth FP. Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. Nat Methods. 2014;12:154–9.
Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS. Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). Proc Natl Acad Sci U S A. 2001;98:9161–6.
Teo YY, Fry AE, Bhattacharya K, Small KS, Kwiatkowski DP, Clark TG. Genome-wide comparisons of variation in linkage disequilibrium. Genome Res. 2009;19:1849–60.
The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.
Trynka G, Hunt KA, Bockett NA, Romanos J, Castillejo G, de la Concha EG, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet. 2012a;43:1193–201.
Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet. 2012b;45:124–30.
Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. Epigenetics Chromatin. 2014;7:33.
Tulah AS, Holloway JW, Sayers I, Yang I, Savarimuthu S, Kim S, et al. Defining the contribution of SNPs identified in asthma GWAS to clinical variables in asthmatic children. BMC Med Genet. 2013;14:100.
VanLiere JM, Rosenberg NA. Mathematical properties of the r2 measure of linkage disequilibrium. Theor Popul Biol. 2008;74:130–7.
VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci. 2009;92:16–24.
Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, et al. Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 2007;8:R39.
Vera DL, Madzima TF, Labonne JD, Alam MP, Hoffman GG, Girimurugan SB, et al. Differential nuclease sensitivity profiling of chromatin reveals biochemical footprints coupled to gene expression and functional DNA elements in maize. Plant Cell. 2014;26:3883–93.
Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010;36:1–48.
Vinkhuyzen AA, Wray NR, Yang J, Goddard ME, Visscher PM. Estimation and partitioning of heritability in human populations using whole genome analysis methods. Annu Rev Genet. 2013;47:75–95.
Visscher PM. Sizing up human height variation. Nat Genet. 2008;40:489–90.
Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2006;2:0316–25.
Visscher PM, Macgregor S, Benyamin B, Zhu G, Gordon S, Medland S, et al. Genome partitioning of genetic variation for height from 11,214 sibling pairs. Am J Hum Genet. 2007;81:1104–10.
Visscher PM, Hill WG, Wray NR. Heritability in the genomics era—concepts and misconceptions. Nat Rev Genet. 2008;9:255–66.
Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90:7–24.
Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101:5–22.
Vitezica ZG, Varona L, Legarra A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics. 2013;195:1223–30.

Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet. 2012;8(8):e1002793.

Waddington CH. The epigenotype. Int J Epidemiol. 2012;41:10–3.

Wallace BC, Schmid CH, Lau J, Trikalinos TA, Lau J, Schmid C, et al. Meta-analyst: software for meta-analysis of binary, continuous and diagnostic data. BMC Med Res Methodol. 2009;9:80.

Wallace JG, Bradbury PJ, Zhang N, Gibon Y, Stitt M, Buckler ES. Association mapping across numerous traits reveals patterns of functional variation in maize. PLoS Genet. 2014;10:e1004845.

Walter A, Liebisch F, Hund A. Plant phenotyping: from bean weighing to image analysis. Plant Methods. 2015;11:14.

Wang L, Uilecan IV, Assadi AH, Kozmik CA, Spalding EP. HYPOTrace: image analysis software for measuring hypocotyl growth and shape demonstrated on Arabidopsis seedlings undergoing photomorphogenesis. Plant Physiol. 2009;149:1632–7.

Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet. 2010;11:843–54.

Wang L, Matsushita T, Madireddy L, Mousavi P, Baranzini SE. PINBPA: cytoscape app for network analysis of GWAS data. Bioinformatics. 2015a;31:262–4.

Wang Q, Yu H, Zhao Z, Jia P. EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles. Bioinformatics. 2015b;31:2591–4.

Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet. 2008;40:575–83.

Weintraub H, Groudine M. Chromosomal subunits in active genes have an altered conformation. Science. 1976;193(4256):848–56.

Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26:2190–1.

Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R, et al. Quality control and conduct of genome-wide association meta-analyses. Nat Protoc. 2014;9:1192–212.

Wise PM, Challagundla KB, Fabbri M. Epigenetics and microRNAs in cancer. In: Rezaei N, editor. Cancer immunology: a translational medical context. New York: Springer; 2015. p. 285–94.

Wolfe MD, Kulakow P, Rabbi IY, Jannink J-L. Marker-based estimates reveal significant non-additive effects in clonally propagated cassava (*Manihot esculenta*): implications for the prediction of total genetic value and the selection of varieties. G3 (Bethesda). 2016;6:3497–506.

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46:1173–86.

Wray NR. Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. Twin Res Hum Genet. 2005;8:87–94.

WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447:661–78.

Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. Genome Biol. 2010;11:R53.

Xenarios I, Salwínski L, Duan XJ, Higney P, Kim S-M, Eisenberg D. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 2002;30:303–5.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42:565–9.

Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011a;88:76–82.

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011b;43:519–25.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014;46:100–6.

Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet. 2015a;47:1114–20.

Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet. 2015b;47:1114–20.

Yang J, Fritsche LG, Zhou X, Abecasis G, International Age-Related Macular Degeneration Genomics Consortium. A scalable Bayesian method for integrating functional information in genome-wide association studies. Am J Hum Genet. 2017;101:404–16.

Yao H, Zhou Q, Li J, Smith H, Yandeau M, Nikolau BJ, et al. Molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. Proc Natl Acad Sci U S A. 2002;99:6157–62.

Yong W-S, Hsu F-M, Chen P-Y. Profiling genome-wide DNA methylation. Epigenetics Chromatin. 2016;9:26.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38:203–8.

Yu C-P, Lin J-J, Li W-H. Positional distribution of transcription factor binding sites in Arabidopsis thaliana. Sci Rep. 2016;6:25164.

Zeggini E, Ioannidis JPA. Meta-analysis in genome-wide association studies. Pharmacogenomics. 2009;10:191–201.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science. 2007;316:1336–41.

Zhang D, Bai G, Zhu C, Yu J, Carver BF. Genetic diversity, population structure, and linkage disequilibrium in U.S. elite winter wheat. Plant Genome J. 2010a;3:117–27.

Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2010b;42:355–60.

Zhang Y, McCord RP, Ho Y-J, Lajoie BR, Hildebrand DG, Simon AC, et al. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell. 2012;148:908–21.

Zhu C, Gore M, Buckler ES, Yu J. Status and prospects of association mapping in plants. Plant Genome J. 2008;1:5–20.

Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci. 2012;109:1193–8.

Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014;111:E455–64.

Genomic Selection

Elisabeth Jonas, Freddy Fikse, Lars Rönnegård, and Elena Flavia Mouresan

Abstract Prediction of phenotypes is not only used for selection and breeding in animal and plant populations but also for the assessment of specific phenotypes, especially predisposition to diseases and disorders in human populations. The use of genetic markers has been shown to be useful for prediction and selection for phenotypic traits. The concept of using genetic markers for prediction of breeding values or phenotypes was suggested many decades ago, but applications of marker-assisted selection were limited due to the low number of markers that could be genotyped and the low number of confirmed quantitative trait loci (QTL) that could be selected upon. Genomic selection, in contrast, utilizes dense genetic markers across the whole genome for the prediction of phenotypes as all QTL can be assumed to be in linkage disequilibrium with at least one marker. Genomic selection allows thereby choosing the genetically best individuals without the need to confirm QTL. The concept of genomic selection, proposed in 2001, has since been further developed and applied. Nowadays, genomic selection is widely applied in breeding populations of plants and animals for the selection of future breeding individuals. The chapter introduces the general concept of genomic selection. It further discusses relevant prerequisites for the application of genomic selection, including genotyping platforms and reference populations. Some of the methods applied today as well as suggested advancements of methods are introduced. The final part of the chapter describes briefly applications in animal, plant, and human populations (status when writing this chapter), before concluding with some general notes on genomic selection.

E. Jonas (✉) · F. Fikse · E. F. Mouresan
Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden
e-mail: elisabeth.jonas@slu.se

L. Rönnegård
Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

Statistics Unit, School of Technology and Business Studies, Dalarna University, Falun, Sweden

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_11,

Keywords Breeding · Genotyping · Parametric methods · Prediction · Reference population · Selection candidates · Semiparametric methods

1 Introduction: Genomic Selection in a Nutshell

The possibility to change the distribution of a trait in animal and plant populations by means of selection has been developed tremendously over the last 100 years. We have gained insight into principles of population genetics and have been able to formulate these principles in terms of statistical models. Quantitative genetic theory is based on the principles of Mendelian inheritance and explains how selection of individuals affects the development of a population in future generations and thereby connects genetics on an individual level to population changes.

Models for genomic selection can be interpreted using quantitative genetic theory, connecting changes on a population level to the set of genotypes observed on an individual level. The idea is that most genes have some effect on a trait and that the sum of all gene effects for an individual can be predicted as genomic breeding values (GEBVs) using markers in linkage disequilibrium with the causative genes. In practice, this is done by first estimating the combined genetic effects for each individual of a reference population and subsequently using this information to predict GEBVs for the selection candidates. This requires that extensive genotype information is available both for the reference population and the selection candidates, which has only been possible for the past two decades.

Modern genotyping technologies enable genotyping of many individuals for a larger number of genome-wide markers at affordable cost. These advances in genotyping technologies have been exploited in genomic selection to compute GEBVs (Fig. 1). Hence, a reference population needs to consist of individuals with both genotype and phenotype information. Marker effect estimates from the reference population are combined with genotypes of selection candidates to predict the genetic potential of the selection candidates. Animals with the most desirable genetic potential are kept for breeding to become parents of the next generation of individuals.

The progress in the field of medium- and high-throughput genotyping platforms along with decreased costs for marker detection via sequencing technologies enhanced the use of genomic information in breeding (Davey et al. 2011). In 2001, Meuwissen et al. (2001) published their landmark paper on the use of genome-wide selection or genomic selection, proposing a marker-based selection methodology that incorporates marker information of many (dense) markers in the prediction model. Only a decade later, this method was already employed and implemented in (dairy cattle) breeding programs, and estimated breeding values (EBVs) based on genomic data (GEBVs) were officially published in a number of countries (Patry 2011). The use of genomic selection has also been a major interest for the breeding of crop species (Heffner et al. 2009; Cabrera-Bosquet et al. 2012). A number of methods for the estimation of effects have been suggested (Meuwissen et al. 2001; Habier et al. 2007; de los Campos et al. 2009; Zhong et al. 2009; Kizilkaya et al. 2010), and further developments are on the way, some of which will be detailed later.

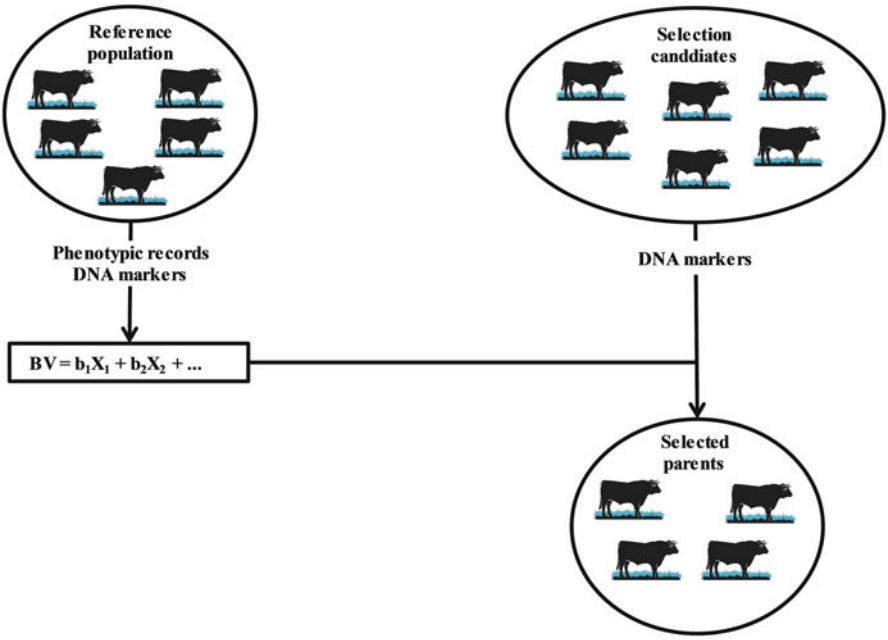

Fig. 1 Schematic overview of the concept of genomic selection including the reference population with both genotype and phenotype individuals and the selection candidates with genotype information. Information on breeding values (BV) will be used to select parents from the population of selection candidates

1.1 How Genomic Selection Really Works

Meuwissen et al. (2001) argue that linkage disequilibrium between markers and quantitative trait nucleotides (QTNs) is the driving force behind genomic prediction. Observing nearby genetic markers supplies information about the QTN, if there is a close association between a marker and the QTN. Given the linkage disequilibrium between QTN and markers as driving force, many expectations and speculations about the behavior of genomic selection have been put forward (e.g., about benefits of whole-genome sequence, across-breed genomic prediction), but more often than not, those expectations and speculations have not been consistent with real data.

Understanding of the mechanisms of genomic prediction became clearer when Habier et al. (2007) showed that accuracies of genomic breeding values were substantially larger than zero even if markers and QTNs were in linkage equilibrium. Genetic markers can capture family relationships and thereby contribute to the accuracy of estimating genomic

breeding values. Habier et al. (2013) described this and investigated the contribution of three information sources to the accuracy of genomic breeding value estimation: markers capturing additive genetic relationships, co-segregation, and linkage disequilibrium.

When the training population is small, much of the accuracy of genomic breeding values is due to markers describing family relationships. A consequence is that the predictive ability rapidly decays over generations. Only if training populations are large, marker estimates reflect more of the effect of actual QTNs nearby instead of additive genetic relationships, and predictions are persistent for more generations.

2 Background

2.1 History of Human-Introduced Genetic Changes to Populations

Selective breeding in both plant and animal species started many thousands of years ago. Genetic change over time was achieved by selecting the best-fit individuals as parents, such that the next generation of individuals was, on average, superior to the parent generation. Selection based on phenotypic criteria has been performed since the domestication of species (Rosenberg and Nordborg 2002; Morrell et al. 2012). While early selection was based on the observation of phenotypes within a group of individuals, more sophisticated tools are used for selection in plant and livestock populations as they are used in farming today. At the beginning of the eighteenth century, Robert Bakewell (1725–1795) established modern breeding by introducing systematic and structured selective breeding (Sweeney and McCouch 2007).

The demonstration of inheritance and the discovery of basic rules of inheritance by Gregor Johann Mendel in the middle of the nineteenth century established the beginning of modern genetics. The complexity of phenotypes and their inheritance could then be explained by their genotypes via rules of allele sharing across generations. Thereafter this had a major impact on animal and plant breeding.

Many of the basic statistical tools used in quantitative genetics were developed in the late nineteenth and early twentieth century by Francis Galton and Karl Pearson. In 1918, Ronald Aylmer Fisher used statistical models to demonstrate the resemblance between relatives and introduced the analysis of variance (Walsh 2001). Further milestones of breeding were laid by Jay Laurence Lush from the 1930s and Charles Roy Henderson from the 1970s and their suggestion of the use of statistical models (Lush 1933, 1947; Henderson 1975a, b). During the twentieth century, many statisticians, quantitative geneticists, and breeders contributed to the development and implementation of different breeding schemes in livestock and plants based on knowledge of trait inheritance and statistical approaches.

2.2 Basic Quantitative Genetics Relevant for Breeding

The basic concept of quantitative genetics, applied in breeding, is that an individual's phenotype P is determined by its genotypic value G and its environment E (Walsh 2001):

$$P = G + E$$

The genotypic value can be decomposed into additive (A), dominance (D), and epistatic (I) values in which A accounts for the average effects, D for the interaction between alleles at one locus, and I for the interaction between alleles at different loci:

$$G = A + D + I$$

A relevant measure applied in quantitative genetics is the narrow-sense heritability, h^2, which is the proportion of the total phenotypic variance P due to the additive genetic effects A:

$$h^2 = \mathrm{Var}(A)/\mathrm{Var}(P)$$

The heritability can be used to describe the phenotypic similarity between relatives or trait variation due to additive genetic effects.

The heritability is also used for the prediction of the response to selection, in the so-called breeder's equation. This equation describes the change of the population mean over one generation or the response to selection ΔZ, with an applied selection differential S (Falconer and Mackay 1996; Lynch and Walsh 1998; Xu and Hu 2010):

$$\Delta Z = h^2 S$$

When the heritability of a trait is close to zero, the response will be very little even if there is strong selection on that trait.

Quantitative genetic analyses initially focused on decomposition of phenotypic variance into underlying components (like A, E) for quantitative traits (i.e., traits involving many genes and influenced by environment). More recently, the possibility to genotype individuals for DNA markers allowed the attention to shift to identification of chromosomal regions with (large) effects on quantitative traits: quantitative trait loci (QTL). Single nucleotide polymorphisms (SNPs) that are causative of the QTL are hereinafter referred to as quantitative trait nucleotides (QTNs).

2.3 Examples of Breeding Programs and Selection Decisions

Breeding programs aim to change certain traits toward a breeding goal in a population. The duration until the genetically improved individuals are available for breeding is expressed as the generation interval. The generation interval is relevant for the genetic

and economic gains in a breeding program: a shorter interval means that improvement can be achieved earlier. Quantitative genetic theory has underpinned the design of selection schemes of plant and livestock populations for many decades. The molecular genetic background of traits has been integrated as a selection tool more recently. The application of these tools in breeding programs has aided selection of the individuals with the best genetic merit for the traits of interest. The true genetic merit, i.e., the true breeding value, of an individual is mostly unknown, and estimated breeding values can be used to predict how well offspring will perform.

Breeding programs aim to identify the best individuals for breeding to produce the next and improved generation. Differences across species regarding the reproduction capacity and the breeding goal traits influence the design of a breeding program. Crossing of lines can be used to create lines with a new combination of characteristics. Crossbreeding often exploits hybrid vigor, or heterosis. Heterosis effects are difficult to predict and are mainly realized in the first generation of crossbreeding. Overcoming cross-incompatibility is relevant in some species in order to allow new trait combinations. The final goal of crossing is to produce a generation with superior traits from each of the parental lines. A conflict exists between the need for diversity within the core breeding population and at least some degrees of uniformity within the production. Nucleus populations or diversity panels can be used to ensure the existence of diverse lines. These nucleus populations are, in many species, kept centrally by few breeding organizations, which define the breeding goals and design the breeding schemes (Figs. 2 and 3).

The breeding goal describes which traits are important for genetic improvement and their relative importance. But not all traits are measured on the selection candidates; some traits are only collected from relatives, and this information can be used for prediction of breeding values for selection candidates based on basic principles of population genetics. The more accurate a trait can be measured, the more accurate selection is. Especially when traits are influenced by other factors, such as the environment, the accuracy of selection might be negatively affected. Evaluation of performance in a controlled environment is one option to reduce the impact of environmental variation on accuracy of selection, and an alternative option is the evaluation of traits under various environmental conditions.

Plant lines with improved traits are still mainly selected based on their phenotypic appearance. However, many generations are needed to produce cultivars with the desired characteristics through conventional breeding, as well as (multilocation) testing (Sharma et al. 2002). There are several constraints in plant breeding including varying outdoor conditions to be accounted for. Also the modes of reproduction influence the possibilities of plant breeding as crossbreeding is restricted in some species.

Improvement schemes for livestock populations are often organized around a nucleus herd. Own performance and performance of relatives (offspring, sibs, parents) are evaluated for the identification of individuals with the best genetic merit. The pedigree plays a major role for developing breeding schemes and mating decisions in animal

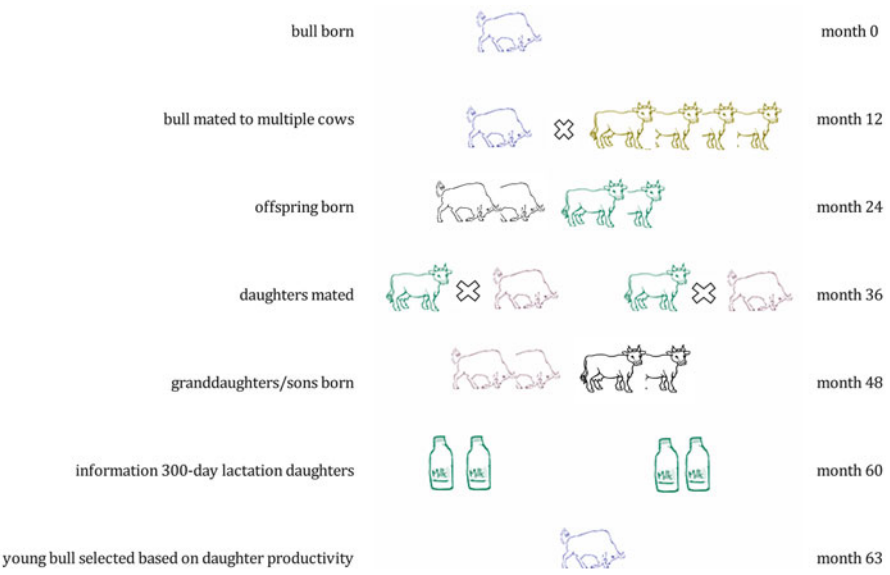

Fig. 2 Example of a traditional breeding scheme in dairy cattle (*Bos taurus*) with the time frame on the right side and the different stages of the breeding cycle on the left. Young bulls for selection are born in month 0 and in month 12 mated to cows. Daughters are born in month 24 and mated in month 36. The granddaughters of the bulls for selection are born, and data on relevant traits (e.g., milk yield, fertility, disease resistance) of the bull's daughters is collected. Information on these traits for one lactation is then available in month 60, and bulls can be selected for breeding based on their daughter's first lactation performance

breeding. The genetics of sires can be distributed widely via artificial insemination. The evaluation of the genetic merit targets therefore mainly the sires in livestock populations, as they have a great genetic impact on their population (Gerrits et al. 2005; Funk 2006).

2.4 Selective Breeding Using Molecular Markers

Selection of breeding stock and lines based on phenotype and pedigree data allowed to improve many breeding populations. But traits are based on changes at the level of the DNA. It has, therefore, been suggested that using genetic information based on the inherited part of the individual, its DNA, will allow a better prediction of the genotypic value and, therefore, the phenotypic value of an individual or its real breeding value. As first suggestions for the use of (molecular) marker information in breeding programs (Dekkers and Hospital 2002), marker-assisted selection (MAS), marker-assisted recurrent selection (MARS), and marker-assisted breeding (MAB) were discussed. For a successful implementation of DNA information in selection scheme information, such

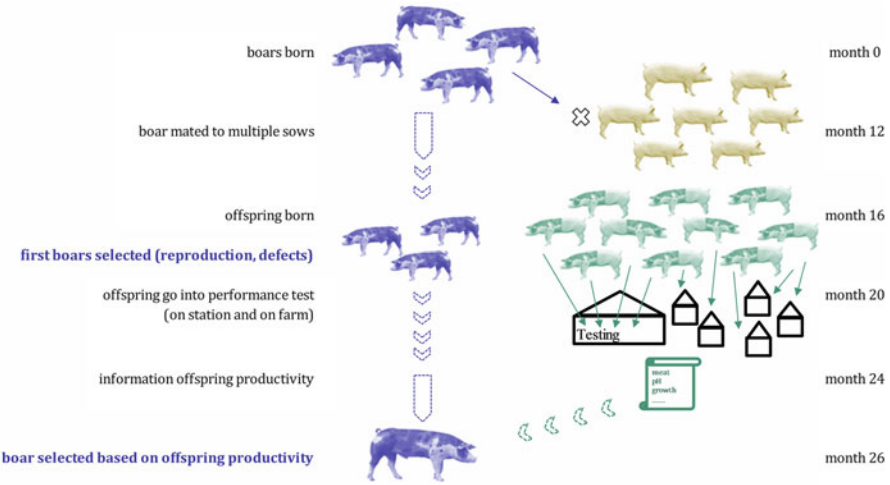

Fig. 3 Example of a traditional breeding scheme in pigs (*Sus scrofa*) with the time frame on the right side and the different stages of the breeding cycle on the left. Young boars from paternal lines for selection are born in month 0 and at month 12 mated to sows from maternal lines. Crossbred offspring are born in month 16 and tested in performance test stations or on farms. The offspring of the boars for selection are slaughtered, and the information from fattening period and slaughterhouse are collected. Information from reproduction and production is available at approximately month 26, and boars can be selected for breeding based on breeding values predicted using the offspring performance

as identification of markers, successful genotyping and validation of genotype and allele frequency in a large amount of individuals are required. Additionally, validated association of the genetic marker with the trait of interest and an assessment of the marker effect in a breeding program are needed. Furthermore, any potential negative effects on other economically important traits need to be excluded.

Selection using genetic markers has been suggested as a preferred method for traits for which phenotypic selection is more difficult, such as traits with low heritability. Other examples are traits for which the assessment of the phenotypes is difficult and cost-extensive or can only be done late in life. It has also been shown that MAS is particularly effective for traits with (one or few) major QTL effects (Gupta et al. 2010; Cabrera-Bosquet et al. 2012). Not only is the increased accuracy of selection due to the use of markers in breeding programs the main advantage of MAS but also the minimization of phenotyping (Bernardo and Yu 2007). In plant breeding, MAS can also be applied in year-round breeding nurseries or greenhouses where phenotypic data are less meaningful as their correlation to field data is low. Marker information can, in such settings, allow a prediction of the phenotypes (Lorenzana and Bernardo 2009). But costs for the identification of useful genetic markers and the implementation of such markers are relatively high compared to the gain achieved. Some markers

are also population- or family-specific. It has been, for this and other reasons, stated that MAS is not well suited for the improvement of crops (Jannink et al. 2010). The identification of useful markers is time-consuming, and as many traits are influenced by multiple genes, also a higher number of markers would be needed for MAS for each single trait (Gupta et al. 2010). Only relatively few causative mutations have been identified (in livestock) and implemented as a routine component in breeding programs.

3 Genomic Selection Designs and Strategies

3.1 Designing Genotyping Platforms

One of the prerequisites of the application of genomic selection is the availability of information from genetic markers across the genome. Many genetic markers were identified using methods such as Sanger sequencing and their information, for example, collected in the NCBI database (https://www.ncbi.nlm.nih.gov/). When next-generation sequencing (NGS) was accessible to a larger number of researchers, the amount of information increased significantly as whole-genome sequences were available. This allowed the discovery of many potential markers, such as SNPs. The time required for the sequencing of the full genome of an individual and the costs for it have been reduced significantly during the last decades (Goodwin et al. 2016). The genome of many species has been sequenced, and genotyping arrays have been developed based on the sequence information (Table 1), but the progress on the assembly of a good reference genome, the availability of full genome sequence information from multiple lines and varieties, and the development of SNP arrays capturing the complex and repetitive genome of plants is slow (Somers et al. 2003; Ganal and Roeder 2007; Trebbi et al. 2011). While genome-wide high-throughput genotyping platforms are available for many livestock species, the progress in the development of such platforms for plants has been slower.

The gene density differs widely between livestock and plant species. However, it is mainly the extent of linkage disequilibrium that plays a major role for the application of molecular genetic tools in breeding and MAS (Chao et al. 2010). Linkage disequilibrium is strongly related to the population history, especially resulting from evolutionary history, mating system, population size, admixture, recombination rate, and selection (Heffner et al. 2009). The decay of linkage disequilibrium varies not only with increasing physical distance of loci between species but also between populations of the same species and across chromosomes (Remington et al. 2001; Maccaferri et al. 2005; Chao et al. 2007; Mather et al. 2007; Tenaillon et al. 2008). Structures of linkage disequilibrium depend also on the breeding scheme; plant breeders, for example, often use full-sib families created from crosses of inbred parents, and linkage disequilibrium will be extensive within each family (Zhong et al. 2009). The marker density required for genomic selection will therefore depend on the population and breeding structure.

Table 1 Genome structure and available information on the genome of selected livestock and plant species including examples of options for genotyping (table restricted from updates in the communities)

Species	Ploidy	Genome structure	Genotyping arrays	Genebuild started/ released[a]	Draft/reference sequence
Barley (*Hordeum vulgare*)	Diploid ($n = 7$)	~5.7 Gb (1.36 GB assembled), complex, repeat-rich, ~1 SNP per 50–189 bases; 79, 379 genes called, 24,287 coding genes; 268 pseudogenes	9K (7,842 SNPs) iSelect SNP[b] chip; planned iSelect 50K SNP[b] chip; DArT markers	Started in 2014	(International Barley Genome Sequencing Consortium et al. 2012; Mascher et al. 2017)
			GoldenGate/BeadArray: OPA1, 1,536 or OPA2, 3,072 SNPs[b]		
Wheat (*Triticum aestivum*)	Hexaploid $n = 7$ (A, B, D)	~13.43 Gb; highly repetitive and duplicated; ~1 SNP per 335 bp; 103, 539 coding genes	9K SNP[b] array, 90K array (in development by consortium), wheat HD (817,000 SNPs)[c] genotyping array; genotyping by sequencing	Started in 2015	(Brenchley et al. 2012)
			Custom chips: GoldenGate[b], iSelect, TaqMan®, KASPar[d]		
Wheat (*Triticum durum*)	Tetraploid ($n = 7$)	Highly repetitive (~85–90%); duplicated	GoldenGate 275 SNPs[b]	No	
Rye (*Secale cereale*)	Tetraploid ($n = 7$)	~7.92 Gb; highly repetitive (~92% repetitive); ~36,000 genes	DArT markers (around 5,000); Rye5K SNP array[b]	No	(Bauer et al. 2017)
Potato (*Solanum tuberosum*)	Tetraploid ($n = 12$)	~844 Mb; 727 Mb assembled; highly heterozygous, inbreeding depression; ~39,021 coding genes	SolSTW array-20K SNP array[b]; potato 12K SolCAP array (in development); Infinium 8303 Potato Array[b]	Started in 2011	(Potato Genome Sequencing Consortium 2011)
Rice (*Oryza sativa L.* ssp. *indica*)	Diploid ($n = 12$)	466 Mb; 412 Mb assembled; 46,022 to 55,615 genes; 40,745 coding genes	GoldenGate, 1,536 or 768 SNPs[b]; BeadXpress, 384 SNPs[b]; GeneChip Rice 44K (44,100 SNPs)[c] array; KASP™ arrays[e]	2010	(Yu et al. 2002)
Rice (*Oryza sativa* ssp. *japonica*)	Diploid ($n = 12$)	~420 Mb; ~1 SNP per 10 kb; 32,000 to 50,000 genes; 35,679 coding genes	GoldenGate[b], 1,536 or 768 SNPs; BeadXpress, 384 SNPs[b]; GeneChip Rice 44K (44,100 SNPs)[c] array; KASP™ arrays[e]	Yes	(Goff et al. 2002)
Maize (*Zea mays*)	Diploid ($n = 10$)	~2.3 Gb; recent genome extension; historical genome duplication; ~1 SNP per 28–214 bp; 37,000 to 63,000 genes; 39,324 coding genes	MaizeSNP50 (56,110 SNPs) BeadChip[b], KASP™ arrays[f]; Maize genotyping array (616,201 variants including 609,442 SNPs)[c]	2015	(Schnable et al. 2009)

Cattle (*Bos Taurus*)	Diploid ($n = 29 + XY$)	~2.65 Gb; 3,000,000 SNPs identified; ~19,994 protein-coding genes; 797 pseudogenes	BovineSNP50 (54,609) SNP[b] BeadChip, BovineHD (777,962 SNPs) BeadChip[b]; BovineLD array[b]; 640,000 SNP[c] array; GGP Bovine LD (v3) (26,000 SNPs) genotyping array[b]	2011	(Elsik et al. 2009)
Pig (*Sus scrofa*)	Diploid ($n = 18 + XY$)	~3.02 Gb; 510,000 SNPs identified; ~21,630 protein-coding genes; 568 pseudogenes	PorcineSNP60 (64,232 SNPs) BeadChip[b]; Porcine genotyping array (658,692 SNPs)[c]	2012	(Groenen et al. 2012)
Chicken (*Gallus gallus*)	Diploid ($n = 33 + ZY$)	~1.29 Gb; 16,362 protein-coding genes; 50 pseudogenes; 1,800,000 SNPs identified	Chicken genotyping[c] array (580,000 SNPs)	2016	(Rubin et al. 2010)
Atlantic salmon (*Salmo salar*)	Diploid ($n = 29$)	~6,000 Mb; many chromosomal rearrangements; 33,709 genes	iSelect Atlantic Salmon 16,500 SNPs[b]; Salmon genotyping array (130,000 SNPs)[c]	No	(Lien et al. 2016)
Sheep (*Ovis aries*)	Diploid ($n = 26 + XY$)	~2.53 Gb; 20,921 protein-coding genes; 290 pseudogenes	OvineSNP50 (64,232 SNPs) BeadChip[b]; OvineHD (700,000 SNPs) chip[g]	2013	(Jiang et al. 2014)
Horse (*Equus caballus*)	Diploid ($n = 31 + XY$)	~2.43Gb; 20,436 protein-coding genes; 4,400 pseudogenes	EquineSNP50 (54,602 SNPs) BeadChip[b]; Axiom® Equine (670,769 SNPs) Genotyping Array[c]	2008	(Wade et al. 2009)
Dog (*Canis lupus familiaris*)	Diploid ($n = 38 + XY$)	~2.39 Mb; 19,856 protein-coding genes; 950 pseudogenes; >2.5 million SNPs; less lineage-specific repeat sequence compared to human; specific haplotype structure due to two bottlenecks	CanineHD (172,115 SNPs), Canine Array Sets A (460,000 SNPs) and B (670,000 SNPs) Axiom genotyping array[c]	2012	(Lindblad-Toh et al. 2005)
Loblolly pine (*Pinus taeda L.*)	Diploid ($n = 12$)	23.2 Gb; 50,172 genes annotated; average intron lengths >2.7 Kb up to 100 Kb		Yes	(Neale et al. 2014; Zimin et al. 2014)
Eucalyptus (*Eucalyptus grandis*)	Diploid ($n = 11$)	Estimated genome size, 640 Mb; approximately 36,376 protein-coding genes (of which 34% in tandem duplications); around 44.5% retrotransposons	DArT markers; EuCHIP60K (60,904 SNPs)[b]	Yes	(Myburg et al. 2014)
Eucalyptus (*Eucalyptus globulus*)	Diploid ($n = 11$)	Estimated genome size, 530 Mb	DArT markers; EuCHIP60K (60,904 SNPs)[b]	Yes	(Myburg et al. 2014)

(continued)

Table 1 (continued)

Species	Ploidy	Genome structure	Genotyping arrays	Genebuild started/ released[a]	Draft/reference sequence
Poplar (*Populus trichocarpa*)	Diploid ($n = 19$)	Estimated genome size, 423 Mb; approximately 36,393 genes		Drafts	(Tuskan et al. 2006)
Domesticated apple (*Malus*×*domestica* Borkh.)	Diploid ($n = 17$)	Estimated genome size, 742.3 Mb; unassembled part of the genome is 98% repetitive (138.4 Mb); diverse ploidy levels within genus and species (two to five times)	IRSC apple 8K (7,867 SNPs) SNP array v1[b]; Axiom_Apple480 (487,249 SNPs)[c]	2015	(Velasco et al. 2010)
European pear (*Pyrus*)	Diploid ($n = 17$)	Estimated genome size, 527–577 Mb depending on species; approximately 43,000 genes; highly heterozygous due to self-incompatibility	Pear Infinium® II 9K SNP array[b]	Draft 2013 for *Pyrus communis*	(Wu et al. 2013; Chagné et al. 2014)
Rubber tree (*Hevea brasiliensis*)	Diploid ($n = 18$)	~1.47 Gb (estimated 2.15 Gb haploid genome); ~43,792 predicted protein-coding genes; repeat-rich (around 71–78% of the genome length)	GBS	Drafts	(Tang et al. 2016)
Norway spruce (*Picea abies*)	Diploid ($n = 12$)	~9.6 Gb; ~29,000 predicted genes	311 iSelect SNP[b] chip for validation	No	(Nystedt et al. 2013)
Black spruce (*Picea mariana*)	Diploid ($n = 12$)		4267 iSelect SNP[b] chip	No	
White spruce (*Picea glauca*)	Diploid ($n = 12 + 1B$)	~20.8 Gb	7338 and 9559 iSelect SNP[b] chip	No	(Birol et al. 2013)

bp base pairs, K, in 1000 base pairs, *MB* megabases, *SNP* single nucleotide polymorphism, KASPar, DArT markers

[a]Date of gene build release on Ensembl (http://www.ensembl.org/index.html), Gramene (http://www.gramene.org/), GDR (https://www.rosaceae.org/), PineRefSeq (http://pinegenome.org/pinerefseq/), or Dendrome (https://dendrome.ucdavis.edu/)

Companies offering the arrays: [b]Illumina (some specific arrays for GeneSeek); [c]Affymetrix; [d]KBioscience; [e]developed for the Generation Challenge Programme and the Integrated Breeding Platform; [f]developed for CIMMYT; [g]International Sheep Genomics Consortium (ISGC)

Table 2 Information on genomic prediction and selection applied in different organisms

Species	Applications	Strategies	Advantages using genomic selection	Challenges
Cattle	Application in purebred Holstein Friesian dairy cattle populations; genomic breeding values used for selection; applications in smaller dairy breeds; test and application in beef cattle	Reduced generation interval and earlier selection of breeding bulls; reduced size of breeding programs with fewer bulls for testing; international collaborations	Reduction of generation interval; reduced costs for keeping bulls (fewer bulls)	Selection for low heritable traits; predictions in crossbred populations such as beef cattle; size of reference population (small cattle breeds); larger effective population in some populations (weaker linkage disequilibrium)
Pig	Studies using simulated and field data; collaborations between public and private sector; some lines on the market	Tests of application in purebred and crossbred populations; most studies within the frame of existing breeding programs	Additional genetic gain expected depending on trait and population	Costs for genotyping relatively high; already short generation interval; applications and accuracies in crossbred populations
Chicken	Advances in private sector in collaboration with public sector in layers and broilers, some lines on the market	Use of a reduced reference population and shorter generation interval	Increased genetic gain depending on trait; advancement especially if no information from relatives is available; improved male selection in layers	Costs for genotyping; refinement of breeding programs; adjustments of breeding goals; restore some variability in the highly specialized commercial lines
Atlantic salmon	Studies using simulated and real data close to applied breeding programs; applications in the private sector	Preselection of candidates for genotyping and within-family selection	Increased genetic gain in within-family selection, generally higher genetic gain with genomic selection	High costs in conventional breeding programs due to large number of selection candidates
Sheep	Some suggestions and applications in larger breeding populations	Test in different breeding populations and resource flocks	Some genetic gain estimated depending on trait (e.g., meat and wool traits); choice of prediction models relevant; possible reduction of generation interval	Value of each animal lower; organization levels of breeding often low; size of the reference population relevant
Horse	Few studies published Some application	International collaborations; selection mainly in sport horses, aim to reduce generation interval	Predicted improvement due to shorter generation interval; improved selection of female horses and young horses; advantages with exchanges genetics	Restricted breeding goals; international collaboration required; little literature

(continued)

Table 2 (continued)

Species	Applications	Strategies	Advantages using genomic selection	Challenges
Dog	Studies done using data from dog breeding originations and simulated data	Test especially for disease traits in pedigree dogs	Accurate selection especially at young age before onset of diseases; genetic gain	Population management; effects of diversity loss; large enough reference populations needed; definition of phenotypes
Barley	Trials and application in the private sector, many suggestions from the public sector	Cross-validation using two generations; inclusion of generations in the greenhouse	Shorter generation interval; selection of strategy will determine if long-term or short-term genetic gain	Right choice of reference population; capture of nonadditive effects such as genotype by environment interaction
Wheat	Trials and application in private sector, many suggestions from public sector using real and simulated data	Test of different strategies including selection schemes different from traditional schemes	High accuracies of prediction but limited to the same environment; improved prediction compared to phenotypic or marker-assisted selection	Selection for disease traits; high costs for genotyping and phenotyping due to large number of breeding lines; inclusion of nonadditive effects; predictions in biparental families
Rice	Suggestions for specific traits, especially drought tolerance; genotyping platform available	Based on marker-assisted approaches, integration of markers for specific traits	Improved yield under drought (markers identified), yield under biotic and abiotic stress	Recording of good phenotypes, understanding the trait architecture; development of new cultivars
Maize	Trials in private sector, many suggestions from public sector using real and simulated data	Different strategies tested to define best reference population, mostly based on hybrid or double haploid lines, aim to reduce phenotyping	Improved selection of diverse lines for crossing; genetic gain depending on traits and models; reduction of generation interval; use of generations in the greenhouse	Improvements needed for multiple family approaches; inclusion of environmental effect especially for disease traits as well as genotype by environment interactions
Perennial ryegrass	Tested in simulations with varying accuracies depending on trait	Changes to current breeding programs to implement genomic selection	Shorter breeding cycle, thereby higher genetic gain	Collection of reliable phenotypes over multiple years to create a good reference population; issues like increased inbreeding and decreased accuracies to be solved

Pine	Use of simulated data and data from structured breeding populations including clonally replicated field trials	Tests of inclusion of different models	Increased accuracy in progeny for some traits, accuracies match phenotypic selection but more efficient due to time decrease	Genotyping costs; complex genome; need of better control of dominance effects; structure of linkage disequilibrium; lacking gene maps
Eucalyptus	Some suggestions made using diverse data (seedling and orchard designs, breeding populations, simulations)	Test especially for diversity and increase genetic gain; test of relationship reference and selection populations; size of reference population and heritability	Accuracies match phenotypic selection; opportunities to decrease generation interval	Genotyping costs; complex genome; accuracy of generation difference between selection candidates and reference population
Spruce	Tested using data from different spruce populations; relevance of genomic selection in recently domesticated/undomesticated species	Single-step methods for breeding programs with simple mating structures and shallow pedigrees	Potential larger gains per unit time compared to traditional breeding; selection of superior individuals within large full-sib families as a short-term goal	Marker densities still low; establishment of reference populations with deep pedigrees; need of solution for open-pollinated mating schemes

Shown is an overview of the applications, strategies, advantages, and challenges in different livestock and crop species and two examples of tree species based on the literature used in the chapter

3.2 Designing Reference Populations

A reference population, also called as discovery or training set, describes the breeding stock for which information on the relevant traits, including data in multiple environments if relevant, is available. The individuals of the reference population are genotyped, and information on their pedigree or relationship is available. Two approaches are generally offered for the reference population: (1) an established reference population before the start of the next breeding cycle (e.g., in multiple-stage selection) and (2) the prediction using a reference set from the same generation as the selection candidates (e.g., in one-stage selection) (Marulanda et al. 2016). The decision on the structure of the reference population and the relationship to the population of selection candidates is relevant in a genomic selection breeding scheme. Deterministic functions to predict the accuracy of genomic breeding values include the size of reference population (Daetwyler et al. 2008): a larger reference population leads to a higher accuracy. When genomic selection was firstly applied and tested in dairy cattle, only bulls were utilized in the reference population. For these bulls, information on the tested progeny were available (as reviewed by VanRaden 2008; Hayes et al. 2009a). As it was assumed that larger training populations result in more reliable predictions, initiatives to pool reference populations across countries emerged, such as the European initiative EuroGenomics (Lund et al. 2011) or a collaboration between the USA and Canada (VanRaden et al. 2009a, b; Muir et al. 2010). It had been pointed out in a review that such international collaborations are desirable (Dürr and Philipsson 2012), because they result in reference populations of tens of thousands progeny-tested dairy bulls. Similar approaches have also been taken in the wheat breeding community to develop universal training populations by merging large phenotype dataset (e.g., by the Wheat Initiative's Expert Working Group on Wheat Breeding Methods and Strategies) (Bassi et al. 2016). Such international connections of data are still less advanced in beef cattle (Berry et al. 2016) and other livestock. The size of the reference population is often restricted by the costs. The increase of the reference population might lead to a shift away from the collection of phenotypes, but collaborations might allow the elaboration of more phenotypes in a larger reference population. A reduction of testing with fewer locations or replications in exchange of more genotyped and phenotyped lines in the reference population has been suggested in plant breeding to balance limited resources when increasing the reference population (Longin et al. 2015).

Implementation of genomic selection led to changes in dairy cattle breeding programs, with less emphasis on progeny testing and selection of fewer bulls, and genotyping of females has become a necessary complement to maintain and update reference populations. It has also been a concern in other breeding schemes that the introduction of genomic selection will reduce the phenotypic evaluation and might have potential drawbacks in the future.

The size of the reference population depends on the resources, and this will determine the accuracy of genomic prediction. Also the structure of the reference population and the relationship to the selection candidates influence the required size of the reference population. One of the first observations of genomic prediction applied to the real data

in dairy cattle was that the accuracy of genomic breeding values was dependent on whether or not the sire of the selection candidate was in the reference population. This observation was further sustained by studies on the distance between reference population and population of selection candidates (e.g., Habier et al. 2010), reporting that accuracy of genomic breeding values decreased with decreasing additive genetic relationship between bulls in reference population and selection candidates. These observations are also influenced by the structures of the populations, such as the linkage disequilibrium and QTL effects. The accuracy of prediction is lower when reference and selection populations are less related. Adding individuals to the reference population will not always lead to gains, as it will largely depend on the relationship to the selection candidates (Calus 2016), thus the ability to cover the linkage disequilibrium between markers and QTL of the selection candidates. Furthermore, if the genetic diversity or the allele frequencies in the selection candidates change, an update of the training population is needed (Bassi et al. 2016). The degree of relatedness within a reference population was also shown to affect the prediction accuracy in livestock, where low relationships among animals in the reference population result in the highest accuracy of genomic breeding values (Pszczola et al. 2012). Such a strategy probably ensures the widest range of possible genotypes present in the reference population. Especially in dairy cattle, where genomic selection has been widely applied, discussions on the actual optimum size of the reference population are ongoing. An example using cows in the reference population suggested that an initial size of 2000 cows would still require that information from 600 cows have to be added every year to keep the accuracies constant (Pszczola and Calus 2015). In Holstein Friesian dairy cattle, the size of the reference population exceeds today more than 30,000 bulls worldwide.

Differences in the design of the reference population in plant breeding do also depend on the mating system of plants. A study using F6 wheat lines showed that a reference population of 700 lines allowed the highest predictive abilities. The tested lines were derived from three different crossing and selfing schemes each based on 60 parental lines (Cericola et al. 2017). Inbreeding plants have higher levels of linkage disequilibrium compared to the population-wide linkage disequilibrium in outbreeding plants. The size of the reference populations has to be larger in outbreeding plants, unless genomic prediction is performed only within families (Lin et al. 2014). Such difficulties can be aligned to multi-breed populations in livestock breeding. The design of the reference population has to follow the criteria stated above also in multi-breed populations. Discussions on the size of the reference population will therefore seldom be concluded in a single number, but the general statement of "the more the better" will be relevant. Bassi et al. (2016) reported that the size of the reference population varies in plant breeding scenarios and can vary from 60 to 10,000 individuals. They also concluded that the size should be as big as possible but that other criteria such as relatedness and trait heritability have to be taken into account (Bassi et al. 2016).

4 Methods and Models Applied in Genomic Selection

A lot of efforts have been devoted to the development of models for genomic prediction, and this section presents an overview of the methods. The methods first proposed, and still commonly used, assume a linear relationship between the phenotype on the one hand and genotypes on the other hand. More recently, nonparametric approaches have been proposed that are less dependent on assumptions like linearity and multivariate normality, among others.

Comparisons of genomic prediction methods have in most cases only identified small differences in the predictive performance from the empirical data, but the differences are expected to increase with larger reference populations. There are several reasons why there might still be small differences in performance between prediction methods. Firstly, the genetic architecture of majority of the traits considered for genomic prediction points toward a polygenic mode of inheritance (i.e., many QTN with relatively small effects), and only a few traits are influenced by a smaller number of QTN with a large effect. Secondly, the validation horizon is often short; methods relying on markers tracing genetic relationships perform well to predict breeding values in the next generation, and advantages of methods exploiting linkage disequilibrium are small. Predictive ability tends to decrease if there are more generations between the selection candidates and the reference population, more so for methods that rely on markers tracing genetic relationships than for methods exploiting linkage disequilibrium. Thus, larger differences between methods can be observed with a longer validation horizon. In addition, we should expect larger differences in performance between prediction methods as the sizes of the reference populations increase in the future.

4.1 Parametric Methods

Consider the linear regression equation where a phenotype is modeled as the sum of additive marker effects:

$$y_i = \mu + \sum_{j=1}^{p} Z_{ij} u_j + e_i \tag{1}$$

Here y_i is the phenotypic observation for individual i, μ is the population mean (ignoring any systematic fixed effects to keep notation simple), p is the number of markers, Z_{ij} is the genotype coding for individual i for marker j, u_j is the additive marker effects, and e_i is the residual effect. The equation is written in matrix form as

$$\boldsymbol{y} = \mu + \boldsymbol{Zu} + \boldsymbol{e} \tag{2}$$

where $\boldsymbol{y}$ is a vector of observations (length n), Z is a matrix with genotypes, $\boldsymbol{u}$ is a vector of marker effects, and $\boldsymbol{e}$ is a vector of random residuals. Genomic breeding values for selection candidates are estimated as

$$\widehat{a_s} = Z_s \widehat{u}, \tag{3}$$

where Z_s is the matrix with genotypes for the selection candidates and $\widehat{u}$ the estimated marker effects.

Treating marker effects as fixed effects yields the ordinary least squares model considered by Meuwissen et al. (2001), but the predictive performance of this model was poor. The number of markers is usually much higher than the number of observations, and the challenge is to obtain estimates of marker effects that yield good predictive performance of genomic breeding values. This can be achieved by elaborated prior distributions of marker effects (i.e., treating marker effects $\boldsymbol{u}$ as random effects) and/or choice of estimation method. Estimation methods that have been evaluated rely on variable selection, shrinkage, or a combination of both. A short summary of the shrinkage methods is given in the sections below; an extensive treatment of all different approaches is outside the scope of this text, and readers are referred to reviews (e.g., de los Campos et al. 2013a; Gianola 2013; Kärkkäinen and Sillanpää 2013).

Shrinkage methods attempt to balance goodness of fit and predictive value by minimizing an objective function consisting of a measure of lack of fit (e.g., residual sum of squares or log likelihood) and a penalty term that causes estimates to be shrunk toward zero. Several options exist for the penalty term: in ridge regression, the penalty is proportional to the sum of squares of estimates of $\boldsymbol{u}$ (L2 norm), and in LASSO (least absolute shrinkage and selection operator), the penalty is proportional to the sum of absolute values of $\boldsymbol{u}$ (L1 norm). The elastic net algorithm uses a weighted combination of the sums of squares and sums of absolute values of $\boldsymbol{u}$ as penalty.

The choice of penalty term corresponds to assuming a specific distribution for the marker effects $\boldsymbol{u}$. For instance, application of ridge regression is equivalent to best linear unbiased prediction (BLUP) of marker effects when marker effects are assumed to follow a normal distribution with mean zero and a variance that is the same for all markers (VanRaden 2008). Other prior distributions for marker effects that have been considered in the context of genomic prediction are the Student distribution (Bayes A) and the Laplace distribution (Bayesian LASSO).

Variable selection methods exploit the assumption that only a small proportion of explanatory variables affect the outcome. The motivation to employ variable selection methods in genomic prediction is that not all genetic markers will be associated with a QTN. The expected effect of markers not associated with the QTN would then be zero. Meuwissen et al. (2001) proposed Bayes B, a variable selection approach where a large portion (π) of the markers was expected to have a zero effect and the remaining proportion ($1–\pi$) an effect drawn from a Student distribution. In their approach, the parameter π had to be specified a priori, but other solutions have been put forward to estimate this parameter from the data (e.g., Habier et al. 2011). The

number of components in the mixture is not restricted to two, and prior distributions consisting of multiple mixtures have been applied [e.g., Bayes R (Erbe et al. 2012)].

Equation (2) is referred to as the SNP model because it models the SNP effects $\boldsymbol{u}$ (length of $\boldsymbol{u}$ is equal to the number of markers, p). Interestingly, the SNP model can be reparametrized by substituting $\boldsymbol{Zu}$ with a vector of genomic breeding values $\boldsymbol{a}$ (length equal to the number of individuals, n). Hence, Eq. (2) can be written as

$$\boldsymbol{y} = \mu + \boldsymbol{a} + \boldsymbol{e} \tag{4}$$

If marker effects are normally distributed $\left(\boldsymbol{u} \tilde{} N\left(0, I\sigma_{\mathrm{u}}^2\right)\right)$, the distribution of $\boldsymbol{a}$ reduces to

$$\boldsymbol{a} \sim N\left(ZZ'\sigma_{\mathrm{u}}^2\right) = N\left(0, G\sigma_{\mathrm{a}}^2\right), \tag{5}$$

where G can be regarded as the realized genetic relationships between individuals. So, the element on row i and column j in the matrix $G\sigma_{\mathrm{a}}^2$ is the covariance between phenotypes of individual i and j.

This approach is commonly referred to as GBLUP. The advantage of this re-parameterization is that genomic breeding values can be predicted using models and software similar to those used for pedigree-based breeding value estimation (with the pedigree-based relationship matrix replaced by a genomic relationship-based matrix). Furthermore, since the number of individuals in the reference population is typically much smaller than the number of markers, the computational demands are much lower.

The variance components σ_{u}^2 and σ_{a}^2 are the same if $ZZ' = G$. By scaling all columns in Z to have zero mean and variance $1/p$, the variance σ_{a}^2 can be interpreted as the additive genetic variance for the trait, and G is the genomic relationship matrix.

The effectiveness of GBLUP depends on how well the genomic relationship (derived from markers) reflects the actual relationships at QTN. This finding (de los Campos et al. 2013b) motivates studies on other approaches to construct the genomic relationship matrices. These differ, for example, in the definition of the base population (Meuwissen et al. 2011), in the age of the relationships they trace (e.g., Sun et al. 2016), or in the weight that is given to chromosomal segments (e.g.., Shen et al. 2013).

4.2 Semiparametric Methods

In this section, we present GBLUP models where the genomic relationship matrix G can either be smoothed, which will decrease the difference in genetic correlations between individuals, or G can be made more rugged to increase the differences in genetic correlations between individuals. These models can be advantageous because they tend to remove noise in the G matrix and give better genomic predictions when, for instance, marker interaction effects are simulated.

The first models presented are geostatistical kriging and reproducing kernel Hilbert space models. The term kriging is used in the geostatistical literature and is equivalent to empirical BLUP. The aim of kriging in geostatistics is to model the correlation between observations located on a map. The pair-wise correlations depend on the distance between the positions where the observation was recorded. By modeling each position as a random effect, the values at positions without any observation can be predicted. This is similar to genomic prediction, but instead of having relatedness (based on genetic markers), the distances between geographical positions are used.

A common family of correlation functions used for kriging is the family of Matérn covariance functions (named after the Swedish statistician Bertil Matérn). It depends on a couple of tuning parameters and the Euclidean distance between geographical positions. In the application for genomic prediction, the covariance function depends on the Euclidean distance between individuals in terms of their additive relationship.

Ober et al. (2011) showed that the kriging model gives better genomic predictions than the standard GBLUP model for simulated interaction effects. They also discuss the similarity and differences of the kriging model with the reproducing kernel Hilbert space (RKHS) approach of Gianola and van Kaam (2008). Similarly as for geostatistical kriging models, RKHS finds a correlation matrix that smooths the genomic correlation matrix (Morota and Gianola 2014).

Both spatial kriging and RKHS models have been shown to outperform GBLUP in genomic predictions when marker interaction effects are included in simulations. However, they are usually based on an additive specification of the marker data with the coding of marker genotypes being evenly spaced integers (such as 0, 1, and 2). Hence, they are not developed specifically for fitting nonadditive marker effects but are nonetheless much more flexible than the standard GBLUP model resulting in better genomic predictions.

4.3 Models Including Nonadditive Effects

Many animal and plant breeding schemes involve crossing of different breeds or lines or genotypes with the goal of harnessing the beneficial effects of breed complementarity and heterosis. The basis of heterosis are nonadditive effects like dominance or even interactions between loci (Falconer and Mackay 1996). It can be useful to include these effects in the statistical models, if these effects contribute substantially to the traits. The basic idea is to decompose the genotypic value into additive (A), dominance (D), and epistatic (I) values.

4.3.1 Models Including Dominance Effects

The SNP model that fits simultaneously additive and dominance effects of SNPs can be written as

$$\boldsymbol{y} = \mu + \boldsymbol{Zu} + \boldsymbol{Xd} + \boldsymbol{e}, \tag{6}$$

where a vector of dominant SNP effects $\boldsymbol{d}$ is included for each of the p SNP markers and an element in the matrix X, x_{ij}, is the indicator variable for the heterozygous genotype of the jth SNP for individual i (Toro and Varona 2010).

In the standard SNP-BLUP, both additive and dominant effects are assumed to have normal distributions:

$$\boldsymbol{u} \sim N\left(0, I\sigma_{\mathrm{u}}^2\right), \quad \boldsymbol{d} \sim N\left(0, I\sigma_{\mathrm{d}}^2\right),$$

The equivalent GBLUP model is

$$\boldsymbol{y} = \mu + \boldsymbol{g} + \boldsymbol{e} \tag{7}$$

Here $\boldsymbol{g}$ is a vector of genomic breeding values (of length n) with
$V(\boldsymbol{g}) = G\sigma_{\mathrm{u}}^2 + D\sigma_{\mathrm{d}}^2$,
where G is the additive and D the dominance genomic relationship matrix.

4.3.2 Models Including Epistatic Effects

The SNP-BLUP model can be extended to include interaction effects between alleles at different loci:

$$\boldsymbol{y} = \mu + \boldsymbol{Zu} + \boldsymbol{Wv} + \boldsymbol{e} \tag{8}$$

where v is the marker interaction effect, a normally distributed random effect, and the matrix W is constructed so that

$$W_j = Z_i \odot Z$$

with subscript giving column index with $j = (i - 1)p + i$ where p is the number of columns in Z and $\odot$ is the direct Hadamard product. Thus, W has n rows and $p \times p$ columns.

The equivalent GBLUP model is

$$\boldsymbol{y} = \mu + \boldsymbol{g} + \boldsymbol{e} \tag{9}$$

with $V(\boldsymbol{g}) = G\sigma_{\mathrm{u}}^2 + H\sigma_{\mathrm{v}}^2$ and $H = G \odot G$ is the epistatic relationship matrix.

However, the extensions of GBLUP in Eqs. (7) and (9) are expected to increase prediction accuracies only if training populations are large, so that marker estimates reflect more the effect of actual QTNs nearby, instead of additive genetic relationships.

5 Accuracies of Genomic Estimated Breeding Values

Accuracies of genomic estimated breeding values are used to quantify the predictive performance and how well can the model predict the real phenotypes of the selection candidates based on information from the reference population. Several attempts have been made to derive deterministic formula to predict accuracy of genomic breeding values (e.g., Goddard et al. 2011). The predictive performance is commonly summarized with two statistics: correlation between genomic breeding values and phenotypes and the coefficient of regression of phenotypes on genomic breeding values. Phenotypes can be actual observations, summary statistics like daughter/progeny deviations (VanRaden and Wiggans 1991), or de-regressed breeding values (Garrick et al. 2009).

Genomic selection studies commonly include an assessment of the predictive performance. This should avoid overfitting of the model that occurs easily as the number of marker effects to estimate is often much larger than the number of observations. Cross-validation is widely used as a technique to assess predictive performance. It divides the reference population into a training and a validation set, estimates marker effects in the training set, and then validates these. Various cross-validation designs have been applied: (1) two-generation scheme, (2) k-fold cross-validation, and (3) repeated subsampling validation. In the two-generation scheme, individuals are assigned to the training or test set based on their generation number or year of birth. The youngest individuals are included in the test set. In a k-fold cross-validation, individuals are divided into k disjoint sets of equal size. In each fold, one set is used for testing, and the other $k-1$ sets are used for training. This splitting is repeated until all sets have been used once for testing. In the repeated subsampling scheme, the reference population is randomly split into a large (e.g., 95%) training and a small testing (e.g., 5%) set. Again the splitting is repeated many times. All these cross-validation schemes have advantages and disadvantages, some of which are discussed by Morota and Gianola (2014), but there is no consensus about which one is the best. The two-generation scheme is the scheme closely resembling a practical genetic evaluation scenario and is the cross-validation scheme most often applied in genomic selection studies.

The purpose of genomic evaluation applied in practice is to predict the performance of future offspring. However, the offspring may be several generations separated from the reference population. In dairy cattle, for example, only the second or third generation of ancestors of selection candidates is included in the reference population in an efficient genomic selection scheme. Nevertheless, most cross-validation studies in dairy cattle have a validation horizon of at most one generation. This is similar in plant populations, especially in the discussed one-stage selection, which is the more common genomic selection scheme (Marulanda et al. 2016). Such a short validation horizon has two main consequences: (1) the predicted accuracy of selection is too optimistic, and (2) the comparison of models may not reflect the actual performance of the models. For example, models better at capturing linkage disequilibrium between markers and QTNs are expected to perform better for a longer validation horizon than models heavily relying on genetic

markers tracing family relationships. A design with a short validation horizon might be a problem in outbreeding plants, as population-wide linkage disequilibrium is large and predictions are more feasible using a family design.

As a concluding remark, the design of a cross-validation study needs to mimic the intended use of genomic breeding values, such that the estimated predictive ability is consistent with the actual application in mind, and several opportunities exist to improve cross-validation studies.

6 Further Advancements of Methods

The methods and models for the use in genomic predictions are continuously advanced. Some of the suggested extensions of the concept of genomic selection are described here. Such extensions are the inclusion of methods for the manipulation of genomes (genome editing), more detailed information (biological information, data on transcriptome or proteome), or improved genotyping tools (use of sequence information and the concept of genomic selection 2.0).

6.1 Integration of Genetically Engineered Individuals

Genetically engineered or genetically modified plants can be found in the food production chain, while the first genetically modified livestock species has only recently been approved for consumption by the FDA after an approximately 20-year approval period. Different techniques can be used for the modification of genomes including transformations, such as microinjection and electroporation. Transformations were the first modifications successfully applied in plant and livestock species. Other modifications include gene knockouts or knock-ins, which are more common in model species like mice, to test the functions and effects of genes. The inhibition of genes for a short time can be done using, for example, RNA interference (RNAi) employing short RNAs. Many of these methods, especially the modification of individuals using transformations, are rather unspecific, and multiple trials need to be done until the modification is successful. Examples of modifications include changes of the product composition [e.g., golden rice (*Oryza sativa*)], introduction of resistance/tolerance against pathogens [e.g., ringspot virus-resistant papaya (*Carica papaya*)], resistance/tolerance against insects [e.g., potato (*Solanum tuberosum*)], resistance/tolerance against herbicides [e.g., soybean (*Glycine max*)], abiotic stress tolerance [maize (*Zea mays*)], and pollination control system (e.g., maize) in plants as well as enhanced growth [AquAdvantage® salmon (*Salmo salar*)], enhanced production [alpha-lactalbumin pigs (*Sus scrofa*)], enhanced metabolism (EnviroPig®), and the production of human drugs [lysozyme goat (*Capra aegagrus hircus*)] in livestock (Forabosco et al. 2013).

A more recently developed method to modify parts of the genome is genome editing, which allows the targeted change of one or few nucleotides at a specified

position in the genome. Genome editing requires programmable nucleases, which were firstly identified in 1996. Methods commonly used for genome editing include zinc finger nuclease (ZFN), transcription activator-like effector nuclease (TALEN), and the Cas9-guide RNA system (CRISPR) (Gaj et al. 2013). Genome editing is of major interest in plants and also in livestock since modifications are more targeted and success rates are higher. While traditional modifications were applied in plants, their applications were more restricted in livestock. The opportunities offered by genome editing have therefore led to a huge interest of the livestock research community. More than 300 edited pigs, cattle (*Bos taurus*), and sheep (*Ovis aries*) have been developed since 2011, using nonhomologous end joining or homology-dependent repair (Tan et al. 2016). Edited animals were produced via zygotes or somatic cells. The technique can be used to produce animals as potential organ donors (pig), disease models (pig), bioreactors (cattle), and founder animals of genetic lines with enhanced productivity (cattle, sheep, goat) and to introduce disease resistance into populations (pig) (Proudfoot et al. 2015). Other traits of interest in livestock are especially the horn phenotype in cattle, mastitis resistance in dairy cattle, and resistance to the African swine fever in pigs. The selection for some of these traits cannot be achieved using other breeding methods since relevant alleles are not present in the population (e.g., resistance to African swine fever). The selection of other traits will require long selection periods with a high risk of inbreeding, for example, if the frequency of the alleles is too low to allow selection without loss of diversity (e.g., selection against horns).

The introduction of such new tools into livestock breeding programs will require that relationships between individuals are taken into account to decrease the risk of higher inbreeding. A simulation study suggested that the application of a combination of genomic selection and "promotion of alleles by genome editing" might lead to substantial improvements of response to selection (Jenko et al. 2015). However, one prerequisite of genome editing is that QTNs are identified. It was furthermore suggested that the breeding programs need to be adapted to avoid a rapid depletion of genetic variation in the population.

6.2 Inclusion of Biological Information

The initial and currently applied idea of genomic selection is that of a black box approach, where knowledge of the function of the markers used for selection is not considered. Nevertheless, incorporating genotypes from whole-genome SNP arrays into existing evaluation systems has been successful in increasing the accuracy of EBV of young animals for commonly recorded traits (Lôbo et al. 2011; Northcutt 2011; Wiggans et al. 2011). However, the applicability of these predictions is limited to selection within breeds as the prediction ability of the estimated marker effects is

highly dependent on the relationship between the reference population and the selection candidates (Boichard et al. 2016).

If the black box approach of genomic selection is overcome and additional biological information is available, genomic evaluation may become more accurate, especially for crossbreed predictions. One of the initiatives to improve accuracies is the "1000 bull genome project" (Daetwyler et al. 2014). The objective of this project is to make the sequence data of over 1000 influential sires available. This should improve imputation, genome-wide association studies (GWAS), and genomic prediction and, more importantly, promote the identification of causal variants.

The availability of accurately annotated genomes, both structural and functional, is essential for the biological insight into traits (Stein 2001). In order to relate markers to genes and phenotypes, a fully assembled genome with known gene locations and structures, information on noncoding RNA, regulatory and repetitive regions is required. Moreover, the functional annotation like gene ontology (GO) classification that describes products of eukaryotic cells in terms of molecular function, biological processes, and cellular components, as well as descriptions of metabolic and signaling pathways and gene regulatory networks, can provide valuable information. Currently, several such databases are available and updated continuously for a variety of species. Some examples are the GO browser agriGO (Du et al. 2010) that represents 45 agricultural species, including plant, fungi, insect pests, and livestock species, and the Reactome (Croft et al. 2011), MetaCyc (Caspi et al. 2014), and KEGG (Kanehisa et al. 2008) databases that integrate genomic, chemical, and systemic functional information.

The incorporation of biological information into the genomic evaluation can be done in various ways. A simple and straightforward approach is the selection of subsets of markers from the whole-genome SNP arrays that are associated with genes or metabolic pathways of interest. This could be extended to include a polygenic component, using pedigree relationships to account for the rest of the genome (Snelling et al. 2011). Moreover, the priors of Bayesian models could be shaped by biological knowledge and become more informative (MacLeod et al. 2016).

6.3 Transcriptome and Proteomic Assisted Selection

High-throughput technology is not only applicable to the information of the genome but also transcriptome, proteome, and metabolome. Information on the transcriptome, such as data from RNA sequencing, does provide information on mutations within the genome and adds knowledge of probable functionality as only expressed genes will contribute to the phenotype. High-throughput platforms such as expression arrays may further allow collecting expression information for many loci and individuals. High-throughput platforms do also exist for the analysis of the proteome (Chawade et al. 2016). Peptide-based selection using mass spectrometry might assist selection for certain phenotypes. Its application had been tested in plants and allowed the selection for traits for which no good genetic markers were available (Chawade et al. 2016). The analysis of metabolites and their variance was suggested as another (post-genomic) tool for

improved selection (Fernie and Schauer 2009). The use of metabolomics-assisted breeding, possibly in combination with sequencing and reverse genetics, might be useful for a number of traits including selection for resistance and tolerance traits in plants (Zamir 2001; Morandini and Salamini 2003; McCouch 2004; Takeda and Matsuoka 2008; Fernie and Schauer 2009). The feasibility of the use of expression profiles or protein signatures in future breeding systems has yet to be explored. A combination of tools based on traits might be a possible scenario for the improved selection especially for the improvement of complex traits.

6.4 Alternative Genotyping Methods

Genotyping of many individuals or lines is a prerequisite for genomic selection. The density of genotyping platforms required for a reliable prediction depends on the population of selection candidates and its genome structure. Genotyping arrays with various densities are available. The use of customized and population-specific arrays with lower marker density to genotype selection candidates and combining these with sequence data of influential ancestors of the selection candidates can reduce costs for genomic selection. Such approaches display alternatives to the use of high-density genotyping arrays and are applied in some breeding programs. The imputation of genotypes can additionally be used to increase the information content. The process of imputation implies that genotypes are predicted, which are not directly assayed in a sample. Ancestors will be genotyped using information on the full genome sequence or high-density marker arrays. If information on such dense genotyping is available, most haplotypes in the populations are covered and thus can be implied in individuals with information on fewer genetic markers (Marchini and Howie 2010). The increase of genotype information in the population might improve the accuracy of genomic selection (Druet et al. 2014). Imputation can also be used for the correction of genotyping errors. However, one essential step to allow accurate imputation is the correct phasing of the genomic information (Hickey 2013; Hickey et al. 2014).

Other alternatives to higher- or lower-density genotyping arrays exist. These should allow to reduce costs by skipping the need to develop genotyping arrays. The use of genotyping by sequencing (GBS) opens opportunities to fill the gap between highly explored lines of major interest and non-reference lines (Spindel et al. 2013; Williams et al. 2014). Genotyping by sequencing is especially of interest when little or no genomic information is available, no dense genotyping platform exists, or genetically highly diverse material is used. This approach is also useful for large genomes. It is therefore of interest especially in plant breeding. However, well-established bioinformatic infrastructures are required to fully explore genotyping by sequencing.

6.5 Genomic Selection 2.0

The term "genomic selection 2.0" was introduced based on the advance of tools for genotyping and sequencing (Hickey 2013). It was suggested that, while progress has been made using genomic selection in general, the large amounts of sequence data generated could be utilized even more. Genomic selection 2.0 is based on the use of big data and the availability of sequence data in combination with imputation methods and aims to integrate new methodologies for the integration of de novo mutations and variants different from SNPs. The genomic information of a huge data set in combination with phenotypic information should be powerful to identify QTNs for many traits. Genomic selection 2.0 intends to avoid sequencing at a high depth, which is not feasible for breeding derived from a large number of male ancestors, as in livestock breeding, or many potential breeding lines, as in plants. The choice of a lower-coverage sequencing for all individuals could assist to discover the haplotypes and allow the imputation to full sequences for all individuals. The power of methodologies will depend on the advancement of applications such as imputation algorithms, technologies for genotyping, and infrastructure of bioinformatic analysis. Furthermore a new generation of genomic selection is an important step to allow higher recombination in populations. It should allow the integration of de novo mutations, occurring from random events during recombination or being introduced by methods such as genome editing (Hickey 2013).

7 Genomic Prediction Applied in Animal, Plant, and Human Populations

7.1 Examples of Genomic Selection in Livestock

7.1.1 Cattle (*Bos Taurus*)

The practical application of genomic selection is of immense interest in dairy cattle, where GEBVs are now common selection criteria across many countries (http://www.interbull.org 2013). This is not the case for other livestock populations, especially in populations where crossbreeding is used, such as pig or beef cattle. Also the lack of phenotypes and genotypes for a reliable prediction of a GEBV or the focus on traits with lower heritability, such as fertility in beef cattle, does restrict the use of genomic selection (Johnston et al. 2012). Genomic selection is expected to be more cost-efficient compared to traditional selection schemes as it allows the reduction of the expensive evaluation of phenotypes and allows an earlier selection of male animals for further breeding. A concern is often the cost for genotyping (of the reference population), but with the ongoing development of genotyping platforms, this might be less of a problem in the near future. Changes might need to be applied to the structure of the breeding industry, which will require a long-term planning (Johnston et al. 2012). But the question is how genomic selection is implemented

in different breeding programs, and which relevant aspects of the species and breeding program have to be taken into account?

Genomic selection is today a selection scheme in dairy cattle, especially Holstein Friesian, in many countries (Patry 2011). Decreased costs for genotyping had progressed genomic selection to a routine in some herds (Hayes et al. 2009a; Hayes et al. 2009b). Jannink et al. (2010) stated that the implementation of SNP information will reduce the costs for genetic evaluation. The cost for obtaining marker information can be equal to the costs to collect phenotypic information from 10 to 20 daughters per bull. Selection in dairy cattle is focused on bulls, and many breeding programs focus on genotyping solely bulls. Costs for genotyping can therefore be kept relatively low. The breeding decisions can be made by preselecting young bulls for further testing in the so-called preselection scheme. An alternative option to make use of the genotypes is the turbo scheme which allows the earlier selection of new breeding bulls (Pryce and Daetwyler 2012; Bouquet and Juga 2013). Additional genotyping of cows will allow a better assessment of additional traits and to increase the size of the reference population. One advantage of genomic selection in dairy cattle is the drop of the length of the generation interval from around 5 to 6 years in traditional dairy cattle breeding programs to around 1.5 years when using genomic selection (Pryce and Daetwyler 2012). The increase of the genetic gain in general might be attached to the risk of higher inbreeding as it might reduce the number of genetically superior breeding animals. Genetic markers should therefore also be used to avoid loss of diversity by carefully observing the remaining of haplotypes and the structure of the population (Young et al. 1988).

The implementation of genomic selection is also of major interest for beef cattle breeding, for which generation intervals are also long. A number of differences compared to dairy cattle exist, including the lower rate of the use of artificial insemination and the use of crossbreeding. The lower rate of artificial insemination, compared with dairy cattle, reduces the contribution of a selected individual to the genetic progress in the population at large and thereby reduces the amount of resources that can be invested in genotyping. The genetic makeup of populations is relevant, and genomic selection would probably be restricted to purebred operations. Predictions in crossbred populations are not as accurate as compared to those in purebred populations. Beef cattle populations are less uniform compared to dairy cattle populations, crossbreeding is common, and both *Bos taurus* and *Bos indicus* populations are a part of breeding schemes (Garrick 2011). The effective population size in many beef cattle population is low as is the number of bulls with reliable EBVs. This will restrict the reference population and the reliability of the estimated GEBVs (Johnston et al. 2012). The combination of data across countries and/or across breeds is an option to overcome the small size of reference populations, but higher-density marker panels might be required to reach reliable predictions when using such datasets (de Roos et al. 2009). Genotypes of cows can also be included to achieve larger reference populations, additionally allowing farmers to select superior cows (Saatchi et al. 2012). This would allow a better selection for fertility, one of the most important traits in cows, which has a low heritability. The inclusion of cows in the selection scheme would require changes to traditional breeding using progeny testing, which focusses largely on bulls. Genomic selection in beef cattle could thereby lead to a more balanced breeding goal via inclusion of animals and traits at the farm level.

Genomic Selection in Dairy Cattle

The seminal publication by Schaeffer (2006) illustrated that adoption of genomic selection could decrease the costs of running a breeding program and increase genetic progress, compared to progeny test schemes that had been in place for many years. The development of a high-density SNP array (Matukumalli et al. 2009) removed a last practical hindrance for the implementation of genomic selection.

The first official release of genomic breeding values was in 2009 in the USA (Wiggans et al. 2011). At that time, just over 5000 progeny-tested bulls were included in the reference population. Using an approach resembling Bayes A, reliabilities of genomic breeding values were on average 50% (VanRaden et al. 2009a, b). This meant an increase of 23% in reliability compared to the reliability of parent averages.

Dairy producers in the USA were quickly to adopt the technology, and by 2012 half of the Holstein service sires were genotyped as young bulls, i.e., bulls with just genotype information and no daughter information (Hutchison et al. 2014). Also breeding companies made changes to their breeding programs and started to use genotyped young bulls as sires of sons. As a consequence, Hutchison et al. (2014) and García-Ruiz et al. (2016) could observe a significant decrease in the generation interval.

Evidence of increased of genetic gain due to genomic selection was presented by García-Ruiz et al. (2016), who reported that the genetic gain for yield increased twofold after the introduction of genomic selection. For fertility, life span, and udder health, even larger increases in genetic gain were observed, in agreement with the prediction that genomic selection would be especially useful for traits with low heritability.

In recent years, focus has been expanded from genotyping predominantly males to genotyping females as well. In July 2017, a new milestone was reached in US dairy genetics with the submission of the two millionth genotyped animal to the US dairy database (Press release, 2017; https://queries.uscdcb.com/News/CDCB%20AGIL%20Two%20Million%20Genotype%20Mark.pdf). Genotyping females will allow commercial dairy farmers to make more informed breeding decisions in their own herd but also provides new opportunities for improved herd management.

7.1.2 Sheep (*Ovis aries*) and Goats (*Capra aegagrus hircus*)

Breeding of small ruminants, sheep and goats, varies as the size and structure of enterprises differ between countries. Small ruminants are especially part of the production system in low-income countries as the resource inputs are low. However, larger breeding cohorts exist in countries with options for higher input and selective

breeding based on performance information (van der Werf 2007). Differences in the management, structure, and size of the populations and breeding programs depend on the product (meat, wool, or milk) and also the location of the farm. The use of local breeds is more common for sheep and goat breeding; thus populations are small and breeding is more often country-specific. The integration of genomic selection into breeding programs is especially discussed for countries with large breeding populations, such as Australia, New Zealand, Great Britain, South Africa, or France. Reference populations have been established in some countries for the collection of reliable phenotypes (Swan et al. 2012). The main restriction of the application of genomic selection in small ruminants is the lack of the data/information on a large number of phenotypes, which is necessary for the creation of a reliable reference population. The shorter generation interval in small ruminants (compared to cattle) and the relatively high genotyping costs will restrict the genetic gain when using genomic selection. Effective population sizes are often large in small ruminants as these populations are usually more heterogeneous compared to other livestock populations. And finally natural service is still more common in many sheep breeding schemes, which will restrict the number of possible fertilizations from each ejaculate to one. More rams are required when natural service is used instead of artificial insemination. Other relevant points to be considered when applying genomic selection in small ruminants are the options for the size of the reference population (Shumbusho et al. 2013); population-specific factors in sheep, including seasonality of the production system; small-scale use of artificial insemination; and low value of individual animals, which will require different approaches of genomic selection compared to dairy cattle (Baloche et al. 2014). Predictions using crossbreed animals are also relevant in sheep. This should allow the application of genomic selection in a larger range of breeding populations and covering existing breed diversities in populations.

7.1.3 Pigs (*Sus scrofa*)

Separate selection schemes at the nucleus level, one for the paternal production-oriented breeds and one for the maternal reproduction-oriented breeds, exist in pig breeding. This structure needs to be considered when using genomic selection. Traditional selection has a larger focus on performance traits with a selection of superior sire lines for improved carcass and meat traits. While genomic selection in male lines can improve selection efficiency/effectiveness, phenotypes of relatives might be needed to increase the reference population (Tribout et al. 2012). Genomic selection could also take better care of the selection for maternal traits. The shift of the focus to maternal traits will especially be feasible when total costs of genotyping are reduced. Simulation studies have shown the improved accuracies of selection for economically important traits also in female purebred lines (Lillehammer et al. 2011; Tribout et al. 2012). However limitations exist for the prediction of performance of crossbred animals, which are usually used in the final stages of the production and often as maternal lines. Genetic correlations between traits in cross- and purebred

animals are less than 1 (Dekkers 2007). Suggestions to overcome the limitation of the less than unity genetic correlation between crossbred and purebred performance have been made, such as the integration of QTL information into breeding decision.

7.1.4 Poultry/Chicken (*Gallus gallus domesticus*)

Poultry has dual use and is therefore bred in different lines to allow for differential selection for egg and meat production. The generation interval in poultry is relatively short with 1 to 1.5 years, and the rate of genetic improvement in traditional breeding is more than double in chicken compared to cattle or pigs. The implementation of genomic selection could reduce the generation interval to only 6 months. But the population sizes need to be carefully evaluated to reduce costs of genotyping on the one hand while not reducing the effective population size. Adequate genotyping platforms have been recently developed, and first predictions showed the potential for genomic selection to increase genetic gain in poultry breeding (Preisinger 2012). However, the advantage and cost-efficiency over conventional breeding have to be proven before genomic selection could be applied as a selection tool in privately owned poultry breeding companies (Preisinger 2012). Hypothetical studies have also suggested the implementation of genomic selection in broiler lines. However, genotyping strategies need to be chosen carefully to reduce costs without the loss of important information on marker-phenotype relationship (Avendaño et al. 2010).

7.1.5 Aquaculture

The introduction of genomic selection has also been discussed in aquaculture, especially fish. Genotyping tools will enable to control inbreeding, but costs are currently the main inhibitor for a quick adoption of genomic selection in fish breeding schemes (Nielsen et al. 2011). Male and female fishes have many offspring; the contribution of male and female individuals to the breeding cohort is therefore high. If the use of genomic markers can improve selection in fish breeding, the expected genetic gain can be twice as high compared to traditional selection using BLUP. The use of genetic markers could also assist controlling inbreeding more effectively (Nielsen et al. 2011). Aquaculture breeding programs might need to be redesigned entirely to accommodate genomic selection. Such changes can be a reduction of number of families or reduced phenotypic evaluation (Sonesson and Meuwissen 2009; Nielsen et al. 2011). A combination of traditional BLUP estimation, preselection of candidates, and low-density genotyping arrays might be one possibility. It could reduce costs for genotyping many potential parents and thus reduce the expected genetic gain only slightly (Lillehammer et al. 2013).

7.2 *Examples of Genomic Selection in Companion Animals*

Estimated breeding values are used successfully for selection in some horse and dog populations, and genomic selection is discussed for further improvement. The lack of large enough reference populations is often the restricting factor for the implementation of genomic selection. Most dog breeds are based on a few founders, and the effective population size is relatively small. The use of genetic markers for selection is feasible. Genotyping arrays are available for both horses and dogs. Genetic markers should improve the predictive ability and lead to a more accurate selection. On the other hand, many of the traits used for selection are complex and not always measureable on a reasonable objective scale. The implementation of genomic selection will require a careful design of an appropriate reference population with reliable and relevant phenotypes.

7.2.1 Dogs (*Canis lupus familiaris*)

Only a few studies have investigated the application of genomic selection in dogs (Sánchez-Molano et al. 2015). Dog breeding is usually done based on pedigrees and phenotypic measurements. Breeding goals include improved health with traits aligned to the breed standard while avoiding inbreeding. Inherited disorders, such as hip dysplasia, heart problems, and certain kinds of cancer, put traditional dog breeding into negative lights and need to be taken into account in breeding programs. Since traits related to health often have a late onset, information from genetic markers using data from a large reference population would therefore be useful. The use of genomic selection or prediction models would also allow a better correction for environmental factors (among which the influence of the breeder). Problems to overcome in dog breeding before genomic selection could be applied are the need of collecting data from many dogs for developing a reference population of appropriate size and the need for reliable phenotypes and for continued phenotype collection after the introduction of genomic breeding values. Genomic selection is a potential tool to improve selection and especially traits related to welfare (such as health or inherited defects) of pedigree dogs.

7.2.2 Horses (*Equus ferus caballus*)

The sport horse industry aims for a more accurate selection to reach high genetic improvements. Generation intervals in horses are around 8–10 years, and earlier selection, for example, by using genetic markers (Haberland et al. 2012; Stock et al. 2016), could increase the rate of genetic improvement. Some of the traits of interest are also related to behavior and temperament, which are difficult to measure objectively. The establishment of international collaborations is not always straightforward. Limited exchange of genetic material leaves many small and (semi-)

isolated populations at risk of decreasing effective population sizes, increased inbreeding, and potential increase in prevalence of inherited diseases. Large reference populations with reliable phenotypes are needed to apply genomic selection with high accuracies. Genomic selection will improve accuracies achieved with EBVs in young animals and also horses imported from other countries for which only scarce information on relatives is available in the importing country. Genomic selection will be especially useful for traits with late onset and low heritability. In a comparison of several selection strategies against osteochondrosis, van Grevenhof (2011) found genomic selection to be a realistic option for the Dutch warmblood population. Similar to dog breeding, a very relevant aspect in horse breeding is the structure of many small-sized studs and fewer large enterprises, compared to livestock breeding, with its challenges to achieve the level of collaboration needed to put in place an organized scheme necessary for a successful implementation of genomic selection. Also international collaborations will be necessary as they have the potential to increase the reference population.

7.3 Examples of Genomic Selection in Crop Plants

The status of the implementation of genomic selection in different crop species varies. Genomic selection is of interest to the public and private crop breeding community. One main reason for the search of improved selection tools is the stable and high costs for phenotyping. As little can be done to reduce costs per line, the only option is a reduction of the number of lines to be phenotyped. The crop breeding community hopes for a significant reduction of costs for the development of new breeding lines when using genomic selection instead of traditional phenotypic selection (Heffner et al. 2009; Heffner et al. 2011; Resende et al. 2012b). But the implementation of genomic selection will depend on costs for genotyping and the availability of whole-genome sequencing and/or genotyping platforms. Crop breeding programs are versatile, and strategies for the implementation of genomic selection will need to be adjusted for each breeding program. Hybrid vigor or heterosis is important in many crop breeding populations, and models for genomic selection should also be able to take nonadditive effects into account (Duvick 1999).

Traditional selection is often based on phenotypic selection. Breeding of inbred lines for the production of hybrids and crossing of diverse parental lines for the production of new inbred lines in successive cycles of selfing are the two main strategies. Phenotypes might differ between the plant materials used for selection in early and advanced cycles of breeding because the number of tested lines in early cycles is often too large for a cost-effective collection of all relevant phenotypes. The use of phenotypes from the final cycles of breeding might therefore reflect more useful data as they will most accurately reflect the final product. The estimation of marker effects based on advanced cycles of selection (Zhao and Xu 2012) needs to be considered carefully. Nonadditive effects due to heterosis or inbreeding effects can change the prediction accuracies. The choice of the reference population has to

consider these effects. Also the structure of the reference populations is important, and one option for a high accuracy of prediction is that individuals in reference and validation subpopulations should show a close relationship (Asoro et al. 2011).

An additional difficulty in crop species is the impact of genotype by environment (GxE) effects on performance of lines (Lorenzana and Bernardo 2009; Heffner et al. 2011). Advanced generation populations in traditional breeding are therefore tested across different environments. The adaptability of crop lines is a relevant criterion for the successful production in the field. The evaluation of genotyped lines across different environments can increase the gain and cost-efficiency of genotyping (Bertin et al. 2010; Xu and Hu 2010; Morrell et al. 2012). Sequencing of selected lines in combination with the repeated collection of phenotypic data has been another suggestion to overcome an inaccurate estimation of genomic breeding values caused by genotype by environment effects (Morrell et al. 2012).

7.3.1 Rice (*Oryza sativa/Oryza glaberrima*)

Reports on genomic selection in rice are rare, and pedigree breeding based on phenotypes is still the predominant breeding method (Li and Zhang 2013). Successes have, for example, been made in increase of yield, but yield potential needs further improvement in the future. Also selective breeding success stories to improve complex traits (such as drought tolerance) are limited. One of the reasons for this limitation is the lack of information on reliable phenotypes especially from hybrid breeding, which is increasingly common in rice (Yan et al. 2011). There is furthermore little genetic variation in the current breeding populations; genotyping might assist to adjust breeding strategies to avoid the loss of important genes due to a more narrow gene pool (Breseghello 2013). Genotyping can also assist to identify more diverse parental lines, which can then be used to achieve high heterosis effects in crossbred populations (Chen et al. 2013). But more research is needed to fully exploit the possibilities of genomic selection in rice.

7.3.2 Maize (*Zea mays*)

Large efforts are underway for the implementation of genomic selection in maize, another important crop in many countries around the globe. Significant improvements have been made since the domestication, and there is little resemblance between the original Balsas teosinte (*Zea mays* ssp. *parviglumis*) before domestication and modern maize plants today. Improvements are especially focused on tassel, ear, cob, and kernel characteristics, flowering traits, as well as resistance to drought and pathogens. Genomic selection will improve the breeding process further as it allows the prediction of untested lines, including testcrosses, in advanced breeding populations (Albrecht et al. 2011). One additional advantage of the application of genomic selection in maize is the reduction of the generation interval. Phenotypic evaluation is not required throughout the entire selection process when genomic selection is used, and generations of lines can be bred in greenhouses (Zhao et al.

2012). The design of the reference population requires good knowledge of the population structure and genetic relationships within and across relevant lines (Albrecht et al. 2011; Windhausen et al. 2012). Biparental or diversity panels and testcrosses are important/useful in maize breeding. Advanced breeding populations are often based on the performance in many testcrosses. But genotyping of all testcrosses will be expensive, and preselection is required. Genetic markers can be used to investigate genetic differences between lines to select more diverse individuals for crossing (Albrecht et al. 2011). Different strategies for genomic selection are being tested in maize. International centers, such as the International Maize and Wheat Improvement Center (CIMMYT), drive research in this sector.

7.3.3 Wheat (*Triticum aestivum*)

It has been shown that the implementation of genomic selection can lead to higher genetic gain per unit time and cost reduction compared to traditional pedigree-based selection in wheat (Burgueno et al. 2012). But the use of related populations in the reference and selection set is a major factor to achieve a reliable accuracy when using genomic selection in wheat (Crossa et al. 2013). Information from the reference population for predictions needs to be collected in environments which are similar or the same to those of the selection candidates as environments play a major role (Crossa et al. 2011, 2013). Accuracies across field trials can be increased when information based on many lines and different environments are included in the modeling of the genetic effects (Burgueno et al. 2012; Dawson et al. 2013). Wheat breeding focusses on a number of traits, including grain yield, quality traits, tolerance to abiotic stresses (drought and heat), and disease resistance, as listed in a review from the CIMMYT breeding scheme (Guzman et al. 2016). A good phenotypic recording of disease traits is important for the selection of more resistant lines. However, evaluation of infection traits is time-consuming and costly. Genomic selection can be used as a strategy to improve the gene pool for resistance (Rutkoski et al. 2011) and also other relevant agronomic traits. It allows the implementation of historic information and the prediction of many traits at the same time. No additional costs may incur if lines are selected based on predicted phenotypes from genomic information as shown in the example of the first wheat line traits of the CIMMYT breeding program (Guzman et al. 2016).

Genomic Selection in Spring Bread Wheat: CIMMYT's Breeding Efforts

The International Maize and Wheat Improvement Center (CIMMYT) has discussed the use of genomic selection for the improvement of their wheat and maize breeding programs early on. The spring bread wheat program is one of the examples in which genomic selection has been tested, and details of the

(continued)

program have been described in Battenfield et al. (2016), and a summary is provided here.

The F7 spring bread wheat lines were derived from F5 lines, which were tested and evaluated for quality traits for 1 year in Mexico. Superior lines from testing were chosen for advanced end-use quality testing. Five plants per lines were used for genotyping using genotyping by sequencing with further imputation of missing genotypes. Marker effects were calculated using a number of models including ridge regression best linear unbiased predictor, reproducing kernel Hilbert space, partial least squares regression, elastic net, and random forest. The efficiency of the models in predicting breeding values was tested using cross-validation on data across multiple years trained at 80% randomly selected data to predict 20% masked data, as well as forward prediction trained on all prior data. The data collection for the described spring bread wheat modeling started with trials harvested in 2010 and included a total of 47,817 lines in the yield trial, of which 7858 lines had been screened for quality. From a total of 5520 of these lines, phenotypes and genotypes were available until 2015.

When comparing the results of predictions using cross-validation and forward prediction, it was concluded that cross-validation will likely lead to an overestimation of the prediction ability of genomic selection. On the other hand, only small differences were observed between the predictive abilities of using different models for genomic selection. Correlations between the observed and predicted phenotypes differed for different traits and varied between years. The response to selection using phenotypic and genomic selection increased between 35% (test weight, kg h L^{-1}) and 147% (alveograph P, tenacity divided by L, extensibility, mm mm^{-1}). One main advantage when implementing genomic selection in the CIMMYT spring bread wheat program is the possibility to select for phenotypes, such as wheat quality, which will, in a phenotypic selection program, only be used as selection criteria during late stages of the breeding program. This advantage is common to many other crop species, most of which evaluate major traits of interest only late in the breeding pipeline. Accuracies from the tested models in the spring bread wheat program were high enough to allow the application of genomic selection and increased with larger training populations. Genomic selection will allow a reduced phenotypic evaluation, which currently requires more seed material and which represents a considerable cost factor. While genomic selection might not replace the collection of phenotypes, it will allow early selection of future breeding material. A 1.4 to 2.7 times greater gain from selection was further predicted when the number of selection candidates increases from 2000 to 10,000. The implementation of genomic selection in the CIMMYT spring bread wheat breeding program has started in 2012, and it is predicted that it will enable the selection for specific end-user traits.

7.3.4 Barley (*Hordeum vulgare*)

A shorter breeding cycle and thereby early selection gain is also the expectation when using genomic selection in addition to phenotypic evaluation in barley breeding. The accuracy of genomic selection is higher if correlations between reference and selection populations are high and/or trait heritabilities are low (Iwata and Jannink 2011). Typical traits used in the breeding goals are yield or yield-related traits (grain dry matter yield or thousand kernel weight), quality traits, and resistances against diseases. A carefully selected reference population may allow an improvement using genomic selection compared to phenotypic selection even in biparental crosses (Jannink et al. 2010). But the predictive ability depends often on the relatedness, the population structure needs therefore to be taken into account, and the number of markers required will depend on population structure and the linkage phase (Thorwarth et al. 2017). A reference population might use more inbred or highly replicated samples, more diverse samples, or lines different from the population used for phenotypic selection. Genomic selection might also help to improve decisions on crossbreeding (Bernardo 2010), if a reference population is well selected. This is especially relevant in self-pollinating plants, such as barley, where time-consuming crossing by hand is required in order to produce biparental crosses. However, some studies express concerns regarding the risks of lower genetic variation due to the loss of favorable alleles (Jannink 2010), especially when breeding cycles are shorter.

7.3.5 Other Crop Species

Other crop species, with complex breeding goals, are forage plants, for which the aim is to increase production as well as maximize perennial persistency. Perennial forage grass [mostly ryegrass (*Lolium perenne*)] plots should be used with a consistent quality and quantity over many years; deployment of hybrid breeding is, therefore, not applicable (Wilkins and Humphreys 2003). Genomic selection should especially improve the prediction when correlations between phenotypic evaluation and performance are low, such as for complex traits or traits that could be recorded only in advanced reproduction cycles. The use of genetic markers might assist the reduction of the lengthy periods for phenotypic selection (Hayes et al. 2013; Resende et al. 2014). However, data and sample management from parental lines, including recordings of pedigree information, need to be improved. Genomic selection should allow a focus on a few traits during the phenotypic evaluation and will enable to control that relevant alleles remain in the breeding cohort (Resende et al. 2014). The use of genetic markers might be more efficient for the introgression of specific genes compared to backcrossing (Wilkins and Humphreys 2003).

Application of genomic selection has, until now, been less discussed for other crop species including examples from the genus *Brassica* (Cowling et al. 2009; Cowling and Balazs 2010; Cullis et al. 2010; Wurschum et al. 2014), oats (*Avena*

sativa) (Asoro et al. 2011), potato (*Solanum tuberosum*) (Barrell et al. 2013), sugar beet (*Beta vulgaris*) (Hofheinz et al. 2012; Wurschum et al. 2013), sugarcane (*Saccharum officinarum*) (Gouy et al. 2013), or soybean (*Glycine max*) (Shu et al. 2013). Some restrictions are the availability of genotyping tools, sizes of possible reference populations, as well as the need for further improvements in evaluation of phenotypes.

It had been suggested that modifications to breeding programs (such as number of lines per breeding cycle, number of test staged in the program, more collaborations between breeders) might be needed to achieve economic gain via genomic selection (Cowling and Balazs 2010; Hayes et al. 2013). It is important to keep in mind that the selection unit is not a single plant but a heterogeneous line, variety or plot. Genomic selection needs to be adapted to address the traits and structure of the distributed product, breeding schemes which are used to produce seeds from inbred or hybrid lines for the use by farmers.

7.4 Examples of Genomic Selection in Trees

The generation interval, breeding cycle, and duration until phenotypes can be evaluated in tree breeding are long. The identification of better estimators for the quality seedlings for the production is therefore a major interest for the forest and fruit tree industry. Advantages of using genomic selection will arise mainly from the shorter selection cycles (Iwata et al. 2011).

7.4.1 Forest Trees

Testing different scenarios of genomic selection in eucalyptus breeding for height and diameter at multiple ages allowed the total breeding cycle to be halved (Resende et al. 2012a). Intensive progeny testing can be eliminated, and a second clonal trial will not be needed allowing for good economic returns (Resende et al. 2012b). Methods to reduce the maturity age (breeding duration) and speed up propagation are already implemented in tree breeding. However, emphasis should be put on reducing the testing phase if the total breeding interval needs to be reduced (Resende et al. 2012a). Even though it had been concluded that genomic selection will, alongside other reproductive methods, decrease the total time of a breeding cycle in conifers, it has also been seen that models predicted early during the breeding cycle, for example, in seedlings, have only limited applicability for the selection of older trees. Also the comparisons of predictions across locations did not lead to high accuracies for all scenarios (Resende et al. 2012a). Additionally, genetic regions explaining trait variation were often population-specific as shown using eucalyptus populations (Resende et al. 2012b). Older data and genotypes within the same breeding scheme including crossings of the same elite trees might therefore be

more useful to create a reference population aiming for high accuracies, as suggested in some of the scenarios in conifers (Iwata et al. 2011).

A number of studies had been conducted to identify markers associated with relevant traits (e.g., wood quality, wood formation, growth, hardiness, drought response, disease resistance) in trees, but not many of those markers are currently being used in breeding programs (Thavamanikumar et al. 2013). Genomic selection has been tested as a theoretic approach in forest trees using simulated datasets; however, many studies show the application of real data (e.g., Resende et al. 2012a; Beaulieu et al. 2014b). Genomic selection has especially been suggested as a useful tool in elite breeding programs where relatively low number of markers are adequate to cover structures of linkage disequilibrium (Thavamanikumar et al. 2013). But the rapid decay of linkage disequilibrium in most tree populations is one of the main problems identified in studies. It was suggested that this limitation could be avoided when using elite trees and thereby introducing a genetic bottleneck. A prediction model built on data from progeny of crosses between elite trees can additionally be used to select elite trees via genomic selection. A study compared estimated breeding values and genomic breeding values using cross-validation within clones from half-sib families of loblolly pine (*Pinus taeda*). Even though derived accuracies were relatively high, this was suggested to be due to family linkage rather than identified historic linkage disequilibrium as only few genetic markers were used (Zapata-Valenzuela et al. 2012). A study in maritime pine (*Pinus pinaster Ait.*) showed good predictive ability for different traits, despite the low marker coverage and low linkage disequilibrium (Isik et al. 2016). A more comprehensive breeding scheme was simulated for a population of conifers (Iwata et al. 2011) for which different scenarios were tested for a 60-year breeding program in a seed orchard. The use of genetic markers in a genomic selection scheme could also provide additional information on parentage since some of the traditional tree breeding programs, for example, in eucalyptus breeding, are open-pollinated (Zelener et al. 2005). Studies have also shown the potential of genomic selection to improve traits in spruce compared to traditional pedigree-based selection (Beaulieu et al. 2014a, b; Ratcliffe et al. 2015; Lenz et al. 2017). The potential of genomic selection over traditional breeding has been shown in recently domesticated or undomesticated populations of trees (e.g., white spruce) but has been suggested for within populations or families due to the low marker coverage (Beaulieu et al. 2014a, b).

Genomic Selection in *Eucalyptus*

Conventional tree breeding is typically characterized by long breeding cycles. Hybrids are often preferred in *Eucalyptus* breeding schemes as they are superior to their parents in the most relevant traits, including growth, wood quality, and biotic and abiotic stress resistance as they inherit relevant characteristics from each of the parents (Tan et al. 2017). The cycle of a conventional breeding scheme in *Eucalyptus* can take between 12 and 18 years;

(continued)

genomic selection does, therefore, offer new opportunities as it might reduce this cycle. However, when selecting superior tree clones in hybrid eucalypt breeding, both additive and nonadditive effects are relevant (Resende et al. 2017). Relatedness of selection and training population can additionally lead to over- or underestimation of the prediction accuracy. It was suggested that a high marker density will be advantageous in such situations (Resende et al. 2017).

Tan et al. (2017) and Resende et al. (2017) used the Illumina Infinium EuCHIP60K, which includes more than 45,000 SNPs to study controlled crossings of *E. urophylla* and *E. grandis* trees. The aim was to test genomic selection for the selection of superior F_2 individuals for traits height, volume, circumference at breast height, basic wood density, and screened pulp yield. Genomic best linear unbiased prediction, ridge regression best linear unbiased prediction, Bayesian LASSO, and reproducing kernel Hilbert space regression were tested in these studies. Predictive abilities of the genomic selection models differed based on the selection scheme, with the highest predictive abilities obtained from cross-validation in a between-family selection including full- and half-sib individuals (Resende et al. 2017). The mean accuracies varied between 0.34 and 0.54 depending on the traits and reached maximums of 0.73 to 0.87 in the best scenario based on relatedness. The predictive ability using different models varied from 0.27 to 0.274, but all models of genomic selection did outperform other pedigree-based predictions. Also this study showed that the relationship between training and selection candidates, as well as the size of the training population, had a large impact on the predictive ability (Tan et al. 2017).

It was concluded from both studies that (a) genomic selection will reduce the time until superior breeding lines are selected and (b) data obtained from genotyping provide additional information on the genomic relationship matrix and can be used for the estimation of heritability. However further issues need to be resolved, such as the selection across generations and environments. The inclusion of nonadditive effects and the estimation in hybrid breeding as purebred parents will not provide information for accurate predictions in hybrid offspring.

7.4.2 Fruit Trees

Traits of interest for breeders of fruit trees are fruit quality (e.g., firmness, astringency, soluble solids, and acidity), precocity, yield, and disease resistance. The selection using traditional methods is difficult as most of these traits are polygenic or complex and controlled by many genes. Information using genetic markers may allow to identify relevant QTL, but methods like MAS are only applicable for traits with a few QTL with major effects, while genomic selection allows the prediction of the total genetic value or phenotype and is thus more applicable for complex traits

(Kumar et al. 2012a). Only a few studies have evaluated the potential of genomic selection in fruit trees, such as apple (*Malus domestica*), grapes (*Vitis vinifera*), or pear (*Pyrus*) (Kumar et al. 2012a; Kumar et al. 2012b; Iwata et al. 2013; Myles 2013). Most of the molecular markers used in apple breeding have focused on resistance traits and applied markers for marker-assisted selection. But such single-gene markers did not provide a method for long-time disease resistance selective breeding because pathogens or pests did develop new strategies to overcome such resistances (Kumar et al. 2012a). Genomic selection is suggested as a possibly better alternative as it incorporates multiple markers and might allow a selection including genes with smaller effects. Two alternative strategies are suggested: the use of genomic selection for parent selection (as in forest trees for the elite parent lines) or for the selection of future cultivars (Kumar et al. 2012a). Preliminary results in an apple and pear tree population have indicated that genomic selection will allow selection prior to expensive phenotypic evaluation and might have the potential to speed up the selection process. However cross-validation within the same generation of trees has been used to derive the accuracies (Kumar et al. 2012a; Iwata et al. 2013). The application of genomic selection in crossbred individuals is relevant in fruit trees. Crossbred scenarios will require the prediction of nonadditive effects. One additional point of consideration is the use of grafts. Full-sib families are commonly used in apple breeding programs, and seedlings are grafted onto clonal rootstocks, a strategy which differs from the cloning used for phenotypic evaluation in forest trees (Kumar et al. 2012a).

If genomic selection in tree breeding can provide similar accuracies as conventional breeding, it will be able to increase genetic gain and reduce sizes and costs for breeding programs significantly. But strategies need to be developed to allow either long-term effects with low decay of accuracy over several generations or options for a cost-efficient regular updating of the prediction model. It has yet to be shown how genomic selection will perform in crossbred situations and across multiple generations, as many of those studies apply their simulation in a single generation only (Grattapaglia and Resende 2011; Kumar et al. 2012a; Zapata-Valenzuela et al. 2012; Iwata et al. 2013).

7.5 *Examples of Genomic Prediction Applied on Human (Homo sapiens) Populations*

Genomic prediction has been suggested as a useful tool in assessing genetic predisposition for human diseases and personalized medicine (de los Campos et al. 2010; Makowsky et al. 2011). However, genomic prediction has not been successfully applied to any great extent in humans yet. Nevertheless, the models for genomic selection have been successful in human studies to estimate the heritability of complex traits (Yang et al. 2010).

The accuracy of genomic predictions in a test population with estimates based on trait values measured in a reference population depends largely on the variance in relatedness between pairs of individuals in the test and reference populations or equivalently the mean linkage disequilibrium over all pairs of loci (Goddard et al. 2011). In humans, linkage disequilibrium is small, and useful genomic prediction would therefore require a very large reference set. Consequently, genomic prediction has not been found to be as useful as in animal and plant populations with larger linkage disequilibrium.

The statistical models developed for genomic selection have been found to be extremely valuable in human genetics for heritability estimation. Separating genetic and environmental effects in humans has been notoriously difficult in the past because human populations generally consist of small families where relatives share many environmental factors. Yang et al. (2010) showed that by combining all SNP information from practically unrelated individuals (i.e., pair-wise genomic correlations between individuals typically smaller than 0.1) in a GBLUP, it is possible to estimate the heritability of complex traits. By using unrelated individuals, any possible confounding of genetic and environmental effects is eliminated.

8 Future Directions and Perspective

Many of the genomic selection research and development efforts focused on improving the accuracy of genomic breeding values, exploring a large range of parametric and nonparametric models for genomic prediction. While the application of genomic selection requires robust machinery for genomic prediction, it is important to realize that the real benefits of genomic prediction can only be harvested when accompanied by changes in the breeding program. Optimizations of breeding strategies that utilize genomic breeding values are thus far underexplored, and much gain can be expected from studies on novel and innovative breeding schemes. Synergies between genomic selection and reproduction techniques and/or genome editing are examples of components of such breeding schemes. Another example of an element to consider in the design of breeding schemes is that strategies for genotyping selection candidates can affect the composition of the future reference population, giving rise to a complex optimization problem if the aim is long-term genetic improvement.

Genotype information can also be used for population management. This relates not only to conservation of populations at risk but also for the maintenance of genetic variability in commercial populations. Genomic selection was believed to have a positive impact on rates of inbreeding, but the first indications of experience from the field report increased rates of inbreeding in genomic breeding schemes. However, there remains much scope for development of genomic tools that consider both genetic progress and maintenance of genetic diversity.

There is much potential to utilize genomic information for prediction of phenotypes of animals, plants, and trees, in order to tailor management, similar to utilizing genomic prediction for personalized medicine discussed in the context of human

genetics. For example, mating schemes can be optimized using genomic information to avoid inbreeding or to capitalize on hybrid vigor and other nonadditive genetic effects. Moreover, knowledge about the predisposition to certain diseases can be used to direct preventive measures to individuals with elevated risk.

In summary, genomic information can in the first place be used to enhance genetic gains and offers also opportunities for improved management, at various levels.

9 Conclusions

Since the first suggestion of genomic selection and prediction in 2001, the development of genotyping methods has allowed the introduction of this advanced selection tool across many populations. Breeders are hoping for an easier and more accurate selection tool, which allows an earlier selection of advanced lines or individuals. Early estimations based on information from dairy populations revealed that the application of genomic selection should increase the rate of genetic gain and that genomic selection has the potential to revolutionize animal breeding (Schaeffer 2006; Hayes et al. 2009a; Thornton 2010; Goddard 2012). Similar improvements have also been predicted for plant breeding. It has been shown in studies using empirical and simulated data that the use of genetic markers will accelerate breeding and reduce the generation interval/time for the development of new varieties (Rudi et al. 2010). Genomic selection in combination with high-throughput phenotyping might revolutionize the selection for complex traits (Cabrera-Bosquet et al. 2012). In Holstein Frisian dairy cattle, the implementation of SNP information was predicted to provide as much information as real data from phenotypes from 10 to 20 daughters per bull (Jannink et al. 2010). Available SNP information would thereby allow to collect phenotypic records from fewer offspring with no loss of accuracy. However, statistical models for different breeding scenarios have to be developed (Heslot et al. 2012). Inclusion of nonadditive effects, such as heterosis or genotype by environment interactions, will be relevant for some traits and in some populations. Improved phenotyping has to be established as the accuracy and throughput of phenotype measurements are currently the main limiting factors (Lorenzana and Bernardo 2009).

There is little doubt that genomic selection is a success in the main dairy cattle breed, Holstein Friesian. Genomic selection is also practiced in other dairy cattle breeds, but not as successful in terms of accuracy of selection as in the Holstein breed, and it remains unclear if the successes can be repeated in other species. Further advancements in technology are needed in situations with complex population compositions and genome structure. Massive sequencing at low coverage (genomic selection 2.0) and better use of biological knowledge as priors in genomic prediction are promising directions of future developments. Good knowledge on the functionality of mutations is imperative, to be able to target the right QTN in selection and avoid unwanted side effects.

The statistical models used for genomic selection in livestock have been proven useful to estimate heritabilities in human genetic studies. Genomic prediction has also been suggested as a tool to predict genetic predisposition of human health disorders, even though not many success stories are documented to date. Similar to prediction phenotypic in humans, genomic prediction has potential to be useful for management purposes at agricultural farms to optimize production processes. Sequencing data is currently used in breeding populations, but reliability of the data and the information acquired from the data have to be questioned: how complex can data be in order to be implemented in prediction models and how much background do we need on the inheritance of genome structures different from polymorphisms. There is little doubt that the inclusion of more information on genotypes will improve predictions. Whether the inclusion of information from molecular genetic markers will be advantageous to other phenotypic and environmental measures is probably a question of costs, rather than results.

The current advances of the methods, some of which introduced here, need further discussion. Methods and models will need to be tested from case to case, and different models might be needed for different traits. Much of the benefits from genomic selection arise from the possibility to determine the outfall of Mendelian sampling as soon as a DNA sample can be taken. The phenotypes can therefore be predicted with higher accuracy as exact genotypes are already known. It thus seems pertinent to determine the accuracy of Mendelian sampling deviations calculated from genomic breeding values, and to consider that statistics in the comparison of models and methods, apart from some exceptions (e.g., Rius-Vilarrasa et al. 2012), this is rarely done.

The validation of prediction models needs careful consideration. Accuracies based on cross-validation might not reflect accuracies of selection achievable in breeding schemes applied in practice. Many of the current selection schemes in plant breeding are based on phenotypes recorded during the first steps of selection, which may be different from those for the final breeding goal. The correlations to final breeding goal might therefore be low. Application of genetic markers will allow a better prediction of early selection. However, accuracies should be calculated based on models applicable to real breeding populations.

Despite the current pitfalls, the concept of genomic selection has led to a number of advances driven by the need for improved selection in plant and livestock populations. It has contributed to the fast application of genotyping and sequencing tools in nonhuman populations. It has also opened new opportunities and advanced options for methods for prediction models. Phenotyping has been put in the spotlight again, as reliable phenotypes are required for accurate predictions. The options of a better use of phenotypes have led to an extension of measurements and inclusion of complex traits, especially such related to health/welfare and sustainability, into selection schemes. While such progress is not solely based on the development of genomic selection, the new opportunity for the use of genome-wide marker sets for the prediction in populations has assisted such new opportunities.

References

http://www.interbull.org. Interbull. 2013. Accessed 31.01.2013.
Albrecht T, et al. Genome-based prediction of testcross values in maize. Theor Appl Genet. 2011;123:339–50.
Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink JL. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. Plant Genome. 2011;4:132–44.
Avendaño S, Watson KA, Kranis A. Genomics in poultry breeding—from utopias to deliverables. In: 9th world congress on genetics applied to livestock production (WCGALP). Germany: Leipzig; 2010
Baloche G, et al. Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. J Dairy Sci. 2014;97:1107–16.
Barrell PJ, Meiyalaghan S, Jacobs JME, Conner AJ. Applications of biotechnology and genomics in potato improvement. Plant Biotechnol J. 2013;11:907–20.
Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J. Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). Plant Sci. 2016;242:23–36.
Battenfield SD, Guzmán C, Gaynor RC, Singh RP, Peña RJ, Dreisigacker S, Fritz AK, Poland JA. Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. Plant Genome. 2016;9(2). https://doi.org/10.3835/plantgenome2016.01.0005.
Bauer E, et al. Towards a whole-genome sequence for rye (Secale cereale L.). Plant J. 2017;89:853–69.
Beaulieu J, Doerksen T, Clement S, MacKay J, Bousquet J. Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity. 2014a;113:343–52.
Beaulieu J, Doerksen TK, MacKay J, Rainville A, Bousquet J. Genomic selection accuracies within and between environments and small breeding groups in white spruce. BMC Genomics. 2014b;15:1048.
Bernardo R. Genomewide selection with minimal crossing in self-pollinated crops. Crop Sci. 2010;50:624–7.
Bernardo R, Yu J. Marker-assisted selection without QTL mapping: prospects for genome-wide selection for quantitative traits in maize. Maize Genet Cooperat Newslett 2007:26.
Berry DP, Garcia JF, Garrick DJ. Development and implementation of genomic predictions in beef cattle. Anim Front. 2016;6:32–8.
Bertin N, Martre P, Génard M, Quilot B, Salon C. Under what circumstances can process-based simulation models link genotype to phenotype for complex traits? Case-study of fruit and grain quality traits. J Exp Bot. 2010;61:955–67.
Birol I, et al. Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics. 2013;29:1492–7.
Boichard D, Ducrocq V, Croiseau P, Fritz S. Genomic selection in domestic animals: principles, applications and perspectives. C R Biol. 2016;339:274–7.
Bouquet A, Juga J. Integrating genomic selection into dairy cattle breeding programmes: a review. Animal. 2013;7:705–13.
Brenchley R, et al. Analysis of the bread wheat genome using whole genome shotgun sequencing. Nature. 2012;491:705–10.
Breseghello F. Traditional and modern plant breeding methods with examples in rice (Oryza sativa L.). J Agric Food Chem. 2013;61:8277–86.
Burgueno J, de los Campos G, Weigel K, Crossa J. Genomic prediction of breeding values when modeling genotype x environment interaction using pedigree and dense molecular markers. Crop Sci. 2012;52:707–19.
Cabrera-Bosquet L, Crossa J, von Zitzewitz J, Serret MD, Araus JL. High-throughput phenotyping and genomic selection: the frontiers of crop breeding converge. J Integr Plant Biol. 2012;54(5):312–20.
Calus MPL. Editorial: genomic selection with numerically small reference populations. Animal. 2016;10:1016–7.

de los Campos G, et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics. 2009;182:375–85.
de los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet. 2010;11:880–6.
de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013a;193:327–45.
de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 2013b;9:e1003608.
Caspi R, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 2014;42:D459–71.
Cericola F, et al. Optimizing training population size and genotyping strategy for genomic prediction using association study results and pedigree information. A case of study in advanced wheat breeding lines. PLoS One. 2017;12:e0169606.
Chagné D, et al. The draft genome sequence of European Pear (Pyrus communis L. "Bartlett"). PLoS One. 2014;9:e92644.
Chao S, Zhang W, Dubcovsky J, Sorrells M. Evaluation of genetic diversity and genome-wide linkage disequilibrium among US wheat (Triticum aestivum L.) germplasm representing different market classes. Crop Sci. 2007;47:1018–30.
Chao SM, et al. Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (Triticum aestivum L.). BMC Genomics. 2010;11:727.
Chawade A, Alexandersson E, Bengtsson T, Andreasson E, Levander F. Targeted proteomics approach for precision plant breeding. J Proteome Res. 2016;15:638–46.
Chen HD, He H, Zhou FS, Yu HH, Deng XW. Development of genomics-based genotyping platforms and their applications in rice breeding. Curr Opin Plant Biol. 2013;16:247–54.
Cowling WA, Balazs E. Prospects and challenges for genome-wide association and genomic selection in oilseed Brassica species. Genome. 2010;53:1024–8.
Cowling WA, Buirchell BJ, Falk DE. A model for incorporating novel alleles from the primary gene pool into elite crop breeding programs while reselecting major genes for domestication or adaptation. Crop Pasture Sci. 2009;60:1009–15.
Croft D, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011;39:D691–7.
Crossa J, et al. Genomic selection and prediction in plant breeding. J Crop Improv. 2011;25:239–61.
Crossa J, et al. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity. 2013;112:48–60.
Cullis BR, Smith AB, Beeck CP, Cowling WA. Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. Genome. 2010;53:1002–16.
Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One. 2008;3:e3395.
Daetwyler HD, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–65.
Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011;12(7):499–510. https://doi.org/10.1038/nrg3012.
Dawson JC, et al. The use of unbalanced historical data for genomic selection in an international wheat breeding program. Field Crop Res. 2013;154:12–22.
Dekkers JCM. Marker-assisted selection for commercial crossbred performance. J Anim Sci. 2007;85:2104–14.
Dekkers JCM, Hospital F. The use of molecular genetics in the improvement of agricultural populations. Nat Rev Genet. 2002;3:22–32.
Druet T, Macleod IM, Hayes BJ. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity. 2014;112:39–47.

Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. 2010. https://doi.org/10.1093/nar/gkx382.
Dürr J, Philipsson J. International cooperation: the pathway for cattle genomics. Anim Front. 2012;2:16–21.
Duvick DN. Heterosis: feeding. People and protecting natural resources. In: Coors JG, Pandey S, editors. The genetics and exploitation of heterosis in crops. Madison, WI: American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc.; 1999. p. 19–29.
Elsik CG, Tellam RL, Worley KC. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science. 2009;324:522–8.
Erbe M, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95:4114–29.
Falconer D, Mackay T. Quantitative genetics. London, UK: Longman, Harrow; 1996.
Fernie AR, Schauer N. Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet. 2009;25:39–48.
Forabosco F, Lohmus M, Rydhmer L, Sundstrom LF. Genetically modified farm animals and fish in agriculture: a review. Livest Sci. 2013;153:1–9.
Funk DA. Major advances in globalization and consolidation of the artificial insemination industry. J Dairy Sci. 2006;89:1362–8.
Gaj T, Gersbach CA, Barbas CF. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. 2013;31(7):397–405.
Ganal MW, Röder MS. Microsatellite and SNP markers in wheat breeding. In: Varshney RK, Tuberosa R, editors. Genomic assisted crop improvement: genomics applications in crops, vol. 2. Dordrecht: Springer; 2007. p. 1–24.
García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, Van Tassell CP. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc Natl Acad Sci U S A. 2016;113(28):E3995–4004. https://doi.org/10.1073/pnas.1519061113.
Garrick DJ. The nature, scope and impact of genomic prediction in beef cattle in the United States. Genet Sel Evol. 2011;43:–17.
Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol. 2009;41:55.
Gerrits RJ, et al. Perspectives for artificial insemination and genomics to improve global swine populations. Theriogenology. 2005;63:283–99.
Gianola D. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics. 2013;194:573–96.
Gianola D, van Kaam JBCHM. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics. 2008;178:2289–303.
Goddard ME. Uses of genomics in livestock agriculture. Animal Production Science. 2012;52:73–7.
Goddard ME, Hayes BJ. Genomic selection. J Anim Breed Genet. 2007;124:323–30.
Goddard ME, Hayes BJ, Meuwissen THE. Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet. 2011;128:409–21.
Goff SA, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science. 2002;296:92–100.
Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.
Gouy M, et al. Experimental assessment of the accuracy of genomic selection in sugarcane. Theor Appl Genet. 2013;126:2575–86.
Grattapaglia D, Resende MDV. Genomic selection in forest tree breeding. Tree Genet Genomes. 2011;7:241–55.
van Grevenhof I. Breeding against osteochondrosis. Wageningen: Wageningen University; 2011.
Groenen MAM, et al. Analyses of pig genomes provide insight into porcine demography and evolution. Nature. 2012;491:393–8.

Gupta PK, Langridge P, Mir RR. Marker-assisted wheat breeding: present status and future possibilities. Mol Breed. 2010;26:145–61.
Guzman C, et al. Wheat quality improvement at CIMMYT and the use of genomic selection on it. Appl Transl Genom. 2016;11:3–8.
Haberland AM, König von Borstel U, Simianer H, König S. Integration of genomic information into sport horse breeding programs for optimization of accuracy of selection. Animal. 2012;6:1369–76.
Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177:2389–97.
Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol. 2010;42:–5.
Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;12:186.
Habier D, Fernando RL, Garrick DJ. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics. 2013;194:597–607.
Hayes B, Bowman P, Chamberlain A, Goddard M. Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2009a;92:433–43.
Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol. 2009b;41:–51.
Hayes BJ, et al. Prospects for genomic selection in forage plant species. Plant Breeding. 2013;132:133–43.
Heffner EL, Sorrells ME, Jannink J-L. Genomic selection for crop improvement. Crop Sci. 2009;49:1–12.
Heffner EL, Jannink JL, Sorrells ME. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome. 2011;4:65–75.
Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometrics. 1975a;31:423–47.
Henderson CR. Use of all relstives in intraherd prediction of breeding values and producing abilities. J Dairy Sci. 1975b;58:1910–6.
Heslot N, Yang H-P, Sorrells ME, Jannink J-L. Genomic selection in plant breeding: a comparison of models. Crop Sci. 2012;52:146–60.
Hickey JM. Sequencing millions of animals for genomic selection 2.0. J Anim Breed Genet. 2013;130:331–2.
Hickey JM, et al. Sequencing millions of animals for genomic selection 2.0. In: Proceedings, 10th world congress of genetics applied to livestock production. Vancouver; 2014.
de Roos APW, Hayes BJ, Goddard ME. Reliability of genomic predictions across multiple populations. Genetics. 2009;183:1545–53.
Hofheinz N, Borchardt D, Weissleder K, Frisch M. Genome-based prediction of test cross performance in two subsequent breeding cycles. Theor Appl Genet. 2012;125:1639–45.
Hutchison JL, Cole JB, Bickhart DM. Short communication: use of young bulls in the United States. J Dairy Sci. 2014;97:3213–120. https://doi.org/10.3168/jds.2013-7525.
International Barley Genome Sequencing Consortium et al. A physical, genetic and functional sequence assembly of the barley genome. Nature. 2012;491:711–6.
Isik F, et al. Genomic selection in maritime pine. Plant Sci. 2016;242:108–19.
Iwata H, et al. Potential assessment of genome-wide association study and genomic selection in Japanese pear Pyrus pyrifolia. Breed Sci. 2013;63:125–40.
Iwata H, Hayashi T, Tsumura Y. Prospects for genomic selection in conifer breeding: a simulation study of Cryptomeria japonica. Tree Genet Genomes. 2011;7:747–58.
Iwata H, Jannink J-L. Accuracy of genomic selection prediction in barley breeding programs: a simulation study based on the real single nucleotide polymorphism data of barley breeding lines. Crop Sci. 2011;51:1915–27.
Jannink J-L, Lorenz AJ, Iwata H. Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics. 2010;9:166–77.
Jannink JL. Dynamics of long-term genomic selection. Genet Sel Evol. 2010;42:–35.
Jenko J, et al. Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs. Genet Sel Evol. 2015;47:55.

Jiang Y, et al. The sheep genome illuminates biology of the rumen and lipid metabolism. Science. 2014;344:1168–73.
Johnston DJ, Tier B, Graser HU. Beef cattle breeding in Australia with genomics: opportunities and needs. Animal Production Science. 2012;52:100–6.
Kanehisa M, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008;36:D480–4.
Kärkkäinen HP, Sillanpää MJ. Fast genomic predictions via Bayesian G-BLUP and Multilocus models of threshold traits including censored Gaussian data. G3 (Bethesda). 2013;3:1511–23.
Kizilkaya K, Fernando RL, Garrick DJ. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J Anim Sci. 2010;88:544–51.
Kumar S, Bink MCAM, Volz RK, Bus VGM, Chagne D. Towards genomic selection in apple (Malus x domestica Borkh.) breeding programmes: prospects, challenges and strategies. Tree Genet Genomes. 2012a;8:1–14.
Kumar S, et al. Genomic selection for fruit quality traits in apple (Malus x domestica Borkh.). PLoS One. 2012b;7(5):e36674.
Lenz PRN, et al. Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (Picea mariana). BMC Genomics. 2017;18:335.
Li Z-K, Zhang F. Rice breeding in the post-genomics era: from concept to practice. Curr Opin Plant Biol. 2013;16:261–9.
Lien S, et al. The Atlantic salmon genome provides insights into rediploidization. Nature. 2016;533:200–5.
Lillehammer M, Meuwissen THE, Sonesson AK. Genomic selection for maternal traits in pigs. J Anim Sci. 2011;89:3908–16.
Lillehammer M, Meuwissen THE, Sonesson AK. A low-marker density implementation of genomic selection in aquaculture using within-family genomic breeding values. Genet Sel Evol. 2013;45
Lin Z, Hayes BJ, Daetwyler HD. Genomic selection in crops, trees and forages: a review. Crop Pasture Sci. 2014;65:1177–91.
Lindblad-Toh K, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005;438:803–19.
Lôbo RB, et al. Implementation of DNA markers to produce genomically—enhanced EPDs in Nellore cattle. Acta Sci Vet. 2011;39(Suppl 1):s23–7.
Longin CFH, Mi X, Würschum T. Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. Theor Appl Genet. 2015;128:1297–306.
Lorenzana RE, Bernardo R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet. 2009;120:151–61.
Lund MS, et al. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet Sel Evol. 2011;43:43.
Lush JL. Linebreeding. Iowa Agric Exp Sta Bull 1933:301.
Lush JL. Family merit and individual merit as bases for selection. Am Nat. 1947;81:241–61.
Lynch M, Walsh B. Genetics and analysis of quantitative traits. Sunderland, MA: Sinauer; 1998.
Maccaferri M, Sanguineti MC, Noli E, Tuberosa R. Population structure and long-range linkage disequilibrium in a durum wheat elite collection. Mol Breed. 2005;15:271–89.
MacLeod IM, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016;17:144.
Makowsky R, et al. Beyond missing heritability: prediction of complex traits. PLoS Genet. 2011;7:e1002051.
Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11:499–511.
Marulanda JJ, et al. Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale. Theor Appl Genet. 2016;129:1901–13.

Mascher M, et al. A chromosome conformation capture ordered sequence of the barley genome. Nature. 2017;544:427–33.
Mather KA, et al. The extent of linkage disequilibrium in rice (Oryza sativa L.). Genetics. 2007;177:2223–32.
Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS, Van Tassell CP. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One. 2009;4:e5350. https://doi.org/10.1371/journal.pone.0005350.
McCouch S. Diversifying selection in plant breeding. PLoS Biol. 2004;2:1507–12.
Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
Meuwissen THE, Luan T, Woolliams JA. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J Anim Breed Genet. 2011;128:429–39.
Morandini P, Salamini F. Plant biotechnology and breeding: allied for years to come. Trends Plant Sci. 2003;8:70–5.
Morota G, Gianola D. Kernel-based whole-genome prediction of complex traits: a review. Front Genet. 2014;5:363.
Morrell PL, Buckler ES, Ross-Ibarra J. Crop genomics: advances and applications. Nat Rev Genet. 2012;13:85–96.
Muir B, Van Doormaal BJ, Kistemaker G. International genomic co-operation—North American perspective. In: Proceedings of the Interbull international workshop, Paris, France; 2010. pp 71–76.
Myburg AA, et al. The genome of Eucalyptus grandis. Nature. 2014;510:356–62.
Myles S. Improving fruit and wine: what does genomics have to offer? Trends Genet. 2013;29:190–6.
Neale DB, et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 2014;15:R59.
Nielsen HM, Sonesson AK, Meuwissen THE. Optimum contribution selection using traditional best linear unbiased prediction and genomic breeding values in aquaculture breeding schemes. J Anim Sci. 2011;89:630–8.
Northcutt SL. Genomic choices. American Angus Association®/AngusGenetics Inc. release. 2011. http://www.angus.org/AGI/GenomicChoice070811.pdf (posted July, 2011)
Nystedt B, et al. The Norway spruce genome sequence and conifer genome evolution. Nature. 2013;497:579–84.
Ober U, et al. Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. Genetics. 2011;188:695–708.
Patry C. Impacts of genomic selection on classical genetic evaluations. Jouy-en-Josas: Institut National de la Recherche Agronomique (INRA); 2011.
Potato Genome Sequencing Consortium, et al. Genome sequence and analysis of the tuber crop potato. Nature. 2011;475:189–95.
Preisinger R. Genome-wide selection in poultry. Animal Production Science. 2012;52:121–5.
Proudfoot C, et al. Genome edited sheep and cattle. Transgenic Res. 2015;24:147–53.
Pryce JE, Daetwyler HD. Designing dairy cattle breeding schemes under genomic selection: a review of international research. Anim Prod Sci. 2012;52:107–14.
Pszczola M, Calus MPL. Updating the reference population to achieve constant genomic prediction reliability across generations. Animal. 2015;10:1018–24.
Pszczola M, Strabel T, van Arendonk JAM, Calus MPL. The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection. J Dairy Sci. 2012;95:5412–21.
Ratcliffe B, et al. A comparison of genomic selection models across time in interior spruce (Picea engelmannii × glauca) using unordered SNP imputation methods. Heredity (Edinb). 2015;115(6):547–55.
Remington DL, et al. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci U S A. 2001;98:11479–84.
Resende M, et al. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytol. 2012a;193

Resende MDV, et al. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol. 2012b;194:116–28.
Resende RMS, Casler MD, Resende MDV. Genomic selection in forage breeding: accuracy and methods. Crop Sci. 2014;54:143–56.
Resende RT, Resende MDV, Silva FF, Azevedo CF, Takahashi EK, Silva-Junior OB, Grattapaglia D. Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. Heredity (Original Article). 2017. https://doi.org/10.1038/hdy.2017.37.
Rius-Vilarrasa E, et al. Influence of model specifications on the reliabilities of genomic prediction in a Swedish–Finnish red breed cattle population. J Anim Breed Genet. 2012;129:369–79.
Rosenberg NA, Nordborg M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet. 2002;3:380–90.
Rubin C-J, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 2010;464:587–91.
Rudi N, Norton GW, Alwang J, Asumugha G. Economic impact analysis of maker-assisted breeding for resistance to pests and post harvest deterioration of cassava. Afr J Agr Res Econ. 2010;4:110–22.
Rutkoski JE, Heffner EL, Sorrells ME. Genomic selection for durable stem rust resistance in wheat. Euphytica. 2011;179:161–73.
Saatchi M, Schnabel RD, Rolf MM, Taylor JF, Garrick DJ. Accuracy of direct genomic breeding values for nationally evaluated traits in US Limousin and Simmental beef cattle. Genet Sel Evol. 2012;44:38.
Sánchez-Molano E, et al. Genomic prediction of traits related to canine hip dysplasia. Front Genet. 2015;6:97.
Schaeffer L. Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet. 2006;123:218–23.
Schnable PS, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326:1112–5.
Sharma HC, Crouch JH, Sharma KK, Seetharama N, Hash CT. Applications of biotechnology for crop improvement: prospects and constraints. Plant Sci. 2002;163:381–95.
Shen X, Alam M, Fikse F, Rönnegård L. A novel generalized ridge regression method for quantitative genetics. Genetics. 2013;193(4):1255–68.
Shu YJ, et al. Genomic selection of seed weight based on low-density SCAR markers in soybean. Genet Mol Res. 2013;12:2178–88.
Shumbusho F, Raoul J, Astruc JM, Palhiere I, Elsen JM. Potential benefits of genomic selection on genetic gain of small ruminant breeding programs1. J Anim Sci. 2013;91:3644–57.
Snelling WM, et al. Partial-genome evaluation of postweaning feed intake and efficiency of crossbred beef cattle12. J Anim Sci. 2011;89:1731–41.
Somers DJ, Kirkpatrick R, Moniwa M, Walsh A. Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome. 2003;46:431–7.
Sonesson AK, Meuwissen THE. Testing strategies for genomic selection in aquaculture breeding programs. Genet Sel Evol. 2009;41:37.
Spindel J, et al. Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theor Appl Genet. 2013;126:2699–716.
Stein L. Genome annotation: from sequence to biology. Nat Rev Genet. 2001;2:493–503.
Stock KF, Jönsson L, Ricard A, Mark T. Genomic applications in horse breeding. Anim Front. 2016;6:45–52.
Sun XC, Fernando R, Dekkers J. Contributions of linkage disequilibrium and co-segregation information to the accuracy of genomic prediction. Genet Sel Evol. 2016;48(1):77.
Swan AA, Johnston DJ, Brown DJ, Tier B, Graser H-U. Integration of genomic information into beef cattle and sheep genetic evaluations in Australia. Animal Production Science. 2012;52:126–32.
Sweeney M, McCouch S. The complex history of the domestication of rice. Ann Bot. 2007;100:951–7.

Takeda S, Matsuoka M. Genetic approaches to crop improvement: responding to environmental and population changes. Nat Rev Genet. 2008;9:444–57.
Tan B, Grattapaglia D, Martins GS, Ferreira KZ, Sundberg B, Ingvarsson PK. Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. BMC Plant Biol. 2017;17(1):110. https://doi.org/10.1186/s12870-017-1059-6.
Tan W, Proudfoot C, Lillico SG, Whitelaw CBA. Gene targeting, genome editing: from Dolly to editors. Transgenic Res. 2016;25:273–87.
Tang C, et al. The rubber tree genome reveals new insights into rubber production and species adaptation. Nat Plants. 2016;2:16073.
Tenaillon MI, Austerlitz F, Tenaillon O. Apparent mutational hotspots and long distance linkage disequilibrium resulting from a bottleneck. J Evol Biol. 2008;21:541–50.
Thavamanikumar S, Southerton SG, Bossinger G, Thumma BR. Dissection of complex traits in forest trees—opportunities for marker-assisted selection. Tree Genet Genomes. 2013;9:627–39.
Thornton PK. Livestock production: recent trends, future prospects. Phil Trans Roy Soc B-Biol Sci. 2010;365:2853–67.
Thorwarth P, et al. Genomic prediction ability for yield-related traits in German winter barley elite material. Theor Appl Genet. 2017;130(8):1669–83.
Toro MA, Varona L. A note on mate allocation for dominance handling in genomic selection. Genet Sel Evol. 2010;42:33.
Trebbi D, et al. High-throughput SNP discovery and genotyping in durum wheat (Triticum durum Desf.). Theor Appl Genet. 2011;123:555–69.
Tribout T, Larzul C, Phocas F. Efficiency of genomic selection in a purebred pig male line. J Anim Sci. 2012;90:4164–76.
Tuskan GA, et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science. 2006;313:1596–604.
van der Werf JHJ. Marker-assisted selection in sheep and goats. In: Guimarães EP, Ruane J, Scherf BD, Sonnino A, Dargie JD, editors. Marker-assisted selection: current status and future perspectives in crops, livestock, forestry and fish. Rome: Food and Agriculture Organization of the United Nations; 2007.
VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
VanRaden PM, Wiggans GR. Derivation, calculation, and use of national animal model information. J Dairy Sci. 1991;74:2737–46.
VanRaden PM, Wiggans GR, Van Tassell CP, Sonstegard TS, Schenkel F. Benefits from cooperation in genomics. In: Proceedings of the Interbull international workshop. Genomic information in genetic evaluations. Uppsala, Sweden; 2009a pp 67–72.
VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS. Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci. 2009b;92:16–24. https://doi.org/10.3168/jds.2008-1514.
Velasco R, et al. The genome of the domesticated apple (Malus [times] domestica Borkh.). Nat Genet. 2010;42:833–9.
Wade CM, et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science. 2009;326:865–7.
Walsh B. Quantitative genetics. In: eLS. John Wiley & Sons Ltd., Chichester. 2001
Wiggans GR, VanRaden PM, Cooper TA. The genomic evaluation system in the United States: past, present, future. J Dairy Sci. 2011;94:3202–11.
Wilkins PW, Humphreys MO. Progress in breeding perennial forage grasses for temperate agriculture. J Agric Sci. 2003;140:129–50.
Williams AV, Nevill PG, Krauss SL. Next generation restoration genetics: applications and opportunities. Trends Plant Sci. 2014;19:529–37.
Windhausen VS, et al. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 (Bethesda). 2012;2:1427–36.
Wu J, et al. The genome of the pear (Pyrus bretschneideri Rehd.). Genome Res. 2013;23:396–408.

Wurschum T, Abel S, Zhao Y. Potential of genomic selection in rapeseed (Brassica napus L.) breeding. Plant Breeding. 2014;133:45–51.
Wurschum T, Reif J, Kraft T, Janssen G, Zhao Y. Genomic selection in sugar beet breeding populations. BMC Genet. 2013;14:85.
Xu S, Hu Z. Methods of plant breeding in the genome era. Genet Res. 2010;92:423–41.
Yan ZB, Yan WG, Deren CW, McClung A. Hybrid rice breeding. B.R. Wells Rice Research Series—Arkansas Agricultural Experiment Station University of Arkansas, vol 591. 2011. pp 61–63.
Yang J, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42:565–9.
Young CW, Bonczek RR, Johnson DG. Inbreeding of and relationship among registered Holsteins. J Dairy Sci. 1988;71:1659–66.
Yu J, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science. 2002;296:79–92.
Zamir D. Improving plant breeding with exotic genetic libraries. Nat Rev Genet. 2001;2:983–9.
Zapata-Valenzuela J, et al. SNP markers trace familial linkages in a cloned population of Pinus taeda-prospects for genomic selection. Tree Genet Genomes. 2012;8:1307–18.
Zelener N, Poltri SNM, Bartoloni N, Lopez CR, Hopp HE. Selection strategy for a seedling seed orchard design based on trait selection index and genomic analysis by molecular markers: a case study for Eucalyptus dunnii. Tree Physiol. 2005;25:1457–67.
Zhao F, Xu S. An expectation and maximization algorithm for estimating G x E interaction effects. Theor Appl Genet. 2012;124:1375–87.
Zhao Y, et al. Accuracy of genomic selection in European maize elite breeding populations. Theor Appl Genet. 2012;124:769–76.
Zhong SQ, Dekkers JCM, Fernando RL, Jannink JL. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics. 2009;182:355–64.
Zimin A, et al. Sequencing and assembly of the 22-Gb loblolly pine genome. Genetics. 2014;196:875–90.

Part IV
Population, Evolutionary and Ecological Genetics Applications and Inferences

Population Genomics Provides Key Insights in Ecology and Evolution

Paul A. Hohenlohe, Brian K. Hand, Kimberly R. Andrews, and Gordon Luikart

Abstract Population genomic tools have revolutionized many aspects of biology, as detailed throughout the chapters of this volume. In particular, population genomics has provided key insights into ecological and evolutionary processes in natural and managed populations. These studies address a wide range of questions, including demography, phylogeny, genetics of ecologically relevant traits, and adaptation. They have also facilitated the conservation and management of biodiversity and harvested populations. Rather than exhaustively document the applications of population genomics in ecology and evolution, in this chapter we provide perspectives on a few key issues confronting researchers seeking to use population genomic tools in non-model systems. A wide variety of molecular and computational genomic approaches are available and have been used in ecological and evolutionary studies. There is no single best approach; rather, the genomic approach used should be tailored to best address the particular study goals and guided by the biology of the system. A large number of trade-offs, costs, and benefits distinguish genomic approaches, which we discuss below. To illustrate these issues, we focus on several published case studies and assess how the research questions were addressed.

Keywords Genetics of adaptation · Inbreeding · Next-generation sequencing · Phylogenomics · Population genetic structure · Population genomics

P. A. Hohenlohe (✉)
Department of Biological Sciences, Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID, USA
e-mail: hohenlohe@uidaho.edu

B. K. Hand · G. Luikart
Flathead Lake Biological Station, Conservation Genomics Group, Division of Biological Sciences, University of Montana, Polson, MT, USA

K. R. Andrews
Department of Fish and Wildlife Sciences, University of Idaho, Moscow, ID, USA

Genetics and Genomics Group, NOAA Pacific Marine Environmental Lab, University of Washington JISAO, Seattle, WA, USA

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_20,

1 Introduction

1.1 Defining Population Genomics in Ecology and Evolution

Population genomic approaches are applied to a wide and growing range of questions in ecology and evolution (Table 1). Some of these questions are long-standing subjects of traditional population genetic studies, but genomic tools provide greatly improved accuracy or the ability to use far fewer sampled individuals. Other questions in ecology and evolution, particularly those that involve identifying specific loci with functional importance, are newly accessible with genomic approaches. The experimental design, molecular techniques, and analytical tools used also vary widely, and a major challenge of applying genomic tools in ecology and evolution is choosing among all of these options. We discuss these considerations in detail below, highlighting a number of published studies that provide illustrative examples of population genomics in ecology and evolution.

There are multiple ways to define the term "genomics" and to distinguish population genomics from population genetics. Traditional population genetics has a long and rich history over the past century, and much of the classical theory of population genetics (e.g., Fisher 1958; Wright 1978) was developed before there was

Table 1 Examples of research issues in ecology and evolution that are addressed with population genomic approaches

Issue in ecology and evolution	Analytical methods and metrics
Broad-sense genomics	
Estimation of genetic diversity	Heterozygosity, allelic diversity, nucleotide diversity
Effective population size	Linkage disequilibrium (LD), two-sample methods
Population structure, admixture	Bayesian clustering, principal component analysis (PCA)
Source population assignment	Clustering methods
Inbreeding	Identity-by-descent methods
Narrow-sense genomics	
Mapping phenotypic traits	Genome-wide association studies (GWAS)
Fine-scale demographic history	Coalescent, diffusion approximation methods
Fine-scale estimates of current historic hybridization	Phylogenetic, haplotype-based methods
Loci for local adaptation	Outlier methods, genotype-environment association (GEA), multilocus covariance
Loci for inbreeding depression	GWAS
Loci for adaptive introgression	Outlier, cline analysis
Defining population units on local adaptation	Outlier, GEA

These are split into "broad-sense" and "narrow-sense" genomic studies (see text for definitions of these terms). Also shown are some of the classes of analytical approaches used to address each issue, illustrated by examples given in the text. For all of these questions, many different genomic approaches may be used, from reduced representation to whole-genome sequencing

a large body of empirical data against which to test it. Molecular population genetic studies in natural populations began in the 1970s with allozyme methods (e.g., Lewontin 1974), and early empirical discoveries led to fundamental changes in our understanding of the forms and amount of genetic variation present in natural populations and the evolutionary forces influencing it (Kimura 1983). From there, the development of new techniques continued to spur the field forward (Allendorf 2017). The advent of PCR (Mullis and Faloona 1987) and Sanger sequencing (Sanger et al. 1977) opened the way to investigating DNA sequence variation at specific loci to address ecological and evolutionary questions, notably the field of molecular systematics (Moritz and Hillis 1996).

A number of other genetic marker types have been developed in recent decades, such as short regions of mitochondrial DNA (Avise 1994) and microsatellites (Selkoe and Toonen 2006), which have become widely used and facilitated studies of genetic variation within natural populations in a wide range of organisms. However, these techniques are limited to a relatively small number of loci, and most require some prior identification of loci, for example, in order to develop PCR primers. In most cases, such as microsatellites, these genetic markers are assumed to represent a random sample of genetic variation across the genome, and are often assumed to reflect neutral evolutionary forces that affect genomes as a whole, such as demography or population structure. Traditional genetic markers like microsatellites have been used to identify functionally important loci (e.g., quantitative trait loci [QTL] in studies of laboratory crosses; Cresko et al. 2004). Nonetheless, because of their sparse distribution across the genome, these loci have had limited utility for addressing a core issue in ecology and evolution: the genetics of adaptation in natural populations.

The current revolution in genomics has been driven by next-generation sequencing technologies that allow heterogeneous pools of DNA fragments – i.e., pools of DNA fragments that differ in sequence and come from multiple locations across the genome – to be sequenced in parallel and in very large numbers (Mardis 2008). This changes the scaling relationship between the number of markers and the workload required for data generation. So, for example, increasing the number of microsatellite or Sanger-sequenced loci in a traditional study may require a concomitant increase in the number of primers to be validated or the number of PCRs to be conducted; in contrast, with next-generation sequencing, large increases in marker number can be achieved simply by adjusting the protocol or increasing the total amount of sequencing (see discussion of these trade-offs below). A simple definition of the term "population genomics" could rely solely on this technological advance, encompassing any study that uses next-generation sequencing and related recent technological advances to assay a large number of loci across the genomes of individuals sampled from one or more populations. This is the "broad-sense" definition of genomics of Garner et al. (2015).

Many population genomic studies under the broad-sense definition address questions that were tractable with traditional markers such as microsatellites, but the increase in number of loci sampled may improve precision and accuracy of the results. We discuss examples of such studies below and also the question of when to

use next-generation sequencing (i.e., broad-sense genomic tools) to address questions that can still be answered with traditional genetic methods. In some cases, the questions addressed with genomic tools are long-standing in ecology and evolution, but the dense sampling of the genome with genomic approaches provides novel insight by revealing much finer-scale patterns. These include estimation of phylogeny, where the different evolutionary histories among regions of the genome can be distinguished, and demographic history, where much finer time scales of inference are possible.

Narrower definitions of population genomics as distinct from genetics emphasize the novel concepts or questions addressed in genomic studies that were previously intractable with traditional methods (Black et al. 2001; Luikart et al. 2003; Allendorf 2017). In ecology and evolution, a central goal is to detect particular loci associated with selection, adaptation, or ecologically relevant traits and to distinguish these from a genome-wide background (Luikart et al. 2003). When a physical or linkage map of the genome is available, sequence or marker data can be placed in a genomic context along chromosomes or linkage groups, and particular regions of the genome that are influenced by evolutionary forces like selection can be identified (Luikart et al. 2018). Even in the absence of a reference map, however, the number of genetic markers possible in studies of non-model organisms allows a qualitative shift in the inferences that can be drawn regarding adaptive processes. As we discuss in more detail below, these inferences do not always require complete sampling of all functionally important parts of the genome.

Here we propose a narrow-sense definition for population genomics in ecology and evolution: *a population genomic study is one in which genetic loci are sampled to a sufficient density across the genome that there is an appreciable likelihood of detecting any genomic regions that are associated with fitness or ecologically relevant traits and distinguishing these factors from background evolutionary forces that affect the genome as a whole*. Below we describe some examples of such "narrow-sense" population genomics.

1.2 Overview of Approaches

Molecular techniques for population genomics in ecology and evolution fall into a few broad categories (Box 1; see also Luikart et al. 2018; Holliday et al. 2018). The range of techniques presents a number of trade-offs in the density and distribution of genetic variation that is sampled across the genome, as well as the number of individual and population samples that may be included given a study's budget, the computational resources required, and the types of inferences that can be made from the data. Importantly, many of the techniques are applicable in cases where little or no prior genomic information is available. This has democratized the field of genomics, opening vast areas of biodiversity to detailed genomic study that was previously impractical.

Box 1 Taxonomy of Methods for Population Genomics in Ecology and Evolution

Traditional genetic methods: These methods include Sanger sequencing of particular loci and any non-sequence-based method for genotyping a set of loci. In some cases, some prior genetic knowledge is required to target specific loci, for instance, to develop PCR primers for amplification. Depending on the loci targeted (e.g., mitochondrial or coding versus noncoding nuclear DNA) and the rate at which it evolves, sequence data can provide insights into a range of time scales from ecological population-level processes to long-term phylogenetic relationships among taxa. Non-sequence-based genotyping methods include allozymes, restriction-fragment analyses, and microsatellites. These techniques are used to produce genotypes for a set of loci across individuals, and these techniques are often most useful for ecological and evolutionary insights within species.

Whole-genome sequencing (WGS): One approach in population genomics is simply to sequence the complete genome of every individual in a sample (e.g., Jones et al. 2012; Ellegren et al. 2012; Robinson et al. 2016). Typically, this is done when a reference genome assembly or physical map is available, so that short-sequence reads from sampled individuals can be aligned against the reference. This approach is also called "whole-genome re-sequencing" because a reference genome has already been sequenced for the species. Samples can either be individually sequenced at high enough coverage to provide individual-level genotype data or pooled to provide population-level allele frequency data. An advantage of WGS is that in addition to identifying single-nucleotide variation, larger-scale genetic variants such as insertion/deletion, copy number variants, and inversions can be identified that may play an important role in adaptation (e.g., Chain et al. 2014; Feulner et al. 2015).

Reduced representation sequencing: While whole-genome sequencing costs continue to decline, making it feasible for ecological and evolutionary studies, it often may not be the most efficient allocation of sequencing effort given the goals of a study, and it imposes substantial bioinformatic burdens. An alternative is to focus sequencing on a reduced representation – a subset – of the genome, so that sequencing effort can be spread across many more individual or population samples. There are several ways to focus on a subset of the genome:

Anonymous reduced representation sequencing includes techniques in which sequencing cannot be targeted at prior-defined loci and may not even be known beforehand. The most common family of such techniques is restriction site-associated DNA sequencing (RADseq; Andrews et al. 2016), a group of techniques united by their use of restriction enzymes to focus

(continued)

Box 1 (continued)

sequencing effort on DNA fragments adjacent to enzyme recognition sites. Restriction enzymes digest DNA at characteristic short (4–8 bp) nucleotide sequences that may occur anywhere in the genome. While the distribution of recognition sites may be biased to some degree (e.g., by GC content or methylation sensitivity), RADseq loci are essentially a random sample across the genome and occur in both coding and noncoding regions.

Transcriptome sequencing focuses sequencing effort on the subset of the genome that is transcribed, by reverse transcribing RNA to DNA during construction of sequencing libraries (Wang et al. 2009; Ekblom and Galindo 2011). In many organisms, such as vertebrates, the transcriptome is a small fraction of the total genome size. To the extent that adaptive variation exists in coding regions (or in regulatory regions tightly linked to coding regions), this approach can increase the chances of identifying adaptive variants, but it also may provide a biased sample of the genome relative to neutral evolutionary processes such as demography.

Sequence capture methods use a prior designed set of probes to focus sequencing effort on a set of hundreds to tens of thousands of loci (Jones and Good 2016). Probes may target genes of interest, putatively neutral loci, or any combination, but they must be designed ahead of time based on prior sequence information. For large, complex genomes, capture methods may allow researchers to avoid repetitive or non-informative genomic regions (McCartney-Melstad et al. 2016). A recent approach combines RADseq and sequence capture, in a protocol called "rapture," to target a subset of previously identified RADseq loci for efficient genotyping across a large number of individuals (Ali et al. 2016).

Multiplex PCR amplicon sequencing is a set of techniques for efficiently amplifying multiple loci with standard PCR primers and then using next-generation sequencing techniques to sequence these loci across many individuals in a single experiment. Like sequence capture methods, multiplex amplicon sequencing requires some prior work to identify loci and design PCR primers, which may target SNPs or other previously identified polymorphisms useful for population genetic studies. An example is the protocol developed by Campbell et al. (2015), called "Genotyping in Thousands by sequencing" (GT-seq), which can target roughly 50–500 loci. GT-seq uses dual barcoding to allow up to thousands of individuals to be multiplexed in a single lane of Illumina sequencing and later separated bioinformatically. Because it targets a relatively modest number of loci, multiplex amplicon sequencing is not suited for conducting, for instance, an initial genome scan for selection, but rather expanding from an initial list of loci of interest to a wider set of populations or individual samples.

How should researchers choose a population genomic approach in an ecological or evolutionary study? The overriding consideration is the goal of the study; the choice of method should be driven by the particular question(s) being addressed and the type of data that would best answer them, given the biology of the system (Andrews and Luikart 2014; Benestan et al. 2016). Methods differ widely in their power to make statistical inferences in natural populations, as well in the cost associated with each method and the trade-offs inherent in sampling design. While no approach is ideal in all cases, the range of options provides flexibility in addressing particular study goals and biological systems and adjusting to constraints of total cost and laboratory or bioinformatics expertise. Within each of the methods in Box 1, there is also wide latitude to adjust technical details, in addition to sampling and experimental design, to tailor genomic techniques to each scientific question. Optimizing these details depends on a large number of considerations (Box 2); a few are discussed in more detail below and illustrated by case studies later in the chapter.

Box 2 Key Questions in Designing a Population Genomic Study

Before embarking on a population genomic study in ecology and evolution, researchers would be well-advised to answer as many of the questions below as possible. These answers will drive the best molecular and bioinformatic approaches to be used, as well as sampling design.

- What are the goals of the study, and what type of data would provide the best statistical power of inference?
- Are genomic techniques necessary at all? Or would a traditional population genetic tool be sufficient and less expensive in time and resources?
- What are the characteristics of the genome? (e.g., total genome size, proportion made up of genic regions, amount of duplicate sequence from whole-genome duplication or transposable elements, etc.)
- What are the prior genomic resources available? (e.g., Is there a genetic map or transcriptome assembly available? Is there a reference genome sequence from the focal species, and how well assembled and annotated is it? Or is there a reference genome from a related species, and if so how divergent?)
- What proportion of the genome, or number of markers, is necessary to cover?
- What are the budget limitations? Total sequencing cost is allocated across several factors: proportion of the genome interrogated, number of markers, number of individuals or populations, length and type of sequencing reads, and depth of coverage.
- What bioinformatics expertise and computational resources would be required to analyze the data?

(continued)

Box 2 (continued)
- How important is reproducibility of the set of loci and compatibility of the data with future studies, for example, applying a similar technique in a related taxon?
- To what extent is the data designed to solely address a particular question or to provide a base of genomic information for multiple future studies?

The first question when designing a study should be: Are genomic approaches appropriate or needed at all? Traditional genetic approaches remain effective tools for addressing a range of questions in ecology and evolution, such as demography, population structure, parentage or sibship, or detection of hybridization. In systems where a technique is established (for instance, where a set of microsatellite loci has been validated), it may be most efficient to avoid the expense and bioinformatic burden of using next-generation sequencing. In addition, this allows newly collected data to be completely compatible with previous studies, for instance, in long-term monitoring studies. However, in the absence of any prior tools or established protocols, genomic techniques like RADseq can be applied to simultaneously identify and genotype a large number of markers across many individuals. For ecological and evolutionary studies of non-model organisms, some genomic techniques are now more cost-efficient than traditional genetic techniques for an initial foray into a new system, even when the focal questions could be addressed with traditional techniques. Furthermore, the use of broad-sense genomics may often improve the accuracy and precision of population genetic estimates and lay the groundwork for further narrow-sense genomic studies.

1.3 How Much of the Genome Should Be Assayed?

Genomic techniques differ widely in what proportion of the genome of each sample is examined. At one extreme, whole-genome sequencing (WGS) provides nearly complete genetic information for each sample, and on the other, reduced representation methods can be dialed down to just a few hundred markers (Andrews et al. 2016; Jones and Good 2016; Ali et al. 2016). As sequencing costs continue to drop, it may seem intuitive to choose the first option – WGS of every sample in a study. However, the costs of WGS still limit most researchers in ecology and evolution to far fewer samples than are optimal to address many research questions, although new techniques may change that in the near future (Therkildsen and Palumbi 2017). There are ways to increase the number of individuals sampled with WGS, for instance, by pooling or low-coverage sequencing. However, a further consideration is that WGS data can impose a substantial computational burden. Researchers in nearly any population genomic study should plan to spend more time on bioinformatics than data generation, and this is certainly true for WGS data. In addition,

growth in computational processing and storage capacity has not kept pace with growth in sequence data-generating capacity. Thus, the bioinformatic costs, beyond the sequencing costs, may outweigh the benefits of WGS data for many ecological and evolutionary studies. Nonetheless, WGS and reduced representation genomic data provide different types of information that are appropriate for addressing different questions in ecology and evolution, as illustrated by case studies below.

As an alternative to WGS, anonymous reduced representation techniques like RADseq can provide a wide range of marker densities across the genome. Some recent discussion has centered around the question of whether the RADseq family of techniques can generate sufficient marker density to address ecological and evolutionary questions (Lowry et al. 2016; McKinney et al. 2017; Catchen et al. 2017). A key consideration is the extent of linkage disequilibrium (LD) across the genome, which effectively scales the density of markers to the proportion of the genome that can be assessed. This is because the signature of evolutionary forces like selection acting at any particular location in the genome will only be measurable if that location is in LD with one or more assayed markers. LD typically decays with distance along a chromosome, although this decay is often far from smooth; in some cases there may be regions of relatively high LD, called "haplotype blocks," punctuated by breakpoints that may reflect locations of elevated recombination rate (Dawson et al. 2002). The extent of LD is not just characteristic of a species but varies among populations due to demographic history, selection, chromosomal structural variation, and other factors (Dunning et al. 2000; Reich et al. 2001). Accordingly, there is vast variation by several orders of magnitude among biological systems in the size of haplotype blocks and thus the density of markers needed to sample a large proportion of them (McKinney et al. 2017).

Under the broad-sense definition of population genomics, many study goals do not require sampling even a majority of haplotype blocks; rather, only a relatively small sampling of the genome is required. Many of these questions could be answered with traditional genetic techniques. However, the increase in markers with genomics can improve accuracy and precision (e.g., below we discuss the relative value of microsatellite loci versus single-nucleotide polymorphism (SNP) loci for statistical inference).

Under the narrower definition of genomics, the proportion of haplotype blocks that are sampled determines the likelihood of detecting functionally important loci (Tiffin and Ross-Ibarra 2014; Catchen et al. 2017). However, even when the goal is to distinguish adaptive variation from the neutral background in a genome scan approach, many scientific questions do not require finding all adaptive loci. Such questions include: Is there a signature of adaptation across the sampled portion of the genome, either within or between populations (Epstein et al. 2016; Funk et al. 2016)? What is the geographic distribution of adaptive variation (White et al. 2013; Ferchaud and Hansen 2016)? Does population structure or phylogeny at adaptive or ecologically relevant loci match that across the rest of the genome (Funk et al. 2012; Wagner et al. 2013)?

In a study addressing these narrow-sense genomic questions in the context of predicted climate change, Bay et al. (2018) identified loci associated with climate

variables across the range of yellow warblers (*Setophaga petechia*) using RADseq. They then assessed genomic vulnerability as the mismatch between current and predicted allele frequencies based on genotype-environment association analyses, in order to predict the species' capacity to adapt to future conditions. They found that populations showing recent declines also tend to be more vulnerable to future selection pressures, potentially informing conservation and monitoring efforts. Studies like Bay et al. (2018), and those addressing the questions above, are narrow-sense genomic ones because they rely on identifying adaptive loci and distinguishing them from the genome-wide background; but they do not require identification of all adaptive loci, let alone functional validation of them. Instead, only a subset of adaptive loci may be detected, but these are still sufficient to address the study goals.

2 Broad-Sense Genomics

2.1 Selectively Neutral Processes

Population genomic approaches can provide more accurate estimates of genetic statistics than traditional techniques. For example, compared to pedigree-based estimates of inbreeding, genomic techniques can provide more accurate estimates of individual and population-level inbreeding. This results from surveying enough markers to determine the actual level of identity by descent within each individual, rather than the expectation based on pedigree relationships (Kardos et al. 2015; Luikart et al. 2018). Population genomic data can also provide greater power to detect inbreeding depression; for example, Hoffman et al. (2014) observed a much higher correlation of fitness and heterozygosity using SNPs compared to microsatellites in harbor seals (*Phoca vitulina*), because their RADseq approach yielded over 14,000 SNP loci.

Several recent studies have directly compared the utility of microsatellites versus genomic SNP data, such as that derived from RADseq or other reduced representation approaches. Because of the number of possible alleles, each microsatellite locus contains potentially much more information than a single SNP locus, which is typically expected to have just two alleles. However, the number of SNP loci commonly available in genomic studies often more than compensates for the lower information content per locus. For example, Malenfant et al. (2015) and Jeffries et al. (2016), studying polar bears (*Ursus maritimus*) and crucian carp (*Carassius carassius*) respectively, found that a RADseq dataset better detected fine-scale population structure than microsatellites. For crucian carp, this was true even when a much smaller sample of individuals per population was used for the RADseq data (Jeffries et al. 2016). Similarly, a study of the Amazonian plant species *Amphirrhox longifolia* using ~4,000 ddRAD loci found that sample sizes of eight were sufficient to estimate diversity when $\geq$1,000 SNPs were used, and sample sizes as low as two provided accurate estimates of F_{ST} when $>$1,500 SNPs were used (Nazareno et al. 2017). These cases illustrate that even when the sample of

individuals is far too small to estimate allele frequency or F_{ST} at any single locus with any accuracy, a large number of loci can still accurately estimate the average F_{ST} across the genome (Nazareno et al. 2017). Similarly, Puckett and Eggert (2016) found that 1,000 SNP loci outperformed 15 microsatellites in assignment of American black bears (*Ursus americanus*) to their natal range. In contrast, Fischer et al. (2017) found that estimates of genetic diversity in *Arabidopsis* populations were not closely aligned between microsatellite and SNP datasets. In fact, heterozygosity at SNP loci was more closely correlated with allelic richness in microsatellite loci than with heterozygosity at microsatellite loci, possibly a result of the different mutation processes in each type of locus.

It is useful to consider the "conversion rate" between microsatellites and SNPs in terms of the information content for different types of analyses. For example, Kaiser et al. (2016) found that a panel of 97 SNPs was equivalent to 6 microsatellite loci in estimating parentage in black-throated blue warblers (*Setophaga caerulescens*). Elbers et al. (2017) found that 100 SNP loci were required to correlate with the results of 10 microsatellite loci in estimating population differentiation (F_{ST}) in the gopher tortoise (*Gopherus polyphemus*), but 800 SNPs were needed to correlate with the same 10 microsatellites in estimating expected and observed heterozygosity. Note that the absolute estimates of F_{ST} and heterozygosity (and other population genetic statistics, like effective number of breeders N_b; Linløkken et al. 2016) may differ between SNPs and microsatellites because of the different mutation rates involved. Overall, these studies put the "conversion rate" between SNPs and microsatellites at anywhere from 10:1 to 80:1, depending on the analysis. However, it is typically feasible to get several orders of magnitude more SNP markers than microsatellites in most cases, in which case the conversion rate no longer matters (Fischer et al. 2017). For example, the study by Elbers et al. (2017) above subsampled their SNP markers from a dataset of nearly 18,000 SNP loci from sequence capture.

2.2 *Neutral Population Genetic Structure and Population Units*

The broad-sense definition of genomics includes the use of genomic tools to improve upon accuracy, precision, and efficiency compared to previous genetic approaches for estimates of, for example, population structure (see below), levels of admixture and inbreeding (Kardos et al. 2015, 2016), or effective population size (N_e). For example, Larson et al. (2014) used RADseq data to estimate N_e in Chinook salmon (*Oncorhynchus tshawytscha*) using the method *N_eEstimator* (Do et al. 2014), which relies on linkage disequilibrium among loci. In this case having a genetic map of the genome can allow the removal of physically linked loci, which can downwardly bias estimates of N_e (Park 2011; Larson et al. 2014). Larson et al. (2014) found that estimates of N_e based on 1,118 RADseq-derived SNPs had far smaller confidence intervals compared to estimates based on 39 previously identified SNP loci.

In the case of identifying genetic structure among populations, the increased precision of genomic tools may identify genetic differentiation that was cryptic to traditional methods. For instance, a phylogeographic study using RADseq for a mosquito (*Wyeomyia smithii*) in eastern North America revealed insights into demographic history that were not identified using traditional markers (Emerson et al. 2010). In this study, the authors used RADseq of pooled population samples to estimate consensus genotypes at a large number of SNP loci for each population and then used these data in a phylogenetic analysis. Because most of the populations are the result of recolonization from refugia following the last Pleistocene glaciation, genetic differentiation among them is relatively recent (beginning 22,000–19,000 years ago). Whereas previous mitochondrial DNA sequence data produced poorly resolved relationships among current populations, the pooled RADseq approach revealed a distinct geographic pattern of recolonization northward and then westward (Fig. 1). One possible factor in this discrepancy is that mitochondrial DNA sequence represents a single locus with different inheritance patterns than nuclear loci, while genomic techniques can sample a large number of loci across the much larger nuclear genome. Particularly in cases like this, with

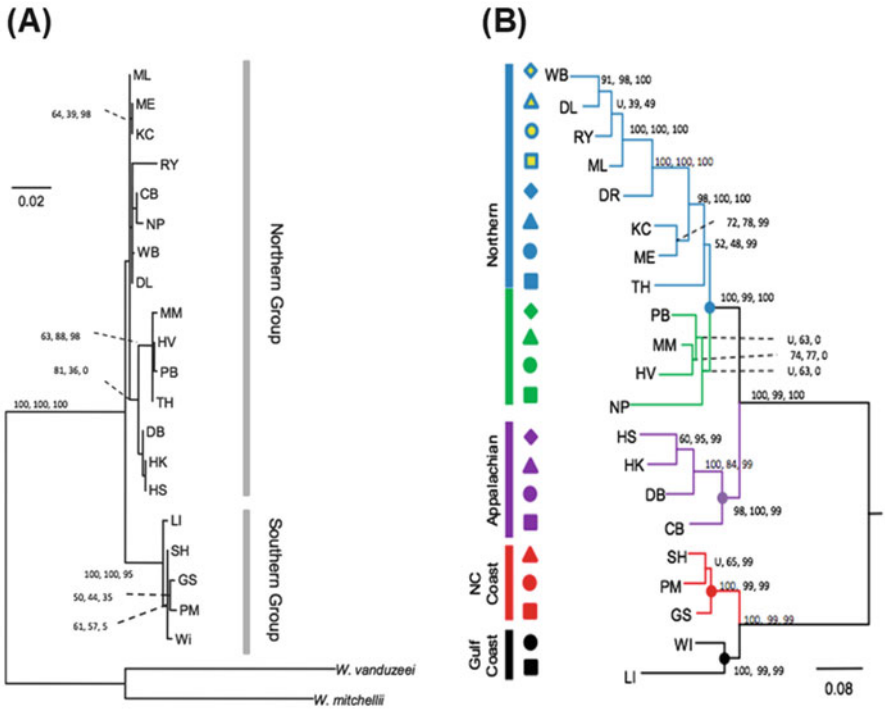

Fig. 1 Improvement of phylogeographic inference in the mosquito *Wyeomyia smithii* using broad-sense population genomic tools. (**a**) Maximum likelihood tree of relationships among populations based on mitochondrial *COI* sequence data. (**b**) Maximum likelihood tree based on 3,741 nuclear SNP loci derived from pooled RADseq data. Modified from Emerson et al. (2010)

recent differentiation among populations within a species, loci across the genome may reflect different phylogenetic histories due to incomplete lineage sorting and migration after formation of the populations. Analytical methods should account for this discrepancy in phylogenetic history among loci.

Genomic approaches also provide great promise for increased power in population assessments and stock identification for managed or harvested species, particularly marine taxa. Stock identification in harvested species is important for conservation and management of populations to avoid overharvest and local population extirpation (Palsbøll et al. 2007). For example, Benestan et al. (2015) used RADseq for American lobster (*Homarus americanus*) to define populations that were previously unresolved using microsatellite markers and to identify a set of loci that could assign individuals to source populations despite the weak genome-wide population structure for this species (mean $F_{ST} = 0.00185$). The authors identified and genotyped 10,000 SNPs using RADseq and then identified a subset of 3,000 high-F_{ST} loci (identified using a training set of samples and validated on an independent set, following Anderson (2010)) that assigned individuals to their source location with 80% success. Low genome-wide values of F_{ST} are expected to be characteristic of wide-ranging taxa with long-distance dispersal and large N_e (Bernatchez 2016). However, functionally important differentiation may occur at a small number of loci, and genomic approaches can identify these loci for ecological and evolutionary inferences. Even if the study goal is not to identify functionally important loci or loci under selection (as it is in "narrow-sense" genomic studies discussed below), the ability of genomic techniques to identify so many markers that a subset of highly differentiated markers can be extracted allows for finer-scale discrimination of population structure.

2.3 Phylogenomics

Genomic tools are increasingly being used for assessing phylogenetic relationships among species and higher taxa (Chan and Ragan 2013; McCormack et al. 2013; Ree and Hipp 2015; Barrett et al. 2016). A major challenge for such phylogenomic studies is that the many parts of the genome sampled by genomic tools may represent different lineage histories, and this has required building on traditional phylogenetic tools that assume a single history. Several different genomic techniques are applied in phylogenomics, including anonymous reduced representation techniques such as RADseq (Ree and Hipp 2015), targeted sequence capture (Bragg et al. 2016), and even whole-genome sequencing (Jarvis 2016). In this latter case, whole-genome data provided a detailed phylogeny and comparative genomic study of an entire vertebrate class, birds (Jarvis 2016). But even for short-sequence techniques such as RADseq, the accessible scale of taxonomic resolution can be quite deep (e.g., over 80-million-year divergence in octocorals, *Paragorgia* spp.; Herrera and Shank 2016). However, Leaché et al. (2015) found conflicting results from sequence capture and RADseq phylogenetic estimates in phrynosomatid lizards. Interestingly,

the best concordance between the sequence capture and RADseq-based SNP trees occurred when less conservative filtering was applied to the RADseq data, providing a large set of SNPs (roughly 16,000) with substantial missing data. This suggests that conservative filtering of genomic SNP data may cause misestimation in some cases.

While conflicting gene trees among loci can be a problem when the goal is to estimate a single species tree, variation among loci may reflect truly different evolutionary histories because of reticulate evolution (Vargas et al. 2017). The power of modern sequencing technology allows for phylogenetic estimation across multiple species or groups on a landscape, so that patterns of reticulate evolution and conflicting gene trees can be examined in a comparative framework (Edwards et al. 2016). While this can challenge the development of new demographic models and phylogenetic analysis tools (Edwards et al. 2016), it can also reveal insights into the adaptive consequences of hybridization and introgression (Keller et al. 2013; Nadeau et al. 2014; further discussion below).

3 Narrow-Sense Genomics

3.1 Detecting Ecologically Relevant and Adaptive Variation

At the heart of many population genomic studies in ecology and evolution is the detection of adaptive or functionally important loci (Luikart et al. 2003, 2018). One way to identify such loci is traditional genetic mapping techniques, made more powerful with the density of loci provided by population genomic approaches. Quantitative trait locus (QTL) mapping is possible for species that can be crossed experimentally (e.g., Miller et al. 2012; Liu et al. 2014) or for which pedigrees are known for natural populations (e.g., Slate et al. 2002; Beraldi et al. 2007; Santure et al. 2013). Genome-wide association studies (GWAS) are also feasible, even in many natural populations of ecological or evolutionary interest, in part because of the "democratization" of genomic techniques to non-model organisms. Some natural systems may be particularly well-suited to this approach; for instance, Nadeau et al. (2014) took advantage of a natural hybrid zone between phenotypically divergent butterfly (*Heliconius* spp.) subspecies to map wing color traits.

A long-standing method to distinguish adaptive loci from the genome-wide background is to identify high-F_{ST} outliers that are suspected to be under divergent natural selection among populations (Lewontin and Krakauer 1973; Beaumont and Nichols 1996). Outlier tests have received some criticism and perhaps been misapplied in some cases (Hermisson 2009; Hohenlohe et al. 2010; Hoban et al. 2016), in part because methods differ in model assumptions. Violations of model assumptions, such as historic demographic fluctuations, can increase variance in F_{ST} among loci and create false positives (Hohenlohe et al. 2010; Whitlock and Lotterhos 2015).

More recently, parallel to the development of landscape genetics and genomic approaches, there is increased interest in directly associating allele frequencies with

environmental variables, through genotype-environment association tests (GEAs; Joost et al. 2007; Coop et al. 2010; Hancock et al. 2011; Fumagalli et al. 2011; Schoville et al. 2012; Frichot et al. 2013; Rellstab et al. 2015; Forester et al. 2016; Hoban et al. 2016). GEAs are conceptually similar and complementary to GWAS approaches in that gene frequencies are associated with environmental factors, whereas in GWAS loci are associated with phenotypic traits. In humans, where very large sample sizes are feasible, several studies have used GEAs to identify important loci linked to environmental factors (Hancock et al. 2011; Fumagalli et al. 2011). In most ecological and evolutionary studies, samples sizes of both individuals and number of markers may be much smaller.

The move toward GEAs has been prompted by greatly increased availability of environmental and genomic data and growing understanding that signatures of adaptive selection can be difficult to distinguish from the selectively neutral genomic background (Schoville et al. 2012). For example, genetic variation underlying polygenic traits may be difficult to detect because the effect size and allele frequency shifts at any single locus may be quite small (Bernatchez 2016). Simulation-based studies have found that, in general, GEAs have more power to detect loci under selection than outlier-based approaches but have higher rates (20–50%) of false positives (De Mita et al. 2013; Frichot et al. 2013; Forester et al. 2016). Recent work has also suggested that multivariate approaches (principal component analysis, redundancy analysis, and population graphs) might help reduce the number of false positives and maintain reasonable power to detect true correlations (Forester et al. 2016; Rajora et al. 2016). Several other methods are also available for detecting loci under selection from population genomic data, and they are appropriate for different population scenarios, data types, types of selection, and time scales (Hohenlohe et al. 2010; Rajora et al. 2016; Luikart et al. 2018).

Bernatchez (2016) outlined a number of factors that can maintain adaptive variation in natural populations and therefore make signatures of adaptation difficult to identify. These include soft selective sweeps, traits with a polygenic basis, epistatic interactions among genes, epigenetics, and various types of balancing selection. Under these conditions, selection does not often drive single beneficial alleles to fixation; rather, the response to selection is relatively slight shifts in allele frequencies. In the potentially large number of cases in which adaptation depends on a large number of loci, detecting selection may be improved by alternative approaches. For instance, a promising recent approach in outlier tests for local adaptation is to focus on allele frequency covariance among loci, rather than allele frequency variation at individual loci (LeCorre and Kremer 2012; Rajora et al. 2016; Lind et al. 2017). Although reliably detecting adaptive loci remains challenging, a large and growing number of studies have detected adaptive variation with population genomic tools and provided insights into multiple aspects of species biology (Luikart et al. 2018). Below we discuss a few case studies.

3.2 Adaptive Population Structure

Advances in the discovery of non-neutral (i.e., candidate adaptive) markers have improved our ability to examine how selection shapes population genetic structure and how and why population structure differs at fitness-related loci compared to the genome-wide background dominated by neutral forces like demography. Recent genomics studies have revealed significant adaptive divergence at outlier loci, even in systems of high gene flow, such as marine organisms and forest trees. This is especially apparent in marine species, presumably because of large effective population sizes and large dispersal differences, which reduce neutral population divergence and allow for selection to act effectively on adaptive loci (Limborg et al. 2012; Corander et al. 2013; Hess et al. 2013; Milano et al. 2014). Genomic patterns of adaptive divergence often vary across spatial scales within a species, and adaptive loci often reveal finer-scale differentiation than neutral loci (Matala et al. 2014; Hand et al. 2016). Simulation-based modeling has further shown that inferences about local adaptation based solely on neutral genetic markers risk incorrectly identifying the underlying mechanisms driving population structure (Landguth and Balkenhol 2012).

In one example, Steane et al. (2015) used genome-wide diversity array technology (DArTseq; Sansaloni et al. 2011) to identify and genotype 16,122 high-quality dominant markers (presence/absence) in gimlet trees (*Eucalyptus salubris*; Steane et al. 2014). *E. salubris* is an obligate seeder that does not survive wildfire; however, it is also a key species for revegetation in a moderate (mesic) to arid region in Southern Australia (Nicolle 2006; Steane et al. 2015). Steane et al. (2015) identified a set of 24 putatively adaptive loci that showed high rates of differentiation ($F_{ST} > 0.7$ and many close to fixation) between two cryptic lineages in *E. salubris*, which appeared to be associated with climate adaptation along a strong aridity gradient. In this case, genome-wide scans were essential in identifying putatively adaptive markers of high differentiation that otherwise would have gone undetected by traditional neutral genetic techniques or phenotypic traits alone.

Killer whales (*Orcinus orca*) provide another example illustrating how neutral and adaptive markers can show different patterns of genetic structure. This species is the most widely distributed marine mammal. Despite the propensity for long-range dispersal and the movement of individual social groups over wide geographic ranges, there appears to be very little ancestral dispersal among sympatric ecotypes that differ in foraging behavior (Moura et al. 2014; Morin et al. 2015). Mitogenomes and 42 independent nuclear loci were found to be in concordance, indicating very limited gene flow among ecotypes (Morin et al. 2015). Moura et al. (2014) further identified a set of putatively adaptive loci (168 of 3,281 variable SNPs). Neutral genetic structure agreed with previous studies in identifying significant differentiation between populations in sympatry. However, adaptive genetic structure differed from neutral patterns and included a reduced set of high-F_{ST} outliers ($F_{ST} > 0.7$) with putative physiologically relevant function related to digestion and reproduction (Moura et al. 2014). The difference in neutral vs. adaptive genetic differentiation offered additional evidence that differentiation among sympatric populations was related to ecological processes more so than genetic drift.

3.3 Adaptive Introgression and Hybridization

Hybridization and introgression have important evolutionary consequences, and understanding these consequences is aided by dense, genome-wide marker coverage. Several powerful tools have been developed for inferring historical patterns of hybridization and introgression from population genomic data (e.g., TreeMix, Pickrell and Pritchard 2012; ALDER, Loh et al. 2013). In systems of hybridizing native and non-native species, introgression can lead to genetic extinction of the native (and often endangered) species (Allendorf et al. 2010).

One of the most well-studied systems involves the Flathead River system in the Northern Rocky Mountains, USA (Boyer et al. 2008; Muhlfeld et al. 2009, 2014; Hohenlohe et al. 2011, 2013; Amish et al. 2012; Hand et al. 2015; Kovach et al. 2015, 2016). Here, native westslope cutthroat trout (*Oncorhynchus clarkii lewisi*) is greatly threatened by hybridization with rainbow trout (*O. mykiss*), the world's most widely introduced fish (Halverson 2010). Hohenlohe et al. (2013) showed improved accuracy in measuring individual admixture proportions when using 3,180 diagnostic SNPs vs. 7 microsatellite loci (Boyer et al. 2008). The use of paired-end RADseq in this study allowed for identification of candidate genes by providing longer contiguous sequence around significant SNPs than previous approaches. Subsequent work included the publication of a reference genome (Berthelot et al. 2014) and the identification of more diagnostic markers for identifying parental ancestry, made possible with a larger sample of individuals and reference-based rather than de novo locus identification (Hand et al. 2015). These technical advances further refined the understanding of the system, revealing that selection in hybridized populations acts primarily against genetic variation from the invasive rainbow trout (Kovach et al. 2016).

Two more illustrative examples of adaptive introgression, and its signature on genomic variation, are from cichlid fish and butterflies. Keller et al. (2013) used RADseq in closely related cichlid taxa (*Pundamilia* and *Mbipia* species) from Lake Victoria. Five taxa were identified by several phenotypic traits, including male coloration. Across much of the genome, the taxa are poorly differentiated, but a subset of loci putatively associated with adaptive differentiation suggests two introgression events among lineages within the group that carried genetic variation for male coloration and opsin alleles (Keller et al. 2013). Similarly, Nadeau et al. (2014) examined striking color pattern differentiation among subspecies of the butterfly *Heliconius melpomene*, using RADseq to identify both loci under divergent selection (high-F_{ST} outliers) and loci associated with phenotypic variation in color pattern (GWAS). They found that signatures from both F_{ST} outlier tests and GWAS converged on a small number of major effect loci, providing evidence that narrow hybrid zones are maintained by strong selection on color pattern.

3.4 Demographic History

Genomic data provide a powerful ability to reconstruct the demographic history of populations. Previous genetic markers, such as microsatellites and mitochondrial DNA sequence, allowed some inference of historical fluctuations in population size using bottleneck tests and approximate Bayesian computation (ABC) methods (e.g., Fontaine et al. 2012; Spurgin et al. 2014). However, genomic data can provide much greater statistical power with ABC methods (Cornuet et al. 2014). For instance, large numbers of SNP loci can be used to estimate the allele frequency distribution, which can be used to test alternative models of demographic history across a set of populations using the method ∂a∂i (Gutenkunst et al. 2009). This method can test for changes in population size (expansion, contraction) as well as migration among populations (Fig. 2). While developed originally for human populations, with the advent of genomic techniques for non-model species, ∂a∂i has been applied widely. For instance, this type of demographic inference can be conducted in a comparative framework, as across six taxon pairs of birds that share similar distribution patterns in disjunct South American dry forest habitats (Oswald et al. 2017).

Estimation of demographic history from SNP data can be combined with detection of outlier loci that show greater differentiation between populations and differential patterns of gene flow among populations. For instance, Leroy et al. (2017) used ∂a∂i and ABC to infer the history of four European white oak species (*Quercus* spp.). They found that a long period of isolation generated some reproductive barriers, but that recent secondary contact due to postglacial warming resulted in

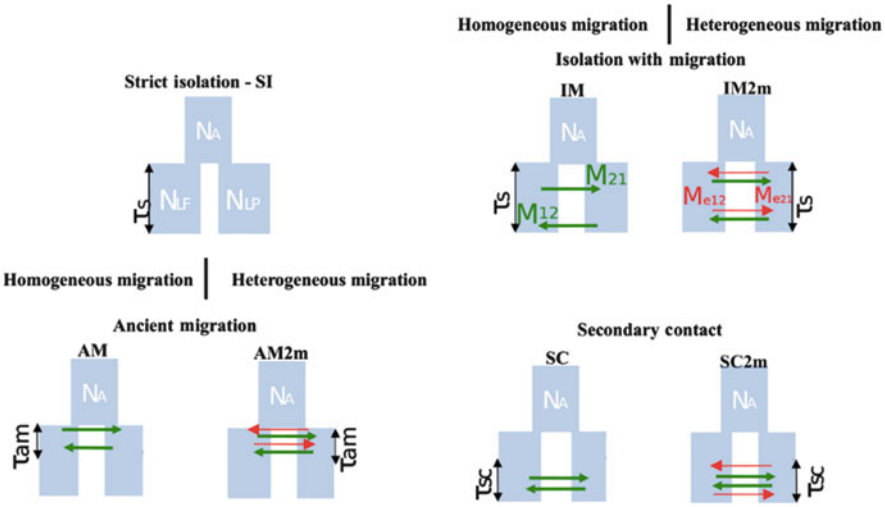

Fig. 2 Demographic scenarios tested in lamprey ecotypes. Four general models are shown for the history of two populations since divergence: strict isolation, isolation with migration, ancient migration, and secondary contact. In each model, parameters are estimated for the population sizes and timing of events. Reproduced with permission from Rougement et al. (2016)

secondary contact and gene flow at some loci, but not uniformly across the genome. Similarly, Rougement et al. (2016) illuminated the effect of migration on divergently selected loci in two lamprey species (*Lampetra* spp.), and Schield et al. (2017) applied a range of tests for migration and selection in western diamondback rattlesnakes (*Crotalus atrox*). All of these studies used reduced representation methods to genotype several thousand SNP loci across dozens to hundreds of individuals, illustrating the ability of this sampling design to be informative about adaptive variation.

Alternatively, the whole-genome sequence of even a single diploid individual can be used to infer the historical course of effective population size using pairwise sequentially Markovian coalescent (PSMC; Li and Durbin 2011). For instance, McManus et al. (2016) reconstructed the demographic history of lowland gorillas (*Gorilla* spp.), identifying a population contraction that appears to correspond with reduction in forest cover at the end of the last glacial maximum. Similarly, Nadachowska-Brzyska et al. (2016) linked climate changes to population fluctuations in *Ficedula* flycatchers (Fig. 3). Both of these studies emphasize that individuals from different populations are likely to exhibit different demographic histories, as might be expected, and a critical assumption in these analyses is that the sequenced individuals are representative of the population unit under study. Additionally, historic population structure can violate assumptions of the model and lead to false signatures of fluctuations in population size (Mazet et al. 2016). A way around these problems may be newer methods that allow analysis of multiple individuals, such as SMC++ (Terhorst et al. 2017). In addition to demographic reconstruction, WGS data can be used in a comparative framework to identify adaptive loci across closely related species, as illustrated in a study of large cats (*Panthera* spp.; Cho et al. 2013).

Demographic fluctuations have important consequences for current levels of genetic diversity and adaptive potential in natural populations. For instance, two studies have addressed this issue in island foxes (*Urocyon littoralis*) using two different genomic methods. Island foxes persist in six populations, each restricted to a separate island off the coast of Southern California, that have historically small population sizes in addition to recent bottlenecks. Robinson et al. (2016) used whole-genome sequencing of a single fox from each island (with the exception of one island represented by two individuals) and found extremely low levels of heterozygosity and the presence of deleterious variants. The approach of using WGS in a very small number of samples is justified here, because populations are likely to be well-mixed within each island and avoid the violation of assumptions mentioned above (Mazet et al. 2016). Funk et al. (2016) addressed the issues of genetic diversity in island foxes using RADseq. This approach assayed far fewer loci but across a total of 188 individuals. This study similarly found low levels of genetic diversity within each population. Because of the larger number of individuals sampled, it was possible to use F_{ST} outlier tests to detect selection, and despite the low overall diversity and differentiation among islands due to drift, there was also

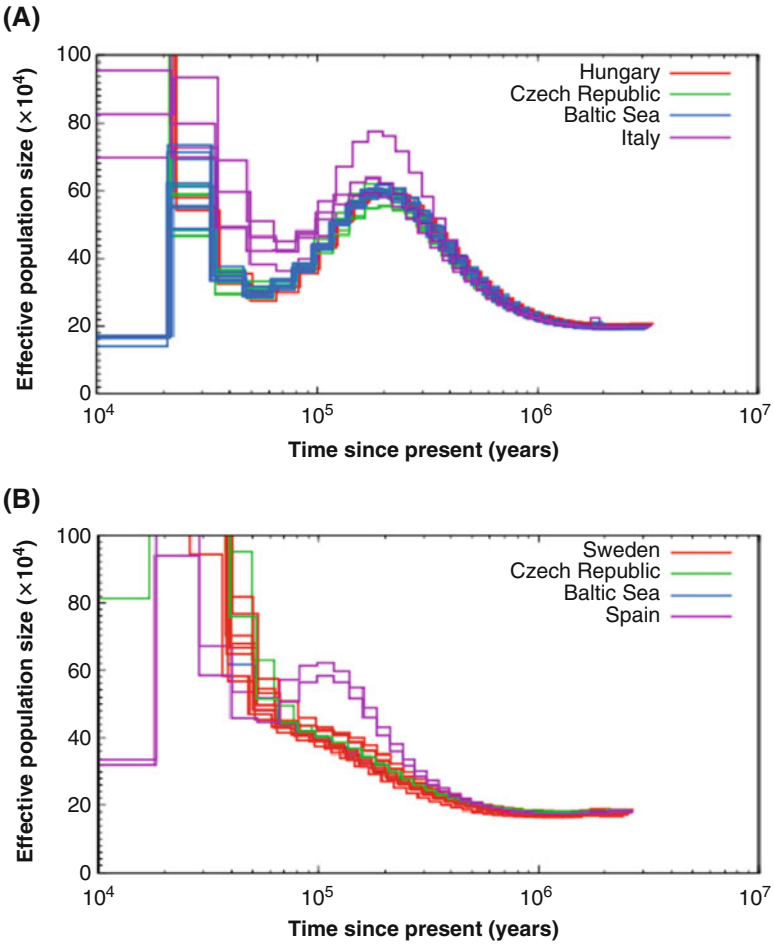

Fig. 3 Reconstruction of the demographic history of populations of (**a**) collared flycatchers (*Ficedula albicollis*) and (**b**) pied flycatchers (*Ficedula hypoleuca*), using the pairwise sequentially Markovian coalescent (PSMC) method (Li and Durbin 2011) on whole-genome sequence data. As illustrated here, PSMC can result in uncertainty at recent time scales, but it allows comparative demographic inference among related taxa occupying the same region. Modified from Nadachowska-Brzyska et al. (2016)

evidence of adaptive divergence among islands. Note that the study by Funk et al. (2016) is a case where not all haplotype blocks, and thus not all potentially adaptive loci, were sampled with the RADseq approach; nonetheless, a narrow-sense genomic question (is there evidence for adaptive differentiation among populations?) was still able to be answered.

4 Conclusions and Future Perspectives

Population genomics has provided numerous insights into ecological and evolutionary processes in natural and managed populations. The wealth of molecular and analytical techniques provides great flexibility in tailoring a population genomic approach to the goals of any particular study and the challenges of any particular biological system. The field is changing rapidly. The cost of acquiring sequence data continues to drop, and novel analytical techniques incorporate improved models of genomic processes and increased statistical power. In particular, whole-genome sequencing may be the best-suited approach to an expanding range of population genomic applications, but nonetheless a variety of reduced representation and targeted sequencing approaches are likely to continue to provide efficient alternatives. It is imperative for researchers in ecology and evolution to educate themselves about the trade-offs involved in designing population genomic studies. With careful consideration of the range of options, population genomics will continue to provide remarkable insights into ecological and evolutionary processes.

Acknowledgments PAH and KRA received support from NSF grant DEB-1316549. BKH was partially supported by funds from NSF grant DOB-1639014 and NASA NNX14AB84G. KRA was supported by the University of Idaho College of Natural Resources, USA. This is PMEL contribution number 4750 and Joint Institute for the Study of the Atmosphere and Ocean (JISAO) and NOAA Cooperative Agreement and NA15OAR4320063, contribution number 2018-0135.

References

Ali OA, O'Rourke SM, Amish SJ, Meek MH, Luikart G, Jeffres C, Miller MR. RAD capture (rapture): flexible and efficient sequence-based genotyping. Genetics. 2016;202:389–400.

Allendorf FW. Genetics and the conservation of natural populations: allozymes to genomes. Mol Ecol. 2017;26:420–30.

Allendorf FW, Hohenlohe PA, Luikart G. Genomics and the future of conservation genetics. Nat Rev Genet. 2010;11:697–709.

Amish SJ, Hohenlohe PA, Painter S, Leary RF, Muhlfeld C, Allendorf FW, Luikart G. RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays. Mol Ecol Res. 2012;12:653–60.

Anderson EC. Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. Mol Ecol Res. 2010;10:701–10.

Andrews KR, Luikart G. Recent novel approaches for population genomics data analysis. Mol Ecol. 2014;23:1661–7.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016;17:81–92.

Avise JC. Molecular markers, natural history and evolution. New York: Chapman & Hall; 1994.

Barrett CF, Bacon CD, Antonelli A, Cano A, Hofmann T. An introduction to plant phylogenomics with a focus on palms. Bot J Linn Soc. 2016;182:234–55.

Bay RA, Harrigan RJ, Underwood VL, Gibbs HL, Smith TB, Ruegg K. Genomic signals of selection predict climate-driven population declines in a migratory bird. Science. 2018;359:83–6.

Beaumont MA, Nichols RA. Evaluating loci for use in the genetic analysis of population structure. Proc R Soc B. 1996;263:1619–26.

Benestan L, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L. RAD-genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species; the American lobster (*Homarus americanus*). Mol Ecol. 2015;24:3299–315.

Benestan LM, Ferchaud AL, Hohenlohe PA, Garner BA, Naylor GJP, Baums IB, et al. Conservation genomics of natural and managed populations: building a conceptual and practical framework. Mol Ecol. 2016;25:2967–77.

Beraldi D, McRae AF, Gratten J, Slate J, Visscher PM, Pemberton JM. Mapping quantitative trait loci underlying fitness-related traits in a free-living sheep population. Evolution. 2007;61:1403–16.

Bernatchez L. On the maintenance of genetic variation and adaptation to environmental change: considerations from population genomics in fishes. J Fish Biol. 2016;89:2519–56.

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun. 2014;5:3657.

Black WC, Baer CF, Antolin MF, DuTeau NM. Population genomics: genome-wide sampling of insect populations. Annu Rev Entomol. 2001;46:441–69.

Boyer MC, Muhlfeld CC, Allendorf FW. Rainbow trout (*Oncorhynchus mykiss*) invasion and the spread of hybridization with native westslope cutthroat trout (*Oncorhynchus clarkia lewisii*). Can J Fish Aquat Sci. 2008;65:658–69.

Bragg JG, Potter S, Bi K, Moritz C. Exon capture phylogenomics: efficacy across scales of divergence. Mol Ecol Resour. 2016;16:1059–68.

Campbell NR, Harmon SA, Narum SR. Genotyping-in-thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. Mol Ecol Res. 2015;15:855–67.

Catchen J, Hohenlohe PA, Bernatchez L, Funk WC, Andrews KR, Allendorf FW. Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. Mol Ecol Res. 2017;17:362–5.

Chain FJJ, Feulner PGD, Panchal M, Eizaguirre C, Samonte IE, et al. Extensive copy-number variation of young genes across stickleback populations. PLoS Genet. 2014;10:e1004830.

Chan CX, Ragan MA. Next-generation phylogenomics. Biol Direct. 2013;8:3.

Cho YS, Hu L, Hou H, Lee H, Xu J, Kwon S, et al. The tiger genome and comparative analysis with lion and snow leopard genomes. Nat Commun. 2013;4:2433.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK. Using environmental correlations to identify loci underlying local adaptation. Genetics. 2010;185:1411–23.

Corander J, Majander KK, Cheng L, Merila J. High degree of cryptic population differentiation in the Baltic Sea herring *Clupea harengus*. Mol Ecol. 2013;22:2931–40.

Cornuet J-M, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, Leblois R, et al. DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. Bioinformatics. 2014;30:1187–9.

Cresko WA, Amores A, Wilson C, Murphy J, Currey M, Phillips P, Bell MA, Kimmel CB, Postlethwait JH. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. Proc Natl Acad Sci U S A. 2004;101:6050–5.

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, et al. A first-generation linkage disequilibrium map of human chromosome 22. Nature. 2002;418:544–8.

De Mita S, Thuillet A-C, Gay L, Ahmadi N, Manel S, Ronfort J, et al. Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. Mol Ecol. 2013;22:1383–99.

Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR. NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data. Mol Ecol Res. 2014;14:209–14.

Dunning AM, Durocher F, Healey CS, Teare MD, McBride SE, Carlomagno F, et al. The extent of linkage disequilibrium in four populations with distinct demographic histories. Am J Human Genet. 2000;67:1544–54.

Edwards SV, Potter S, Schmitt CJ, Bragg JG, Moritz C. Reticulation, divergence, and the phylogeography–phylogenetics continuum. Proc Natl Acad Sci. 2016;113:8025–32.

Ekblom R, Galindo J. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity. 2011;107(1):11.

Elbers JP, Clostio RW, Taylor SS. Population genetic inferences using immune gene SNPs mirror patterns inferred by microsatellites. Mol Ecol Resour. 2017;17:481–91.

Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, et al. The genomic landscape of species divergence in *Ficedula* flycatchers. Nature. 2012;491:756–60.

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, et al. Resolving postglacial phylogeography using high-throughput sequencing. Proc Natl Acad Sci U S A. 2010;107:16196–200.

Epstein B, Jones M, Hamede R, Hendricks S, McCallum H, Murchison EP, et al. Rapid evolutionary response to a transmissible cancer in Tasmanian devils. Nat Commun. 2016;7:12684.

Ferchaud AL, Hansen MM. The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: three-spine sticklebacks in divergent environments. Mol Ecol. 2016;25:238–59.

Feulner PGD, Chain FJJ, Panchal M, Huang Y, Eizaguirre C, Kalbe M, et al. Genomics of divergence along a continuum of parapatric population differentiation. PLoS Genet. 2015;11:e1004966.

Fischer MC, Rellstab C, Leuzinger M, Roumet M, Gugerli F, Shimizu KK, et al. Estimating genomic diversity and population differentiation – an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. BMC Genomics. 2017;18:69.

Fisher RA. The genetic theory of natural selection. New York: Dover; 1958.

Fontaine MC, Snirc A, Frantzis A, Koutrakis E, Öztürk B, Öztürk AA, Austerliz F. History of expansion and anthropogenic collapse in a top marine predator of the Black Sea estimated from genetic data. Proc Natl Acad Sci U S A. 2012;109:E2569–76.

Forester BR, Jones MR, Joost S, Landguth EL, Lasky JR. Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. Mol Ecol. 2016;25:104–20.

Frichot E, Schoville SD, Bouchard G, François O. Testing for associations between loci and environmental gradients using latent factor mixed models. Mol Biol Evol. 2013;30:1687–99.

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, et al. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. 2011;7:e1002355.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW. Harnessing genomics for delineating conservation units. Trends Ecol Evol. 2012;27:489–96.

Funk WC, Lovich RE, Hohenlohe PA, Hofman CA, Morrison SA, Sillett TS, et al. Adaptive divergence despite strong genetic drift: genomic analysis of the evolutionary mechanisms causing genetic differentiation in the island fox (*Urocyon littoralis*). Mol Ecol. 2016;25:2176–94.

Garner BA, Hand BK, Amish SJ, Bernatchez L, Foster JT, Miller KM, et al. Genomics in conservation: case studies and bridging the gap between data and application. Trends Ecol Evol. 2015;31:81–2.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5:e1000695.

Halverson A. An entirely synthetic fish: how rainbow trout beguiled America and overran the world. New Haven: Yale University Press; 2010.

Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, et al. Adaptations to climate-mediated selective pressures in humans. PLoS Genet. 2011;7:e1001375.

Hand BK, Hether TD, Kovach RP, Muhlfeld CC, Amish SJ, Boyer MC, O'Rourke SM, Miller MR, Lowe WH, Hohenlohe PA, Luikart G. Genomics and introgression: discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. Curr Zool. 2015;61:146–54.

Hand BK, Muhlfeld CC, Wade AA, Kovach RP, Whited DC, Narum SR, Matala AP, Ackerman MW, Garner BA, Kimball JS, Stanford JA, Luikart G. Climate variables explain neutral and adaptive variation within salmonid metapopulations: the importance of replication in landscape genetics. Mol Ecol. 2016;25:689–705.

Hermisson J. Who believes in whole-genome scans for selection? Heredity. 2009;103:283–4.

Herrera S, Shank TM. RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. Mol Phylogenet Evol. 2016;100:70–9.

Hess JE, Campbell NR, Close DA, Docker MF, Narum SR. Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. Mol Ecol. 2013;22:2898–916.

Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, et al. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. Am Nat. 2016;188:379–97.

Hoffman JI, Simpson F, David P, Rijks JM, Kuiken T, Thorne MAS, et al. High-throughput sequencing reveals inbreeding depression in a natural population. Proc Natl Acad Sci U S A. 2014;111:3775–80.

Hohenlohe PA, Phillips PC, Cresko WA. Using population genomics to detect selection in natural populations: key concepts and methodological considerations. Int J Plant Sci. 2010;171:1059–71.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G. Next-generation RAD sequencing identified thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Mol Ecol Res. 2011;11:117–22.

Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps-Hughes N, Boyer MC, Muhlfeld CC, Allendorf FW, Johnson EA, Luikart G. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. Mol Ecol. 2013;22:3002–13.

Holliday JA, Hallerman EM, Haak DC. Genotyping and sequencing technologies in population genetics and genomics. Cham: Springer; 2018.

Jarvis ED. Perspectives from the avian phylogenomics project: questions that can be answered with sequencing all genomes of a vertebrate class. Annu Rev Anim Biosci. 2016;4:45–59.

Jeffries DL, Copp GH, Lawson Handley L, Olsén KH, Sayer CD, Hänfling B. Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. Mol Ecol. 2016;25:2997–3018.

Jones MR, Good JM. Targeted capture in evolutionary and ecological genomics. Mol Ecol. 2016;25:185–202.

Jones FC, Grabherr MG, Chan YF, Russell P, Maucell E, Johnson J. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484:55–61.

Joost S, Bonin A, Bruford W, Després CC, Erhardt G, Taberlet P. A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. Mol Ecol. 2007;16:3955–69.

Kaiser SA, Taylor SA, Chen N, Sillett TS, Bondra ER, Webster MS. A comparative assessment of SNP and microsatellite markers for assigning parentage in a socially monogamous bird. Mol Ecol Resour. 2016;17:183–93.

Kardos M, Luikart G, Allendorf FW. Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. Heredity. 2015;115:63–72.

Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW. Genomics advances the study of inbreeding depression in the wild. Evol Appl. 2016;9:1205–18.

Keller I, Wagner CE, Greuter L, Mwaiko S, Selz OM, Sivasundar A, et al. Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. Mol Ecol. 2013;22:2848–63.

Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press; 1983.

Kovach RP, Muhlfeld CC, Boyer MC, Lowe WH, Allendorf FW, Luikart G. Dispersal and selection mediate hybridization between a native and invasive species. Proc R Soc B. 2015;282:20142454.

Kovach RP, Hand BK, Hohenlohe PA, Cosart TF, Boyer MC, Neville HH, Muhlfeld CC, Amish SJ, Carim K, Narum SR, Lowe WH, Allendorf FW, Luikart G. Vive la résistance: genome-wide selection against introduced alleles in invasive hybrid zones. Proc R Soc B. 2016;283:20161380.

Landguth EL, Balkenhol N. Relative sensitivity of neutral versus adaptive genetic data for assessing population differentiation. Conserv Genet. 2012;13:1421–6.

Larson WA, Seeb LW, Everett MV, Waples RK, Templin WD, Seeb JE. Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). Evol Appl. 2014;7:355–69.

Leaché AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD, Linkem CW. Phylogenomics of Phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biol Evol. 2015;7:706–19.

LeCorre V, Kremer A. The genetic differentiation at quantitative trait loci under local adaptation. Mol Ecol. 2012;21:1548–66.

Leroy T, Roux C, Villate L, Boldénès C, Romiguier J, Paiva JAP, et al. Extensive recent secondary contacts between four European white oak species. New Phytol. 2017;214:865–78.

Lewontin RC. Genetic basis of evolutionary change. New York: Columbia University Press; 1974.

Lewontin RC, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics. 1973;74:175–95.

Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nat Genet. 2011;475:493–6.

Limborg MT, Helyar SJ, DeBruyn M, Taylor MI, Nielsen EE, Ogden R, Carvalho GR, Bekkevold D. Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). Mol Ecol. 2012;21:3686–703.

Lind BM, Friedline CJ, Wegrzyn JL, Maloney PE, Vogler DR, Neale DB, Eckert AJ. Water availability drives signatures of local adaptation in whitebark pine (*Pinus albicaulis* Engelm.) across fine spatial scales of the Lake Tahoe Basin, USA. Mol Ecol. 2017;26:3168–85.

Linløkken AN, Haugen TO, Mathew PK, Johansen W, Lien S. Comparing estimates of number of breeders Nb based on microsatellites and single nucleotide polymorphism of three groups of brown trout (*Salmo trutta* L.). Fish Manag Ecol. 2016;23:152–60.

Liu J, Shikano T, Leinonen T, Cano JM, Li M-H, Merilä J. Identification of major and minor QTL for ecologically important morphological traits in three-spined sticklebacks (*Gasterosteus aculeatus*). G3 Genes Genomes Genet. 2014;4:595–604.

Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. Inferring admixture histories of human populations using linkage disequilibrium. Genetics. 2013;193:1233–48.

Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, et al. Breaking RAD: an evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. Mol Ecol Resour. 2016;17:142–52.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet. 2003;4:981–94.

Luikart G, Kardos M, Hand BK, Rajora OP, Aitken SN, Hohenlohe PA. Population genomics. Cham: Springer; 2018.

Malenfant R, Coltman DW, Davis CS. Design of a 9K Illumina BeadChip for polar bears (*Ursus maritimus*) from RAD and transcriptome sequencing. Mol Ecol Resour. 2015;15:587–600.

Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008;24:133–41.

Matala AP, Ackerman MW, Campbell MR, Narum SR. Relative contributions of neutral and non-neutral genetic differentiation to inform conservation of steelhead trout across highly variable landscapes. Evol Appl. 2014;7:682–701.

Mazet O, Rodríguez W, Grusea S, Boitard S, Chikhi L. On the importance of being structured: instantaneous coalescence rates and a re-evaluation of human evolution. Heredity. 2016;116:362–71.

McCartney-Melstad E, Mount GG, Shaffer HB. Exon capture optimization in amphibians with large genomes. Mol Ecol. 2016;16:1084–94.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol. 2013;66:526–38.

Mckinney GJ, Larson WA, Seeb LW, Seeb JE. RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on breaking RAD by Lowry et al. (2016). Mol Ecol Resour. 2017;17:356–61.

McManus KF, Kelley JL, Song S, Veeramah KR, Woerner AE, Stevison LS, et al. Inference of gorilla demographic and selective history from whole-genome sequence data. Mol Biol Evol. 2016;32:600–12.

Milano I, Babbucci M, Cariani A, Atanassova M, Bekkevold D, Carvalho GR, et al. Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (*Merluccius merluccius*). Mol Ecol. 2014;23:118–35.

Miller MR, Brunelli JP, Wheeler PA, Liu S, Rexroad CE, Palti Y, et al. A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. Mol Ecol. 2012;21:237–49.

Morin PA, Parsons KM, Archer FI, Ávila-Arcos MC, Barrett-Lennard LG, Dalla Rosa L, et al. Geographic and temporal dynamics of a global radiation and diversification in the killer whale. Mol Ecol. 2015;24:3964–79.

Moritz C, Hillis DM. Molecular systematics: context and controversies. In: Hillis DM, Moritz C, Mable BK, editors. Molecular systematics. 2nd ed. Sunderland: Sinauer; 1996. p. 1–16.

Moura AE, Kenny JG, Chaudhuri R, Hughes MA, Welch AJ, Reisinger RR, et al. Population genomics of the killer whale indicates ecotype evolution in sympatry involving both selection and drift. Mol Ecol. 2014;23:5179–92.

Muhlfeld CC, Kalinowski ST, McMahon TE, Taper ML, Painter S, Leary RF, Allendorf FW. Hybridization rapidly reduces fitness of a native trout in the wild. Biol Lett. 2009;5:328–31.

Muhlfeld CC, Kovach RP, Jones LA, Al-Chokhachy R, Boyer MC, Leary RF, et al. Invasive hybridization in a threatened species is accelerated by climate change. Nat Clim Change. 2014;4:620–4.

Mullis KB, Faloona FA. Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. Methods Enzymol. 1987;155:335–50.

Nadachowska-Brzyska K, Burri R, Smeds L, Ellegren H. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. Mol Ecol. 2016;25:1058–72.

Nadeau NJ, Ruiz M, Salazar P, Counterman B, Medina JA, Ortiz-Zuazaga H, et al. Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. Genome Res. 2014;24:1316–33.

Nazareno AG, Bemmels JB, Dick CW, Lohmann LG. Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. Mol Ecol Resour. 2017;17:1136–47.

Nicolle D. A classification and census of regenerative strategies in the eucalypts (*Angophora, Corymbia* and *Eucalyptus* – Myrtaceae), with special reference to the obligate seeders. Aust J Bot. 2006;54:391–407.

Oswald JA, Overcast I, Mauck WM, Anderson MJ, Smith BT. Isolation with asymmetric gene flow during the nonsynchronous divergence of dry forest birds. Mol Ecol. 2017;26:1386–400.
Palsbøll PJ, Bérubé M, Allendorf FW. Identification of management units using population genetic data. Trends Ecol Evol. 2007;22:11–6.
Park L. Effective population size of current human population. Genet Res. 2011;93:105–14.
Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8:e1002967.
Puckett EE, Eggert LS. Comparison of SNP and microsatellite genotyping panels for spatial assignment of individuals to natal range: a case study using the American black bear (*Ursus americanus*). Biol Conserv. 2016;193:86–93.
Rajora OP, Eckert AJ, Zinck JWR. Single-locus versus multilocus patterns of local adaptation to climate in eastern white pine (*Pinus strobus,* Pinaceae). PLoS One. 2016;11:e0158691.
Ree RH, Hipp AL. Inferring phylogenetic history from restriction site associated DNA (RADseq). In: Hörandl E, Appelhans MS, editors. Next-generation sequencing in plant systematics. Königstein: International Association for Plant Taxonomy, IAPT; 2015. p. 1–24.
Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, et al. Linkage disequilibrium in the human genome. Nature. 2001;411:199–204.
Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R. A practical guide to environmental association analysis in landscape genomics. Mol Ecol. 2015;24:4348–70.
Robinson JA, Ortega-Del Vecchyo D, Fan Z, Kim BY, Vonholdt BM, Marsden CD, et al. Genomic flatlining in the endangered island fox. Curr Biol. 2016;26(9):1183.
Rougement Q, Gagnaire P-A, Perrier C, Genthon C, Besnard A-L, Launey S, Evanno G. Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and nonparasitic lamprey ecotypes. Mol Ecol. 2016;26:142–62.
Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74:5463–7.
Sansaloni C, Petroli C, Jaccoud D, Carling J, Detering F, Grattapaglia D, et al. Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. BMC Proc. 2011;5:P54.
Santure AW, De Cauwer I, Robinson MR, Poissant J, Sheldon BC, Slate J. Genomic dissection of variation in clutch size and egg mass in a wild great tit (*Parus major*) population. Mol Ecol. 2013;22:3949–62.
Schield DR, Adams RH, Card DC, Perry BW, Pasquesi GM, Jezkova T, et al. Insight into the roles of selection in speciation from genomic patterns of divergence and introgression in secondary contact in venomous rattlesnakes. Ecol Evol. 2017;7:3951–66.
Schoville SD, Bonin A, François O, Lobreaux S, Melodelima C, Manel S. Adaptive genetic variation on the landscape: methods and cases. Annu Rev Ecol Evol Syst. 2012;43:23–43.
Selkoe KA, Toonen RJ. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecol Lett. 2006;9:615–29.
Slate J, Visscher PM, MacGregor S, Stevens D, Tate ML, Pemberton JM. A genome scan for quantitative trait loci in a wild population of red deer (*Cervus elaphus*). Genetics. 2002;162:1863–73.
Spurgin LG, Wright DJ, van der Velde M, Collar NJ, Komdeur J, Burke T, Richardson DS. Museum DNA reveals the demographic history of the endangered Seychelles warbler. Evol Appl. 2014;7:1134–43.
Steane DA, Potts BM, McLean E, Prober SM, Stock WD, Vaillancourt RE, et al. Genome-wide scans detect adaptation to aridity in a widespread forest tree species. Mol Ecol. 2014;23:2500–13.
Steane DA, Potts BM, McLean E, Collins L, Prober SM, Stock WD, Vaillancourt RE, Byrne M. Genome-wide scans reveal cryptic population structure in a dry-adapted eucalypt. Tree Genet Genomes. 2015;11:33.
Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. 2017;49:303–9.

Therkildsen NO, Palumbi SR. Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Mol Ecol Resour. 2017;17:194–208.

Tiffin P, Ross-Ibarra J. Advances and limits of using population genetics to understand local adaptation. Trends Ecol Evol. 2014;29:673–80.

Vargas OM, Ortiz EM, Simpson BB. Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: *Diplostephium*). New Phytol. 2017;214:1736–50.

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, et al. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol Ecol. 2013;22:787–98.

Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.

White TA, Perkins SE, Heckel G, Searle JB. Adaptive evolution during an ongoing range expansion: the invasive bank vole (*Myodes glareolus*) in Ireland. Mol Ecol. 2013;22:2971–85.

Whitlock MC, Lotterhos KE. Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F_{ST}. Am Nat. 2015;186:S24–36.

Wright S. Evolution and the genetics of populations. Chicago: University of Chicago Press; 1978.

Inferring Demographic History Using Genomic Data

Jordi Salmona, Rasmus Heller, Martin Lascoux, and Aaron Shafer

Abstract Characterizing population histories has been a major focus in evolutionary and conservation biology for decades. Driven by a desire to understand population histories, researchers have been modeling simple demographic scenarios with genetic data since the 1970s. In the last decade, the availability of genomic data and the number of demographic inference methods have dramatically increased and constitute a continuously evolving sub-discipline within population genetics. Genome sequences—both reduced representation and whole-genome sequencing and re-sequencing—contain a trove of information related to population histories and permit reconstructing complex demographic scenarios. In combination with new powerful and flexible analytical methods, population demographic inference from genomic data has revealed surprising, dynamic, and conservation-relevant histories. This chapter discusses recent advancements in demographic inference made possible by genome sequence and new analytical tools. As the theory and models of demographic inference have matured, and data sets have grown, likewise has the recognition of limitations and confounding effects. We caution that the increasing

J. Salmona
Laboratoire Evolution and Diversité Biologique, UMR 5174, CNRS/Université Toulouse III Paul Sabatier, 118, route de Narbonne, Toulouse, Cedex 9 31062, France

Université de Toulouse, UMR 5174 EDB, Toulouse, France

R. Heller
Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, Copenhagen N 2200, Denmark

M. Lascoux
Department of Ecology and Genetics, Uppsala University, Uppsala 75 236, Sweden

A. Shafer (✉)
Forensic Science and Environmental and Life Sciences, Trent University, Peterborough, ON, Canada, K9J 7B8
e-mail: aaronshafer@trentu.ca

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2017_1,

sophistication of methods should not override the critical evaluation of the researcher. Demographic inferences with genomic data offer powerful windows into the past but we encourage users to recognize inherent limitations of model assumptions, use simulations to identify potential biases, and include complementary and supporting analyses.

Keywords Approximate-Bayesian computation • Coalescent • Effective population size • Genealogy • Haplotypes • Migration

1 Introduction and a Brief History

There is a long-standing and ongoing interest in understanding population histories. Stories about human expansion and dispersal patterns have captivated audiences for decades, while quantifying changes in effective population size (N_e) and migration rates are vital information for understanding organismal biology and informing resource conservation and management. On a broad scale the two most commonly invoked drivers of population change are climate or environmental change and anthropogenic disturbance (Barnosky et al. 2004; Nelson et al. 2006). Environmental changes are known to be a major factor in shaping biodiversity, both among and within species. For example, climate change and the ensuing habitat and vegetation changes are recognized to have a strong impact on biodiversity and populations (Parmesan and Yohe 2003; Thuiller 2007). In other cases, human activities have been shown to be the main driver of fluctuations in wild species (Fahrig 2003; Vitousek et al. 1997) and will continue to be for the foreseeable future (McKee et al. 2004). Consequently, from both an applied and basic research standpoint, there is considerable interest in characterizing demographic histories with the ultimate goal of identifying the factors shaping species distributions and population fluctuations over time.

Inferring demographic histories presents a practical challenge because the time frames addressed generally exceed human documentation, are patchily or sparsely represented in the fossil and pollen records, or cannot be inferred from recent trends. This leads to a considerable level of uncertainty or only enables broad inference surrounding population parameter estimates in lieu of alternative methods and data. In such cases, population genetics provides a powerful framework for addressing questions related to demography because it uses data from contemporary individuals to infer historical events. Population genetic approaches capitalize on the fact that historical demography strongly affects the genetic variation observed across the genomes of contemporary individuals.

The idea that molecular data contains information on the evolutionary history of populations traces back to the beginning of the twentieth century (e.g., Hirschfeld and Hirschfeld 1919). But not until the 1970s did population geneticists begin to develop statistical tools and summary statistics that could be used to infer

demographic history from genetic polymorphism data. These methods (Table 1) can be classified using three basic criteria:

1. The type of information used (e.g., summary statistics, site frequency spectrum, mutation and recombination rates);
2. The class and inheritance of genetic markers (e.g., cpDNA, mtDNA vs. nuclear, microsatellites loci vs. DNA sequences);
3. The model assumptions and inferable demographic scenarios.

The first genetic attempts to understand the demographic history of populations were reached through the use of summary statistics, specifically assessing deviations from expected values under an equilibrium model (e.g. Ewens 1972; Watterson 1974, Tajima 1989), measuring heterozygosity excess (Luikart and Cornuet 1998), and comparing the number of microsatellite alleles to the allelic size range (Garza and Williamson 2001; Table 1). These summary statistics were eventually packaged in easy-to-use software, and have become very popular notwithstanding their known limited statistical power (Hoban et al. 2014; Peery et al. 2012). Despite recent increase in the application of genome-scale data, many of these methods and approaches remain the backbone of demographic inference.

A key breakthrough came from the development of the coalescent theory (Box 1; Hudson 1983; Kingman 1982; Tajima 1983) that played a prominent role in demographic inference using population genetics. This was followed by an integration of coalescent theory and likelihood approaches describing the probability of the data given a particular model. Spurred on by the availability of sequence and microsatellite data and increased computing power, approaches that derived likelihoods based on classical population genetics or coalescent theory started to appear in the 1990s. These coalescent-based methods allowed estimating parameters, such as changes in effective population size (N_e), mutation (μ) and migration rates (M), and split times (T) under a specific demographic model (Beaumont 1999, 2003; Storz and Beaumont 2002; Wakeley and Hey 1997).

Box 1. The Coalescent

Until the 1980s the forward in time Wright-Fisher (WF) model (see Glossary) was the main population genetics model. The WF model, together with diffusion approximations, yielded results on, for example, the probability of fixation of new mutations and their expected time to fixation (or loss) under various conditions. While the WF model was a natural choice when trying to

(continued)

Table 1 Non-exhaustive list of demographic inference methods for (population) genomic data

Method	Used marker	Used information	Assumed model	Estimated parameters	Reference(s)
*HH*n	Mapped SNPs	Identity by descent/state segments	N_e fluctuations	N_i	(MacLeod et al. 2009, 2013)
PSMC	Long DNA strands	Coalescence time and recombination	N_e fluctuations	N_i, for $T > 10^4$	(Li and Durbin 2011)
MSMC	Long DNA strands	Coalescence time and recombination	N_e fluctuations	N_i, T	(Schiffels and Durbin 2014)
fastNeutrino	SNPs	Folded and unfolded SFS	N_e fluctuations	N_i	(Bhaskar et al. 2015)
Doris	Long range haplotypes	Identity by descent/state segments	N_e fluctuations and IM model	N_0, N_i, θ, T, m	(Palamara and Pe'er 2013; Palamara et al. 2012)
Inferring demography from IBS	Long range haplotypes	Identity by descent/state segments	IM model	N_0, N_i, θ, T, m	(Harris and Nielsen 2013)
TNSFS	SNPs	Segregating site/sample size = TNSFS	N_e recent growth	N_0, N_1, θ, T	(Chen et al. 2015)
TRACTS	Phased SNPs	Local ancestry tract	Non-admixed populations	m over time	(Gravel 2012)
diCal	Phased SNPs	Coalescence time and recombination	N_e fluctuations	N_i	(Sheehan et al. 2013)
PopSizeABC	Mapped SNPs	Folded 1D SFS + LD	N_e fluctuations	N_i	(Boitard et al. 2016)
Stairway plot	SNPs	Unfolded or folded 1D SFS	N_e fluctuations	N_i	(Liu and Fu 2015)
Linkage disequilibruim	SNPs	Runs of homozygosity and coalescence time	N_e fluctuations	N_i	(MacLeod et al. 2013)
bSFS	SNPs	SFS	N_e fluctuations and IM model	N_0, N_i, θ, T, m	(Lohse et al. 2016)
#	SNPs	2DSFS	IM model	N_0, N_i, θ, T, m	(Kern and Hey 2016)
Jaatha	SNPs	SFS, 2DSFS	N_e fluctuations and IM model	N_i	(Naduvilezhath et al. 2011)
aSFS	SNPs	SFS from several co-distributed species	N_e fluctuations	N_i	(Xue and Hickerson 2015)

(continued)

Table 1 (continued)

Method	Used marker	Used information	Assumed model	Estimated parameters	Reference(s)
δaδi	SNPs	SFS, 2DSFS, 3DSFS	N_e fluctuations and IM model	N_0, N_i, θ, T, m	(Gutenkunst et al. 2009)
fastsimcoal2	SNPs	SFS, 2DSFS, nDSFS	Any given model	Any given parameter	(Excoffier et al. 2013)
ABC	Any kind of marker	User defined	Any given model	Any given parameter	(Beaumont et al. 2002)

N_e effective population size, N_i effective population sizes across time, N_0 current effective population size, N_t ancient effective population size, T divergence time, θ $4N_e\mu$, m migration rate among populations, IM isolation with migration, LD linkage desequilibrium, SFS site frequency spectrum, SNP single nucleotitic polymorphism, ABC approximate Baeysian computation

Box 1 (continued)

predict future changes in genetic diversity, it was limited when it came to historical inferences. Here, the *n*-coalescent offered a simple and elegant solution and has become one of the most widely used population genetics models. The coalescent was proposed by Kingman (1982) as a continuous-time approximation to the Wright-Fisher model when the population size is large. Hudson (1983) and Tajima (1983) gave more intuitive derivations and there are excellent reviews (Hudson 1990; Nordborg 2001; Marjoram and Joyce 2010) and books (Hein et al. 2005; Wakeley 2008) on the coalescent but we briefly explain important aspects.

The *n*-coalescent models the ancestry of a sample of n sequences (from a single locus) backwards in time. It assumes that the sample size n is small compared to the total size of the population $2N$, with $2N$ reflecting the number of chromosomes in a diploid population. Coalescent events, where two contemporary sequences have a common ancestor, link the *n*-sequences in the genealogy. If the number of sequences is $2N$, the probability that two gene sequences have a common ancestor in the previous generation is $1/2N$. The probability that two genes among a sample of size n have a common ancestor in the previous generation is:

$$[n(n-1)]/2 \times 1/2N = n(n-1)1/4N.$$

So the probability that two of the n genes have a common ancestor t generations back in time is given by the geometric distribution:

$$[1-(n(n-1)/4N)]^{t-1} \times n(n-1)/4N.$$

As N is large the geometric distribution will converge to an exponential distribution: if one rescales the process by $2N$, then the coalescence times are exponentially distributed with mean $n(n-1)/2$. Figure 1 shows the

(continued)

Box 1 (continued)

coalescent process for a sample of six sequences taken from a population of size 2*N*. In this example it takes on average 2*N*/15 generations to go from six lineages to five lineages, 2*N*/10 to go from five lineages to four lineages, and so on until we reach the Most Recent Common Ancestor (MRCA) of the sample. Strikingly, the genealogy is dominated by the last event: roughly, half of the time is spent waiting for the last coalescent event! An important consequence of the coalescent model is the realization that gene genealogies are highly variable, hence, everything else being equal, two independent genes might have very different genealogies. This has a profound impact on the way we carry out population genetics inferences: in a single, random mating, constant size population following the standard coalescent, it is generally more informative to increase the number of loci rather than the number of individuals. Most populations are structured or have gone through population size changes so having a large, carefully sampled set of sequences is crucial to obtaining reliable demographic inferences in these cases.

To be of general interest and not simply a mathematical curiosity, a model must be robust to departures from its basic assumptions. A large number of studies have shown this to be a case of the *n*-coalescent. In particular,

(continued)

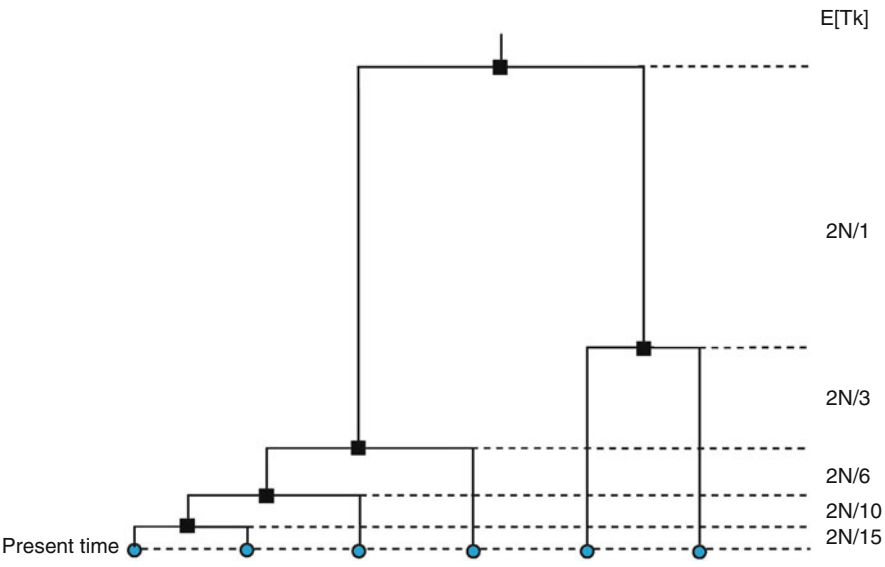

Fig. 1 Illustration of the standard coalescent model in a diploid population 2*N*. The expected coalescence times from *k* to *k*−1 lineages (for instance, 2*N*/15 is the expected time to go from 6 to 5 lineages) are on the *right*. $E(T)$ denotes the expected age of the most recent common ancestor given *k* lineages

Box 1 (continued)

processes occurring on two time scales, one fast and one slow, for example, partial selfing (Nordborg and Donelly 1997), seed banks (Kaj et al. 2001) or population structure with high migration (Wakeley 2008) will lead to gene genealogies with a topology very similar to that of the standard coalescent model. The gene genealogies are simply obtained by a rescaling of the effective population size. For example, selfing is a very fast process (with a probability of coalescence of ½) and outcrossing a slow one (the coalescent rate is now ½N) that therefore dominates the coalescence process. The coalescent with partial selfing is then the standard coalescent with $N_e = N(2-s)$ where s is the selfing rate (Nordborg and Donelly 1997). Selection or recombination cannot be easily accommodated and led to new processes, namely the ancestral selection graph (Krone and Neuhauser 1997) or the ancestral recombination graph (Griffiths and Marjoram 1997). The latter is at the root of new models of demographic inferences such as PSMC (Li and Durbin 2011) discussed later on in this chapter.

In parallel to recent computational power and data availability increase, models have grown in complexity and are becoming more realistic. For example, Whitlock and McCauley (1999) showed that most assumptions related to migration rate estimates under the island model were violated in nature. Recent developments have made possible the joint estimation of several populations' size changes (Gutenkunst et al. 2009; Hey and Nielsen 2004; Kuhner 2006) and/or inferences extended to multiple mutation models (Heled and Drummond 2008; Leblois et al. 2014; Nikolic and Chevalet 2014; Wu and Drummond 2011). Importantly, as we transition from genetic to genomic-based inference of population history and demography, many of these methods can simply be scaled up to accommodate the larger data sets.

Here we provide an overview of demographic inference in the age of genomics. Our focus will be on methods and interpretations and we draw on examples from wildlife and simulation studies. Many methodological advances are directly attributable to studies focusing on human history; however, genomic inference of human historical demography is a complex and controversial subject deserving a chapter of its own. We therefore focus on non-human examples. The remaining chapter is divided into three parts: (1) the application of traditional demographic approaches to genomic data; (2) the development of demographic approaches specific to large-scale genomic data; and (3) a discussion of the key parameters and limitations. The goal of this chapter is to provide a bridge linking the large number of available methods and growing number of genomic data sets. We hope this chapter provides a useful resource for both beginners and experienced researchers interested in understanding demographic and population histories.

2 The Site Frequency Spectrum (SFS) and Demographic Inference

Perhaps the most fundamental yet under-appreciated aspect of genomic data is the site frequency spectrum (SFS—Fig. 2), also referred to as the allele frequency spectrum. The SFS is the distribution of the allele frequencies of a given set of loci (often SNPs) in a population or sample (Evans et al. 2007; Fisher 1930; Kimura 1964; Wright 1938). Many important population genetic statistics such as Tajima's D and F_{ST} can be derived from the SFS (Nielsen et al. 2009; Wakeley 2008). With genome-wide data most of the methods used for estimating population genetic parameters from a few loci are not applicable or become computationally intractable when methods rely on the explicit representation of a coalescent tree for each site or locus (Nielsen and Slatkin 2013). Several methods based on the SFS allow for the consideration of thousands of sites simultaneously, without assuming that they all have the same tree (Nielsen and Slatkin 2013). The SFS represents the *distribution of the alleles* by classes of frequency (i.e. singletons, doubletons, etc.) and its statistical properties in terms of population demographic history are generally well understood under the coalescent and the diffusion models of neutral evolution (Fu 1995; Griffiths 2003; Griffiths and Tavaré 1998; Kimura 1955; Polanski et al. 2003; Živković and Stephan 2011). Two key advantages of SFS-based methods are that they can correctly estimate demographic models with a small number of polymorphisms (e.g., Adams and Hudson 2004; Shafer et al. 2015) and missing data can be accounted for (Gutenkunst et al. 2009).

Several types of SFS can be calculated leading to a few important terms and concepts. The number of populations or species considered is reflected in the terminology: the SFS which is denoted as 1dSFS for a single population can be extended for two populations (2dSFS or jointSFS) or even more than two populations, in which case we use the term joint-SFS. Both the 1dSFS and 2dSFS are easily visualized (Figs. 2 and 3) while the presentation of multiple (>2) dimensions limits visual presentation of the joint-SFS. The 1dSFS is simply the distribution of minor allele frequencies. The prior knowledge of the ancestral allelic state is further factored in with unknown ancestral allelic states termed folded SFS and known ancestral states termed unfolded SFS (Fig. 2). The unfolded SFS contains more information (see Fig. 3) but is dependent on knowing the ancestral allelic state that by definition is unknown. Using data from one to several closely related outgroups can *polarize* the mutations and be a proxy for inferring the ancestral and derived nucleotide states (Gutenkunst et al. 2009; Hernandez et al. 2007). Multiple species alignments together with ad-hoc tools (e.g., Paten et al. 2008a, b) are helpful in this regard and several approaches have been proposed to aid in the correct identification of ancestral states (Hernandez et al. 2007; Hwang and Green 2004; Matsumoto et al. 2015). One can also include an error term in the model to quantify the uncertainty in the ancestral state inference (Burgarella et al. 2015). Because identifying ancestral states is still subject to several limitations,

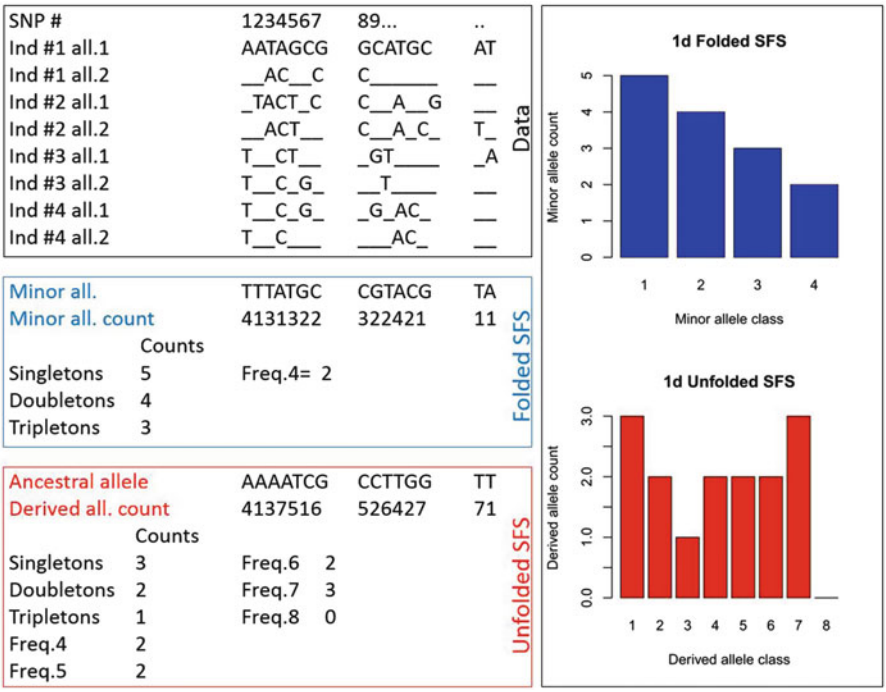

Fig. 2 Calculating and representing 1-dimensional site frequency spectrum (1dSFS) with a simple dataset. The data are presented by 15 variants (SNPs) typed for four diploid individuals. The folded 1dSFS is based on the counts of the minor alleles' frequency in the sample (*blue bar plot*). The unfolded 1dSFS calculated based on the counts of the derived alleles' frequency in the sample (*red bar plot*). *All* allele, *Freq* frequency

sources of bias and uncertainties, analyzing the folded-SFS is quite common and often incurs fewer sources of errors (e.g., Qiu et al. 2015).

For "n" diploid samples, the unfolded 1dSFS will be a vector ($2n + 1$) of the proportion of sites carrying "k"-mutations and can include counts of ancestral monomorphic state (not shown in Fig. 1; see Fig. 3). In other words, the derived allelic class no. 8 in Fig. 2 represents homozygous derived alleles, while class no. 0 (not shown) represents homozygous ancestral alleles. Changes in N_e will distort the gene genealogies and impact the frequency distributions across the SFS, and this is what is used to infer population size changes. For example, an expanding population (say a species colonizing recently deglaciated terrain) will have more singletons in the 1dSFS than a population of constant size. Conversely, a declining population will show a deficit of singletons. By fitting the frequency spectrum expected under a particular demographic model to the observed SFS, relative changes in N_e can be inferred (Boitard et al. 2016; Excoffier et al. 2013; Gutenkunst et al. 2009; Liu and Fu 2015).

For a 2dSFS with N individuals, the plot (Fig. 3) reflects the total number of segregating sites in which the derived allele (or minor allele if folded SFS) is

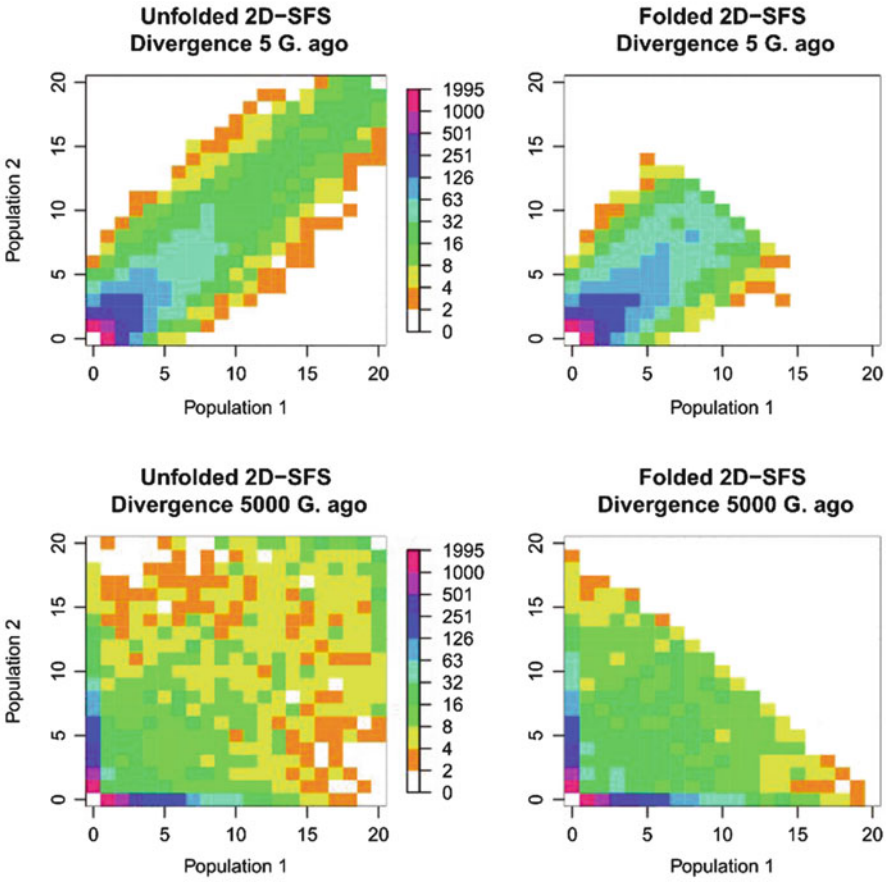

Fig. 3 Two dimensional site frequency spectrum (2D-SFS) obtained from simulated populations diverging recently (5 generations ago) and historically (5,000 generations). Plots based on 10,000 SNPs and 10 diploid samples per population

observed with the corresponding frequency in each population. Interpreting the 2dSFS is somewhat less intuitive as each cell—which corresponds to $2n \times 2n$ combination of alleles with counts in population 1 and counts in population 2—must be connected to a scale bar. Visual interpretations aside, the 2dSFS contains an incredible amount of information about population size fluctuations, population divergence, migration, and selection. For example, in recently diverged populations, the density of alleles in the 2dSFS will be concentrated along the diagonal reflecting a recent, shared history (Fig. 3, upper plots). In contrast, for highly divergent populations, most of the SFS density is concentrated along the axes as most alleles are private (Fig. 3, lower plots). Similarly, migration from one population to another will result in increased shared alleles (see Fig. 2; Gutenkunst et al. 2009).

Calculating SFS is highly sensitive to the SNP calling approach and quality of the data (e.g., coverage, error rate, bioinformatics pipeline; Han et al. 2013; Nielsen et al. 2012; Shafer et al. 2017), and the outgroup species or method used to polarize the derived allele. To tackle the biases originating from genotype calling, the SFS can be estimated directly from a genotype likelihood considering both coverage and base quality (Korneliussen et al. 2014; Nielsen et al. 2012). Several approaches have been proposed to infer relaxed demographic history from SFS (Table 1). The Stairway plot (Liu and Fu 2015) and PopSizeABC (Boitard et al. 2016) make use of 1dSFS to infer the population size fluctuations over time of a single population. Although using very different approaches, both methods allow preliminary analysis of the populations' demographic histories and are useful for developing hypotheses and priors for building and comparing more complex scenarios. These methods have in particular shed light on the ancestry of killer whale (*Orcinus orca*) ecotypes (Foote et al. 2016) and cattle breeds (Boitard et al. 2016). Several approaches have been proposed to resolve speciation and divergence events under the isolation with migration model (Naduvilezhath et al. 2011; Kern and Hey 2016). Furthermore, model-based approaches implemented in $\partial a \partial i$ (Gutenkunst et al. 2009) and fastsimcoal2 (Excoffier et al. 2013) offer the possibility to compare complex scenarios from 1d and joint-SFS, including population size changes, population splits, and divergence, and enable estimating parameters of interest like migration. Methods analyzing the SFS have gained recent popularity—particularly due to their fast computing ability—allowing the comparison of very complex scenarios for large range of organisms. For instance, Schubert et al. (2014) successfully applied $\partial a \partial i$ in combination with other approaches to date horse (*Equus ferus* ssp. *caballus*) domestication, Arnold et al. (2015) unraveled the evolution of tetraploidy in *Arabidopsis arenosa*, and Malaspinas et al. (2016) resolved population history from rather complex models. Recently developed, the aggregate SFS of several co-distributed species (aSFS) provides a useful framework to test for simultaneous demographic changes in response to environmental changes (Xue and Hickerson 2015).

3 Approximate Bayesian Computation and Demographic Inference

We have introduced coalescent theory (Box 1) and alluded to Approximate Bayesian Computation (ABC) in the previous section. A key development in population genetics has been the merger of these two areas. While coalescent simulations are extremely fast, the likelihood estimates are computationally taxing (Marjoram and Tavaré 2006). ABC methods rely on summary statistics (but see Sousa et al. 2009) and simulations, not likelihoods, thus circumventing the need for likelihood calculations in population genetics (Beaumont et al. 2002). In the case of demographic inference, a common approach has been to use a coalescent sampler (e.g., ms;

Hudson 2002) under a wide distribution of parameter values (referred to as a prior in the Bayesian framework). For example, in a simple isolation model one might choose to run one million coalescent simulations with each simulation reflecting a different split time taken from a distribution within the prior. Each simulation has a series of summary statistics calculated, for example heterozygosity and F_{ST}, that are then compared to the same summary statistics generated from the observed data. A subset of simulations close to the observed value(s) are retained and used to infer the demographic parameters of interest. Model selection via Bayesian posterior probabilities can then be conducted, but this is an area of active debate (Robert et al. 2011).

The above is a cursory overview of ABC and there are many ABC variants and considerations users need to be aware of (see Beaumont 2010; Bertorelle et al. 2010; Csilléry et al. 2010). However, both whole-genome and reduced representation sequencing approaches, along with SNPs are compatible with ABC and the coalescent (Li and Jakobsson 2012; Shafer et al. 2015). Virtually any demographic scenario can be tested, provided it can be simulated. Collectively, this has led to the growing popularity and increased availability of user-friendly software devoted to demography (Cornuet et al. 2014; Pavlidis et al. 2010; Wegmann et al. 2010; Csilléry et al. 2012; Boitard et al. 2016). Another attractive aspect is that the critical coalescent parameters Rho (ρ) and Theta (Θ), whose distribution is generally poorly known, can be given priors and estimated in the ABC framework. In experiments where sequencing error might be a problem, error models can be used to transform simulated data and this appears to improve ABC parameter estimates (Veeramah et al. 2015). This versatility is a key attribute of ABC-based demographic inference.

Turning millions of sequence reads into a handful of summary statistics does have its limitations. For short-read genomic data and SNPs, while ρ can effectively be ignored, coalescent simulations based on a reasonable Θ produce primarily monomorphic sites, wasting considerable computer resources or forcing shortcuts that lead to biased estimates (Shafer et al. 2015). The available independent summary statistics are limited with these data as the more powerful linkage-based statistics are not available. Consequently, historical changes in N_e such as bottlenecks are difficult to infer (Shafer et al. 2015, 2017), but it should be noted that contemporary N_e estimates are tractable with linkage-based approaches (Waples et al. 2016). For large-scale re-sequencing data, despite the speed of the coalescent, the number of replicate simulations does become a limiting factor. For example, Li and Jakobsson (2012) required 7,000 computer hours to complete 50,000 simulations reflecting 10,000 genome regions of size 100 kb. Despite the availability of standalone packages, no general rules apply and each data type and set of models requires its own unique exploration (Bertorelle et al. 2010). ABC is relatively quick and easy to run, but this to some extent blurs the underlying intricacies of the model and data, and thus as in all models, researchers should be aware of potential biases and limitations and customize the ABC to their data and system.

4 The Development of Demographic Approaches Specific to Large-Scale Genomic Data

The use of the SFS and ABC existed prior to the large-scale generation of genomic data, and many of the standard methods were simply scaled up. However, new methods dependent on genomic data have emerged, often with very exciting results. Most notably was the publication of pairwise sequentially Markovian coalescent (PSMC) method (Li and Durbin 2011). It was built upon the theoretical work of Wiuf and Hein (1999), McVean and Cardin (2005), and Marjoram and Wall (2006) that developed sequential Markovian coalescent algorithms to approximate the coalescent model for large chromosome fragments undergoing recombination. This method was revolutionary as it uses information from just one diploid genome. The key feature of PSMC is that it relies on the coalescence to estimate the time to the most recent common ancestor of two alleles at a given locus. Because the rate of coalescent events is inversely proportional to N_e, N_e can be estimated. Conceptually the method simply moves along chromosome encountering older and younger tracts that reflect different population histories resulting in estimates of N_e over time. Looking at two examples in Fig. 4, in example 1, Nadachowska-Brzyska et al. (2015) applied PSMC to 38 different bird whole genomes, showing expansions and contractions coinciding with climate cycles have been a common feature of many bird species during the Quaternary period (Fig. 4a). Importantly, the divergent N_e paths are consistent with speciation and lineage splitting and corroborated conservation listings based on long-term population declines. Similarly, using ancient DNA and PSMC, Palkopoulou et al. (2015) documented the demographic decline of wooly mammoth.

Subsequently, Schiffels and Durbin (2014) tweaked the method to allow multiple genomes, known as multiple sequentially Markovian coalescent (MSMC). One of the nice aspects of MSMC is that it allows for preliminary dating of population (Warren et al. 2015) and species splits (Wang et al. 2016) because the N_e estimates converge when the sampled individuals share a common ancestor. In example 2 (Fig. 4b), Warren et al. (2015) showed N_e trajectories of five vervet subspecies based on whole-genome re-sequencing consistent with their known geographic ranges and histories of isolation, ultimately supporting relative stable populations over time. Although robust SMC inferences require relatively high coverage (18×) and low missing data (low 25%; Nadachowska-Brzyska et al. 2016), MSMC requires phased data (Schiffels and Durbin 2014) and sudden and recent estimates in N_e are difficult to recover with the PSMC (Li and Durbin 2011). In addition, all estimates assuming panmixia (thus impacting more than MSC) are strongly influenced by population structure (Box 2). That said, these methods have really galvanized the population genetics community, allowing estimates of the long-term population and species history via analysis of genetic variation across the entire genome.

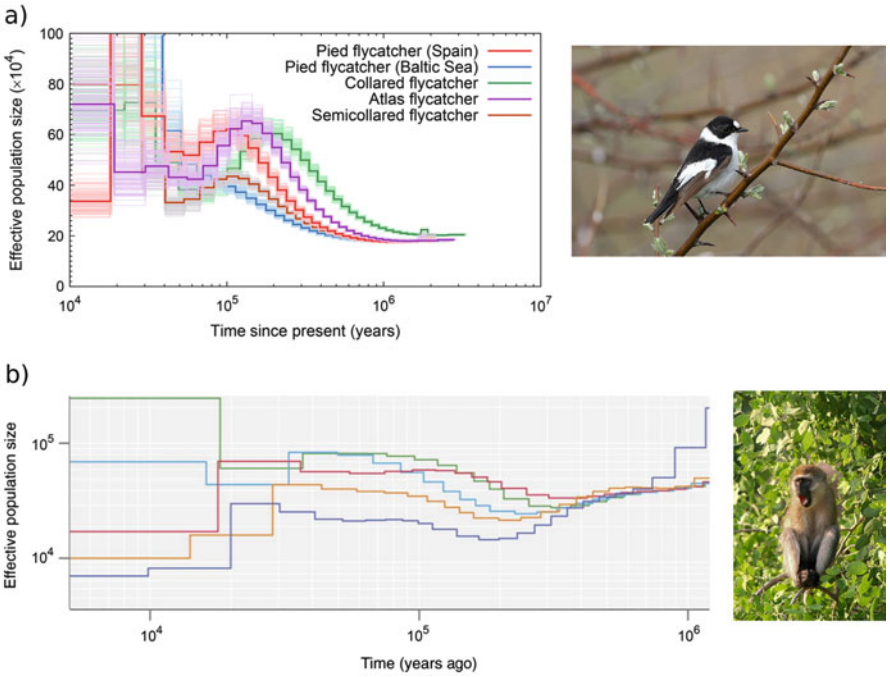

Fig. 4 Two example studies using the pairwise sequentially Markovian coalescent (PSMC) and multiple sequentially Markovian coalescent (MSMC) methods. (**a**) Multiple PSMC plots from flycatcher species from (Nadachowska-Brzyska et al. 2016). (**b**) MSMC plots of vervet subspecies from (Warren et al. 2015)

Box 2. The Problem of Population Structure

Population structure can confound analyses of population size changes over time. The basic problem is illustrated in Fig. 5. In a structured population, the probability of coalescence changes dynamically over time as a function of the migration process. Therefore, structure forces the genealogy of a sample to deviate from a standard, non-structured coalescent process.

Even though the potentially confounding effect of population structure on demographic inference was identified as early as Wakeley (1999), the majority of methods developed to detect population size change rely on the assumption that samples were obtained from populations that can be approximated by a Wright–Fisher model (i.e., assuming panmixia). In the real world, population structure is nearly ubiquitous and most natural populations form spatial networks interconnected through gene flow.

An attempt to assess the effect of a very common type of population structure (Isolation-by-distance; IBD) came relatively recently (Leblois

(continued)

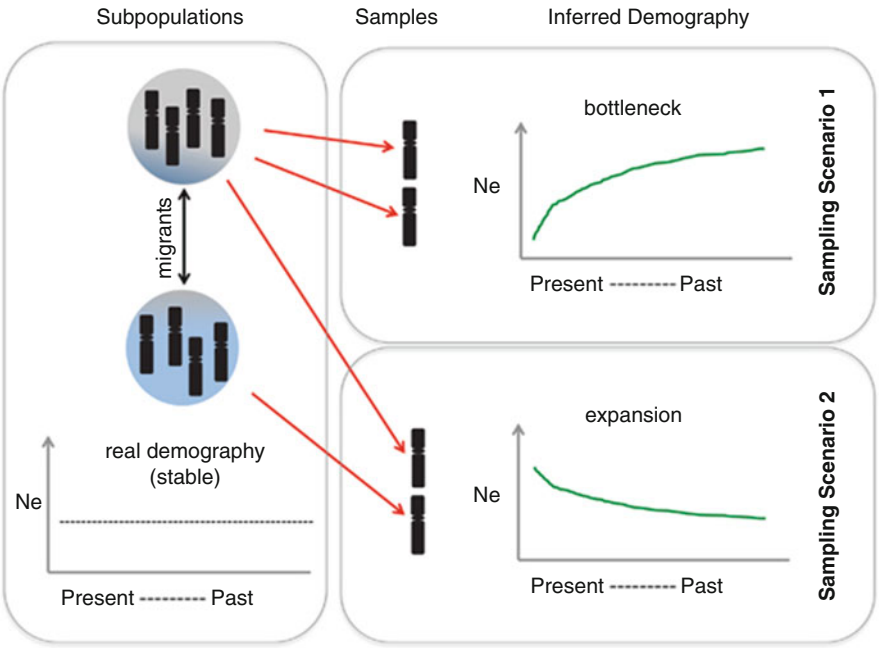

Fig. 5 A schematic representation of how population structure and sampling influences demographic inference. Figure taken from Orozco-terWenge (2016)

Box 2 (continued)

et al. 2006). Using a simulation-based approach, Leblois et al. (2006) demonstrated that IBD frequently leads to spurious demographic inferences. Städler et al. (2009) showed that population structure influences the expectations of summary statistics and could lead to artifacts of population contraction or growth. The development of full-likelihood coalescent-based methods and later the Bayesian skyline plot methods (Drummond et al. 2006) led to widespread interest in inferring population size change from genetic data, and many studies have applied these methods with varying degrees of attention to the possible confounding effect of structure. These methods share the assumption of panmixia and no immigration outlined above (but see Kuhner 2006). The persistent structure effect in these more sophisticated methods was only recently revisited by (Nielsen and Beaumont 2009; Chikhi et al. 2010) and Heller et al. (2013) using simulated data under various stationary but structured population scenarios. These studies quantified the structure effect and identified parameter ranges and sampling schemes under which false bottleneck signals are likely. Similarly, Li and Durbin (2011) were

(continued)

Box 2 (continued)

aware that the N_e estimates from the PSMC were strongly influenced by population structure. Despite having been known for a long time and repeatedly demonstrated, the structure effect is frequently ignored in PSMC applications and others, possibly because it is mistakenly believed that structure is not an issue when a single genome is used.

The structure effect in PSMC has been investigated in a series of recent studies by Mazet et al. (2015, 2016). These authors developed a statistical approach to disentangle structure from population size change using the information of one individual in the case of two simple models (Mazet et al. 2015). In addition they introduced the IICR (inverse instantaneous coalescence rate), equivalent to effective population size in panmictic models, but potentially "misleading in structured models" (Mazet et al. 2016). Mazet et al. (2016) showed that any PSMC result showing changes in population size has an analogous scenario of structured populations with changes in migration.

An important pattern that has repeatedly emerged is that the sampling scheme interplays in complex, yet readily explained ways with population structure to create confounding demographic signals. As a general observation, sampling only one deme tends to increase the strength of spurious signals (Chikhi et al. 2010; Heller et al. 2013; Peter et al. 2010; Städler et al. 2009). Conversely, sampling individuals from several demes tended to reduce the occurrence of spurious signals. However, when comparing pooled (several individuals from several demes, but not all demes) and scattered sampling (one individual from each deme) of structured populations, it is hard to make direct and general recommendations. A recent study showed that scattered sampling is less prone to falsely detecting bottlenecks, but also has reduced power to detect those (Heller et al. 2013). Several strategies have been proposed to circumvent or reduce the confounding effect of structure. (Chikhi et al. 2010; Heller et al. 2013; Städler et al. 2009) all suggested population sampling strategies to minimize structure confounding effects. Further, alternative approaches offer the possibility to specifically test the relative fit of structured and nonstructured models using ABC (Csilléry et al. 2010; Peter et al. 2010; Wegmann et al. 2010; Heller et al. 2012) or the joint-SFS (Excoffier et al. 2013; Gutenkunst et al. 2009). Including unsampled populations (Beerli 2004) in the model is also advisable when an influx of genetic material is suspected. In summary, structure is ubiquitous in nature yet poorly reflected in the history of methods for inferring population size changes. As a bare minimum, all studies undertaking demographic inference should apply some means of testing the sensitivity of their results to population structure.

Prior to PSMC and MSMC, animal breeders had been interested in using whole-genome sequence and large-scale SNPs to identify regions of the genome that are identical-by-descent. Such regions are manifested in long stretches of identical sequence termed runs of homozygosity (ROH). Mating between individuals with a more recent common ancestor results in longer ROH (Kardos et al. 2015). ROH are of particular interest in studying inbreeding depression and its genetic basis; in the demographic context, however, many short ROH are indicative of small N_e in the past, whereas long ROH are indicative of a recent decrease in N_e (Kardos et al. 2016). Simply quantifying the size and distribution of ROH gives clear insights into demographic history (Kirin et al. 2010). Macleod et al. (2013) extended this and developed a coalescent model relying on ROH. This model summarizes linkage disequilibrium among ROH that is used to infer N_e. The model works on unphased single and multiple diploid genomes, and works well for more recent estimates of N_e. This method lends itself well to the use of SNPs, but in natural populations $>$100,000 SNPs and SNPs from a closely related species, if not conspecific, are required to accurately recover the ROH landscape (Shafer et al. 2016).

5 Identifying and Recognizing Limitations

"Essentially, all models are wrong, but some are useful" (Box et al. 1987). Population genetics models are not an exception to this rule. Models typically assume panmixia and non-overlapping generations while this is seldom true in nature (but see Moran 1958). Furthermore most methods developed to detect population size change assume a single population with no structure (see Box 2), a unique population size change, and a typical Poisson distribution of reproductive success (random mating). Moreover, there is a confluence of genetic signatures that could be the result of different evolutionary processes such as population structure, admixture events, selection, mutation and recombination, and population size change. Thus, to conclude that a specific demographic history has shaped the analyzed genomic data, one must first reject reasonable alternative explanations based on the model's known limitations.

An important take home message is that the use of genomic data does not necessarily lead to a more accurate demographic inference (Mazet et al. 2016). If the model assumed is fundamentally miss-specified, it will lead to a misleading result, and increasing the amount of data simply provides increased precision for an inaccurate parameter. Best-practice would be to first address the following questions:

1. Is the experimental design appropriate for addressing a demographic question?
2. To what extent do the genomic data or system violate the key assumptions of the model?
3. Is the assumed model *somewhat* realistic given the sample and the species biology?

Sampling should be planned considering the major impact it has on statistical power and the putative structure of the population. Similarly, consideration of the impact of unsampled or "ghost populations" (Beerli 2004) on the SFS and summary statistics should be factored into the model. Unsampled populations are likely the case for many natural populations and including such a ghost population in the model would be advisable (Excoffier et al. 2013). Simulating data to test the model is generally a good idea as it will allow for the robustness of results to be fully assessed (Excoffier et al. 2013; Excoffier and Foll 2011; Li and Jakobsson 2012; Shafer et al. 2015). These two later points—modeling unsampled populations and simulating data—are likely not appealing for research groups interested in quick turnaround times, but are critical for proper evaluation and could easily become standard practice with only adding a few additional weeks to studies.

Earlier we alluded to *misspecification of models*. It is important to recognize that demographic models require parameter values that are generally provided by the user. This might be prior to running the model, or after when converting coalescent units to biological values. Inaccurate generation times, mutation rates or specified demographic scenario will impede accurate inferences at they are factored into Θ and ρ that are central to most demographic models. While we have saved this section to the end, in many ways it is the most important.

Generation time. Despite having an understandable definition (see Glossary), generation time is typically an unknown parameter in natural populations of long-lived animals and plants whose reproduction span can extend over many decades. Estimating generation time requires long-term behavioral and demographic studies and often equate to best guesses. A range of values taken from ecological parameters or from a closely related species can be applied with various formulas proposed (e.g., Bienvenu et al. 2013). It should be noted that generation time is used as a scaling parameter, meaning it simply gets factored into the estimate post-hoc, but can alter time estimates if inaccurate.

Mutation rate and models. Mutation models are generally unknown, but likelihood-based approaches devoted to their estimation are available (Posada and Crandall 1998). Many of these substitution models have been scaled up (Carvajal-Rodríguez 2008) and are most applicable to forward-in-time simulations. The uncertainties associated with mutation rate (μ) subject to large variation across lineages and marker type are numerous (Ho 2014; Moorjani et al. 2016). Therefore, mutation rate uncertainty is often a main driver behind large confidence intervals associated with time estimates. For most species and type of markers, mutation rate is unknown, and can be approximated using values of closely related species (e.g., Heller et al. 2012) or subjected to a large prior (Storz and Beaumont 2002). When possible mutation rate can be estimated using dated ancient DNA to reduce dating uncertainties (e.g. Allentoft et al. 2015; Orlando et al. 2013). Dating results should thus always be interpreted cautiously, and presented together with mutation model and rate parameter priors.

Recombination rate. Recombination rate (r) can be effectively ignored with unlinked SNP and RADseq data (Shafer et al. 2015). With long contiguous sequences, improper recombination rates will adversely impact parameter estimates

(Li and Jakobsson 2012). Similar to mutation rates, having a large prior or treating r as a nuisance parameter are appropriate courses of action. While debatable, r is more system and genome specific than μ and thus requires special attention as it will impact parameter estimates and accuracy.

Simulations and iterations. Both SFS and ABC methods will require users to decide as to the number of simulations or iterations to perform. Assessing model convergence and prior distributions should be regularly conducted, even at the cost of added computation time. There is no golden rule as to the number of simulations but they will likely number in the tens of thousands if not more (Li and Jakobsson 2012; Robinson et al. 2014; Shafer et al. 2015).

6 Concluding Remarks and Future Prospects

Methods aimed at inferring demographic history can generally not retrieve demographic events older than $2N_e$ generations ($4N_e$ for diploid species) as power fades rapidly upon moving back in time. This is because most alleles will coalesce in the recent past: in the standard coalescent all alleles except two coalesce before $2N_e$ generations ago, but it takes more than $2N_e$ generations to reach the MRCA, with only a few independent lineages probing deep the past (Li and Durbin 2011; Mazet et al. 2015). The advent of genomic sequences, and for example, development of the SMC approaches (Li and Durbin 2011; Schiffels and Durbin 2014; Sheehan et al. 2013) has taken advantage of the fact that hundreds of thousands of independent loci have independent coalescent histories, allowing reconstruction of demographic histories up a million years ago. The increasing panel of approaches have gained in complexity and realism (Table 1), now allowing for multiple size changes that fit more closely to the population fluctuations expected over climate oscillations (e.g. Boitard et al. 2016; Nikolic and Chevalet 2014; Schiffels and Durbin 2014; Wu and Drummond 2011). Moreover, several methods now allow for estimation of demographic parameters within the framework of complex models integrating population splits, multiple population, gene-flow, admixture events, bottleneck and growth (e.g., Excoffier et al. 2013; Gutenkunst et al. 2009).

This chapter attempts to provide a bridge linking the large number of available methods and growing number of genomic data sets. The SFS is a fundamental summary statistics that is often not fully understood by practitioners; therefore, we focused a lot on the SFS as it is a central aspect of demographic inference using genomic data. Both the coalescent and ABC methods also featured prominently and are reflective of current applications and popularity. Approaches like forward-in-time simulations were not examined, simply because of computational limitations and lack of widespread use although that is likely to change in the near future. We hope in subsequent versions of this chapter they will feature more prominently.

Glossary

Approximate Bayesian computation (ABC) compares summary statistics from observed and simulated data to make demographic and statistical inferences. ABC does not rely on computing a likelihood-function.

Bottleneck a massive and temporary reduction in (effective) population size that results in an associated reduction of genetic diversity.

Genetic drift changes in the frequency of alleles due to random mating (and allele segregation in diploids). Changes are more pronounced in small populations.

Coalescent theory mathematical model governing the expected distribution of coalescence times back to a common ancestor in a population sample.

Diffusion approximation approximation of the Wright-Fisher (WF) model that leads to a continuous time stochastic process that is easier to study mathematically. It is used to derive useful formulas such as the expected time to fixation of a mutation.

Divergence time (T) estimated divergence time between two populations measured as the number of generations, typically divided by $2N_e$.

Effective population size (N_e) the size of an idealized (Wright-Fisher) population with the same amount of genetic drift as the given real population. In most organisms, effective size is less than census size because of factors such as overlapping generations, reproductive inequality, and sex bias.

Genealogy the ancestral relationship, for a particular segment of the genome, among sampled chromosomes. This takes the form of a branching tree for non-recombining data, but becomes a tangled graph (the "ancestral recombination graph") with recombination.

Generation time is the average interval between identical life history stages across successive generations. Generation time is often expressed in years.

Migration (M) is the average number of migrants entering each population per generation defined as $4N_em$ where m is the proportion of individuals per generation in each population that are immigrants.

Recombination the process of exchanging genetic material between homologous chromosomes during meiosis resulting in new combinations of alleles in the resulting gametes.

Rho (ρ) is the population-scaled recombination rate defined as $4N_er$ in diploid organisms.

Panmictic population a population in which all pairs of individuals are equally likely to mate.

Site frequency spectrum (SFS) also called the allele frequency spectrum, is the distribution of the allele frequencies of a given set of loci in a sample, and is often visualized as a histogram.

Tajima's D a summary statistic that compares two estimators of the population-scaled mutation rate Θ to detect departures from the standard coalescent model. Departures can reflect demography or selection.

Theta (Θ) is the population-scaled mutation rate equal to $4N_e\mu$ in diploid organisms. It is the product of the N_e and mutation rate μ and measures the capacity of a population to maintain genetic variability. Among organisms of similar μ, it functions as a measure of relative effective population size.

Wright-Fisher model is a discrete-time model of stochastic reproduction (see also genetic drift) that assumes a population of size N, random mating, and non-overlapping generations.

References

Adams AM, Hudson RR. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics. 2004;168:1699–712.

Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, et al. Population genomics of Bronze Age Eurasia. Nature. 2015;522:167–72.

Arnold B, Kim S-T, Bomblies K. Single geographic origin of a widespread autotetraploid Arabidopsis arenosa lineage followed by interploidy admixture. Mol Biol Evol. 2015;32:1382–95.

Barnosky AD, Koch PL, Feranec RS, Wing SL, Shabel AB. Assessing the causes of late Pleistocene extinctions on the continents. Science. 2004;306:70–5.

Beaumont MA. Detecting population expansion and decline using microsatellites. Genetics. 1999;153:2013.

Beaumont MA. Estimation of population growth or decline in genetically monitored populations. Genetics. 2003;164:1139–60.

Beaumont MA. Approximate Bayesian computation in evolution and ecology. Annu Rev Ecol Evol Syst. 2010;41:379–406.

Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002;162:2025–35.

Beerli P. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. Mol Ecol. 2004;13:827–36.

Bertorelle G, Benazzo A, Mona S. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol Ecol. 2010;19:2609–25.

Bhaskar A, Wang YXR, Song YS. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Res. 2015;25(2):268–79. doi:10.1101/gr.178756.114.

Bienvenu F, Demetrius L, Legendre S. A general formula for the generation time. ArXiv Prepr. 2013:ArXiv13076692.

Boitard S, Rodriguez W, Jay F, Mona S, Austerlitz F. Inferring population size history from large samples of genome-wide molecular data-an approximate Bayesian computation approach. PLoS Genet. 2016;12:e1005877.

Box GE, Draper NR, et al. Empirical model-building and response surfaces. New York: Wiley; 1987.

Burgarella C, Gayral P, Ballenghien M, Bernard A, David P, Jarne P, et al. Molecular evolution of freshwater snails with contrasting mating systems. Mol Biol Evol. 2015;32:2403–16.

Carneiro M, Afonso S, Geraldes A, Garreau H, Bolet G, Boucher S, Tircazes A, Queney G, Nachman MW, Ferrand N. The genetic structure of domestic rabbits. Mol Biol Evol. 2011;28:1801–16.

Carvajal-Rodríguez A. GENOMEPOP: a program to simulate genomes in populations. BMC Bioinforma. 2008;9(1):223.

Chen H, Hey J, Chen K. Inferring very recent population growth rate from population-scale sequencing data: using a large-sample coalescent estimator. Mol Biol Evol. 2015;32 (11):2996–3011. doi:10.1093/molbev/msv158.
Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. Genetics. 2010;186:983.
Cornuet J-M, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, Leblois R, Marin J-M, Estoup A. DIYABC v2. 0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. Bioinformatics. 2014;30:1187–9.
Csilléry K, Blum MG, Gaggiotti OE, François O. Approximate Bayesian computation (ABC) in practice. Trends Ecol Evol. 2010;25:410–8.
Csilléry K, François O, Blum MG. abc: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol. 2012;3:475–9.
Drummond AJ, et al. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006;4(5): e88.
Evans SN, Shvets Y, Slatkin M. Non-equilibrium theory of the allele frequency spectrum. Theor Popul Biol. 2007;71:109–19.
Ewens WJ. The sampling theory of selectively neutral alleles. Theor Popul Biol. 1972;3:87–112.
Excoffier L, Foll M. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics. 2011;27:1332–4.
Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS Genet. 2013;9:e1003905.
Fahrig L. Effects of habitat fragmentation on biodiversity. Annu Rev Ecol Evol Syst. 2003;34:487–515.
Fisher RA. The distribution of gene ratios for rare mutations. Proc Roy Soc Edinburgh. 1930;50:205–22.
Foote AD, Vijay N, Ávila-Arcos MC, Baird RW, Durban JW, Fumagalli M, Gibbs RA, Hanson MB, Korneliussen TS, Martin MD, et al. Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. Nat Commun. 2016;7:11693.
Fu Y-X. Statistical properties of segregating sites. Theor Popul Biol. 1995;48:172–97.
Garza JC, Williamson EG. Detection of reduction in population size using data from microsatellite loci. Mol Ecol. 2001;10:305–18.
Gravel S. Population genetics models of local ancestry. Genetics. 2012;191:607–19.
Griffiths RC. The frequency spectrum of a mutation, and its age, in a general diffusion model. Theor Popul Biol. 2003;64:241–51.
Griffiths RC, Marjoram P. An ancestral recombination graph. In: Donnelly P, Tavar'e S, editors. Progress in population genetics and human evolution, IMA volumes in mathematics and its applications, vol 87. New York: Springer; 1997. p. 100–117.
Griffiths RC, Tavaré S. The age of a mutation in a general coalescent tree. Stoch Models. 1998;14:273–95.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5:e1000695.
Han E, Sinsheimer JS, Novembre J. Characterizing bias in population genetic inferences from low coverage sequencing data. Mol Biol Evol. 2013;31(3):723–35. doi:10.1093/molbev/mst229.
Harris K, Nielsen R. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 2013;9:e1003521.
Hein J, Schierop MH, Wiuf C. Gene genealogies, variation and evolution. A primer in coalescent theory. Oxford, UK: Oxford University Press; 2005.
Heled J, Drummond AJ. Bayesian inference of population size history from multiple loci. BMC Evol Biol. 2008;8:289.

Heller R, Bruniche-Olsen A, Siegismund HR. Cape buffalo mitogenomics reveals a Holocene shift in the African human–megafauna dynamics. Mol Ecol. 2012;21:3947–59.
Heller R, Chikhi L, Siegismund HR. The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. PLoS One. 2013;8:e62992.
Hernandez RD, Williamson SH, Bustamante CD. Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol. 2007;24:1792–800.
Hey J, Nielsen R. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics. 2004;167:747–60.
Hirschfeld L, Hirschfeld H. Serological differences between the blood of different races: the results of researches on the Macedonian front. Lancet. 1919;194:675–9.
Ho SY. The changing face of the molecular evolutionary clock. Trends Ecol Evol. 2014;29:496–503.
Hoban S, Arntzen JA, Bruford MW, Godoy JA, Rus Hoelzel A, Segelbacher G, Vilà C, Bertorelle G. Comparative evaluation of potential indicators and temporal sampling protocols for monitoring genetic erosion. Evol Appl. 2014;7:984–98.
Hudson RR. Properties of a neutral allele model with intragenic recombination. Theor Popul Biol. 1983;23:183–201.
Hudson RR. Gene genealogies and the coalescent process. Oxf Surv Evol Biol. 1990;7(1):44.
Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic 9variation. Bioinformatics. 2002;18:337–8.
Hwang DG, Green P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci U S A. 2004;101:13994–4001.
Kaj I, Krone SM, Lascoux M. Coalescent theory for seed bank models. J Appl Prob. 2001;38:285–300.
Kardos M, Luikart G, Bunch R, Dewey S, Edwards W, McWilliam S, Stephenson J, Allendorf FW, Hogg JT, Kijas J. Whole-genome resequencing uncovers molecular signatures of natural and sexual selection in wild bighorn sheep. Mol Ecol. 2015;24:5616–32.
Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW. Genomics advances the study of inbreeding depression in the wild. Evol Appl. 2016:n/a-n/a. doi: 10.1111/eva.12414.
Kern AD, Hey J. Exact calculation of the joint allele frequency spectrum for generalized isolation with migration models. BioRXiv. 2016. doi: http://dx.doi.org/10.1101/065003.
Kimura M. Solution of a process of random genetic drift with a continuous model. Proc Natl Acad Sci. 1955;41:144–50.
Kimura M. Diffusion models in population genetics. J Appl Probab. 1964;1:177–232.
Kingman JFC. The coalescent. Stoch Process Their Appl. 1982;13:235–48.
Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. Genomic runs of homozygosity record population history and consanguinity. PLoS One. 2010;5:e13996.
Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014;15:356.
Krone SM, Neuhauser C. Ancestral processes with selection. Theor Popn Biol. 1997;51:210–37.
Kuhner MK. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics. 2006;22:768–70.
Kuhner MK. Coalescent genealogy samplers: windows into population history. Trends Ecol Evol. 2009;24:86–93.
Leblois R, Estoup A, Streiff R. Genetics of recent habitat contraction and reduction in population size: does isolation by distance matter? Mol Ecol. 2006;15:3601–15.
Leblois R, Pudlo P, Néron J, Bertaux F, Beeravolu CR, Vitalis R, Rousset F. Maximum likelihood inference of population size contractions from microsatellite data. Mol Biol Evol. 2014;31(10):2805–23. doi:10.1093/molbev/msu212.
Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–6.

Li S, Jakobsson M. Estimating demographic parameters from large-scale population genomic data using approximate Bayesian computation. BMC Genet. 2012;13:22.
Liu X, Fu Y-X. Exploring population size changes using SNP frequency spectra. Nat Genet. 2015;47:555–9.
Lohse K, Chmelik M, Martin SH, Barton NH. Efficient strategies for calculating blockwise likelihoods under the coalescent. Genetics. 2016;202:775–86.
Luikart G, Cornuet J-M. Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. Conserv Biol. 1998;12:228–37.
MacLeod IM, Hayes BJ, Goddard ME, et al. A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. Genet Res. 2009;91:413–26.
MacLeod IM, Larkin DM, Lewin HA, Hayes BJ, Goddard ME. Inferring demography from runs of homozygosity in whole genome sequence, with correction for sequence errors. Mol Biol Evol. 2013;30(9):2209–23. doi:10.1093/molbev/mst125.
Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of aboriginal Australia. Nature. 2016;538:207–14.
Marjoram P, Joyce P. Practical implications of coalescent theory. Chapter 5. In: Heath LS, Ramakrishnan N, editors. Problem solving handbook in computational 63 biology and bioinformatics. New York: Springer; 2010.
Marjoram P, Tavaré S. Modern computational approaches for analysing molecular genetic variation data. Nat Rev Genet. 2006;7:759–70.
Marjoram P, Wall JD. Fast "coalescent" simulation. BMC Genet. 2006;7:16.
Matsumoto T, Akashi H, Yang Z. Evaluation of ancestral sequence reconstruction methods to infer nonstationary patterns of nucleotide substitution. Genetics. 2015;200:873–90.
Mazet O, Rodríguez W, Chikhi L. Demographic inference using genetic data from a single individual: separating population size variation from population structure. Theor Popul Biol. 2015;104:46–58.
Mazet O, Rodriguez W, Grusea S, Boitard S, Chikhi L. On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference. Heredity. 2016;116:362–71.
McKee JK, Sciulli PW, Fooce CD, Waite TA. Forecasting global biodiversity threats associated with human population growth. Biol Conserv. 2004;115:161–4.
McVean GAT, Cardin NJ. Approximating the coalescent with recombination. Philos Trans R Soc B. 2005;360:1387–93.
Moorjani P, Gao Z, Przeworski M. Human germline mutation and the erratic evolutionary clock. PLoS Biol. 2016;14(10):e2000744. doi:10.1371/journal.pbio.2000744.
Moran PAP. Random processes in genetics. In: Proceedings of the Cambridge Philosophical Society. 1958. p. 60.
Nadachowska-Brzyska K, Li C, Smeds L, Zhang G, Ellegren H. Temporal dynamics of avian populations during pleistocene revealed by whole-genome sequences. Curr Biol. 2015;25:1375–80.
Nadachowska-Brzyska K, Burri R, Smeds L, Ellegren H. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. Mol Ecol. 2016;25:1058–72.
Naduvilezhath L, Rose LE, Metzler D. Jaatha: a fast composite-likelihood approach to estimate demographic parameters. MolEcol. 2011;20:2709–23.
Nelson GC, Dobermann A, Nakicenovic N, O'Neill BC. Anthropogenic drivers of ecosystem change: an overview. Ecol Soc. 2006;11.
Nielsen R, Beaumont MA. Statistical inferences in phylogeography. Mol Ecol. 2009;18:1034–47.
Nielsen R, Slatkin M. An introduction to population genetics: theory and applications. Sunderland, MA: Sinauer Associates; 2013.
Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, Indap A, Bustamante CD, Clark AG. Darwinian and demographic forces affecting human protein coding genes. Genome Res. 2009;19:838–49.

Nielsen R, Korneliussen TS, Albrechtsen A, Wang J. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PLoS One. 2012;7(7): e37558.
Nikolic N, Chevalet C. Detecting past changes of effective population size. Evol Appl. 2014;7:663–81.
Nordborg M. Coalescent theory. In: Balding DJ, Bishop MJ, Cannings C, editors. Handbook of statistical genetics. New York: Wiley; 2001. p. 179–208
Nordborg M, Donelly P. The coalescent process with selfing. Genetics. 1997;146(3):1185–95.
Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature. 2013;499:744–8.
Orozco-terWengel P. The devil is in the details: the effect of population structure on demographic inference. Heredity. 2016;116:349–50.
Palamara PF, Pe'er I. Inference of historical migration rates via haplotype sharing. Bioinformatics. 2013;8:i180–8.
Palamara PF, Lencz T, Darvasi A, Pe'er I. Length distributions of identity by descent reveal fine-scale demographic history. Am J Hum Genet. 2012;91:1150.
Palkopoulou E, Mallick S, Skoglund P, Enk J, Rohland N, Li H, Omrak A, Vartanyan S, Poinar H, Götherström A, Reich D, Dalén L. Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. Curr Biol. 2015;25:1395–400.
Parmesan C, Yohe G. A globally coherent fingerprint of climate change impacts across natural systems. Nature. 2003;421:37–42.
Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 2008a;18:1814–28.
Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, Birney E. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res. 2008b;18:1829–43.
Pavlidis P, Laurent S, Stephan W. msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. Mol Ecol Resour. 2010;10:723–7.
Peery MZ, Kirby R, Reid BN, Stoelting R, Coucet-Beer E, Robinson S, Vasquez-Carillio C, Pauli JN, Palsboll PJ. Reliability of genetic bottleneck tests for detecting recent population declines. Mol Ecol. 2012;21:3403–18.
Peter BM, Wegmann D, Excoffier L. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. Mol Ecol. 2010:4648–60.
Polanski A, Bobrowski A, Kimmel M. A note on distributions of times to coalescence, under time-dependent population size. Theor Popul Biol. 2003;63:33–40.
Posada D, Crandall KA. Modeltest: testing the model of DNA substitution. Bioinformatics. 1998;14:817–8.
Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. Great ape genetic diversity and population history. Nature. 2013;499:471–5.
Qiu Q, Wang L, Wang K, Yang Y, Ma T, Wang Z, Zhang X, Ni Z, Hou F, Long R, et al. Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. Nat Commun. 2015;6:10283.
Robert CP, Cornuet J-M, Marin J-M, Pillai NS. Lack of confidence in approximate Bayesian computation model choice. Proc Natl Acad Sci. 2011;108:15112–7.
Robinson JD, Bunnefeld L, Hearn J, Stone GN, Hickerson MJ. ABC inference of multi-population divergence with admixture from unphased population genomic data. Mol Ecol. 2014;23 (18):4458–71.
Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014;46:919–25.
Schubert M, Jónsson H, Chang D, Der Sarkissian C, Ermini L, Ginolhac A, Albrechtsen A, Dupanloup I, Foucal A, Petersen B, et al. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. Proc Natl Acad Sci. 2014;111:E5661–9.

Shafer ABA, Gattepaille LM, Stewart REA, Wolf JBW. Demographic inferences using short-read genomic data in an approximate Bayesian computation framework: in silico evaluation of power, biases and proof of concept in Atlantic walrus. Mol Ecol. 2015;24:328–45.
Shafer ABA, Miller JM, Kardos M. Cross-species application of SNP chips is not suitable for identifying runs of homozygosity. J Hered. 2016;107:193–5.
Shafer ABA, Peart CR, Tusso S, Maayan I, Brelsford A, Wheat CW, Wolf JBW. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. Methods Ecol Evol. 2017. doi: 10.1111/2041-210X.12700.
Sheehan S, Harris K, Song YS. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics. 2013;194:647–62.
Sousa VM, Fritz M, Beaumont MA, Chikhi L. Approximate Bayesian computation (ABC) without summary statistics: the case of admixture. Genetics. 2009;181(4):1507–19.
Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics. 2009;182:205–16.
Storz JF, Beaumont MA. Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. Evolution. 2002;56:154–66.
Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105:437–60.
Tajima F. The effect of change in population size on DNA polymorphism. Genetics. 1989;123:597.
Thuiller W. Biodiversity: climate change and the ecologist. Nature. 2007;448:550–2.
Veeramah KR, Woerner AE, Johnstone L, Gut I, Gut M, Marques-Bonet T, Carbone L, Wall JD, Hammer MF. Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate bayesian computation approach. Genetics. 2015;200:295–308.
Vitousek PM, Mooney HA, Lubchenco J, Melillo JM. Human domination of earth's ecosystems. Science. 1997;277:494–9.
Wakeley J. Nonequilibrium migration in human history. Genetics. 1999;153:1863.
Wakeley J. Coalescent theory: an introduction. San Francisco: W.H. Freeman; 2008.
Wakeley J, Hey J. Estimating ancestral population parameters. Genetics. 1997;145:847–55.
Wang J, Street NR, Scofield DG, Ingvarsson PK. Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American aspens. Mol Biol Evol. 2016;33(7):1754–67. doi:10.1093/molbev/msw051.
Waples RK, Larson WA, Waples RS. Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. Heredity. 2016;117 (4):233–40.
Warren MJ, Thomas GWC, Hahn MW, Raney BJ, Aken B, Nag R, Schmitz J, Churakov G, Noll A, Stanyon R, Webb D, Thibaud-Nissen F, Nordborg M, Marques-Bonet T, Dewar K, Weinstock GM, Wilson RK, Freimer NB. The genome of the vervet (Chlorocebus aethiops sabaeus). Genome Res. 2015;25:1921–33.
Watterson GA. The sampling theory of selectively neutral alleles. Adv Appl Probab. 1974:463–88.
Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics. 2010;11:116.
Whitlock MC, McCauley DE. Indirect measures of gene flow and migration: FST|[ne]|1/(4Nm+1). Heredity. 1999;82:117–25.

Wiuf C, Hein J. Recombination as a point process along sequences. Theor Popul Biol. 1999; (55):248–59.
Wright S. The distribution of gene frequencies under irreversible mutation. Proc Natl Acad Sci. 1938;24:253–9.
Wu C-H, Drummond AJ. Joint inference of microsatellite mutation models, population history and genealogies using transdimensional Markov Chain Monte Carlo. Genetics. 2011;188:151–64.
Xue AT, Hickerson MJ. The aggregate site frequency spectrum (aSFS) for comparative population genomic inference. Mol Ecol. 2015;24:6223–40.
Zhao S, Zheng P, Dong S, Zhan X, Wu Q, Guo X, Hu Y, He W, Zhang S, Fan W, et al. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nat Genet. 2013;45:67–71.
Živković D, Stephan W. Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. Theor Popul Biol. 2011;79:184–91.

Advancing Biogeography Through Population Genomics

Jeremy S. Johnson, Konstantin V. Krutovsky, Om P. Rajora, Keith D. Gaddis, and David M. Cairns

Abstract Biogeography is a multifaceted field that integrates geography, geology, ecology, and biology to investigate both historical and ecological questions of how spatial and temporal patterns of varying environmental factors impact the distribution of species and their evolutionary history. Genomes contain imprints of these impacts and, when such genomic imprints are rightly deciphered and interpreted, can help us address these questions. In the past 10 years, incredible advances have been made with respect to acquiring and deciphering genome sequences. The advances in genomics and bioinformatics and the decreasing costs of nucleotide sequencing have reduced many of the barriers to using genomics in biogeography. Here, we introduce some of the strategies and approaches from population genomics that can be integrated into biogeography research. First, we introduce the field of biogeography and define the two well-established broad subdisciplines of ecological and historical biogeography along with the traditional methods that they use. Next, we present examples of how population genomics approaches can be used to address

J. S. Johnson (✉)
School of Forestry, Northern Arizona University, Flagstaff, AZ, USA
e-mail: jeremy.johnson@nau.edu

K. V. Krutovsky
Department of Forest Genetics and Forest Tree Breeding, Georg-August University of Göttingen, Göttingen, Germany

Department of Ecosystem Science and Management, Texas A&M University, College Station, TX, USA

Laboratory of Population Genetics, N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia

Genome Research and Education Center, Siberian Federal University, Krasnoyarsk, Russia

O. P. Rajora
Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, NB, Canada

K. D. Gaddis · D. M. Cairns
Department of Geography, Texas A&M University, College Station, TX, USA

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_39,

biogeographic questions. To illustrate how both ecological and historical biogeography can benefit from adopting a population genomics approach, we outline our own research on mountain hemlock as a case study. We also briefly discuss the application of biogeography in biological conservation. We conclude the chapter by discussing some of the remaining challenges and future research avenues that become possible by integrating population genomics into biogeography research.

Keywords Biogeography · Dendrogenomics · Ecological biogeography · Historical biogeography · Next-generation sequencing · Paleogenomics · Phylogenomics · Reduced representation genomics · SNPs · Species distribution modeling

1 Introduction

Biogeography is a multifaceted discipline that integrates the fields of geography, geology, ecology, and biology. The discipline has a long history (nearly two centuries) focused on understanding the origin, abundance, distribution, and evolutionary history of species along with the processes that structure it (Mayr 1942; Andrewartha and Birch 1954; Hutchinson 1959; MacArthur 1960). Biogeographers tackle questions that seek to unravel patterns of biological diversity, focusing on the physical (geographic) and ecological (evolutionary) processes responsible for structuring these patterns (Fosberg 1976).

More broadly, a unique aspect of any multifaceted integrative discipline, such as biogeography, is the capacity of its members to conduct research on the periphery of the field, allowing interdisciplinary or multidisciplinary research to flourish (Baerwald 2010; Millington et al. 2011a). Case in point, the technological advances of the past decade in genomics, associated with high-throughput sequencing and the availability of bioinformatics approaches, have provided a unique opportunity to advance and unify biogeography research by assimilating population genomics concepts and approaches (Johnson et al. 2016).

Ecological and historical biogeography have long been distinguished as separate components of plant geography (Candolle 1820). And today, the modern core in biogeography is still characterized by both historical and ecological approaches, but interest in human-environment interactions and applied conservation has grown within the field (Millington et al. 2011b). The application of biogeographical principles, theories, and analysis in biological conservation has been recently termed as conservation biogeography (Whittaker et al. 2005). Ecological biogeographers (Hengeveld 1993; Cox et al. 2016; Blumler et al. 2011) and geographical ecologists (Veblen 1989) study contemporary ecological processes acting on current patterns of species distribution, while historical biogeographers study the origin and evolutionary history of species and the long-term changes in the distribution of organisms in conjunction with past processes (Veblen 1989; MacDonald 2003; Cox et al. 2016; Lomolino et al. 2017; Blumler et al. 2011). Biogeographers studying human-environment interactions focus on the interrelationships between people and their

environment, investigating how humans adapt to and change their environment (MacDonald 2003; Cowell and Parker 2004). Conservation biogeography, a subdiscipline of conservation biology, incorporates both ecological and historical biogeography to address questions of biodiversity conservation (Whittaker et al. 2005). In all of these subdisciplines, inclusion of genomic variation can further serve as a proxy for process/pattern relationships, and its inclusion will advance and unify the discipline of biogeography.

The aim of this chapter is to provide context and encouragement for biogeography scholars to adopt and incorporate population genomics concepts and approaches into their research in order to increase the resolution at which they address fundamental biogeographical research questions. We begin by discussing the questions of interest in biogeography and the current methods used to address these questions in both the ecological and historical contexts. Following this discussion, we elaborate on how population genomics can assist biogeography in gaining better insights into spatial patterns of organisms. Because biogeographers approach their research from different perspectives, depending on whether they are focused on historical or ecological processes and patterns, in this chapter, we treat these two research avenues separately with the caveat that a combination of both ecological and historical approaches is needed for a holistic treatment of an organism's biogeography. To demonstrate how population genomics can be integrated into biogeography in a unified way, we outline our own research on mountain hemlock (*Tsuga mertensiana*) as a case study example. Finally, we conclude the chapter with a discussion of challenges and future research directions, with an emphasis on the importance of interdisciplinary and multidisciplinary research and theoretical considerations. While this chapter focuses mostly on plant biogeography, as nearly all geographic biogeography has over the last century (Millington et al. 2011b), the population genomics approaches we discuss are applicable to zoogeography and animal biogeography. We direct readers interested in a more detailed treatment of biogeography as it is treated in geography departments to MacDonald (2003) and Millington et al. (2011a).

2 Biogeography Questions and Subdisciplines

Within the field of geography, biogeography falls into the sub-domain of physical geography. Broadly defined, biogeography assesses how geology, climate, the physical environment, biotic interactions (including humans), and evolutionary processes have shaped and continue to shape the distribution of life. For instance, there are well-known relationships between the distribution of organisms and patterns in temperature, light, and moisture (Schimper 1903), often along subtle environmental gradients. These associations can be strongly correlated with the species distribution and are codified in the concept of the fundamental niche (Hutchinson 1959). When the factors linked to the physical environment are taken together with

biological interactions, such as competition, and patterns of disturbance, we can infer the species niche, though rarely it is possible to completely characterize it.

Biogeographers are interested in the processes that result in changing species' distributions. Approaches used to answer these questions vary widely between theoretical and empirical approaches and span over population- and community-level analysis. And as all good geographers do, biogeographers make maps focusing on the distributions of species, with an emphasis on the size, shape, and continuity of the distributions through time. The focus of the analysis may be on historical patterns, such as vicariance and speciation, or it may emphasize explaining current patterns as well as forecasting the future changes in distribution or community composition resulting from changing climate and disturbance regimes. Indeed, currently, one of the most intriguing questions in biogeography is how ongoing climate change will affect the distribution of species and their populations. All of these questions have, in one form or another, been addressed over the past 200 years.

The temporal scale of the question will usually dictate the type of analysis. Temporally, microscale analysis, as an example, on the order of 10–100s of years can usually be performed by sampling vegetation directly, while longer macroscale questions, on the order of 100–100,000s of years, will rely on various proxy methods, such as sediment and ice cores, pollen, or tree rings (see Delcourt et al. 1982, Fig. 2). Below, we briefly discuss several (but certainly not all) of the common population-based approaches currently used in both historical and ecological biogeography in order to contextualize the questions that are addressed in biogeography. As you will notice, many of the methods can be applied to both sub-domains of biogeography. Here our focus is on population-level analysis as it most closely aligns with the tools that population genomics can contribute to the field of biogeography (Fig. 1).

2.1 Ecological Biogeography

The primary focus of ecological biogeography is to understand the ecological processes that influence the current patterns of species distribution. The causes of these patterns are repeatedly related to different temporally varying aspects of their spatial location, including the abiotic physical environment, positive and negative biological interactions, and patterns of periodic disturbances and stresses.

2.1.1 Physical Environment

The physical environment acts as the abiotic template for which patterns of vegetation are reflected. Landscape- and regional-scale analyses of population distributions are partially influenced by varying patterns of temperature, moisture, nutrients, soil, and their topographic position. And because these factors fluctuate spatially, they are often spatially autocorrelated. We know that individuals found close together usually

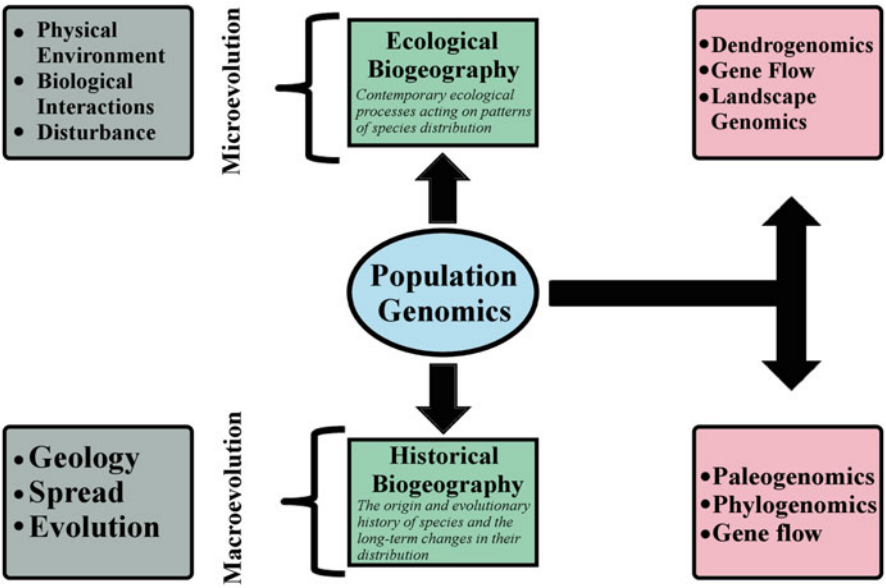

Fig. 1 A conceptual model of biogeography and its unification through population genomics and an evolutionary underpinning. Ecological biogeography is concerned with the contemporary ecological processes acting on patterns of species distribution, while historical biogeography addresses questions about the origin and evolutionary history of species and their long-term changes in their distribution. Both subfields of biogeography use a diverse set of methods, approaches, and theory to address their respective questions. Ecological biogeography is more concerned with biotic and abiotic interactions and the role of short-term disturbances and most closely aligns with microevolutionary processes, while historical biogeography focuses on longer temporal scale process associated with geology and macroevolution. Both ecological and historical biogeographies will benefit from the inclusion of population genomics approaches, and, in fact, under the theory of evolution, population genomics can help to unify the two subdisciplines of biogeography

experience the same abiotic environmental conditions compared to those that are found more distant. We can identify some of these patterns by mapping the distribution of a species along with measurements of their physical environment to identify correlations and associations. Classic ecological biogeography studies have done just this. An obvious example is that of Alexander von Humboldt (1769–1859), considered as the father of biogeography, and his observations on Mount Chimborazo. It is there he noted that vegetation type changed predictably with altitude. Importantly, he observed that the transition of vegetation types mirrored that of changes in the latitudinal distributions of vegetation globally (Von Humboldt and Bonpland 1805). Prominently, von Humboldt noted that these changes corresponded to changes in the physical environment with an emphasis on temperature, climate, and atmospheric pressure. von Humboldt inspired two centuries of ecological biogeographers to investigate patterns of vegetation-environment associations. A striking example of the physical environment's role

in generating biogeographical patterns is that of the alpine tree line position. At a global scale, there is a thermal growth constraint that occurs when mean growing season temperature is below 6.7°C, which limits the advance of forests and results in the spatial location of the alpine tree line ecotone (Körner 1998; Körner and Paulsen 2004). Correlative analysis of this type has benefited greatly by the revolution in geographic information systems and ecoinformatics.

2.1.2 Biological Interactions

While abiotic factors and the physical environment have an important role in spatial and temporal patterns of species distributions, biotic factors are also important. The role of interspecific competition was demonstrated by Connell (1961) when he showed that the distribution of barnacles (*Cirripedia* spp.) in the intertidal zone was controlled by abiotic factors (desiccation) at their upper distribution, but, surprisingly, interspecific competition reduced their distribution at their lower distribution. Since Connell's study, a great deal of research on biological interactions, such as competition, predation, mutualism, amensalism, parasitism, and commensalism, has shown that their non-mutually exclusive interactions are important influences of species distributions. Moreover, in plants, the role of biological vectors for pollination and seed dispersal is critically important (Potts et al. 2010; Ashman et al. 2004).

2.1.3 Disturbance

Lastly, ecological biogeography studies the changing patterns of disturbance and its role in the distribution of plant populations. In some cases, biological interactions are characterized as a disturbance, for example, pest outbreaks, but other disturbances are the result of wind, fire, water, and ice in addition to humans. Disturbance is any discrete event in space and time that "disrupts" a population and changes the composition and/or configuration of components of the physical and biotic environment (White and Pickett 1985). Natural disturbances are, in fact, an important component of plant regeneration and population health and contribute to complex landscape heterogeneity. There are several types of disturbances that are frequently studied in ecological biogeography. Fire has been studied rather extensively in biogeography because of its widespread occurrence and higher return frequency, which has provided a way to investigate patterns of vegetation dynamics and successional replacement in forests (Veblen et al. 1994) and grasslands (Bond and Keeley 2005). Additional examples of studies of disturbance include wind (Kleinman and Hart 2017), drought (Woodhouse et al. 2010), and ice (Lafon and Speer 2002) disturbances.

2.2 *Historical Biogeography*

In contrast to ecological biogeography, historical biogeography addresses questions over longer temporal scales ranging from thousands to millions of years in the past. Furthermore, historical biogeography often addresses larger spatial scales consisting of regions to global scales. The focus of many historical biogeography questions centers on understanding the role of geological processes on species distributions, the spread of organisms, and species evolution.

2.2.1 Geology

George-Louis Leclerc, Comte de Buffon (1701–1788) noted how distant but environmentally similar locations had distinct assemblages of biological life (Buffon 1791). This observation, appropriately titled "Buffon's Law," is also known as the first principle of biogeography and has generated many questions concerning the observed patterns of vegetation globally. Not only did ecologically similar locations often have different assemblages of flora and fauna, but sometimes distant locations had surprisingly similar assemblages. Shortly thereafter, Charles Lyell (1797–1875) championed James Hutton's (1726–1797) influential theory of uniformitarianism which led to the realization that Earth is much older than had previously been thought. In spite of the great advances of the forefathers of biogeography, they had no knowledge of the dynamic nature of the Earth and so could not explain many of their observations on the distribution of life. Why, in fact, did Buffon see different species in ecologically similar locations? Why were fossil remains of similar species found in geographically distant regions? At this point biogeographers predominantly invoked long-distance dispersal or the rise of unlikely land bridges as the standard explanation for biogeographical patterns (Schickhoff et al. 2014), in many cases without any reasonable evidence to the contrary. Alfred Wegener's (1880–1930) (1924) *Pangea* supercontinent and the theory of continental drift explained many of the disjunct biogeographical patterns, but his theory lacked a viable mechanism that would allow continents to move around. It wasn't until the 1960s, when suboceanic observations of mid-oceanic ridges and the theory of seafloor spread became known (Hess 1962; Heezen 1960) along with the support from research in paleomagnetism (Vine and Matthews 1963), which provided evidence that continents did in fact move. The theory of plate tectonics (Dietz 1961) was accepted, and this revelation provided an alternative explanation to dispersal in explaining biogeographic distributions and patterns. Plate tectonics ushered in a new chapter in biogeography, one where alternative theories of the causes of historic biogeographic patterns could be tested (Wiley 1988).

Vicariance is a process whereby patterns of isolated disjunct populations result from a continuous range splitting, either through orogenic events or tectonics (Giller et al. 2004). But shifting continents alone could not explain all of the patterns of biological diversity. Climate also changed through time. Louis Agassiz (1807–1873)

postulated that a great ice age had occurred. The expansion and retraction of ice sheets throughout the Quaternary had a profound impact on the distribution of life. By combining an understanding of the timing of continental drift with the distribution of fossils and historic climates, biogeography began to pinpoint the origin of taxa based on physical characteristics of extant and extinct species and their observed spatial location.

2.2.2 Spread

While geology provides a theoretical basis and an alternative mechanism to the Darwinian dispersalist explanations for the historic distribution of species, dispersal itself was still an important process shaping biogeographic patterns at historical scales. The spread and radiation of species via dispersal has been one of the primary foci of historical biogeography. In particular, long-distance dispersal, or jump dispersal as it is often referred to, is of the greatest interest in historical biogeography, though diffusion is also an important aspect of understanding past colonization processes.

There are several reasons that plants disperse. At ecological scales, plants disperse to avoid direct competition for resources with their parents and close neighbors and also to avoid inbreeding and the deleterious effects that come with inbreeding depression (Howe and Smallwood 1982; Nathan and Muller-Landau 2000). Secondly, and at much longer time scales, plants disperse to expand their range and take advantage of new suitable habitat or, when needed, as a means to escape deteriorating habitat (Willson and Traveset 2000). Historic climate change throughout the Quaternary, associated with alternating glacial interglacial periods, was a significant factor impacting the distribution of plants in the northern hemisphere (Davis 1981). One of the more intriguing historical biogeographical observations was made by Clement Reid (1899). Reid, a paleobotanist, understood that, based on estimated rates of seed dispersal at the end of the nineteenth century, in order for the oak (*Quercus* spp.) trees in northern Britain to reach their current geographic position, it would have taken nearly a million years to travel the distance from their glacial refugia. This of course was not true, and based on the data gleaned from the paleorecord, Reid's observation led him to speculate about the role of long-distance dispersal in recolonization of northern landscapes following Pleistocene glaciations (discussed in Skellam 1951; Clark et al. 1998). In this case, the other alternative was that small populations of species survived in situ in microrefugia and did not, in fact, disperse long distance to recolonize northern environments (Hylander et al. 2015; Rull 2009). Biogeographers are now starting to generate empirical datasets, mostly thanks to genetic analysis, that can begin to capture long-distance dispersal events, despite the extreme difficulty in measuring and observing these dispersal events.

Another line of evidence in support of long-distance dispersal as a process generating biogeographic patterns stems from observations of biological life on islands – a research area termed island biogeography. The diversity of life on islands can be explained by the equilibrium theory of island biogeography (MacArthur and

Wilson 1963, 1967), which states that species richness is a function of island distance to a colonizing source and the size (area) of the island. In this theory, large close islands should harbor the highest species richness, while small distant islands should be species poor in comparison. These patterns emerge from species dispersal capabilities and extinction rates on the island. However, the fact alone that life exists on islands, especially when new islands rise or following defaunating volcanic events sensu Krakatau (Whittaker et al. 1989), is a testament to the dispersal capabilities of many organisms (Diamond 1974). Thanks in part to the development of phylogeography approaches, this hypothesis can be explored. Today both dispersalist and vicariance biogeographies remain vital research areas in historical biogeography.

2.2.3 Evolution

The history of evolutionary thinking is long and well documented, and there are many volumes dedicated to its study. There are also many excellent popular treatments of the subject that we direct readers to (Mayr 1942, 1982; Wilson 1992). It is not our intention to review the vast literature on the roles of evolution in the creation of biodiversity and biogeogrpahic patterns, and some other chapters in this book provide excellent examples and discussions on this topic. However, it is worth mentioning several of the key theories that are important to historical biogeography. The names of Charles Darwin (1809–1882) and Alfred Russel Wallace (1823–1913) are undoubtedly familiar to all, and the importance of the theory of evolution by natural selection cannot be understated. It is important that dispersal and vicariance processes, though Darwin and Wallace were unaware of the latter, resulted in the historic separation of species. As a product of physical separation, reproductive isolation occurs, and natural selection through survival of the fittest rewards reproductively superior individuals resulting in the shifts of the gene pools (allele frequencies) of populations with the effect of a gradual (or sometimes not so gradual) divergence of species and ultimately speciation.

Many questions of interest in historical biogeography arise from a desire to understand how species come into existence, why they are able to live in a specific environment, and ultimately what cause them to go extinct. The large number of species concepts is a testament to the fact that we still have an incomplete understanding of the definition of a species. Despite this fact, most definitions of species propose that some form of isolation (either reproductive or geographic) is usually required for species to diverge and evolve into a new lineage (but there is a growing body of research exploring the idea of speciation with ongoing gene flow (Christe et al. 2017; Menon et al. 2018; Yang et al. 2017)). Novel genetic variation and sometimes beneficial traits arise through random genetic mutation. Allopatric speciation, in its simplest form, results from geographic isolation where separation by physical distance or physical barriers arrests the spread of novel variation from reaching isolated populations of the species (Mayr 1947). Over time, the populations

diverge and become reproductively isolated and evolve new traits leading to new species formation.

In contrast to allopatric speciation, some speciation events occur without the aid of physical separation. In this case, variation within the geographical range leads to speciation and is referred to as sympatric speciation. Plant divergence and sympatric speciation can result from temporal shifts in phenology and thus reproductive isolation, for example, the timing of flowering (Silvertown et al. 2005; Hancock et al. 2011). In either case, some form of isolation can lead to genetic drift and natural selection for genetic variation that improves reproductive success or against variation that reduces reproductive success and ultimately contributes to the rise of new species. Within the field of population genetics, some of the primary explanations for evolutionary divergence and genome reorganization include founder effects, bottlenecks, selection, and the effects of small population size (Endler 1977; Hewitt 1996).

Our discussion above has provided a brief look at some of the classic questions that both ecological and historical biogeography addresses. Although our discussion is far from exhaustive, it is clear that multiple interacting factors, dispersal, vicariance, and speciation, are responsible for the historical spatial pattern of biological life, while ongoing changes to the physical environment, biological interactions, and patterns of disturbance are influencing ongoing changes in species distributions.

3 Traditional Approaches in Biogeography

Within the framework of both ecological and historical biogeography discussed above, we will explore some of the traditional approaches that both subdisciplines of biogeography use to address their respective questions with an eye toward demonstrating how population genomics can improve the resolution and performance of these approaches. Again, the separate subdisciplines of biogeography are split to focus on ecological and historical biogeography.

3.1 Ecological Biogeography

In order to understand the contemporary distributions of plants, ecological biogeography must combine an understanding of the roles of the biotic and abiotic environment with the process of disturbance. Practitioners of ecological biogeography use many different approaches to investigate biogeographical questions. Often a combination of field-based methods, physiological experimentation, spatial analysis, and simulation modeling are used to test hypotheses about the current spatial patterns of species. Ultimately ecological biogeography strives to understand the causes of species distributions and to predict how these processes will change future distributional patterns. Here, we explore some of the methods and approaches frequently used in ecological biogeography to address the questions outlined above.

3.1.1 Species Distribution/Ecological Niche Modeling

Though the terms can be confusing, ecological niche modeling (ENM) or species distribution modeling (SDM) is a method that allows biogeographers to assess what ecological (usually topoclimatic) factors are important predictors of a species' actual or potential distribution (Franklin and Miller 2009). In brief, a species' range is predicted based on the correlation between the presence or absence of a species at a specific geographic location and the unique ecological and geographic factors that occur at that location. The implementation of this approach has been facilitated by the development of Maximum Entropy (MaxEnt) modeling (Phillips et al. 2006) and other multiple regression-based methods (Franklin and Miller 2009). Despite the wide use of SDMs and ENMs in biogeography, there are still many problems associated with sampling biases and correct parameterization of the models that must be addressed (Phillips et al. 2009; Kramer-Schadt et al. 2013).

The original article introducing MaxEnt (Phillips et al. 2006) has been cited more than 8,000 times at the time of writing this chapter and over 1,500 times since 2017 serving as a testament to its popularity and ease of use in creating SDMs. More advanced SDMs are now being created using machine learning approaches, such as boosted regression trees and random forests as well as generalized additive and linear mixed models (Shirk et al. 2018). Shirk et al. (2018) generated a suite of SDM of southwestern white pine (*Pinus strobiformis*) to identify the roles of the climate, soil, and topography and predicted the current and future range of this pine. Their findings showed that different areas of the range would either expand or contract with the predicted future climate change. Their findings are important as they illustrate the vulnerability of the species to multiple hazards, including their susceptibility to the non-native invasive pathogen *Cronartium ribicola*, responsible for the white pine blister rust disease, which has spread into the species range. The study also illustrates the need for inclusion of population genomics data into SDM as the range of southwestern white pine consists of unique genetic populations that are likely locally adapted and will respond differentially to the multiple threats.

SDM approaches are being used to understand how species' ranges will change in the future and to inform management and conservation decisions (Ferrarini et al. 2016). Biogeography has focused mostly on the climatic relationships between species and their distributions (Franklin 2010), but a growing number of models are starting to incorporate life history traits and biological interactions (Pöyry et al. 2008; Iverson et al. 2011; Araújo et al. 2005; Fordham Damien et al. 2013).

3.1.2 Landscape Ecology

Often ecological biogeographers found in geography departments today have adopted a landscape ecological approach. Landscape ecology spawned from a desire to understand how the spatial composition and configuration of the landscape (matrix) influenced the processes responsible for patterns of species distributions,

and not solely the inverse relationship of processes and its influence on the pattern (Kupfer 1995, 2011; Turner 1989; Forman and Godron 1981). Changing patterns of landscape heterogeneity thus have cascading effects on processes governing the distributions of species, in particular patterns of energy and material fluxes, dispersal, and functional connectivity in plants (Pickett and Cadenasso 1995).

The first step in landscape ecology is to quantify landscape patterns. This involves quantifying both the composition of the landscape and its configuration. The composition of the landscape is related to different types or kinds of landscape elements (forests, farms, towns), while the configuration is described by their spatial arrangement at a specific scale (Kupfer 2011; Wiens 1989). The process of quantifying landscape patterns is facilitated by advances in remote sensing and geographic information systems. For the study of plants, the role of landscape fragmentation, habitat loss, and broadscale land-cover change is of particular interest (Haines-Young and Chopping 1996).

The basic approach after identifying the analytical objective (e.g., what environmental processes determine the landscape pattern of species distribution) is to quantify the landscape based on the appropriate spatial extent, thematic content, and resolution (Cushman et al. 2016). The analytical step then involves statistically relating the landscape patterns to various indices and metrics designed to help explain landscape structure (Kupfer 2012). Ecological biogeographers with a landscape ecology research program often focus on the role of humans and their role in biogeographic patterns (Chhetri et al. 2017) and oftentimes focus on applied conservation biogeography and management priorities such as reserve design (Kupfer 1995). Several ecological biogeography studies have used landscape ecological approaches to study the role of global change on spatial patterns of vegetation. For example, the impacts of landscape structure change on southern pine beetle outbreaks were assessed using landscape simulation models finding that patterns of host aggregation in addition to landscape structure were important determinants of pest outbreak severity (Cairns et al. 2008). Naito and Cairns (2011), using a remote sensing approach, studied the spatial patterns of shrub expansion in the arctic finding that topographic position and hydrologic characteristics were important factors associated with the amount of expansion, specifically showing that shrubs were expanding preferentially into wetter environments. The role of environmental feedbacks has been explored by combining simulation and observational analysis to test the role of positive feedbacks on spatial patterns of alpine tree line finding that directional feedbacks associated with wind contributed to observed patterns at tree line (Alftine and Malanson 2004).

3.1.3 Dendrochronology-Based Approaches

Patterns of disturbance are often reflected in the annual growth of trees. Forest dynamics associated with disturbance can be assessed because a disturbance event can be precisely dated based on the location of a disturbance scar relative to annual tree rings. One aspect of dendrochronology is concerned with identifying specific

events, usually some sort of disturbance, recorded in the life of trees and shrubs and is used to reconstruct the history of the environment (Harley et al. 2018). A common biogeographical analysis is to reconstruct fire or pest history and their role in stand dynamics. This approach identifies fire scars or insect galleries within cross sections of trees and then dates them to outbreak cycles or reconstructs fire return intervals. Examples are plentiful and have in the past explored the impact of insect outbreaks on forest structure (Veblen et al. 1991), the relationship between fire regime and climate (Grissino-Mayer and Swetnam 2000), and the role of management and policy (fire suppression) on vegetation structure (Flatley et al. 2015; Lafon et al. 2017), and identified ice storm disturbance in forests (Lafon and Speer 2002).

The data obtained from these studies has led to a better understanding of disturbance regimes in a variety of systems and can be used to better inform forest management activities. Dendrochronology, as we will discuss, is also an important approach in unraveling historical biogeographic patterns and climate events.

3.1.4 Genetics Approaches

Simply finding correlations with the physical environment may not be enough to determine how patterns of species distributions will change in the future. All biological life has at least one life history stage, where an individual moves. For example, trees do not move for the vast majority of their life; however seed and pollen dispersals do allow the movement of propagules prior to establishment and growth. The capacity of an individual to move in response to changing environmental conditions or shifts in resource availability ultimately will impact the acclimation and adaptability of the population and the species as a whole. It also will influence the patterns of the species distribution in the future. Many aspects of the physical environment constrain the movement of individuals and organisms. If movement is restricted by barriers, and populations become geographically isolated, then they may become genetically depauperate, and their capacity to evolve novel responses to ecological change will be reduced. When this happens local populations may go extinct, or, worse, entire species may blink out of existence. For this and other reasons, ecological biogeography often assesses population-level demographic processes to understand the role of the physical environment on distribution patterns. In plants, it has been very difficult to determine the patterns of individual propagule movement (Cain et al. 2000; Nathan et al. 2002; Nathan 2006), and many early studies relied on mathematical models to reconstruct patterns of dispersal (Hamilton and May 1977; Skellam 1951; Clark 1998). Methods of tracking biological movement have most often been observational in nature and include the use of tagging and tracking (Kays et al. 2015), seed traps (McCaughey and Schmidt 1987; Bullock and Moy 2004), or capture mark and release methods (Levey and Sargent 2000; Xiao et al. 2006). All of these methods have provided useful insight into biological movement but suffer from challenges associated with quantifying the tail of the dispersal distribution, specifically measuring long-distance dispersal. In some cases, the use of genetic markers has provided an alternative to measuring plant movement.

For instance, inference of pollen and propagule movement and dispersal has been carried out in many plant species using genetic markers (Markwith and Scanlon 2006; Gaddis et al. 2016; Degen et al. 2004). More direct measures of dispersal have also been conducted using genetic markers and parentage analysis (Piotti et al. 2009; Dow and Ashley 1996; Johnson et al. 2017c; Ismail et al. 2017; Robledo-Arnuncio and Gil 2004; González-Martínez et al. 2002). In some instances, both approaches have been combined and compared to better understand plant movement (Oddou-Muratorio and Klein 2008). Recently, Bullock et al. (2017) assessed 168 empirically derived dispersal kernels spanning 144 vascular plant species and classified them into different dispersal modes. They showed that because of variation in dispersal mode, long-distance dispersal (at the 95th percentile of the kernel) varied quite a lot. By combining data on dispersal distance with species distribution models, ecological biogeographers can better predict the future range of species and rate of movement (Hamann et al. 2015; Loarie et al. 2009). O'Connell et al. (2007), using allozyme genetic markers, provided empirical evidence for extensive pollen-mediated gene dispersal between natural stands of a widespread northern temperate and boreal conifer, *Picea glauca*, in a landscape fragmented by agriculture and lake water in north-central Ontario, Canada. The average minimum pollen dispersal distance in outcrossed matings was found to be 619 m.

3.2 Historical Biogeography

Throughout our discussion of historical biogeography, it is clear that the role of earth history, dispersal, and evolution is intertwined. Many of the questions of interest in historical biogeography are thus related in terms of answering these high-level questions. Because we cannot go back in time and observe various biological patterns and historical events, we must rely on proxy data to reconstruct the past and infer patterns and processes. Several different types of proxy data have been used to address historical patterns, and each one provides a different degree of spatial and temporal resolution.

3.2.1 Dendrochronology-Based Approaches

At shorter temporal scales (1,000s of years), many historical biogeographers focus on understanding species-climate relationships using dendrochronological approaches. This approach, similar to the use of dendrochronology in ecological biogeography, is concerned with assessing growth patterns reflected in the interannual growth rings of trees and shrubs. The growth rings of several trees can be compared and used to infer the climate record using cross dating techniques. Finally, these data can be used to reconstruct historic climates and their role in biogeographic patterns (Fritts 1976). Growth patterns in trees often reflect landscape-scale climate variation (temperature, precipitation, herbivory)

(Speer 2010) and indicate how past conditions influenced growth at the species level. Tree rings provide an indication of how physiological mechanisms constrain growth as a result of, for example, moisture and/or temperature (Park Williams et al. 2012). The ring width, density, and pattern of tree rings allow a time series of climate growth relationships to be developed and indicate how changes in environmental stress relate to a tree's ability to grow well (Speer 2010). Case in point, ring width indices (RWI) in several subalpine species have been correlated with temperature and precipitation to inform our understanding of the past and future vegetation dynamics within the system and yield data on both spatial and temporal factors limiting growth at a species' range limit (Taylor 1995; MacDonald et al. 1998; Young et al. 2011). Moreover, dendrochronology-based approaches have been used to assess vegetation dynamics along environmental gradients. Patterns of growth within the alpine tree line ecotone are tightly linked with climate, and reconstruction of demographic patterns allows patterns of regeneration and ecotone movement to be inferred (Elliott 2012; Chhetri and Cairns 2016; Danby and Hik 2007).

3.2.2 Paleo-Based Approaches

Reconstructing patterns of plant colonization and range dynamics and identifying the climatic constraints on these distributions are an important research area in historical biogeography. Palynology is a paleo-based approach where fossil pollen or plant micro-/macrofossils (seeds, leaves, cones, plant pieces, and charcoal) are used as proxies to reconstruct prior distributions and environmental conditions. By combining maps of past distributions with radiocarbon dating and analysis of oxygen isotope from ice and sediment cores, species arrival dates and estimates of climate can be derived. In essence, this method tracks the accumulation of pollen (and microfossils) within peat, lake, and marine sediment cores and dates them. In pollen analysis, at first, pollen will be limited in quantity, but as a species colonizes closer to a core sampling site, the fraction of pollen should increase allowing the timing and geographic position of the range to be estimated by studying their stratigraphy (MacDonald 2003). Though not a perfect representation of vegetation assemblages and their distribution, it does provide synoptic information about expansion and contraction of populations in response to broadscale historic climate changes (usually at the scale of the Quaternary) and provide a means to inform general community composition and historic vegetation dynamics. Importantly, this allowed not only the spatial pattern of biodiversity to be explored, but it also allowed researchers to infer what climatic conditions species were adapted to.

Global reconstructions of climate from proxy data have improved our understanding of past climates on the order of millennia (Mann and Jones 2003) to the entire Holocene (Marcott et al. 2013). By combining paleo-based methods with modeling approaches, researchers are beginning to more accurately map range positions (Bruening et al. 2018). Much of this research has focused on northern migration of plants in North America and Europe following Quaternary glaciations. Several studies have used palynological and paleo-based approaches to reconstruct

historic migration of forest species. Patterns of postglacial spread of European beech (*Fagus sylvatica*) following the last glacial maximum were assessed using over 400 pollen records allowing both the location of refugia and the pathways of recolonization to be reconstructed in Europe (Magri 2008). Mountain hemlock (*Tsuga mertensiana*) pollen and macrofossil data were extracted from sediment cores in the Rocky Mountains of Idaho, USA, and used to reconstruct Holocene long-distance dispersal of the species from its coastal refugia to isolated populations in Idaho (Herring et al. 2017). A combination of pollen and oxygen-isotope analysis was used to reconstruct vegetation dynamics in the Swiss Alps during the late glacial interstadial. Interestingly, this analysis showed that species composition has no modern analog and that the assemblages of species changed through immigration and extinction over the historic record (Ammann et al. 2013). Classic biogeography studies have used pollen analysis to reconstruct forest movement following the last glacial maximum (Webb 1987; Delcourt and Delcourt 1981; Davis 1981).

3.2.3 Phylogeography and Molecular Population Genetics Approaches

Within historical biogeography, the use of molecular genetic markers and population genetics concepts and approaches has been extensive. Cladistics and phylogenetic analyses have provided historical biogeography the ability to assess the history and evolution of lineages. Cladistics was developed so that morphological traits or endemism patterns could be compared among taxa and evolutionary history could be inferred (Hennig 1979). Cladistics is used in historical biogeography to place taxon together based on their trait differences or similarities, assuming that species with similar traits share a common ancestor from which the traits were inherited. The problem here is that it can be difficult to assess if similar traits are the result of convergent or parallel evolution (Cox et al. 2016). This problem was solved when cytoplasmic DNA (mitochondrial or chloroplast DNA in plants), with a known rate of mutation (Kimura 1968), was used to identify the timing of evolutionary divergence. Knowing the timing of these historic events helped identify if the processes of divergence were related to dispersal, vicariance, or even sympatric diversification (Avise 2000; Avise et al. 1987). For example, several different molecular markers, including allozyme (Wheeler and Guries 1982; Parker and Hamrick 1996), amplified fragment length polymorphisms (AFLPs) (Despres et al. 2002), and microsatellites (Mayol et al. 2015) have contributed much to our understanding of the role of Quaternary climate change on patterns of range expansion and contraction (Parducci et al. 2012; Hewitt 2004). In fact, much of the current distribution of vegetation in North America has been influenced by the glacial cycles of the Quaternary (Shafer et al. 2010). Moreover, microsatellite analysis has helped to sharpen our understanding of evolution and adaptation through oscillating glacial cycles (Provan and Bennett 2008; Hewitt 1996; Petit et al. 2003). In forest trees, phylogeographic methods have provided an avenue to assess population divergence (Gugger et al. 2010) and postglacial migration and phylogeographic patterns (e.g., Zinck and Rajora 2016). Combining SDM, phylogography and fossil improve the identification

of the locations of climate refugia during the last glacial maximum (Shafer et al. 2010; Parducci et al. 2012). Zinck and Rajora (2016) studied the range-wide genetic diversity and population structure of 33 eastern white pine (*Pinus strobus*) populations using 12 nuclear and three chloroplast microsatellite markers and applied approximate Bayesian computation approach to test various evolutionary scenarios. Their results supported the presence of two main postglacial recolonization routes originating from a single southern refugium in the mid-Atlantic plain. One route gave rise to populations at the western margin of the species' range in Minnesota and western Ontario, and the second route gave rise to central-eastern populations, which branched into two subgroups: central and eastern. The phylogeographic patterns were consistent with the fossil pollen data.

4 Population Genomics Approaches as Applied in Biogeography

Population genomics assesses population data obtained from whole (or nearly whole)-genome sequences in both nuclear and organelle genomes (or genomic DNA constituting the combined nuclear and organelle DNA) to determine the origin, amount, and outcome of genetic variation within a spatial framework as well as the factors contributing to the variation. Genetic variation is often categorized as either adaptive or selectively neutral. The latter form of variation is affected mostly by selectively neutral factors, such as genetic drift and gene flow (e.g., Krutovsky et al. 2012), as the affected genomic regions have no apparent effect on the fitness of the organism through changes in survival and reproduction as suggested by Kimura's neutral theory of molecular evolution (Kimura 1979). Because natural selection is mostly absent from the portions of the genome that are selectively neutral, they serve as indicators of demographic patterns, population ancestry, and the influence of ecological and geographic heterogeneity on gene flow (Selkoe and Toonen 2006) and are of particular interest in studies of large-scale patterns of species distribution. Nevertheless, neutral genetic variation cannot easily be associated with environmental factors unless those factors impact population-level dynamics, such as population size, gene flow, or the system of mating (Holderegger et al. 2010). In contrast, adaptive genetic variation, by definition, has been under selection. Studying spatial patterns of adaptive genetic variation is critical to our understanding of the effects of past climate change, landscape and land use change, and human-environment interactions and may inform our predictions of future evolutionary potential and local adaptation in the event that the selecting agents remain the same or similar to the past (Holderegger and Wagner 2008; Stapley et al. 2010; Savolainen et al. 2013). Understanding the past provides an opportunity to improve future predictions of biodiversity. On balance, improving our ability to quantify how species are adapting to changing environmental conditions will help us to prioritize management strategies and allocate resources for conservation goals (Hoban 2018). Combining both

historical and ecological biogeography approaches will provide the most useful insights into future patterns of biodiversity. It is within this context that population genomics can best contribute to biogeography research.

Population genomics is already making important contributions to biogeography research. Here we provide examples of population genomics approaches that can be, and are being, used to address many of the questions laid out in Sect. 2. We break down biogeography into its two primary components: ecological biogeography and historical biogeography.

4.1 Ecological Biogeography

Population genomics can improve the understanding of ecological biogeography by providing an additional layer of biological information that is useful in assessing adaptive evolution and local adaptation to changing environments as well as allow ecologically relevant traits to be considered. Moreover, population genomic data will contribute to better predictions of species distributions in the face of multiple threats, such as climate change and invasion by novel pathogens. Here we discuss how population genomics can provide benefits to several aspects of ecological biogeography.

4.1.1 Species Distribution Modeling

Understanding how species will respond to climate change, invasive pathogens, and shifting disturbance regimes will become an increasingly important goal for resource managers and conservationists. Both ecological and historical biogeography (discussed in Sect. 3.1.1) use SDMs to understand past and future distributions. In order to improve hindcasts of past ranges and forecast future distributions, and avoid the assumption of niche uniformity across the species range, traditional SDMs and ENMs must begin to include genomic variation as a predictor of these patterns (Razgour 2015; Gotelli and Stanton-Geddes 2015). By including genomic variation, a species can be clustered into genetically distinct populations across its range and patterns of ancestry, and admixture can be considered. SDMs can then account for spatial patterns of local adaptation and assess the potential for species niche divergence or convergence (Gotelli and Stanton-Geddes 2015).

New approaches that combine population genomics, SDMs, and genotype-environment associations are beginning to provide this level of prediction. For example, Ikeda et al. (2016) included population genetic structure of Fremont cottonwood (*Populus fremontii*) to test if including genetic data would improve ecological niche models. Their results showed that inclusion of demographic genetic variation improved the predictability of the models 12-fold. Moreover, each of the three genetic ecotypes identified was associated with different climatic factors suggesting that under future climate change, niche divergence may occur. Similar

work using the plant model *Arabidopsis thaliana* found that SDMs that included neutral genomic variation could improve spatial predictions of conservation-relevant genetic units (Marcer et al. 2016).

An alternative approach, and one that uses population genomic data more specifically, is to view genomic variation within a sampling site in a way similar to species assemblages in community ecology. In this way, tens to hundreds of thousands of SNP markers and their site-level allele frequencies can be compared between several sample sites along an environmental or climatic gradient, and their association with environmental measures can be identified. This multivariate approach, thus, facilitates prediction and mapping of genomic variation across the landscape and improves predictions of adaptive potential. In essence, a genomically informed ecological niche model (gENM) can be created. This approach was demonstrated across the range of balsam poplar (*Populus balsamifera*), where multidimensional genomic variation was mapped by reducing its dimensionality using the principal components analysis and then the first three PCs were used to map the genomic variation in geographic space (Fitzpatrick and Keller 2015). In addition to predicting the spatial pattern of genomic variation, Fitzpatrick and Keller (2015) showed that adaptive variation, related with phenology, was strongly associated with temperature.

At the population level, intraspecific genomic variation determines how sensitive a given population is to environmental change and thus their capacity to adapt locally. The combination of exposure to changing climate and decreased levels of genomic diversity can reduce adaptive capacity in a population. This is especially true in species with reduced dispersal capabilities. By identifying the diversity and type (neutral or adaptive) of population genomic variation and also quantifying the exposure to climate change, the vulnerability of a population can be assessed, and management actions can be prioritized. Razgour et al. (2017) introduced a framework that combined genomics and SDM to identify populations under threat from climate change. In their study, Razgour et al. (2017) combined genome-environment associations and outlier tests with SDM to assess range shift potential along with the level of genomic diversity available to respond to climate change in the bat *Plecotus austriacus*. They showed that changing niche suitability would likely limit dispersability of the bat and reduce its evolutionary potential due to geographic isolation, drift, and small population size.

Clearly ecological biogeography can better predict the outcome of global environmental change by including genomic variation in predictive distribution models as opposed to assuming genetic uniformity as has been the custom. Spatial genotype-environment associations allow predictions of adaptive capacity to be considered when considering how to manage populations.

4.1.2 Landscape Ecology and Landscape Genomics

Landscape ecological analysis is one area of ecological biogeography that has greatly benefited from the inclusion of genomic data and population genomics

approaches more specifically. Over the past 15 years, an analytical approach has emerged to address both biological movement and adaptation and its relationship to landscape heterogeneity, a field coined as landscape genetics (Manel et al. 2003). The topic of landscape genetics and genomics is covered by Balkenhol et al. (2017) in this book. However, we would be remiss if we did not mention one aspect of landscape genomics that is of great interest in ecological biogeography employing landscape ecology. Specifically, we will discuss landscape connectivity.

Biological movement occurs in all organisms at one time or another and is essential for the survival of a species in the presence of shifting niches and biological interactions. Our landscape ecology discussion (Sect. 3.1.2) noted how the spatial arrangement of landscape elements (forests, farms, grasslands) and patterns of topography and climate influence a species niche. Understanding how the composition and configuration of the landscape influence effective movement (seed and pollen dispersal) of propagules and genes between discrete patches, populations, or metapopulations is described as the species functional connectivity (Auffret et al. 2017). There have been major advances in population connectivity studies by adding genomic variation and genetic distance measures into landscape ecology models (Dyer and Nason 2004; Dyer 2015). Circuit and graph theory-based approaches have been the most popular. Graph approaches are based on a model where populations or individuals are treated as nodes and the length of the connection between nodes is measured as genetic distance. A heuristic approach is used to remove connections between nodes and produce an optimal graph with the least total number of connections having the greatest power to explain overall connectivity. Using this framework, each population can be assessed for its importance in maintaining genetic connectivity and gene flow. Moreover, the graph can be superimposed onto the landscape in order to identify areas with greater or lesser connectivity based on the relative distance, which can inform identification of landscape features that facilitate or impede movement (Dyer and Nason 2004; Dyer 2007, 2015; Dyer et al. 2012; Garroway et al. 2008). Also, the population graph approach could be used for understanding the multilocus architecture of local adaptation in plants (see Rajora et al. 2016).

Circuit theory-based approaches improve upon classic least-cost path approaches by examining all possible pathways across a landscape in lieu of identifying a single optimal route across a region (McRae and Beier 2007; McRae et al. 2008; Shah and McRae 2008). A specific landscape or climatic layer can be examined as a cost surface (called a resistance surface) with areas ranked as allowing for greater or lesser movement. All potential pathways between two target populations are examined relative to this resistance surface (McLean et al. 2016; Franklin and Krueger 1968; Etterson et al. 2016; Sork et al. 2010), estimating the pairwise resistance, a value that accounts for geographic distance between locations and the ease with which an individual may move across the landscape via all possible routes (McRae and Beier 2007). This method compares alternative resistance hypotheses to explain genetic distance between sampling locations and identify those variables that best explain movement in a given area (Orsini et al. 2013; Sexton et al. 2014; Ruiz-Gonzalez et al. 2015). The most popular program currently used for this

approach is Circuitscape (Shah and McRae 2008) and assesses electrical nodes that are connected by a series of conductors. More numerous or larger connections enhance electrical current flow. Electric circuit theory can be translated into individuals and populations being represented by the nodes and the resistance of the landscape being measured as the degree of connectivity between nodes (McRae 2006; McRae and Beier 2007; McRae et al. 2008, 2013; Shah and McRae 2008).

Several studies have combined SDMs with connectivity modeling and genetic variation to identify the level of connectivity in both plants and animals. This approach was used in the riparian species *Populus angustifolia*. Genetic connectivity was assessed using resistance modeling and SDMs finding that riparian corridors facilitated connectivity, while terrestrial uplands were two and half times more resistant to gene flow (Bothwell et al. 2017). Moreover, historic migration and landscape connectivity of a long-lived conifer *Tsuga mertensiana* were assessed using a combination of SDMs, electric circuit theory, and genomic data consisting of 6124 SNPs. This study found that patterns of genomic diversity were correlated with recent climate resistance and not mid-Holocene climate resistance (Johnson et al. 2017b). Studies of animals, too, have demonstrated the utility of using genomic variation and landscape characteristics to assess functional connectivity. Razgour et al. (2017) used thousands of anonymous SNPs in *Plecotus austriacus*, the gray long-eared bat, to identify connectivity between populations and to identify populations that may be threatened by future climate change due to reduced connectivity and isolation. One of the important insights from the analysis of landscape functional connectivity is that many barriers to gene flow are cryptic across heterogeneous landscapes and require high-resolution genomic data to detect (Micheletti and Storfer 2017). This insight means that including genomic resolution data into connectivity analysis will vastly improve our knowledge of gene flow and biological movement.

Lastly, most plants have a wide range of variation in traits associated with fitness and are locally adapted to the environment in which they are found (Savolainen 2011). Though some variation in traits can be associated with genetic drift and gene flow, much of the trait variability is the product of natural selection from spatially varying environmental factors on traits beneficial to a specific environment, suggesting that the variation will differ across the species range (Sork 2018). Understanding the genomic basis of local adaptation will improve our predictions of the future spatial distribution of species under scenarios of rapid climate change. And population genomics approaches are being widely used to understand the genetic architecture of local adaptation in plants, animal, and other organisms (see the chapter "Landscape Genomics: Understanding Relationships Between Environmental Heterogeneity and Genomic Characteristics of Populations" in this book).

4.1.3 Dendrogenomics

Dendrogenomics is a combination of dendrochronology and population genomics, whereby phenotypes related to annual incremental growth (dendrophenotypes) can be associated with a plant's genome using genome-wide association techniques

(Evans et al. 2018). Past measures of growth in common gardens have been mostly restricted to height and diameter, which, as a single snapshot in time, may not accurately reflect the genetic architecture controlling fitness and local adaptation (Alberto et al. 2013). Dendrogenomics and the use of dendrophenotypes provide a longitudinal collection of phenotypes that will provide a higher-resolution picture of how trees have responded to past disturbance and climate fluctuations.

The dendrogenomic approach has only recently been introduced, but it has already yielded important support for local adaptation and the role of climatic constraints (Housset et al. 2018). Some early examples of dendrogenomics include the work by Johnson et al. (2017a) who tested hypotheses about variation in growth along an environmental gradient and its association with individual genomic diversity and by Heer et al. (2018) who test for associations between growth and adaptive genes during stress and disturbance events. In particular, Heer et al. (2018) identified dendrophenotypes in *Abies alba*, including the slope of the standardized tree ring index to identify the start of a period of stress linked with air pollution and drought and their association with fitness genes. They identified 15 genes associated with different dendrophenotypes that were related to drought stress and photosynthesis.

This is an area of population genomics that shows great promise for biogeography. Biogeographers, as we have already shown, are quite comfortable using dendroecological approaches, and by including genomics an additional layer of biological inference can be made, improving our understanding of adaptability and resilience under multiple environmental hazards and shifting disturbance regimes.

4.2 Historical Biogeography

It is important to mention again that because natural selection does not apparently affect large portions of the genome that are supposedly selectively neutral, they serve as indicators of demographic patterns, population ancestry, and the influence of ecological and geographic heterogeneity on gene flow (Selkoe and Toonen 2006). The historic conditions associated with founder and bottleneck events, small population size, and genetic drift leave a long-lasting mark on the gene pool of a population. Historical biogeography is benefiting from population genomics in several areas. Because of this, putatively neutral genomic variation can be used to assess historical biogeographic patterns. Some of the most important research avenues include demography, historic species distribution, colonization, dispersal, and evolution of lineages. Here we discuss how population genomics can benefit several aspects of historical biogeography.

4.2.1 Population Genomic Structure and Gene Flow

Demographic processes can be inferred by assessing the level of genetic distinctness or subdivision that occurs between spatially coherent populations. Distinct patterns

in population genomic variation can result from the conflicting effects of genetic drift and gene flow. Genetic drift will cause populations to become more genetically different, while gene flow will make them more similar. Because of ecological and geographic heterogeneity and isolation, gene flow is usually spatially restricted. Thus, based on the permeability of the landscape and the degree of physical separation, after successive generations, genetic drift will lead to an identifiable differentiation among populations. The high ability to discriminate between populations using population genomics information can often provide an improved understanding of the causes of historic changes in species distributions.

Population genomics allows one to assess population genetic structure by assessing numerous variable loci spread across the entire genome. There are several metrics in population genetics developed to look at the partitioning of genetic variation between subpopulations and the whole population based on neutral genomic regions. For instance, the most common metric used to determine the level of population substructure or subdivision is F_{ST} (Wright 1949), wherein larger F_{ST} values indicate that two or more distinct populations of a species are more differentiated and, if selection is not considered, supposedly experienced less gene flow between them than populations with lower values. By comparing genetic differences between populations, measured as F_{ST}, with landscape and ecological variables separating sites, it is possible to identify the spatial and ecological processes that are contributing to population connectivity. The literature is full of population genomic studies assessing patterns of genetic structure and admixture and have been evaluated in forest trees (Menon et al. 2018; Johnson et al. 2017b; Geraldes et al. 2014), shrubs (Xu et al. 2017; Lee et al. 2018), and wild flowers (Puzey et al. 2016; Barker et al. 2016). One of the advantages of using a population genomics approach to investigate population structure over inferences made using previous methods is that fine-scale patterns of introgression, hybridization, and speciation can be elucidated more easily and more precisely. For example, Menon et al. (2018) found in a range-wide analysis of southwestern white pine (*Pinus strobiformis*) and limber pine (*Pinus flexilis*) that gene flow was ongoing during species divergence and that species boundaries were a result of disruptive selection. Another example is the use of mitochondrial genomic markers to infer past colonization history of forests during Quaternary glaciations. Semerikov et al. (2018) sequenced 3 genomic mitochondrial markers in 90 populations of Scots pine (*Pinus sylvestris*) and identified 7 mitotypes of the species that reflect the split between western and eastern populations. Five of the mitotypes were found in the species western range with one mitotype in the east and one shared between east and west. Their findings provide support for a European and Ural refugia and recolonization to the east following glaciations. Both of these examples illustrate the level of insight into demographic processes that heretofore was difficult to obtain without genomic-level data. By using a genomics approach, historical biogeography will enhance our understanding of the history and spread of biodiversity. Case in point, *Populus alba* and *Populus tremula* admixture and backcrossing in Europe was explored using allozyme markers, finding that natural *P. alba* x *P. tremula* hybrids were backcrossing more with *P. alba* than with *P. tremula* (Rajora and Dancik 1992).

However, the more detailed recent studies using genome-wide approaches (RAD-seq and whole-genome resequencing) provided a higher resolution of the admixture patterns between *P. alba* and *P. tremula*, shedding light on their ancestral admixture, fine-scale chromosomal ancestry, pre- and post-zygotic barriers and selection maintaining reproductive isolation, genomic divergence and identification of speciation genes, and biogeography (Christe et al. 2016, 2017). Such information and resolution were not possible to obtain using allozymes. Population genomics research also revealed that in natural interspecific hybrids between *P. balsamifera* and *P. trichocarpa*, there was more introgression from *P. balsamifera* to *P. trichocarpa* than the reverse (Suarez-Gonzalez et al. 2018). These two examples illustrate that using genomic resolution data can provide more precise and detailed information than previous genetic approaches.

4.2.2 Paleogenomics

Paleogenomics, a term coined by Birnbaum et al. (2000), is concerned with the study of ancient DNA (aDNA) as a way to untangle the historical patterns of species distributions, paleopopulation dynamics, and evolution. Paleogenomics can aid in reconstruction of historic distributions of floras and patterns of colonization following the glacial events of the Quaternary using new genomic tools, which allow the analysis of aDNA (including paleoenvironmental aDNA and paleodietary aDNA). Ancient DNA can be obtained from samples collected from lake sediments, peats, permafrost soils, preserved megafaunal gut contents, coprolites, and other sources of preserved DNA with the potential to reconstruct floristics from the last 800 kyr (Birks and Birks 2015). In particular, this line of analysis may provide support identifying the locations of micro and macro glacial refugia.

To date, most paleogenomic studies have combined palynological analysis (pollen and macrofossil) with aDNA to reconstruct flora. Using ancient mtDNA extracted from lake sediment, Parducci et al. (2012) explored the historic distribution of *Picea abies* finding that the species survived glaciation in glacial refugia showing that they were present earlier than indicated by pollen analysis alone. Another study sequenced the *trn*L plastid region and part of the ITS1 spacer region from aDNA in 242 permafrost samples across the arctic representing 50 kyr of plant diversity and tracked the changes in plant composition through time (Willerslev et al. 2014). Other paleogenomic studies have used lake sediment (Boessenkool et al. 2014; Epp et al. 2015; Alsos et al. 2015), peats (Parducci et al. 2015), soils (Wilmshurst et al. 2013), and with aDNA alone in preserved middens (Murray et al. 2012). In general, aDNA is used in combination with pollen or other fossil data to both expand and validate reconstructions. As an exemplar of this approach, Parducci et al. (2015) reconstructed a European flora from the past 40 kyr and found that by including aDNA they could include five additional species that were undetected by either pollen or macrofossils alone and they were able to improve their reconstruction of glacial flora.

4.2.3 Phylogenomics

While much of population genomics has been firm in the wheelhouse of biologists with a focus on microevolutionary processes, phylogenetic and phylogeographic research have had a longer relationship with historical biogeography as well as an association with phylogenetic biology and macroevolutionary processes (Avise 2009). Phylogenetic studies explicitly deal with spatial and temporal dimensions of evolution and genealogy among taxa, while phylogeography tends to focus on untangling conspecific evolutionary lineages. While phylogenetics has relied on generating cladograms of taxa based on similarities or differences in phenotypic characters and more recently molecular characters, phylogeography has, from the start, relied on molecular-level variation to generate gene trees. In either case, the use of genetic markers in phylogenetic or phylogeographic research to date has relied heavily on cytoplasmic DNA, either mitochondrial or chloroplast sequences due to their known mutation rates based on the molecular clock (Kimura 1968). The emergence of high-throughput nucleotide sequencing is allowing a far more detailed analysis of evolutionary relationships to be resolved. This is an important shift in understanding the evolutionary history of species because when assessing gene trees from cytoplasmic DNA, usually only a small snapshot of a species genealogical history is represented (Avise 2010). Although challenges exist, generation of nuclear gene trees will be a major step forward in phylogenetics and phylogeography as they transition into phylogenomics.

The use of reduced representation genomic approaches, such as restriction-associated DNA sequencing (RAD-seq), and the incorporation of SNP markers have proven to be one of the most beneficial approaches in phylogenomics (McCormack et al. 2013; Leaché and Oaks 2017) because detailed patterns of phylogeography and insipient speciation can be discovered compared to past approaches (Emerson et al. 2010).

A comparison of SNP versus traditional microsatellite markers found that use of SNPs increased phylogeographic resolution and was better able to reconstruct past divergence events in red mangrove (*Rhizophora mangle*) (Hodel et al. 2017). This study showed that although microsatellites resulted in higher values of F_{ST}, RAD-seq, and a genomic approach, was able to resolve a classic phylogeographic break, where microsatellites could not.

Another classic example of combining phylogeography and RAD-seq involved the pitcher plant mosquito (*Wyeomyia smithii*), where Emerson et al. (2010) used the approach to resolve genetic structure and evolutionary direction in the species following Pleistocene glaciation in the southern Appalachian Mountains, USA. They found a phylogeographic separation of *W. smithii* into both a northern and southern group, a finding that agrees with the current distribution of the species. It is likely that the northern clade originated from a southern glacial refugium. Importantly, using only genomic data, Emerson et al. (2010) showed that high-throughput genomics could be used in a phylogeographic framework to resolve evolutionary history of a species. The RAD-seq approach has been used successfully to resolve

phylogeographic relationships in several tree species (Hodel et al. 2017; Zhou et al. 2018; Deng et al. 2018; Fitz-Gibbon et al. 2017; Hipp et al. 2018), as well as other diverse taxa (Herrera and Shank 2016; Wagner et al. 2013).

Another genomic approach is the use of ultraconserved genomic elements (UCEs) and conserved ortholog sets (COS) (Krutovsky et al. 2006; Faircloth et al. 2012). This approach is allowing comparable genomic regions to be analyzed across species and across millions of years of evolutionary history. UCEs, in particular, allow comparable DNA fragments from distantly related species to be aligned and compared so that evolutionary history can be inferred. Faircloth et al. (2012) demonstrated the utility of this approach by sequencing 2020 UCEs in 10 Amniota genomes and then using the UCE-anchored loci successfully recovered the known phylogeny of 9 non-model avian species.

Challenges, of course, still exist for phylogenomic studies using genomic data (Leaché and Oaks 2017). However, the exponential increase in phylogenomic studies using SNP data is evidence that historical biogeographers have a powerful new tool for assessing the evolutionary history of organisms.

Lastly, a growing literature is championing the inclusion of SDMs in phylogenomic studies (Alvarado-Serrano and Knowles 2014; Scoble and Lowe 2010). In particular, the use of SDMs allows for the development and evaluation of phylogeographic hypothesis by hindcasting or forecasting a species niche (Richards et al. 2007) and allowing researchers to identify hypothesized locations of past populations (Knowles et al. 2007) or to infer the future distribution and divergence of species. Biogeography, as mentioned in Sect. 3.1.1, has been at the forefront of developing SDMs, and a more refined understanding of the origin, spread, and diversity of life can be achieved by combining population genomics approaches and SDMs into historical biogeography.

5 Population Genomics Inference of Mountain Hemlock (*Tsuga mertensiana* Bong (Carr)) Biogeography: A Case Study

To illustrate how a population genomics perspective can be integrated into biogeography research, we provide a case study using an example of our own research on mountain hemlock (*Tsuga mertensiana* Bong (Carr)) on the Kenai Peninsula, Alaska (Johnson et al. 2017a, b, c). Global ecological change is a serious threat to biodiversity at all scales. Changes in temperature and precipitation along with the introduction of invasive pathogens, shifts in disturbance regimes, and changes in land cover and land use will have unforeseen consequences for biological life. Knowing how plants have responded to changes in the past climate and accounting for alternative future responses will help managers and policy makers plan for the future. We conducted a multiscale genomic study to understand how a long-lived

conifer tree, mountain hemlock, has responded to past climate variability and to assess its potential to respond to future changes.

We used a biogeographic and population genomics approach based on double-digest restriction-associated DNA sequencing (ddRAD-seq) (Peterson et al. 2012) to address both historical and ecological research questions. Specifically we sought (1) to determine if isolated stands of mountain hemlock on the Kenai Peninsula of Alaska were glacial relicts or the product of rare long-distance dispersal following glacial retreat (Johnson et al. 2017b), (2) to test if local seed production or dispersal from beyond the ecotone is driving tree line dynamics (Johnson et al. 2017c), and (3) to test if individual trees with higher average genomic diversity were better able to grow at alpine tree line (Johnson et al. 2017a). This combination of hypotheses allowed us to test ecological biogeography hypothesis about patterns of contemporary seed dispersal, potential rates of migration in response to climate change, and patterns of adaptability and also allowed us to address historical biogeography questions related to gene flow and forest response to Pleistocene glaciation.

Mountain hemlock is a highly outcrossed, monoecious, wind-pollinated species with large winged seeds and pollen (Owens and Molder 1975; Means 1990; Ally et al. 2000). The tree species is usually found in cool wet environments along the Kenai coastal, alpine, and subalpine zones and is a major component of the forest along the Gulf of Alaska coast. Mountain hemlock stand expansion and migration are related to length of growing season, a function of winter temperature and snowpack, and summer temperatures and moisture availability (Peterson and Peterson 2001; Taylor 1995). High elevation mountain hemlock growth correlates negatively to spring snowpack depth and positively to summer growing season temperature (Taylor 1995; Peterson and Peterson 2001). Additionally, warm July temperatures result in increased seed production (Woodward et al. 1994). These combinations of factors suggest that the species will migrate to higher elevations and latitudes as climate change advances.

Climate on the Kenai is boreal maritime with both temperature and precipitation gradients from east to west. Nearly all of the Kenai was covered by the late Wisconsin Cordilleran ice sheet approximately 26,000–14,500 years ago (Rymer and Sims 1982; Ager 2007). A few microrefugia have been proposed to have harbored species during this glaciation in the northwest Kenai Mountains and the eastern Kenai Lowlands (Jones et al. 2009). Though there is no palynological evidence of mountain hemlock being present in these purported microrefugia during the past glaciation, their survival there cannot currently be ruled out.

We sampled needle tissue from eight populations across the Kenai Peninsula consisting of ten individuals per population to assess intraspecific genomic variation. At one site, we conducted an exhaustive sampling of 168 individuals to assess both seed movement within the alpine tree line ecotone and to assess the relationship of dendrophenotypes with observed individual heterozygosity. We extracted genomic DNA from all individuals and used ddRAD-seq to generate a dataset of 6124 SNPs for our analysis.

We created both contemporary and mid-Holocene SDMs and generated a resistance surface based on the species climate niche. We then calculated population

genetic distance and landscape resistance matrices to test our alternative hypotheses related to our expectations of glacial microrefugia or long-distance dispersal using circuit analysis. Additionally we conducted an exclusion-based parentage analysis to quantify seed dispersal distance and direction and to estimate immigration of seeds into the alpine tree line ecotone.

Lastly, it has been hypothesized that trees with higher levels of genomic diversity (measured as average individual heterozygosity (IndHet) should have more stable growth patterns (Babushkina et al. 2016; Mitton 1978). We tested this hypothesis related to local adaptation by combining dendroecology with our genomic dataset. We assessed if the dendrophenotypes of average tree ring width (AvrTRW) and variance in tree ring width (VarTRW) were correlated with IndHet.

Our historical analysis showed that genomic diversity and population genomic structure differed between isolated stands of mountain hemlock and those found across the rest of the peninsula. An isolation-by-resistance approach based on electrical circuit analysis identified high landscape connectivity and conductance across the peninsula. Genetic variation was primarily explained by landscape resistance and not geographic distance (isolation-by-distance) based on redundancy analysis (Fig. 2a). These findings suggest that mountain hemlock colonized the peninsula via long-distance dispersal and repeated founding events accompanied by high levels of ongoing gene flow. To address ecological biogeographic questions, we studied patterns of dispersal and growth at the local scale (Fig. 2b). To better understand contemporary patterns of dispersal, an exclusion-based parentage analysis identified seed dispersal events ranging from 1.44 m to 326.85 m with a mean dispersal distance of 73 m (Fig. 2d). Most seeds arrived as immigrants from beyond the tree line. Overall direction of dispersal was downslope with the longest dispersal events occurring in that direction. However, a few dispersal events did allow seeds to move to higher elevations suggesting a capacity to advance tree lines to higher elevations (Fig. 2c). Long-distance dispersal was quantified at the 99th percentile of the dispersal curve and accounted for dispersal at distances greater than 450 m (Fig. 2d). This analysis indicates that mountain hemlock tree line stability is not necessarily controlled by local seed production but via seed immigration from beyond the study area and at distances greater than 0.5 km.

Lastly, our analysis of genomic adaptability found that there was no significant correlation between IndHet and either AvrTRW or VarTRW. However, AvrTRW and VarTRW were significantly correlated ($P < 0.01$) suggesting that under poor growing conditions, trees grow poorly regardless of the level of individual genomic diversity.

This combination of research represents a novel integration of genomics and geography to answer a pertinent set of questions allowing us to have a deeper understanding of how plants may respond to shifting climate conditions. Moreover, this analysis demonstrates how using population genomics allows biogeography to holistically address questions spanning both historical and ecological subfields of the discipline.

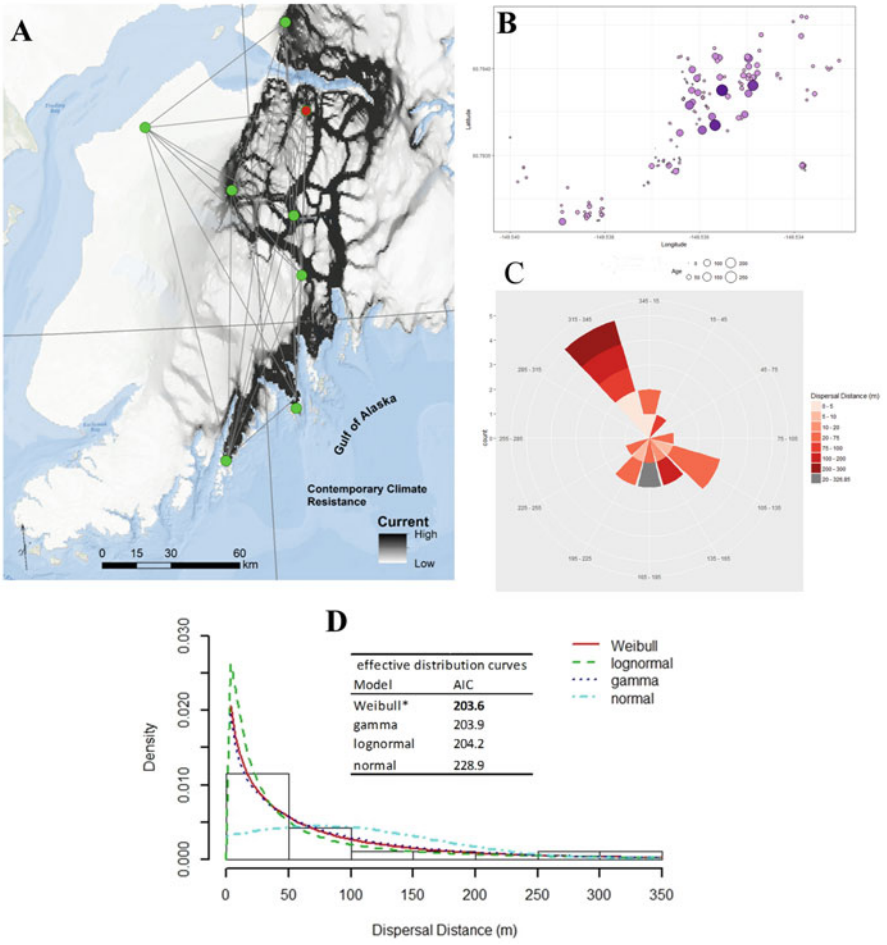

Fig. 2 Population genomic analysis of mountain hemlock at historical (**a**) and ecological time scales (**b–d**) and landscape (**a**) to local spatial scales (**b–d**). (**a**) Kenai Peninsula, Alaska, with eight sampling populations (green dots). Species distribution modeling (SDM) was used to generate a contemporary resistance surface that was used to generate a matrix of resistance distances (current) between each population, characterizing the ease at which propagules could move across the landscape as a function of the climatic niche. Additionally, pairwise F_{st} was calculated to test for isolation-by-geographic distance (gray lines). Redundancy analysis found that contemporary climate was the best explanation for observed genomic variation ($P < 0.05$). (**b**) The site-level analysis located in the north-central portion of the Kenai Peninsula (red dot). The spatial arrangement of trees cored and sampled along a single transect were used to conduct both a parentage analysis and an assessment of adaptive growth potential. The size of the purple circle is a function of the age of the tree. (**c**) The wind rose shows the magnitude and direction of dispersal events identified from our exclusion-based parentage analysis. Color of the wind rose corresponds to dispersal distance, and each segment represents a dispersal event. Most dispersal events were downslope; however, several dispersal events did move seeds to higher elevations over relatively long distances (20–200 m). (**d**) The empirical dispersal events were used to parameterize a dispersal kernel, so that the 99th percentile of the tail could be quantified to calculate long-distance dispersal. Of the probability distribution functions fit, the Weibull distribution best fit the empirical data based on Akaike information criterion (AIC), and long-distance dispersal was quantified as seed movement greater than 0.4 km (from Johnson et al. 2017c)

5.1 Application of Biogeography in Biodiversity Conservation: Conservation Biogeography

Concerns about the future of biodiversity globally have been growing over the past few decades as a result of human-modified landscapes and climate change. Species extinctions have been occurring at a rapid pace, and it has been proposed that biogeographic research can contribute to biodiversity conservation through the vehicle of conservation biogeography (Whittaker et al. 2005). This biogeography research uses many of the tools that we have discussed in this chapter. Conservation biogeography is an applied biogeography that aims to address questions of biodiversity conservation and distributional dynamics. In particular, conservation biogeography relies on the use of SDMs to identify ecological factors that correlate with species distribution (niche) in order to better manage biodiversity and set conservation priorities (Franklin 2010). As we have already demonstrated, using population genomics and landscape genomics methods in conjunction with SDMs will allow the identification of cryptic barriers to dispersal and gene flow as well as the locations of potential refugia (Scoble and Lowe 2010). The inclusion of genomic-scale data will lend additional rationality to conservation decisions.

Incorporating population genomic variation can improve conservation decision support models that are frequently used to prioritize landscapes and species for protection by further identifying variation in locally adapted populations so that conservation models are not static across the range of the species. Population genomics approaches that identify portions of a species range with higher adaptive capacity associated with projections of future environmental conditions would be candidates for conservation. As an example of conservation biogeography incorporating molecular tools, Kraft et al. (2010) used intergenic transcribed spacer (ITS)-based estimates of age for 337 neoendemics in the California flora and data on the range size to identify high areas of biodiversity (hotspot). They found that contrary to the prevailing thoughts on biodiversity hotspots in California, their combined molecular and distributional data shifted the estimates of high endemism from the coast toward the desert and Great Basin regions of the state. Importantly, they showed that many of the areas of high endemism fall outside of the current protected areas and illustrate how inclusion of molecular-level data can improve our decision on how to prioritize conservation areas.

Biogeography theory, specifically island biogeography, has been used extensively in reserve design oscillating between the single large or several small (SLOSS) model as it applies to size and geographic isolation of reserve patches (Diamond 1975). Understanding the population genomic ramifications of these designs including the functional connectivity and degree of reproductive isolation and its effect on genomic diversity is important. However, the intervening matrix characteristics must also be considered (Kupfer et al. 2006). Additionally, our current underestimation of global biodiversity, in particular at the microscales,

known colloquially as the Linnean shortfall (Lomolino et al. 2017), and our lack of knowledge about the corresponding geographic distribution of their ranges, known as the Wallace shortfall (Lomolino 2004), mean that we do not have a complete picture of global biodiversity. Here, again, genomic approaches such as environmental DNA (eDNA) sequencing and metagenomics can help improve our cataloging of species and their distribution by identifying the community composition of microscopic organisms that have been difficult to identify.

6 Challenges and Future Research Avenues

The toolbox that biogeography has at its disposal is quite diverse as we have demonstrated. This list of tools, concepts, and approaches is by no means complete, and geographers have long adapted and adopted techniques and theories from many other fields of science. By incorporating a population genomics approach, biogeography will be able to address many of the fundamental questions that have long been intractable due to costs and limited biological resolution. For instance, identifying how organisms move across the landscape and respond to fragmentation and climate change or testing how geographic isolation and the possible reductions in genetic diversity will lead to speciation, extinction, or adaptation. We can begin to address these questions using population genomics and geographic techniques.

Tobler's first law of geography states that "everything is related to everything else, but near things are more related than distant things" (Tobler 1970). This is a testable hypothesis using genetics and genomics techniques. One example is the process of dispersal in forest trees; gene flow between populations occurs in two ways, either by pollen or seed dispersal. In both cases, in order for gene flow to take place, successful establishment of a seedling must occur, known as effective dispersal (Cain et al. 2000). So, to evaluate Tobler's first law, biogeographers must first ask the question: Do the relationships of trees to each other decrease with increasing distance as would be expected by chance (Degen et al. 2004)? In general, gene flow is expected to demonstrate less differentiation between neighboring populations than distant ones (Muir and Schlötterer 2005). Ally and Ritland (2007) tested the spatial genetic structure of *Tsuga mertensiana* and found that relatedness of individual trees decreased with increasing pairwise distance, thus supporting Tobler's first law. Similar pattern was observed for *Thuja occidentalis* (Pandey and Rajora 2012).

There are a number of challenges that affect biogeography and its incorporation of population genomics. Though the costs of genomics and genetics analysis have decreased substantially, and specialized expensive equipment has become easily available via sequencing services and genome centers, a functional theoretical understanding of molecular and population genetics is required to interpret results appropriately (Allendorf et al. 2010, 2012; Benestan et al. 2016). Currently, a few biogeographers have appropriate experience with the genetics theory, but this must

change. Without a base understanding of the genetics theory, it will be difficult to design a research project and interpret the results in a meaningful way.

MacDonald (2000) realized that biogeographers have tremendous opportunities to contribute to research that will make transformative advances to the well-being of the planet and suggested that in order to take advantage, interdisciplinary research teams will be essential. Assembling teams of researchers with unique specialties and in-depth expertise can allow for quite productive research. We agree with MacDonald (2000) that participating in interdisciplinary teams that include geneticists, bioinformaticians, and biogeographers and allow groups to take advantage of new technological advances will allow biogeography to flourish.

Population genomics challenges associated with study design and interpretation have been discussed in this book. In addition to the problems associated with study design and interpretation, there are also issues of data overload. At present, the amount of genomic and environmental data generated far outstrips our ability to store and analyze it (Parisod and Holderegger 2012). The sequencing capabilities of high-throughput massively parallel sequencing techniques are on the order of billions of base pairs. Improved bioinformatics techniques are critical to future advancement in the use of genetic and environmental data (Pop and Salzberg 2008) in ecological biogeography.

The acquisition of geographic and spatial data has also exploded, similarly to genomic data, and is on track to outstrip the analytical capabilities of researchers without proper informatics system and algorithm improvements. Much of the rapid increase in environmental spatial data acquisition is a result of automated sensor development and deployment and decreased costs of use over the last couple of decades (Guillot et al. 2009). Porter et al. (2012) suggested a sister field to bioinformatics termed ecoinformatics. Developments in sensor design, signal processing algorithms, wideband communication systems, and new storage techniques are among the problems that must be overcome if research is to optimize the vast quantities of new data (Baraniuk 2011). If next-generation massively parallel sequencing can generate billions of nucleotide records, high-capacity sensors also have the ability to collect billions of records a year. Baraniuk (2011) pointed out that a bottleneck exists and that, in 2011, the amount of data generated exceeded twice the global storage capacity.

In both bio- and ecoinformatics, it will be important in the future to improve on the quality assurance and quality control of the vast quantities of data and the methods used for analysis. From system design to data transmission, analysis, and storage advances in automated methods will be essential. We believe, as Porter et al. (2012) stressed in their conclusions, that it will be imperative for future ecologists and biologists, and we add biogeographers, to have a basic understanding of informatics allowing them to navigate the technological tools that will enable them to analyze ecological and sensor data helping them to answer interdisciplinary questions.

7 Future Direction in Biogeography

From a biogeography perspective, adopting a population genomics approach, facilitated by a dramatic increase in the number of variable genome-wide markers used, will advance the field by leaps. Adoption of this approach should improve the precision of estimating species distributions, rates of movement, functional connectivity, historical evolution of lineages, and population demographic parameters, such as effective population size. Variations on whole-genome sequencing and reduced representation sequencing of populations to assess their intraspecific variation will permit the analysis of genomic diversity at a much higher resolution compared to past genetic markers, such as microsatellites, AFLPs, and allozymes. High-throughput sequencing approaches can allow rapid marker identification that can be compared and combined across a diverse set of collaborators and laboratories. Population genomics approaches will further open up the possibility of screening individuals and populations for adaptive loci, which will improve our assessment of populations that may be vulnerable to rapidly changing environments. Biogeography research has long integrated methods and approaches from a diverse array of research fields, and the rewards of adopting a population genomics perspective will propel both ecological and historical biogeography forward, as well as unify the two subdisciplines under an overarching theory of evolution. Indeed, the most important discoveries are likely to come about by combining these two approaches. As population genomics methods and models improve, the depth to which biogeography inquiry can be addressed will flourish, and questions, such as what will the spatial pattern and configuration of future species distributions look like as a result of climate change, that have heretofore been beyond our grasp will finally be addressed. Novel technologies and bioinformatics pipelines in conjunction with high-throughput phenotyping and open-source analytical packages are still emerging. As these technologies mature and costs are further reduced, biogeography can move from assessing gene flow and landscape connectivity using neutral genetic markers to identifying putatively adaptive loci as targets of selection. Moreover, an increase in the availability of full-genome sequences will make possible whole-genome comparisons across geographic regions allowing for the identification of a broad array of ecological and climatic factors influencing biological processes.

As we move forward, it will be important to move beyond an assessment of population genetic structure to infer potential future responses to climate and landscape change and to begin to predict how species will respond based on our knowledge of adaptive capacity (Holderegger et al. 2010; Storfer et al. 2010; Neale and Kremer 2011; Manel and Holderegger 2013; Sork et al. 2013). Finding general responses across multiple species will continue to be an important research goal (Ćalić et al. 2015). A recent adoption of landscape genomics approaches in community ecology, known as landscape community genomics (Hand et al. 2015), could be incorporated into biogeography to improve our understanding of species assemblages and their responses to global change.

8 Conclusions

Population genomics will greatly enhance the resolution at which biogeography questions can be addressed through both ecological and historical lenses. Incorporating the evolutionary perspective that population genomics approaches provide has the potential to unify the discipline of biogeography in all of its various disciplinary configurations (e.g., geography, biology, geology, and ecology). The rapid advancement in genomics technologies and the affordability of sequencing will allow various approaches outlined here to be tailored to a wide range of research endeavors. In the future, we may see field-based sequencing, affordable whole-genome sequencing, and a wave of epigenetic approaches that will allow very precise and detailed investigation of the origin, spread, and distribution of species as well as their potential future patterns.

Acknowledgments JSJ received support from a National Science Foundation Grant [BCS-1333527]. We would like to thank Aaron Shafer for constructive feedback on an earlier version of this chapter.

References

Ager TA. Vegetation response to climate change in Alaska: examples from the fossil record. Vol Open-file report 2007–1096. US Geological Survey. 2007.

Alberto FJ, Aitken SN, Alía R, González-Martínez SC, Hänninen H, Kremer A, et al. Potential for evolutionary responses to climate change – evidence from tree populations. Glob Chang Biol. 2013;19(6):1645–61. https://doi.org/10.1111/gcb.12181.

Alftine KJ, Malanson GP. Directional positive feedback and pattern at an alpine tree line. J Veg Sci. 2004;15:3–12.

Allendorf FW, Hohenlohe PA, Luikart G. Genomics and the future of conservation genetics. Nat Rev Genet. 2010;11(10):697–709. https://doi.org/10.1038/nrg2844.

Allendorf FW, Luikart GH, Aitken SN. Conservation and the genetics of populatins. 2nd ed. Chichester: Wiley; 2012.

Ally D, Ritland K. A case study: looking at the effects of fragmentation on genetic structure in different life history stages of old-growth mountain hemlock (*Tsuga mertensiana*). J Hered. 2007;98(1):73–8. https://doi.org/10.1093/jhered/esl048.

Ally D, El-Kassaby YA, Ritland K. Genetic diversity, differentiation and mating system in mountain hemlock (*Tsuga mertensiana*) across British Columbia. For Genet. 2000;7:97–108.

Alsos IG, Sjögren P, Edwards ME, Landvik JY, Gielly L, Forwick M, et al. Sedimentary ancient DNA from Lake Skartjørna, Svalbard: assessing the resilience of arctic flora to Holocene climate change. The Holocene. 2015;26(4):627–42. https://doi.org/10.1177/0959683615612563.

Alvarado-Serrano DF, Knowles LL. Ecological niche models in phylogeographic studies: applications, advances and precautions. Mol Ecol Resour. 2014;14(2):233–48. https://doi.org/10.1111/1755-0998.12184.

Ammann B, van Leeuwen JFN, van der Knaap WO, Lischke H, Heiri O, Tinner W. Vegetation responses to rapid warming and to minor climatic fluctuations during the Late-Glacial Interstadial (GI-1) at Gerzensee (Switzerland). Palaeogeogr Palaeoclimatol Palaeoecol. 2013;391:40–59. https://doi.org/10.1016/j.palaeo.2012.07.010.

Andrewartha HG, Birch LC. The distribution and abundance of animals. Chicago: University of Chicago Press; 1954.

Araújo MB, Whittaker RJ, Ladle RJ, Erhard M. Reducing uncertainty in projections of extinction risk from climate change. Glob Ecol Biogeogr. 2005;14(6):529–38. https://doi.org/10.1111/j.1466-822X.2005.00182.x.

Ashman T-L, Knight TM, Steets JA, Amarasekare P, Burd M, Campbell DR, et al. Pollen limitation of plant reproduction: ecological and evolutionary causes and consequences. Ecology. 2004;85(9):2408–21. https://doi.org/10.1890/03-8024.

Auffret AG, Rico Y, Bullock JM, Hooftman DAP, Pakeman RJ, Soons MB, et al. Plant functional connectivity – integrating landscape structure and effective dispersal. J Ecol. 2017;105:1648–56. https://doi.org/10.1111/1365-2745.12742.

Avise JC. Phylogeography: the history and formation of species. Cambridge: Harvard University Press; 2000.

Avise JC. Phylogeography: retrospect and prospect. J Biogeogr. 2009;36(1):3–15. https://doi.org/10.1111/j.1365-2699.2008.02032.x.

Avise JC. Perspective: conservation genetics enters the genomics era. Conserv Genet. 2010;11:665–9.

Avise JC, Arnold J, Ball RM, Bermingahm E, Lamb T, Neigel JE, et al. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu Rev Ecol Syst. 1987;18:489–522.

Babushkina EA, Vaganov EA, Grachev AM, Oreshkova NV, Belokopytova LV, Kostyakova TV, et al. The effect of individual genetic heterozygosity on general homeostasis, heterosis and resilience in Siberian larch (*Larix sibirica* Ledeb.) using dendrochronology and microsatellite loci genotyping. Dendrochronologia. 2016;38:26–37. https://doi.org/10.1016/j.dendro.2016.02.005.

Baerwald TJ. Prospects for geography as an interdisciplinary discipline. Ann Assoc Am Geogr. 2010;100:493–501.

Balkenhol N, Dudaniec RY, Krutovsky KV, Johnson JS, Cairns DM, Segelbacher G, et al. Landscape genomics: understanding relationships between environmental heterogeneity and genomic characteristics of populations. In: Rajora OP, editor. Population genomics concepts, strategies and approaches. Cham: Springer; 2017.

Baraniuk RG. More is less: signal processing and the data deluge. Science. 2011;331:717–9.

Barker BS, Andonian K, Swope SM, Luster DG, Dlugosch KM. Population genomic analyses reveal a history of range expansion and trait evolution across the native and invaded range of yellow starthistle (*Centaurea solstitialis*). Mol Ecol. 2016;26(4):1131–47. https://doi.org/10.1111/mec.13998.

Benestan LM, Ferchaud AL, Hohenlohe PA, Garner BA, Naylor GJP, Baums IB, et al. Conservation genomics of natural and managed populations: building a conceptual and practical framework. Mol Ecol. 2016;25(13):2967–77. https://doi.org/10.1111/mec.13647.

Birks HJB, Birks HH. How have studies of ancient DNA from sediments contributed to the reconstruction of quaternary floras? New Phytol. 2015;209(2):499–506. https://doi.org/10.1111/nph.13657.

Birnbaum D, Coulier F, Pébusque MJ, Pontarotti P. "Paleogenomics": looking in the past to the future. J Exp Zool. 2000;288(1):21–2. https://doi.org/10.1002/(sici)1097-010x(20000415)288:1<21::aid-jez2>3.0.co;2-q.

Blumler MA, Cole A, Flenley J, Schickhoff U. History of biogeographical thought. In: Millington AC, Blumler MA, Schickhoff U, editors. The Sage handbook of biogeography. Los Angeles: SAGE; 2011.

Boessenkool S, McGlynn G, Epp LS, Taylor D, Pimentel M, Gizaw A, et al. Use of ancient sedimentary DNA as a novel conservation tool for high-altitude tropical biodiversity. Conserv Biol. 2014;28(2):446–55. https://doi.org/10.1111/cobi.12195.

Bond WJ, Keeley JE. Fire as a global 'herbivore': the ecology and evolution of flammable ecosystems. Trends Ecol Evol. 2005;20(7):387–94. https://doi.org/10.1016/j.tree.2005.04.025.

Bothwell HM, Cushman SA, Woolbright SA, Hersch-Green Erika I, Evans LM, Whitham TG, et al. Conserving threatened riparian ecosystems in the American west: precipitation gradients and river networks drive genetic connectivity and diversity in a foundation riparian tree (*Populus angustifolia*). Mol Ecol. 2017;26(19):5114–32. https://doi.org/10.1111/mec.14281.

Bruening JM, Bunn AG, Salzer MW. A climate-driven tree line position model in the White Mountains of California over the past six millennia. J Biogeogr. 2018;45:1067–76. https://doi.org/10.1111/jbi.13191.

Buffon GLL. Histoire Naturelle, Général et Particuliére. Paris: Imprimerie Nationale; 1791.

Bullock JM, Moy IL. Plants as seed traps: inter-specific interference with dispersal. Acta Oecol. 2004;25(1):35–41. https://doi.org/10.1016/j.actao.2003.10.005.

Bullock JM, González LM, Tamme R, Götzenberger L, White SM, Pärtel M, et al. A synthesis of empirical plant dispersal kernels. J Ecol. 2017;105(1):6–19. https://doi.org/10.1111/1365-2745.12666.

Cain ML, Milligan BG, Strand AE. Long-distance seed dispersal in plant populations. Am J Bot. 2000;87(9):1217–27.

Cairns DM, Lafon CW, Waldron JD, Tchakerian M, Coulson RN, Klepzig KD, et al. Simulating the reciprocal interaction of forest landscape structure and southern pine beetle herbivory using LANDIS. Landsc Ecol. 2008;23(4):403–15. https://doi.org/10.1007/s10980-008-9198-7.

Ćalić I, Bussotti F, Martínez-García PJ, Neale DB. Recent landscape genomics studies in forest trees—what can we believe? Tree Genet Genomes. 2015;12(1):1–7. https://doi.org/10.1007/s11295-015-0960-0.

Candolle APD. Essai elementaire de géographie botanique. Dictionnaire des sciences naturelles, vol. 18. Paris: Levrault; 1820.

Chhetri PK, Cairns DM. Dendroclimatic response of Abies spectabilis at treeline ecotone of Barun Valley, eastern Nepal Himalaya. J For Res. 2016;27(5):1163–70. https://doi.org/10.1007/s11676-016-0249-7.

Chhetri PK, Shrestha KB, Cairns DM. Topography and human disturbances are major controlling factors in treeline pattern at Barun and Manang area in the Nepal Himalaya. J Mt Sci. 2017;14(1):119–27. https://doi.org/10.1007/s11629-016-4198-6.

Christe C, Stolting KN, Bresadola L, Fussi B, Heinze B, Wegmann D, et al. Selection against recombinant hybrids maintains reproductive isolation in hybridizing *Populus* species despite F_1 fertility and recurrent gene flow. Mol Ecol. 2016;25:2482–98.

Christe C, Stölting KN, Paris M, Fraïsse C, Bierne N, Lexer C. Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. Mol Ecol. 2017;26(1):59–76. https://doi.org/10.1111/mec.13765.

Clark JS. Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. Am Nat. 1998;152(2):204–24. https://doi.org/10.1086/286162.

Clark JS, Fastie C, Hurtt G, Jackson ST. Reid's paradox of rapid plant migration. Bioscience. 1998;48:13–24.

Connell JH. The influence of interspecific competition and other factors on the distribution of the barnacle *Chthamalus stellatus*. Ecology. 1961;42(4):710–23. https://doi.org/10.2307/1933500.

Cowell CM, Parker AJ. Biogeography in the annals. Ann Assoc Am Geogr. 2004;94(2):256–68. https://doi.org/10.1111/j.1467-8306.2004.09402002.x.

Cox CB, Moore PD, Ladle R. Biogeography an ecological and evolutionary approach. 9th ed. Hoboken: Wiley; 2016.

Cushman SA, McRae BH, McGarigal K. Basics of landscape ecology: an introduction to landscapes and population processes for landscape geneticists. In: Balkenhol N, Cushman Samuel A, Storfer AT, Waits LP, editors. Landscape genetics: concepts, methods, applications. West Sussex: Wiley; 2016. https://doi.org/10.1002/9781118525258.ch02.

Danby RK, Hik DS. Variability, contingency and rapid change in recent subarctic alpine tree line dynamics. J Ecol. 2007;95(2):352–63. https://doi.org/10.1111/j.1365-2745.2006.01200.x.

Davis MB. Quaternary history and the stability of forest communities. In: West DC, Shugart HH, Botkin DB, editors. Forest succession, Springer advanced texts in life sciences. New York: Springer; 1981. p. 132–53. https://doi.org/10.1007/978-1-4612-5950-3_10.

Degen B, Bandou E, Caron H. Limited pollen dispersal and biparental inbreeding in *Symphonia globulifera* in French Guiana. Heredity. 2004;93:585–91.

Delcourt PA, Delcourt HR. Vegetation maps for eastern North America: 40,000 Yr BP to the present. In: Romans R, editor. Geobotany II. New York: Plenum; 1981. p. 123–66.

Delcourt HR, Delcourt PA, Webb T III. Dynamic plant ecology: the spectrum of vegetational change in space and time. Quat Sci Rev. 1982;1(3):153–75.

Deng M, Jiang X-L, Hipp AL, Manos PS, Hahn M. Phylogeny and biogeography of east Asian evergreen oaks (*Quercus* section *Cyclobalanopsis*; *Fagaceae*): insights into the Cenozoic history of evergreen broad-leaved forests in subtropical Asia. Mol Phylogenet Evol. 2018;119:170–81. https://doi.org/10.1016/j.ympev.2017.11.003.

Despres L, Loriot S, Gaudeul M. Geographic pattern of genetic variation in the European globeflower *Trollius europaeus* L. (Ranunculaceae) inferred from amplified fragment length polymorphism markers. Mol Ecol. 2002;11(11):2337–47. https://doi.org/10.1046/j.1365-294X.2002.01618.x.

Diamond JM. Colonization of exploded volcanic islands by birds: the supertramp strategy. Science. 1974;184(4138):803–6.

Diamond JM. The island dilemma: lessons of modern biogeographic studies for the design of natural reserves. Biol Conserv. 1975;7(2):129–46. https://doi.org/10.1016/0006-3207(75)90052-X.

Dietz R. Continent and ocean basin evolution by spreading of the sea floor. Nature. 1961;190:854–7.

Dow BD, Ashley MV. Microsatellite analysis of seed dispersal and parentage of saplings in bur oak, *Quercus macrocarpa*. Mol Ecol. 1996;5(5):615–27. https://doi.org/10.1111/j.1365-294X.1996.tb00357.x.

Dyer RJ. The evolution of genetic topologies. Theor Popul Biol. 2007;71:71–9.

Dyer RJ. Population graphs and landscape genetics. Annu Rev Ecol Evol Syst. 2015;46(1):327–42. https://doi.org/10.1146/annurev-ecolsys-112414-054150.

Dyer RJ, Nason JD. Population graphs: the graph theoretic shape of genetic structure. Mol Ecol. 2004;13(7):1713–27.

Dyer R, Chan D, Gardiakos V, Meadows C. Pollination graphs: quantifying pollen pool covariance networks and the influence of intervening landscape on genetic connectivity in the north American understory tree, *Cornus florida* L. Landsc Ecol. 2012;27(2):239–51. https://doi.org/10.1007/s10980-011-9696-x.

Elliott G. The role of thresholds and fine-scale processes in driving upper treeline dynamics in the Bighorn Mountains, Wyoming. Phys Geogr. 2012;33(2):129–45. https://doi.org/10.2747/0272-3646.33.2.129.

Emerson KJ, Clayton RM, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, et al. Resolving postglacial phylogeography using high-throughput sequencing. Proc Natl Acad Sci. 2010;107(37):16196–200. https://doi.org/10.2307/20779639.

Endler JA. Geographic variation, speciation, and clines. Princeton: Princeton University Press; 1977.

Epp LS, Gussarova G, Boessenkool S, Olsen J, Haile J, Schrøder-Nielsen A, et al. Lake sediment multi-taxon DNA from North Greenland records early post-glacial appearance of vascular plants and accurately tracks environmental changes. Quat Sci Rev. 2015;117:152–63. https://doi.org/10.1016/j.quascirev.2015.03.027.

Etterson JR, Schneider HE, Gorden NLS, Weber JJ. Evolutionary insights from studies of geographic variation: contemporary variation and looking to the future. Am J Bot. 2016;103(1):5–9. https://doi.org/10.3732/ajb.1500515.

Evans MEK, Gugger PF, Lynch AM, Guiterman CH, Fowler JC, Klesse S, et al. Dendroecology meets genomics in the common garden: new insights into climate adaptation. New Phytol. 2018;218(2):401–3. https://doi.org/10.1111/nph.15094.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst Biol. 2012;61:717–26. https://doi.org/10.1093/sysbio/sys004.

Ferrarini A, Selvaggi A, Abeli T, Alatalo JM, Orsenigo S, Gentili R, et al. Planning for assisted colonization of plants in a warming world. Sci Rep. 2016;6:28542. https://doi.org/10.1038/srep28542. https://www.nature.com/articles/srep28542#supplementary-information

Fitz-Gibbon S, Hipp AL, Pham KK, Manos PS, Sork VL. Phylogenomic inferences from reference-mapped and de novo assembled short-read sequence data using RADseq sequencing of California white oaks (*Quercus* section *Quercus*). Genome. 2017;60(9):743–55. https://doi.org/10.1139/gen-2016-0202.

Fitzpatrick MC, Keller SR. Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. Ecol Lett. 2015;18(1):1–16. https://doi.org/10.1111/ele.12376.

Flatley WT, Lafon CW, Grissino-Mayer HD, LaForest LB. Changing fire regimes and old-growth forest succession along a topographic gradient in the great Smoky Mountains. For Ecol Manag. 2015;350(Supplement C):96–106. https://doi.org/10.1016/j.foreco.2015.04.024.

Fordham Damien A, Akçakaya HR, Araújo Miguel B, Keith David A, Brook Barry W. Tools for integrating range change, extinction risk and climate change information into conservation management. Ecography. 2013;36(9):956–64. https://doi.org/10.1111/j.1600-0587.2013.00147.x.

Forman RTT, Godron M. Patches and structural components for a landscape ecology. Bioscience. 1981;31(10):733–40. https://doi.org/10.2307/1308780.

Fosberg FR. Geography, ecology and biogeography. Ann Assoc Am Geogr. 1976;66(1):117–23. https://doi.org/10.1111/j.1467-8306.1976.tb01075.x.

Franklin J. Moving beyond static species distribution models in support of conservation biogeography. Divers Distrib. 2010;16(3):321–30. https://doi.org/10.1111/j.1472-4642.2010.00641.x.

Franklin JF, Krueger K. Germination of true fir and mountain hemlock seed on snow. J For. 1968;66:416–7.

Franklin J, Miller JA. Mapping species distributions. Cambridge: Cambridge University Press; 2009.

Fritts HC. Tree rings and climate. New York: Academic Press; 1976.

Gaddis KD, Thompson PG, Sork VL. Dry-washes determine gene flow and genetic diversity in a common desert shrub. Landsc Ecol. 2016;31(10):2215–29. https://doi.org/10.1007/s10980-016-0393-7.

Garroway CJ, Bowman J, Carr D, Wilson PJ. Applications of graph theory to landscape genetics. Evol Appl. 2008;1(4):620–30. https://doi.org/10.1111/j.1752-4571.2008.00047.x.

Geraldes A, Farzaneh N, Grassa CJ, McKown AD, Guy RD, Mansfield SD, et al. Landscape genomics of Populus trichocarpa: the role of hybridization, limited gene flow, and natural selection in shaping patterns of population structure. Evolution. 2014;68(11):3260–80. https://doi.org/10.1111/evo.12497.

Giller PS, Myers AA, Riddle BR. Earth history, vicariance, and dispersal. In: Lomolino MV, Sax DF, Brown JH, editors. Foundations of biogeography. Chicago: University of Chicago Press; 2004. p. 267–76.

González-Martínez S, Gerber S, Cervera M, Martínez-Zapater J, Gil L, Alía R. Seed gene flow and fine-scale structure in a Mediterranean pine (*Pinus pinaster* Ait.) using nuclear microsatellite markers. Theor Appl Genet. 2002;104(8):1290–7. https://doi.org/10.1007/s00122-002-0894-4.

Gotelli NJ, Stanton-Geddes J. Climate change, genetic markers and species distribution modelling. J Biogeogr. 2015;42(9):1577–85. https://doi.org/10.1111/jbi.12562.

Grissino-Mayer HD, Swetnam TW. Century-scale climate forcing of fire regimes in the American Southwest. The Holocene. 2000;10(2):213–20.

Gugger PF, Sugita S, Cavender-Bares J. Phylogeography of Douglas-fir based on mitochondrial and chloroplast DNA sequences: testing hypotheses from the fossil record. Mol Ecol. 2010;19(9):1877–97. https://doi.org/10.1111/j.1365-294X.2010.04622.x.

Guillot G, Leblois R, Coulon A, Frantz AC. Statistical methods in spatial genetics. Mol Ecol. 2009;18:4734–56.
Haines-Young R, Chopping M. Quantifying landscape structure: a review of landscape indices and their application to forested landscapes. Prog Phys Geogr. 1996;20(4):418–45.
Hamann A, Roberts DR, Barber QE, Carroll C, Nielsen SE. Velocity of climate change algorithms for guiding conservation and management. Glob Chang Biol. 2015;21(2):997–1004. https://doi.org/10.1111/gcb.12736.
Hamilton WD, May RM. Dispersal in stable habitats. Nature. 1977;269:578–81.
Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, et al. Adaptation to climate across the *Arabidopsis thaliana* genome. Science. 2011;334:83–6.
Hand BK, Lowe WH, Kovach RP, Muhlfeld CC, Luikart G. Landscape community genomics: understanding eco-evolutionary processes in complex environments. Trends Ecol Evol. 2015;30(3):161–8. https://doi.org/10.1016/j.tree.2015.01.005.
Harley G, Baisan C, Brown P, Falk D, Flatley W, Grissino-Mayer H, et al. Advancing dendrochronological studies of fire in the United States. Fire. 2018;1(1):11.
Heer K, Behringer D, Piermattei A, Bässler C, Brandl R, Fady B, et al. Linking dendroecology and association genetics in natural populations: stress responses archived in tree rings associate with SNP genotypes in silver fir (*Abies alba* Mill.). Mol Ecol. 2018;27(6):1428–38. https://doi.org/10.1111/mec.14538.
Heezen BC. The rift in the ocean floor. Sci Am. 1960;203(4):98–114.
Hengeveld R. Ecological biogeography. Prog Phys Geogr. 1993;17(4):448–60. https://doi.org/10.1177/030913339301700403.
Hennig W. Phylogenetic systematics. 3rd ed. Urbana: University of Illinois Press; 1979.
Herrera S, Shank TM. RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. Mol Phylogenet Evol. 2016;100:70–9. https://doi.org/10.1016/j.ympev.2016.03.010.
Herring EM, Gavin DG, Dobrowski SZ, Fernandez M, Hu FS. Ecological history of a long-lived conifer in a disjunct population. J Ecol. 2017;106(1):319–32. https://doi.org/10.1111/1365-2745.12826.
Hess HH. History of ocean basins. In: Engel AEJ, James HL, Leonard BF, editors. Petrological studies: a volume to Honor A. F. Buddington. Boulder: Geological Society of America; 1962. p. 599–620.
Hewitt GM. Some genetic consequences of ice ages, and their role in divergence and speciation. Biol J Linn Soc. 1996;58(3):247–76. https://doi.org/10.1111/j.1095-8312.1996.tb01434.x.
Hewitt GM. Genetic consequences of climatic oscillations in the quaternary. Philos Trans R Soc Lond Ser B Biol Sci. 2004;359(1442):183.
Hipp AL, Manos PS, González-Rodríguez A, Hahn M, Kaproth M, McVay JD, et al. Sympatric parallel diversification of major oak clades in the Americas and the origins of Mexican species diversity. New Phytol. 2018;217(1):439–52. https://doi.org/10.1111/nph.14773.
Hoban S. Integrative conservation genetics: prioritizing populations using climate predictions, adaptive potential and habitat connectivity. Mol Ecol Resour. 2018;18(1):14–7. https://doi.org/10.1111/1755-0998.12752.
Hodel RGJ, Chen S, Payton AC, McDaniel SF, Soltis P, Soltis DE. Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: comparing microsatellites and RAD-Seq and investigating loci filtering. Sci Rep. 2017;7(1):17598. https://doi.org/10.1038/s41598-017-16810-7.
Holderegger R, Wagner HH. Landscape genetics. Bioscience. 2008;58:199–207.
Holderegger R, Buehler D, Gugerli F, Manel S. Landscape genetics of plants. Trends Plant Sci. 2010;15(12):675–83. https://doi.org/10.1016/j.tplants.2010.09.002.
Housset JM, Nadeau S, Isabel N, Depardieu C, Duchesne I, Lenz P, et al. Tree rings provide a new class of phenotypes for genetic associations that foster insights into adaptation of conifers to climate change. New Phytol. 2018;218:630–45. https://doi.org/10.1111/nph.14968.
Howe HF, Smallwood J. Ecology of seed dispersal. Annu Rev Ecol Syst. 1982;13:201–28.

Hutchinson GE. Homage to Santa Rosalia or why are there so many kinds of animals? Am Nat. 1959;92:145–59.

Hylander K, Ehrlén J, Luoto M, Meineri E. Microrefugia: not for everyone. Ambio. 2015;44(1):60–8. https://doi.org/10.1007/s13280-014-0599-3.

Ikeda DH, Max TL, Allan GJ, Lau MK, Shuster SM, Whitham TG. Genetically informed ecological niche models improve climate change predictions. Glob Chang Biol. 2016;23(1):164–76. https://doi.org/10.1111/gcb.13470.

Ismail SA, Ghazoul J, Ravikanth G, Kushalappa CG, Shaanker RU, Kettle CJ. Evaluating realized seed dispersal across fragmented tropical landscapes: a two-fold approach using parentage analysis and the neighbourhood model. New Phytol. 2017;214(3):1307–16. https://doi.org/10.1111/nph.14427.

Iverson LR, Prasad AM, Matthews SN, Peters MP. Lessons learned while integrating habitat, dispersal, disturbance, and life-history traits into species habitat models under climate change. Ecosystems. 2011;14(6):1005–20. https://doi.org/10.1007/s10021-011-9456-4.

Johnson JS, Gaddis KD, Cairns DM, Lafon CW, Krutovsky KV. Plant responses to global change: next generation biogeography. Phys Geogr. 2016;37:93–119. https://doi.org/10.1080/02723646.2016.1162597.

Johnson J, Chhetri P, Krutovsky K, Cairns D. Growth and its relationship to individual genetic diversity of mountain hemlock (*Tsuga mertensiana*) at alpine treeline in Alaska: combining dendrochronology and genomics. Forests. 2017a;8(11):418. https://doi.org/10.3390/f8110418.

Johnson JS, Gaddis KD, Cairns DM, Konganti K, Krutovsky KV. Landscape genomic insights into the historic migration of mountain hemlock in response to Holocene climate change. Am J Bot. 2017b;104(3):439–50. https://doi.org/10.3732/ajb.1600262.

Johnson JS, Gaddis KD, Cairns DM, Krutovsky KV. Seed dispersal at alpine treeline: an assessment of seed movement within the alpine treeline ecotone. Ecosphere. 2017c;8(1):e01649. https://doi.org/10.1002/ecs2.1649.

Jones MC, Peteet DM, Kurdyla D, Guilderson T. Climate and vegetation history from a 14,000-year peatland record, Kenai Peninsula, Alaska. Quat Res. 2009;72(2):207–17. https://doi.org/10.1016/j.yqres.2009.04.002.

Kays R, Crofoot MC, Jetz W, Wikelski M. Terrestrial animal tracking as an eye on life and planet. Science. 2015;348(6240):aaa2478.

Kimura M. Evolutionary rate at the molecular level. Nature. 1968;217:624. https://doi.org/10.1038/217624a0.

Kimura M. The neutral theory of molecular evolution. Sci Am. 1979;241(5):98–129.

Kleinman JS, Hart JL. Response by vertical strata to catastrophic wind in restored *Pinus palustris* stands. J Torrey Bot Soc. 2017;144(4):423–38. https://doi.org/10.3159/TORREY-D-16-00046.1.

Knowles LL, Carstens BC, Keat Marcia L. Coupling genetic and ecological-niche models to examine how past population distributions contribute to divergence. Curr Biol. 2007;17(11):940–6. https://doi.org/10.1016/j.cub.2007.04.033.

Körner C. A re-assessment of high elevation treeline positions and their explanation. Oecologia. 1998;115(4):445–59. https://doi.org/10.1007/s004420050540.

Körner C, Paulsen J. A world-wide study of high altitude treeline temperatures. J Biogeogr. 2004;31(5):713–32. https://doi.org/10.1111/j.1365-2699.2003.01043.x.

Kraft NJB, Baldwin BG, Ackerly DD. Range size, taxon age and hotspots of neoendemism in the California flora. Divers Distrib. 2010;16(3):403–13. https://doi.org/10.1111/j.1472-4642.2010.00640.x.

Kramer-Schadt S, Niedballa J, Pilgrim JD, Schröder B, Lindenborn J, Reinfelder V, et al. The importance of correcting for sampling bias in MaxEnt species distribution models. Divers Distrib. 2013;19(11):1366–79. https://doi.org/10.1111/ddi.12096.

Krutovsky K, Elsik C, Matvienko M, Kozik A, Neale D. Conserved ortholog sets in forest trees. Tree Genet Genomes. 2006;3(1):61–70. https://doi.org/10.1007/s11295-006-0052-2.

Krutovsky KV, Burczyk J, Chybicki IJ, Finkeldey R, Pyhajarvi T, Robledo-Arnuncio JJ. Gene flow, spatial structure, local adaptation, assisted migration in trees. In: Schnell RJ, Priyadarshan PM, editors. Genomics of tree crops. New York: Springer; 2012. p. 71–116.
Kupfer JA. Landscape ecology and biogeography. Prog Phys Geogr. 1995;19:18–34.
Kupfer JA. Theory in landscape ecology and its relevance to biogeography. In: Millington AC, Blumler MA, MacDonald G, Schickhoff U, editors. The Sage handbook of biogeography. London: SAGE; 2011.
Kupfer JA. Landscape ecology and biogeography: rethinking landscape metrics in a post-FRAGSTATS landscape. Prog Phys Geogr. 2012;36:400–20.
Kupfer JA, Malanson GP, Franklin SB. Not seeing the ocean for the islands: the mediating influence of matrix-based processes on forest fragmentation effects. Glob Ecol Biogeogr. 2006;15(1):8–20. https://doi.org/10.1111/j.1466-822X.2006.00204.x.
Lafon CW, Speer JH. Using dendrochronology to identify major ice storm events in oak forests of southwestern Virginia. Clim Res. 2002;20(1):41–54.
Lafon CW, Naito AT, Grissino-Mayer HD, Horn SP, Waldrop TA. Fire history of the Appalachian region: a review and synthesis. USDA United States forest service general technical report SRS-219. 2017.
Leaché AD, Oaks JR. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. Annu Rev Ecol Evol Syst. 2017;48(1):69–84. https://doi.org/10.1146/annurev-ecolsys-110316-022645.
Lee S-R, Jo Y-S, Park C-H, Friedman JM, Olson MS. Population genomic analysis suggests strong influence of river network on spatial distribution of genetic variation in invasive saltcedar across the southwestern United States. Mol Ecol. 2018;27(3):636–46. https://doi.org/10.1111/mec.14468.
Levey DJ, Sargent S. A simple method for tracking vertebrate-dispersed seeds. Ecology. 2000;81(1):267–74. https://doi.org/10.1890/0012-9658(2000)081[0267:ASMFTV]2.0.CO;2.
Loarie SR, Duffy PB, Hamilton H, Asner GP, Field CB, Ackerly DD. The velocity of climate change. Nature. 2009;462:1052–5.
Lomolino MV. Conservation biogeography. In: Lomolino MV, Heaney LR, editors. Frontiers of biogeography: new direction in the geography of nature. Sunderland: Sinauer Associates; 2004. p. 1–3.
Lomolino MV, Riddle BR, Whittaker RH. Biogeography: biological diversity across space and time. 5th ed. Sunderland: Sinauer; 2017.
MacArthur R. On the relative abundance of species. Am Nat. 1960;94(874):25–36. https://doi.org/10.2307/2458395.
MacArthur RH, Wilson EO. An equilibrium theory of insular zoogeography. Evolution. 1963;17:373–87.
MacArthur RH, Wilson EO. The theory of island biogeography. Princeton: Princeton University Press; 1967.
MacDonald GM. Preparing biogeographers for the third millennium. J Biogeogr. 2000;27:49–50.
MacDonald GM. Biogeography. New York: Wiley; 2003.
MacDonald GM, Szeicz JM, Claricoates J, Dale KA. Response of the Central Canadian treeline to recent climatic changes. Ann Assoc Am Geogr. 1998;88(2):183–208. https://doi.org/10.1111/1467-8306.00090.
Magri D. Patterns of post-glacial spread and the extent of glacial refugia of European beech (*Fagus sylvatica*). J Biogeogr. 2008;35(3):450–63.
Manel S, Holderegger R. Ten years of landscape genetics. Trends Ecol Evol. 2013;28(10):614–21. https://doi.org/10.1016/j.tree.2013.05.012.
Manel S, Schwartz MK, Luikart G, Taberlet P. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol Evol. 2003;18(4):189–97. https://doi.org/10.1016/s0169-5347(03)00008-9.
Mann ME, Jones PD. Global surface temperatures over the past two millennia. Geophys Res Lett. 2003;30(15). https://doi.org/10.1029/2003GL017814.

Marcer A, Méndez-Vigo B, Alonso-Blanco C, Picó FX. Tackling intraspecific genetic structure in distribution models better reflects species geographical range. Ecol Evol. 2016;6(7):2084–97. https://doi.org/10.1002/ece3.2010.

Marcott SA, Shakun JD, Clark PU, Mix AC. A reconstruction of regional and global temperature for the past 11,300 years. Science. 2013;339(6124):1198.

Markwith SH, Scanlon MJ. Characterization of six polymorphic microsatellite loci isolated from *Hymenocallis coronaria* (J. LeConte) Kunth (Amaryllidaceae). Mol Ecol Notes. 2006;6(1):72–4.

Mayol M, Riba M, González-Martínez Santiago C, Bagnoli F, Beaulieu JL, Berganzo E, et al. Adapting through glacial cycles: insights from a long-lived tree (*Taxus baccata*). New Phytol. 2015;208(3):973–86. https://doi.org/10.1111/nph.13496.

Mayr E. Systematics and the origin of species. New York: Columbia University Press; 1942.

Mayr E. Ecological factors in speciation. Evolution. 1947;1(4):263–88. https://doi.org/10.1111/j.1558-5646.1947.tb02723.x.

Mayr E. The growth of biological thought. Cambridge: Harvard University Press; 1982.

McCaughey WW, Schmidt WC. Seed dispersal of Engelmann spruce in the intermountain west. Northwest Sci. 1987;61:1–6.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol. 2013;66(2):526–38. https://doi.org/10.1016/j.ympev.2011.12.007.

McLean N, Lawson CR, Leech DI, van de Pol M. Predicting when climate-driven phenotypic change affects population dynamics. Ecol Lett. 2016;19(6):595–608. https://doi.org/10.1111/ele.12599.

McRae BH. Isolation by resistance. Evolution. 2006;60:1551–61.

McRae BH, Beier P. Circuit theory predicts gene flow in plant and animal populations. Proc Natl Acad Sci. 2007;104(50):19885–90. https://doi.org/10.1073/pnas.0706568104.

McRae BH, Dickson BG, Keitt TH, Shah VB. Using circuit theory to model connectivity in ecology, evolution, and conservation. Ecology. 2008;89(10):2712–24. https://doi.org/10.1890/07-1861.1.

McRae BH, Shirk A, Platt JT. Gnarly landscape utilities: resistance and habitat calculator. Seattle: The Nature Conservancy; 2013.

Means JE. Tsuga mertensiana. In: Burns RM, Honkala BH, editors. Silvics of North America, vol. 1. Washington: Forest Service; 1990. p. 1279–306.

Menon M, Bagley JC, Friedline CJ, Whipple AV, Schoettle AW, Leal-Sàenz A, et al. The role of hybridization during ecological divergence of southwestern white pine (*Pinus strobiformis*) and limber pine (*P. flexilis*). Mol Ecol. 2018;27(5):1245–60. https://doi.org/10.1111/mec.14505.

Micheletti SJ, Storfer A. An approach for identifying cryptic barriers to gene flow that limit species' geographic ranges. Mol Ecol. 2017;26(2):490–504. https://doi.org/10.1111/mec.13939.

Millington AC, Blumler MA, MacDonald G, Schickhoff U. The SAGE handbook of biogeography. London: SAGE; 2011a.

Millington AC, Blumler MA, Schickhoff U. Situating contemporary biogeography. In: Millington AC, Blumler MA, Schickhoff U, editors. The SAGE handbook of biogeography. Los Angeles: SAGE; 2011b.

Mitton JB. Relationship between heterozygosity for enzyme loci and variation of morphological characters in natural populations. Nature. 1978;273(5664):661–2.

Muir G, Schlötterer C. Evidence for shared ancestral polymorphism rather than recurrent gene flow at microsatellite loci differentiating two hybridizing oaks (*Quercus* spp.). Mol Ecol. 2005;14(2):549–61.

Murray DC, Pearson SG, Fullagar R, Chase BM, Houston J, Atchison J, et al. High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. Quat Sci Rev. 2012;58:135–45. https://doi.org/10.1016/j.quascirev.2012.10.021.

Naito AT, Cairns DM. Relationships between Arctic shrub dynamics and topographically derived hydrologic characteristics. Environ Res Lett. 2011;11(4):1–8.

Nathan R. Long-distance dispersal of plants. Science. 2006;313(5788):786–8. https://doi.org/10.1126/science.1124975.

Nathan R, Muller-Landau HC. Spatial patterns of seed dispersal, their determinants and consequences for recruitment. Trends Ecol Evol. 2000;15(7):278–85. https://doi.org/10.1016/s0169-5347(00)01874-7.

Nathan R, Katul GG, Horn HS, Thomas SM, Oren R, Avissar R, et al. Mechanisms of long-distance dispersal of seeds by wind. Nature. 2002;418(6896):409–13.

Neale DB, Kremer A. Forest tree genomics: growing resources and applications. Nat Rev Genet. 2011;12(2):111–22. https://doi.org/10.1038/nrg2931.

O'Connell LM, Mosseler A, Rajora OP. Extensive long-distance pollen dispersal in a fragmented landscape maintains genetic diversity in white spruce. J Hered. 2007;97:640–5.

Oddou-Muratorio S, Klein EK. Comparing direct vs. indirect estimates of gene flow within a population of a scattered tree species. Mol Ecol. 2008;17(11):2743–54. https://doi.org/10.1111/j.1365-294X.2008.03783.x.

Orsini L, Vanoverbeke J, Swillen I, Mergeay J, De Meester L. Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization. Mol Ecol. 2013;22(24):5983–99. https://doi.org/10.1111/mec.12561.

Owens JN, Molder M. Sexual reproduction of mountain hemlock (*Tsuga mertensiana*). Can J Bot. 1975;53(17):1811–26. https://doi.org/10.1139/b75-211.

Pandey M, Rajora OP. Genetic diversity and differentiation of core vs. peripheral populations of eastern white cedar, *Thuja occidentalis* (*Cupressaceae*). Am J Bot. 2012;99(4):690–9. https://doi.org/10.3732/ajb.1100116.

Parducci L, Jørgensen T, Tollefsrud MM, Elverland E, Alm T, Fontana SL, et al. Glacial survival of boreal trees in northern Scandinavia. Science. 2012;335(6072):1083.

Parducci L, Väliranta M, Salonen JS, Ronkainen T, Matetovici I, Fontana SL, et al. Proxy comparison in ancient peat sediments: pollen, macrofossil and plant DNA. Philos Trans R Soc B. 2015;370(1660):20130382.

Parisod C, Holderegger R. Adaptive landscape genetics: pitfalls and benefits. Mol Ecol. 2012;21:3644–6.

Park Williams A, Allen CD, Macalady AK, Griffin D, Woodhouse CA, Meko DM, et al. Temperature as a potent driver of regional forest drought stress and tree mortality. Nat Clim Chang. 2012;3:292. https://doi.org/10.1038/nclimate1693. https://www.nature.com/articles/nclimate1693#supplementary-information

Parker KC, Hamrick JL. Genetic variation in sand pine (*Pinus clausa*). Can J For Res. 1996;26(2):244–54. https://doi.org/10.1139/x26-028.

Peterson DW, Peterson DL. Mountain hemlock growth responds to climatic variability at annual and decadal time scales. Ecology. 2001;82(12):3330–45. https://doi.org/10.1890/0012-9658(2001)082[3330:mhgrtc]2.0.co;2.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: an inexpensive method for De Novo SNP discovery and genotyping in model and non-model species. PLoS One. 2012;7(5):1–11.

Petit RJ, Aguinagalde I, de Beaulieu J-L, Bittkau C, Brewer S, Cheddadi R, et al. Glacial refugia: hotspots but not melting pots of genetic diversity. Science. 2003;300(5625):1563–5. https://doi.org/10.1126/science.1083264.

Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. Ecol Model. 2006;190(3–4):231–59. https://doi.org/10.1016/j.ecolmodel.2005.03.026.

Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl. 2009;19(1):181–97. https://doi.org/10.1890/07-2153.1.

Pickett STA, Cadenasso ML. Landscape ecology: spatial heterogeneity in ecological systems. Science. 1995;269(5222):331–4.

Piotti A, Leonardi S, Piovani P, Scalfi M, Menozzi P. Spruce colonization at treeline: where do those seeds come from. Heredity. 2009;103(2):136–45.
Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet. 2008;24:142–9.
Porter JH, Hanson PC, Lin C. Staying afloat in the sensor data deluge. Trends Ecol Evol. 2012;27:121–9.
Potts SG, Biesmeijer JC, Kremen C, Neumann P, Schweiger O, Kunin WE. Global pollinator declines: trends, impacts and drivers. Trends Ecol Evol. 2010;25(6):345–53. https://doi.org/10.1016/j.tree.2010.01.007.
Pöyry J, Luoto M, Heikkinen RK, Saarinen K. Species traits are associated with the quality of bioclimatic models. Glob Ecol Biogeogr. 2008;17(3):403–14. https://doi.org/10.1111/j.1466-8238.2007.00373.x.
Provan J, Bennett KD. Phylogeographic insights into cryptic glacial refugia. Trends Ecol Evol. 2008;23(10):564–71. https://doi.org/10.1016/j.tree.2008.06.010.
Puzey JR, Willis JH, Kelly JK. Population structure and local selection yield high genomic variation in *Mimulus guttatus*. Mol Ecol. 2016;26(2):519–35. https://doi.org/10.1111/mec.13922.
Rajora OP, Dancik BP. Genetic characterization and relationships of *Populus alba*, *P. tremula*, and *P.* x *canescens*, and their clones. Theor Appl Genet. 1992;84(3):291–8. https://doi.org/10.1007/BF00229485.
Rajora OP, Eckert AJ, Zinck JWR. Single-locus versus multilocus patterns of local adaptation to climate in eastern white pine (*Pinus strobus, Pinaceae*). PLoS One. 2016;11(7):e0158691. https://doi.org/10.1371/journal.pone.0158691.
Razgour O. Beyond species distribution modeling: a landscape genetics approach to investigating range shifts under future climate change. Ecol Inform. 2015;30:250–6. https://doi.org/10.1016/j.ecoinf.2015.05.007.
Razgour O, Taggart John B, Manel S, Juste J, Ibáñez C, Rebelo H, et al. An integrated framework to identify wildlife populations under threat from climate change. Mol Ecol Resour. 2017;18(1):18–31. https://doi.org/10.1111/1755-0998.12694.
Reid C. The origin of the British Flora. London: Dulau; 1899.
Richards CL, Carstens BC, Knowles LL. Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. J Biogeogr. 2007;34(11):1833–45.
Robledo-Arnuncio JJ, Gil L. Patterns of pollen dispersal in a small population of *Pinus sylvestris* L. revealed by total-exclusion paternity analysis. Heredity. 2004;94(1):13–22.
Ruiz-Gonzalez A, Cushman SA, Madeira MJ, Randi E, Gómez-Moliner BJ. Isolation by distance, resistance and/or clusters? Lessons learned from a forest-dwelling carnivore inhabiting a heterogeneous landscape. Mol Ecol. 2015;24(20):5110–29. https://doi.org/10.1111/mec.13392.
Rull V. Microrefugia. J Biogeogr. 2009;36(3):481–4. https://doi.org/10.1111/j.1365-2699.2008.02023.x.
Rymer MJ, Sims JD. Lake-sediment evidence for the date of deglaciation of the hidden Lake area, Kenai Peninsula, Alaska. Geology. 1982;10(6):314–6. https://doi.org/10.1130/0091-7613(1982)10<314:leftdo>2.0.co;2.
Savolainen O. The genomic basis of local climatic adaptation. Science. 2011;334(6052):49–50. https://doi.org/10.1126/science.1213788.
Savolainen O, Lascoux M, Merilä J. Ecological genomics of local adaptation. Nat Rev Genet. 2013;14(11):807–20.
Schickhoff U, Blumler MA, Millington AC. Biogeography in the early twenty-first century: a science with increasing significance for Earth's changes and challenges. Geogr Pol. 2014;87(2):221–40.
Schimper AFW. Plant-geography upon a physiological basis (Pflanzen-geographie auf physiologischer Grundlage). Oxford: Clarendon Press; 1903.

Scoble J, Lowe AJ. A case for incorporating phylogeography and landscape genetics into species distribution modelling approaches to improve climate adaptation and conservation planning. Divers Distrib. 2010;16(3):343–53. https://doi.org/10.1111/j.1472-4642.2010.00658.x.

Selkoe KA, Toonen RJ. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecol Lett. 2006;9(5):615–29. https://doi.org/10.1111/j.1461-0248.2006.00889.x.

Semerikov VL, Semerikova SA, Putintseva YA, Tarakanov VV, Tikhonova IV, Vidyakin AI, et al. Colonization history of scots pine in Eastern Europe and North Asia based on mitochondrial DNA variation. Tree Genet Genomes. 2018;14(1):8. https://doi.org/10.1007/s11295-017-1222-0.

Sexton JP, Hangartner SB, Hoffmann AA. Genetic isolation by environment or distance: which pattern of gene flow is most common? Evolution. 2014;68(1):1–15. https://doi.org/10.1111/evo.12258.

Shafer ABA, Cullingham CI, CÔTÉ SD, Coltman DW. Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. Mol Ecol. 2010;19(21):4589–621. https://doi.org/10.1111/j.1365-294X.2010.04828.x.

Shah VB, McRae BH. Circuitscape: a tool for landscape ecology. In: Proceedings of the 7th python in science conference, 2008, p. 62–66.

Shirk AJ, Cushman SA, Waring KM, Wehenkel CA, Leal-Sáenz A, Toney C, et al. Southwestern white pine (*Pinus strobiformis*) species distribution models project a large range shift and contraction due to regional climatic changes. For Ecol Manag. 2018;411:176–86. https://doi.org/10.1016/j.foreco.2018.01.025.

Silvertown J, Servaes C, Biss P, Macleod D. Reinforcement of reproductive isolation between adjacent populations in the park grass experiment. Heredity. 2005;95:198. https://doi.org/10.1038/sj.hdy.6800710.

Skellam JG. Random dispersal in theoretical populations. Biometrika. 1951;38(1/2):196–218. https://doi.org/10.2307/2332328.

Sork VL. Genomic studies of local adaptation in natural plant populations. J Hered. 2018;109(1):3–15. https://doi.org/10.1093/jhered/esx091.

Sork VL, Davis FW, Westfall R, Flint A, Ikegami M, Wang H, et al. Gene movement and genetic association with regional climate gradients in California valley oak (*Quercus lobata* Née) in the face of climate change. Mol Ecol. 2010;19(17):3806–23. https://doi.org/10.1111/j.1365-294X.2010.04726.x.

Sork VL, Aitken SN, Dyer RJ, Eckert AJ, Legendre P, Neale DB. Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. Tree Genet Genomes. 2013;9(4):901–11. https://doi.org/10.1007/s11295-013-0596-x.

Speer JH. Fundamentals of tree-ring research. Tucson: University of Arizona Press; 2010.

Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, et al. Adaptation genomics: the next generation. Trends Ecol Evol. 2010;25(12):705–12. https://doi.org/10.1016/j.tree.2010.09.002.

Storfer A, Murphy MA, Spear SF, Holderegger R, Waits LP. Landscape genetics: where are we now? Mol Ecol. 2010;19(17):3496–514.

Suarez-Gonzalez A, Hefer CA, Lexer C, Cronk Quentin CB, Douglas CJ. Scale and direction of adaptive introgression between black cottonwood (*Populus trichocarpa*) and balsam poplar (*P. balsamifera*). Mol Ecol. 2018;27:1667–80.

Taylor AH. Forest expansion and climate change in the mountain hemlock (*Tsuga mertensiana*) zone, Lassen Volcanic National Park, California, U.S.A. Arct Alp Res. 1995;27:207–16.

Tobler WR. A computer movie simulating urban growth in the Detroit region. Econ Geogr. 1970;46:234–40.

Turner MG. Landscape ecology: the effect of pattern on process. Annu Rev Ecol Syst. 1989;20:171–97.

Veblen TT. Biogeography. In: Gaile GL, Willmott CJ, editors. Geography in America at the dawn of the twenty-first century. Columbus: Merrill; 1989. p. 28–46.
Veblen TT, Hadley KS, Reid MS, Rebertus AJ. The response of subalpine forests to spruce beetle outbreak in Colorado. Ecology. 1991;72(1):213–31. https://doi.org/10.2307/1938916.
Veblen TT, Hadley KS, Nel EM, Kitzberger T, Reid M, Villalba R. Disturbance regime and disturbance interactions in a Rocky Mountain subalpine forest. J Ecol. 1994;82(1):125–35. https://doi.org/10.2307/2261392.
Vine FJ, Matthews DH. Magnetic anomalies over oceanic ridges. Nature. 1963;199:947. https://doi.org/10.1038/199947a0.
Von Humboldt A, Bonpland A. Essai sur la géographie des plantes. Paris: Schoell; 1805.
Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, et al. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol Ecol. 2013;22(3):787–98. https://doi.org/10.1111/mec.12023.
Webb T III. The appearance and disappearance of major vegetational assemblages: long-term vegetational dynamics in eastern North America. Vegetatio. 1987;69(1–3):177–87. https://doi.org/10.1007/bf00038699.
Wheeler NC, Guries RP. Biogeography of lodgepole pine. Can J Bot. 1982;60(9):1805–14. https://doi.org/10.1139/b82-227.
White PS, Pickett STA. Natural disturbance and patch dynamics: an introduction. In: Pickett STA, White PS, editors. The ecology of disturbance and patch dynamics. Orlando: Academic Press; 1985.
Whittaker RJ, Bush MB, Richards K. Plant recolonization and vegetation succession on the Krakatau islands, Indonesia. Ecol Monogr. 1989;59(2):59–123. https://doi.org/10.2307/2937282.
Whittaker RJ, Araujo MB, Jepson P, Ladle RJ, Watson JEM, Willis KJ. Conservation biogeography: assessment and prospect. Divers Distrib. 2005;11(1):3–23.
Wiens JJ. Spatial scaling in ecology. Funct Ecol. 1989;3:385–97.
Wiley EO. Vicariance biogeography. Annu Rev Ecol Syst. 1988;19:513–42.
Willerslev E, Davison J, Moora M, Zobel M, Coissac E, Edwards ME, et al. Fifty thousand years of Arctic vegetation and megafaunal diet. Nature. 2014;506(7486):47–51. https://doi.org/10.1038/nature12921.
Willson MF, Traveset A. The ecology of seed dispersal. In: Fenner M, editor. Seeds: the ecology of regeneration in plant communities. 2nd ed. New York: CABI; 2000.
Wilmshurst JM, Moar NT, Wood JR, Bellingham PJ, Findlater AM, Robinson JJ, et al. Use of pollen and ancient DNA as conservation baselines for offshore islands in New Zealand. Conserv Biol. 2013;28(1):202–12. https://doi.org/10.1111/cobi.12150.
Wilson EO. The diversity of life. Cambridge: Harvard University Press; 1992.
Woodhouse CA, Meko DM, MacDonald GM, Stahle DW, Cook ER. A 1,200-year perspective of twenty-first century drought in southwestern North America. Proc Natl Acad Sci. 2010;107(50):21283.
Woodward A, Silsbee DG, Schreiner EG, Means JE. Influence of climate on radial growth and cone production in subalpine fir (*Abies lasiocarpa*) and mountain hemlock (*Tsuga mertensiana*). Can J Forest Res. 1994;24:1133–43.
Wright S. The genetical structure of populations. Ann Hum Genet. 1949;15(1):323–54. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x.
Xiao Z, Jansen PA, Zhang Z. Using seed-tagging methods for assessing post-dispersal seed fate in rodent-dispersed trees. For Ecol Manag. 2006;223(1):18–23. https://doi.org/10.1016/j.foreco.2005.10.054.
Xu B, Sun G, Wang X, Lu J, Wang IJ, Wang Z. Population genetic structure is shaped by historical, geographic, and environmental factors in the leguminous shrub *Caragana microphylla* on the Inner Mongolia Plateau of China. BMC Plant Biol. 2017;17(1):200. https://doi.org/10.1186/s12870-017-1147-7.

Yang M, He Z, Shi S, Wu CI. Can genomic data alone tell us whether speciation happened with gene flow? Mol Ecol. 2017;26(11):2845–9. https://doi.org/10.1111/mec.14117.

Young AB, Cairns DM, Lafone CW. Dendroclimatic relationships and possible implications for mountain birch and scots pine at treeline in northern Sweden through the twenty-first century. Can J Forest Res. 2011;41:450–9.

Zhou W, Ji X, Obata S, Pais A, Dong Y, Peet R, et al. Resolving relationships and phylogeographic history of the *Nyssa sylvatica* complex using data from RAD-seq and species distribution modeling. Mol Phylogenet Evol. 2018;126:1–16. https://doi.org/10.1016/j.ympev.2018.04.001.

Zinck JWR, Rajora OP. Post-glacial phylogeography and evolution of a wide-ranging highly-exploited keystone forest tree, eastern white pine (*Pinus strobus*) in North America: single refugium, multiple routes. BMC Evol Biol. 2016;16(1):56. https://doi.org/10.1186/s12862-016-0624-1.

Adaptation Without Boundaries: Population Genomics in Marine Systems

Marjorie F. Oleksiak

Abstract From the surface, the world's oceans appear vast and boundless. Ocean currents, which can transport marine organisms thousands of kilometers, coupled with species that spend some or all of their life in the pelagic zone, the open sea, highlight the potential for well-mixed, panmictic marine populations. Yet these ocean habitats do harbor boundaries. In this largely three-dimensional marine environment, gradients form boundaries. These gradients include temperature, salinity, and oxygen gradients. Ocean currents also form boundaries between neighboring water masses even as they can break through barriers by transporting organisms huge distances. With the advent of next-generation sequencing approaches, which allow us to easily generate a large number of genomic markers, we are in an unprecedented position to study the effects of these potential oceanic boundaries and can ask how often and when do locally adapted marine populations evolve. This knowledge will inform our understanding of how marine organisms respond to climate change and affect how we protect marine diversity. In this chapter I first discuss the major boundaries present in the marine environment and the implications they have for marine organisms. Next, I discuss the how genomic approaches are impacting our understanding of genetic connectivity, ocean fisheries, and local adaptation, including the potential for epigenetic adaptation. I conclude with considerations for marine conservation and management and future prospects.

Keywords Adaptation · Conservation · Genomic diversity · Genomics · Genotyping by sequencing, GBS · Next-generation sequencing, NGS · Population genetic structure and differentiation · SNPs

M. F. Oleksiak (✉)
Marine Biology and Ecology, Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL, USA
e-mail: moleksiak@miami.edu

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_32,

1 Introduction

The world's oceans are seemingly without boundaries. This vast environment covers 70% of the world and contains approximately 2.2 million eukaryotic marine species (Mora et al. 2011). The world's oceans' apparent boundlessness results from this vast environment combined with the large distances ocean currents can transport the eggs, larvae, and juveniles of many marine species. The combination of this dispersal potential with highly mobile adults in many marine species results in species occupying large areas encompassing diverse marine ecosystems. Not surprisingly, many genetic studies have shown apparent panmixia in ocean populations.

Historically, genetic studies using a few presumably neutral markers show little genetic differentiation among many marine species' populations (Waples 1998; Conover et al. 2006). Population genetic studies use F_{ST} values (Wright 1949), which measure the genetic variance among populations relative to the total variance (within plus between), to determine interpopulation genetic differentiation. Indeed, the average F_{ST} value across 57 marine fish species was 0.062, while the median was only 0.02 (Ward et al. 1994; Waples 1998) (though note that in very large populations, even low F_{ST} values can be statistically significant), and many marine species with dispersive life stages show limited genetic differentiation (Palumbi 2003). Based on selectively neutral markers, it has been shown that even rare, long-distance dispersal can maintain genetic homogeneity between populations (Waples 1998). Additionally, many marine species have large population sizes, which also tend to minimize genetic divergence due to genetic drift, because the amount of change due to neutral processes decreases as the population size increases (Kliman et al. 2008). Overall, many large marine populations show minimal among-population genetic differentiation at neutral loci.

This historical perspective is changing. Recently, biologists have gone from looking at a few targeted genes or a few genetic markers to looking at whole genomes, and this is rapidly changing our perspective from the idea of well-mixed populations to that of intraspecific differences – some due to local adaptations – that reflect the ecological settings of local populations (Hand et al. 2015; Rudman et al. 2015; Barabas and D'Andrea 2016; Messer et al. 2016; Wood and Brodie 2016). This challenges the assumptions about marine species dispersal and raises the questions: how connected are marine populations, what physical and biological factors affect this connectivity, how rampant is local adaptation, and what are the biological, evolutionary, and conservation implications?

The change in our perspective about the adaptive potential of marine organisms despite large dispersal in large part reflects the availability of genome-wide information facilitated by next-generation sequencing (NGS) technologies that allow us to sequence many thousands of genes in any organism (Crawford and Oleksiak 2016). Population biologists have gone from looking at target genes or a few genetic markers to looking at whole genomes using these high-throughput sequencing technologies. For marine species, these recent genomic approaches have opened up the world of marine population genomic studies because now one can quantify

the nucleotide variation at thousands of loci without needing a complete reference genome. Instead, by sampling loci across the genome, NGS can be applied and used to analyze any species, from those with small genomes (million of base pairs, bp) to those with very large genomes (100 s of billions of bp). Yet, even though NGS can be applied to most marine species, many marine species are logistically difficult to study because they are often difficult to observe, collect, identify, and study either in situ or in the laboratory.

For marine species that can be studied, NGS offers two different approaches for conducting genomic studies, each with different challenges: (1) sequencing whole genomes or (2) sequencing selected portions of genomes (selected fragments, expressed sequences, or targeted sequences). Sequencing the complete or whole genome of marine organisms offers the advantage of identifying nearly all informative DNA sequence changes. Yet, whole genome sequencing for non-model species is not trivial because genome assembly requires high-performance computing and extensive bioinformatics (Willette et al. 2014). This problem becomes much more severe for population genomic studies where many hundreds of individuals need to be compared to determine fisheries stocks, demographic parameters, or adaptive changes. Thus, until whole genome bioinformatics methods (e.g., starting with whole genome assembly and annotation followed by variant detection and analyses) for hundreds of individuals are possible, a more effective approach for many studies on marine species is likely to use the second approach of selective sequencing (though see Therkildsen and Palumbi (2016) for progress on whole genome sequencing for multiple individuals and Reid et al. (2016, 2017) for a population genomics approach using whole genome sequencing). The exceptions are commercially important species, especially fish species, with an abundance of resources (Nielsen et al. 2009b).

For marine species without whole genome resources, population genomic studies are more likely to sequence only a portion of the genome or a reduced subset of a species' genome, for example, transcriptomes (transcriptome or gene expression studies are not discussed here though they have been used with a wide variety of marine organisms to infer adaptation and response to climate change (Oleksiak 2010; Stillman and Armstrong 2015)), a selected subset of the genome (e.g., exome capture hybridization followed by sequencing (Ng et al. 2009)) or a reduced representation of the genome (e.g., Rad-seq (Baird et al. 2008; Etter et al. 2011) or genotyping by sequencing, GBS (Elshire et al. 2011)), where tens of thousands of restriction endonuclease fragments are sampled and sequenced). These genomic approaches allow researchers to identify and quantify single nucleotide polymorphisms (SNPs) for marine population genomics studies. The development of thousands of polymorphic DNA markers provided by high-throughput sequencing approaches has given researchers the unprecedented capability to interrogate across the entire genomes of virtually any species. Researchers can use these partial genomic approaches on just about any intractable marine species, as long as the appropriate samples can be collected. The advantage of these genomic approaches is that they require little development. That is, unlike microsatellites that require extensive marker development, including identification and optimization, these

genomic approaches remove one of the bottlenecks in ecological genetic and genomic studies. For difficult-to-study marine species, this capability opens the door to studies that identify populations, population parameters, population structure, and the processes that affect these attributes.

2 Major Population Boundaries in the Oceans

2.1 Environmental Temperature Changes

Many marine populations are large and inhabit environments without obvious physical barriers. Coupled with high dispersal of different life stages, this suggests that populations should be panmictic. Yet physical barriers do exist in marine environments, and these barriers can isolate populations and potentially drive both random (genetic drift) and adaptive variation among populations. What are these physical barriers?

One of the most important physical barriers is temperature differences among population locations. Indeed, the fronts and major currents defined by Dana's temperature boundaries (Dana 1853) continue to define the major pelagic habitats in the world's oceans (Spalding et al. 2012). Although temperature changes in the oceans are not as extreme as those on land, water's large heat capacity, which is approximate 4 times that of air, and large heat conduction, which is approximately 25 times that of air (Ricklefs 1979), means that most marine organisms are ectotherms and have body temperatures equal to the water temperature. Thus because of water's heat conduction and capacity, water temperature differences directly affect body temperatures and metabolic processes that rely on temperature: there is a two- to threefold change in chemical and physiological processes with every 10°C change in temperature. Consequently, temperature clines have long been recognized as important features for marine organism distributions (Dana 1853) (Fig. 1), with temperatures affecting either survival or reproduction (Hutchins 1947).

A classic example of temperature affecting fitness and causing natural selection occurs in the saltmarsh minnow, *Fundulus heteroclitus*. *F. heteroclitus* is distributed along a steep thermocline, the East Coast of the United States of America. Here, for every degree latitude change, there is approximately a 1°C temperature change. Thus, *F. heteroclitus* populations living in Maine experience approximately >12°C colder mean annual temperatures than *F. heteroclitus* populations living in Georgia. This thermal cline is thought to be responsible for the adaptive divergence in enzyme biochemistry and expression (Powers et al. 1991; Oleksiak et al. 2002).

A well-characterized example of adaptive divergence is lactate dehydrogenase B (*LDH-B*) in *F. heteroclitus* (Fig. 2). *LDH-B* has two common alleles and is expressed in the heart, liver, and red blood cells (Powers et al. 1991). The two common *LDH-B* allele frequencies are strongly differentiated with latitude. Elegant enzymatic studies showed that not only do *LDH-B* alleles vary with latitude, so too do catalytic efficiencies (Place and Powers 1979, 1984). One allele (*LDH-B*^{a}), which is fixed

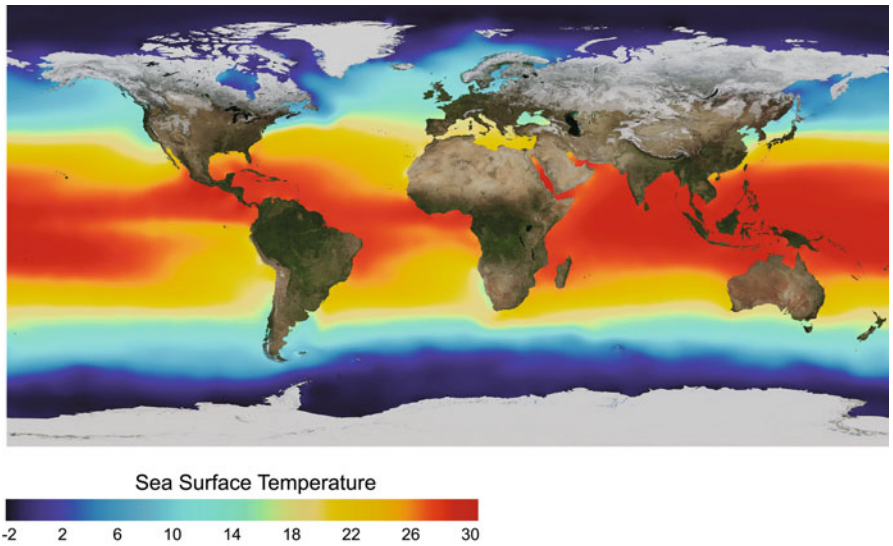

Fig. 1 Average sea surface temperature. Note the variation of the temperature by latitude, from the warm region along the equator to the cold regions near the poles. Image courtesy of NASA/Goddard Space Flight Center

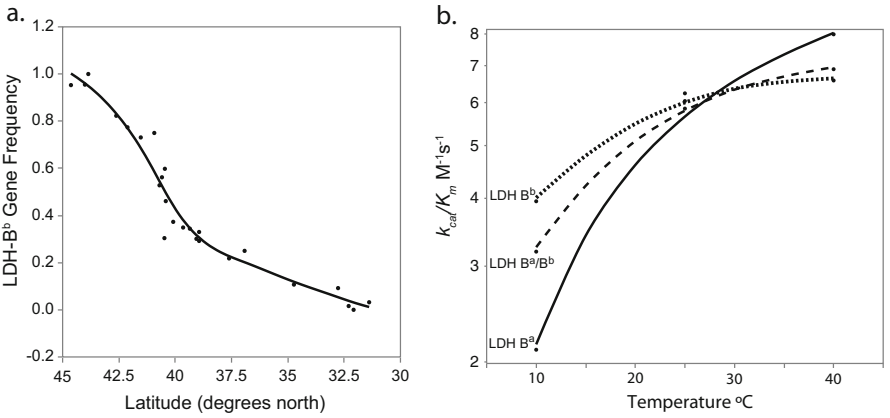

Fig. 2 Ldh-B. (**a**) Variation in the LDH-Bb allele frequencies with latitude. (**b**) Reaction rates (km/kcat) for three LDH-B genotypes (aa, ab, and bb). Notice that at 10°C the bb genotype had a greater reaction rate than the other genotypes, while at temperatures above 25°C, this was reversed. Adapted from Powers and Place (1978) and Place and Powers (1979)

in southern populations but nearly absent from northern populations (Place and Powers 1978), has the highest catalytic efficiency near 40°C. In contrast, the alternative allele (*LDH-B^b*), which is common in northern *F. heteroclitus* populations, has the highest catalytic efficiency near 10°C. In fact, *LDH-B* allele

frequency is a function of latitude between northern Maine and South Carolina, a distance of over 1,600 km, and is consistent with kinetic variation among the *LDH-B* genotypes (Powers et al. 1991). The differences in LDH-B enzyme kinetics have important biological effects. The *LDH-B* genotypes affect hemoglobin-oxygen affinity, hatching time, and adult swimming performance (DiMichele and Powers 1982a, b; DiMichele et al. 1991). In addition to biochemical differences between *LDH-B* alleles, there are adaptive differences in LDH-B enzyme levels, which compensate for colder northern temperature and affect cardiac metabolism (Crawford and Powers 1989; Pierce and Crawford 1997; Podrabsky et al. 2000). These studies, with functional enzyme biochemistry tied to evolved gene expression, provide one of the clearest examples of natural selection occurring in natural populations.

This adaptive divergence in enzyme kinetics and expression occurs in *F. heteroclitus*, which do not have a pelagic larval stage and have a relatively small home range (Lotrich 1975; Able et al. 2006, 2012). With limited dispersal, one might expect natural selection to affect genotypes along a temperature cline. In contrast, during sexual reproduction the sea anemone (*Metridium senile*) releases sperm and eggs into the water column. Fertilized gametes drift in the plankton for 1–6 months before settling and metamorphosing into juveniles. Due to this relatively long pelagic larval duration (pld), *M. senile* can spread hundreds of kilometers from its origins (Hoffmann 1981). Yet, *M. senile* phosphoglucose isomerase (*GPI*) allele frequencies vary along a steep thermal gradient (Hoffmann 1981), and similar to *LDH-B* alleles, the different alleles differ in their kinetic properties, with greater differences evident at low temperatures. Consistent with temperature maintaining this allelic variation, the allelic variants showed the highest pentose-shunt metabolic flux differences at low temperatures (Zamer and Hoffmann 1989).

A final example of temperature effects on marine community structure is exemplified by enzymatic studies across closely related barracuda species (genus *Sphyraena*), including north temperate, subtropical, and south temperate species. Lactate dehydrogenase-A (LDH-A) proteins in six barracuda species have different apparent Kms for substrate and cofactor. For all species, Km increases with increasing assay temperature. However, the Kms for the six species are all the same when measured at the fish's normal temperature (Holland et al. 1997). This conservation of Kms arises from ~1.7-fold differences in the Kms when they are measured at a common temperature. These barracuda species have evolved different LDH-A proteins, yet unlike *F. heterclitus* populations that inhabit waters that differ by up to 12°C, the barracuda only inhabit waters that differ by 3–4°C. This suggests that even small temperature changes can drive natural selection, and thus global climate change might have significant effects on ectotherm survival and evolution.

Temperature clearly affects the population structure in the above examples. Indeed, temperature alone can predict 53–99% of the present day population structure along coastlines for shallow benthic faunas (Belanger et al. 2012). Yet because temperature differences often fall along a latitudinal cline, care must be taken to differentiate adaptive responses from demographic ones (Vasemagi 2006; Strand et al. 2012).

2.2 Salinity Changes

Salinity is another potential barrier in the marine environment, especially near coastlines, which can be significantly affected by freshwater input and evaporation from tidal pools and estuaries. Dealing with changing salinities can be energetically costly due to the need to either osmoregulate or osmoconform to maintain homeostasis. The blue mussel, *Mytilus edulis*, is an osmoconformer and accumulates intracellular organic osmolytes to match the ambient osmotic pressure in response to increased salinities. *M. edulis*' life history suggests that *M. edulis* populations should exhibit little genetic structure. They release fertilized eggs into the water column, and pelagic larvae remain in the water column for 3–7 weeks and can travel several hundred kilometers before settling (Newell 1989). Indeed, many loci exhibit little differentiation in protein polymorphisms (Levinton 1976). However, *M. edulis LAP* (leucine aminopeptidase I) alleles and LAP activities are associated with changing salinity. LAP's importance for osmoregulation is the production of amino acid osmolytes: LAP cleaves neutral or hydrophobic amino acids from N-terminal polypeptide ends, and *M. edulis* release these free amino acids into the cytosol to balance increased osmotic pressure due to increased salinity. Although adult *M. edulis* populations have altered *LAP* allele frequencies dependent on salinity, different *LAP* allele frequencies are not found in the settling larvae suggesting that differential juvenile mortality establishes the allelic difference in response to the salinity cline (Hilbish and Koehn 1985). This adaptive divergence occurs despite high gene flow.

The above four enzymatic studies characterizing allele frequency differences in targeted genes (*LDH-B*, *GPI*, *LDH-A*, and *LAP*) reveal biochemical differences that can be related to environmentally dependent, whole organism physiology. Thus, they illustrate how environmental variation can shape and maintain allele frequency differences between populations even in populations with high gene flow. Such local adaptation is dependent on gene flow and selection and often involves a genotype by environment interaction (Conover et al. 2006).

2.3 Ocean Currents

In contrast to temperature and salinity clines, ocean currents tend to homogenize populations by increasing gene flow between populations. Ocean currents are continuous, directed movement of seawater generated by forces, such as wind combined with the Coriolis effect, temperature and salinity differences, and breaking waves. Winds plus the Coriolis effect drive ocean surface currents. They can move huge volumes of water in well-defined, predictable patterns. Many currents are fast with strong thermal boundaries between the surrounding ocean water. These surface currents can transport marine species' eggs and larvae long distances. They also can form fronts, where two currents or water masses collide or where eddies shoot off.

Because many marine species have long pelagic larval durations (pld), marine populations thousands of kilometers apart can be well-connected (Kinlan and Gaines 2003).

Currents can enhance the connectivity among populations, yet currents and the resulting fronts formed between distinct water masses also can form important physical barriers in the world's oceans. These water mass differences can impose selection pressures or gene flow barriers, resulting in genetic differentiation between continuously distributed marine organisms (Saunders 1986). The genetic differentiation of marine organisms based on geographic distributions can reveal phylogeographic patterns where populations diverge and there is similar population structure for one or many species. Similar phylogeographic patterns among independent species suggest similar vicariant histories potentially related to periodic environmental changes during the Pleistocene as well as the species' life history patterns and dispersal capabilities (Avise 1992). Many of these marine phylogeographic patterns are found near land, for example, Cape Canaveral, Florida, on the East Coast of the United States (Avise 1992) and the Indo-West Pacific marine environment (Mcmillan and Palumbi 1995; Williams and Benzie 1998; Barber et al. 2000). Cape Mendocino off the California coast of the United States is another region where range limits of multiple species suggest a sharply delimited transition zone, yet here, intraspecific genetic divergences are not common (Burton 1998).

In contrast to these near shore oceanographic features, the Antarctic Polar Front has been proposed as a biogeographical barrier in an open-ocean environment (Clarke et al. 2005). The Antarctic Polar Front (Fig. 3) forms a barrier where cold, northward-flowing *Antarctic* waters meet the relatively warmer waters of the subantarctic and prevent a free north-south water exchange. The Antarctic Polar Front is large and deep. It has strong prevailing currents and a steep (3–4°C) temperature cline (Eastman 1993). Even if organisms do traverse this front, the temperature difference across the front likely limits many Antarctic and subantarctic species from establishing viable populations: it is too hot or too cold on the other side. A number of taxa show genetic divergence between South American and Antarctic locations. These include a variety of species, many with long-lived larvae, such as ribbon worms (Thornhill et al. 2008), bivalves (Page and Linse 2002), brittle stars (Hunter and Halanych 2008), krill (Patarnello et al. 1996), fish (Shaw et al. 2004), and colonial alga (Medlin et al. 1994). These data show that even though these species can disperse over large geographic areas, the Antarctic Polar Front and associated Antarctic Circumpolar Current form a physical oceanographic barrier that restricts such dispersal over evolutionary time (Thornhill et al. 2008).

Another potential oceanographic barrier to population connectivity is the Eastern Pacific Barrier. The Eastern Pacific Barrier is an ~5,000 km stretch of uninterrupted water with depths between 5,000 and 8,000 km (Grigg and Hey 1992) that separates the central from the eastern Pacific Ocean. While the Eastern Pacific Barrier is not a barrier to fish (Rosenblatt and Waples 1986; Lessios et al. 1998; Lessios and Robertson 2006), sea urchins, *Tripneustes* sp. (Lessios et al. 2003), and crown-of-thorns seastar, *Acanthaster planci* (Nishida and Lucas 1988), it does form an almost

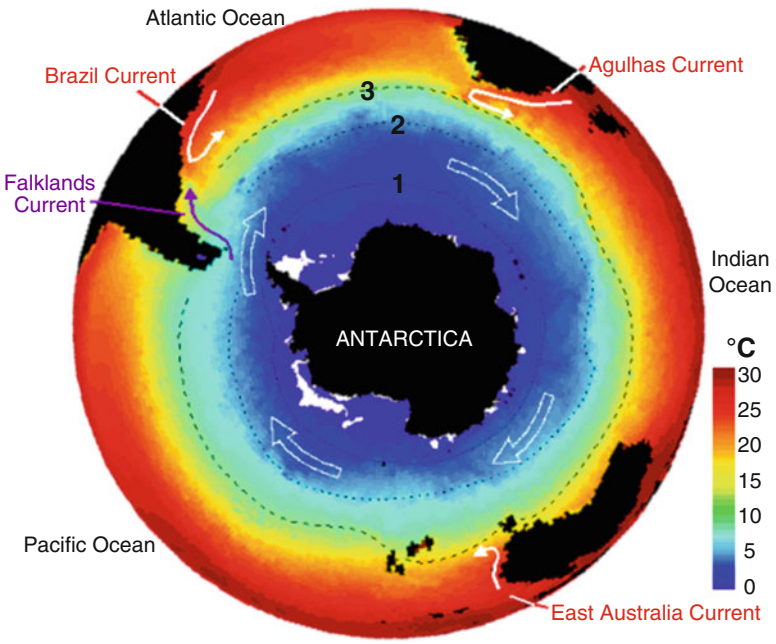

Fig. 3 Sea surface temperature (SST) map of the Southern Ocean in summer. Three fronts can be seen as areas where the temperature change from north to south is particularly fast. (1) The polar front, (2) the subantarctic front, and (3) the subtropical front – the northern boundary of the Southern Ocean. White outlined arrows indicate the flow of the Antarctic Circumpolar Current. Source: NOC from SST climatology data. http://www.seos-project.eu/modules/oceancurrents/oceancurrents-c02-s03-p03.html. This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA

complete barrier for the coral *Porites lobata*, a coral species with potentially long-lived larvae (Baums et al. 2012). Interestingly, on an island just north and west of the Galapagos, there is a Pacific *P. lobata* population that is more genetically similar to western populations than the geographically closer eastern Pacific populations. However, this population did not migrate further south and east to the Galapagos, ~3,000 km closer than the western populations and on the same side of the Eastern Pacific Barrier, suggesting that other factors, potentially temperature, were also limiting its dispersal. This highlights the importance of interacting factors, both biotic and abiotic, in affecting population differentiation.

2.4 Other Potential Barriers

A variety of other potential physical barriers exist in the world's oceans. Many of these are exacerbated by human activities. These include suitable habitat availability

(Burton et al. 1979; Riginos and Nachman 2001), hypoxia (Breitburg et al. 2009), and pollution (Bozinovic and Oleksiak 2011; Hamilton et al. 2016). Further, it is often not just one factor that promotes genetic differentiation in marine populations but instead is the combined effects of multiple factors. For instance, genetic differentiation of the subtidal fish *Axoclinus nigricaudus*, which has benthic eggs but planktonic larvae, cannot be explained by a single factor and instead is correlated with the combined effects of biogeography, geographical distance, and habitat availability (Riginos and Nachman 2001). The dispersion potential of many if not all of the physical barriers can be affected by biological factors as well. Certainly, the variety of different species' dispersal patterns across the same geographic range involving the same currents suggests the importance of life history and dispersal capability in shaping population structure. Thus, in addition to physical processes, biological processes such as local adaptation, reproductive strategy, and larval behavior (e.g., Swearer et al. 2002; Jones et al. 2005; Almany et al. 2007; Shulzitski et al. 2016) can influence the genetic structuring of marine organisms despite their long-distance dispersal potentials. For many organisms inhabiting the marine environment, it is the interaction between the physical and biological processes that drives their population structures. The challenge remains to understand biological processes in the context of the physical environment, and for questions of population structure, genomic approaches now give us an incredibly powerful toolbox to address this challenge.

3 Population Genomics in the Oceans

The studies described above targeted specific genes or a small number of putatively neutral markers (e.g., mitochondrial sequences or microsatellites). Yet now with NGS approaches, we can interrogate across whole genomes, which provides us with readily available genetic markers spread across the entire genome. These hundreds to thousands of widely distributed genetic markers provide two benefits for population genetic studies. First, the large number of markers and broad genomic coverage provide greater confidence in estimating neutral population genetic parameters, such as effective population size and migration rate, and allow us to more confidently establish neutral expectations (Allendorf et al. 2010). Second, because loci under selection should be affected by demography and evolutionary history differently than neutral loci, population genomic approaches also enhance the ability to identify adaptive or functionally important loci and genes (Luikart et al. 2003). In large marine populations with generally low F_{ST} values for neutrally evolving genes, loci under directional selection may be easier to distinguish from neutral expectations (Nielsen et al. 2009b) and are indicative of hidden population structure, which may be important for conservation or to enhance our understanding of biological oceanographic processes.

The idea that restricted gene flow, with resulting non-neutral DNA polymorphism patterns, occurs between marine populations is converging with the idea of much

more nuanced environmental gradients, and knowledge of the rapidity and breadth of these non-neutral changes. It will greatly enhance our understanding of how life adapts to global change. Yet, this same population genomics approach has the potential drawback that while we can identify loci apparently under selection, we often lack functional data to "prove" adaptation or natural selection. Unlike earlier, targeted gene studies, follow-up functional studies will often be lacking, leaving us to simply postulate functional effects. This is especially true when natural selection acts on polygenic traits. That is, if selection is acting on many genes, each with a small effect, then functional assays need to take into consideration many different allelic combinations across many loci. For example, a biallelic trait affected by 2 loci will have 9 genotypes, 1 affected by 3 loci will have 27, and 1 affected by 4 loci will have 81. Even with today's gene editing capabilities, which work across species (Jinek et al. 2012), it is incredibly difficult to study the functional outcomes of polygenic effects, especially in non-model, multicellular organisms, and both ecological and physiological knowledge of the organism under study will be crucial for interpreting genomic data. Thus, linking potentially adaptive loci to biologically important phenotypes remains challenging, especially for many marine species, which often cannot be reared or even maintained in the laboratory (Oleksiak 2016). Even so, genomic results provide a powerful starting point to complement and direct functional approaches to better understand adaptive variation in marine populations.

3.1 Genetic Connectivity and Phylogeography

High-throughput genomic approaches have made it relatively easy to study the genetic connectivity of any marine species (but see Waples and Gaggiotti (2006) and Gagnaire et al. (2015) for problems defining marine populations) and also provide an excess of genetic markers with which to explore adaptive variation in both space and time (Crawford and Oleksiak 2016). Most of these plentiful genetic markers support conclusions of previous research using fewer, neutral markers. For example, a genotyping by sequencing approach – where next-generation sequencing is used to genotype hundreds of individuals at a time at many loci – identified approximately 1,320 SNPs in populations of the estuarine fish, sailfin molly (*Poecilia latipinna*). Fish inhabiting three geographically close, salt marsh flats (within 10 km of each other) in the Florida Keys showed little genetic differentiation (F_{ST} values <0.0125 for most SNPs [less than 1.25% of the variance among populations relative to the total]) (Nunez et al. 2015). These data support previous studies using allozymes and microsatellites (Trexler et al. 1997; Apodaca et al. 2013), also suggesting few differences among South Florida sailfin molly populations. In addition to confirming the genetic connectivity of these populations, these data also identify a small percentage of loci (~1.4%) that are potentially adaptive. These loci show F_{ST} values that are unlikely to occur relative to random permutations of loci with similar heterozygosities and identify population structure

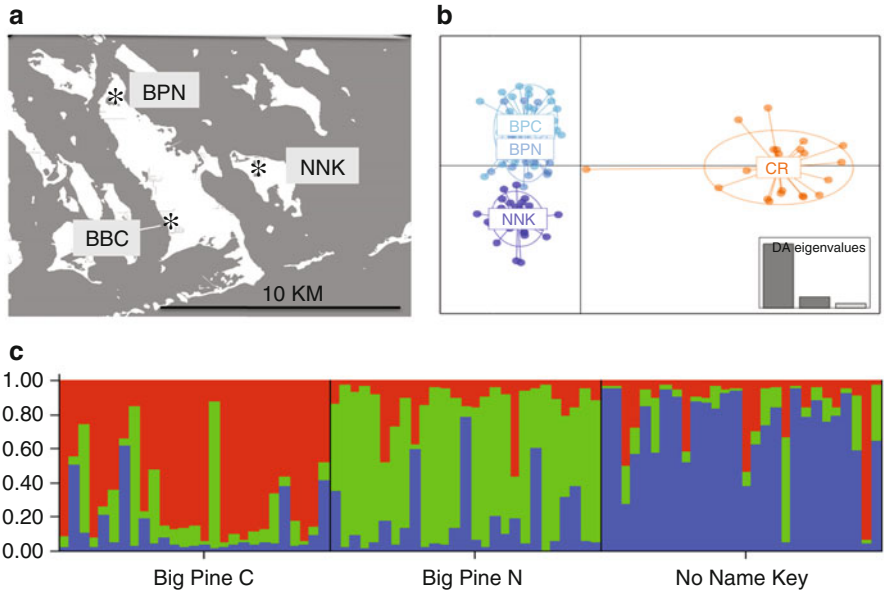

Fig. 4 Sailfin molly population genomics. (**a**) Florida Key sailfin molly populations that are within 10 km of each other. (**b**) DAPC (discrimination analyses of principal components) showing demographic structure between NoName Key (NNK) and the two Big Pine Key (BPC and BPN) populations. Crandon Park (CR) is a population in Miami Dade approximately 180 km away. (**c**) Structure analysis of the three Key populations, which discriminates among all three populations. Adapted from Nunez et al. (2015)

not visible using neutral loci (Fig. 4). Furthermore, these excessive F_{ST} values suggest adaptive divergence due to local environments.

Similar results have been found for the threatened staghorn coral, *Acropora cervicornis*, the fastest growing Caribbean coral. Staghorn coral populations between Florida and sites in the Caribbean show significant genetic structure only across large geographic distances in both nuclear and mitochondrial genes (Vollmer 2007), suggesting restricted gene flow reflecting ocean currents that potentially isolate populations. Yet, within the Florida Reef Tract, the third largest barrier reef system in the world, extending from Biscayne Bay to the Tortugas Banks (nearly 240 km) and bounded by the Florida Current to the east, analysis of *A. cervicornis* using microsatellites showed little population differentiation and no significant population structure (Baums et al. 2010). Recently, genotyping by sequencing was used to genotype *A. cervicornis* individuals along the Florida Reef Tract at ~4,700 loci. While most of the genetic diversity (>90%) was found to reside within populations similar to previous studies, the genomic analyses showed significant variation along the Florida Reef Tract, including 300 SNPs with significant F_{ST} values and significant divergence relative to distance even over small spatial scales (Drury et al. 2016). These studies highlight the ability of population genomic approaches to identify previously unresolved population structure. While this

might simply reflect the higher number of genomic markers and thus not be biologically relevant (Hedrick 1999), diagnostic genetic markers allow for population discrimination and source population identification, which are important aspects for regulating protected populations or for defining marine protected areas (MPAs).

The greater resolution of genetic differences based on population genomic approaches has been found in an increasing number of other studies for many diverse marine taxa. For instance, population genomic analyses using high-confidence SNPs identified highly resolved phylogeographic relationships for natural populations of the sea anemone (*Nematostella vectensis*), a developing cnidarian model for comparative and ecological genomics. This resolution was not achieved in previous studies using traditional markers (Reitzel et al. 2013). Similarly for Pacific lamprey (*Entosphenus tridentatus*), a highly dispersive anadromous fish with high gene flow, a genotyping by sequencing approach using ~4,000 genetic markers showed that although neutral variation identified some evidence of more than one population, similar to previous studies using fewer genetic markers, analyses of adaptive variation, which was associated with geography and life history, found a much finer genetic structure scale within the broad regions sampled (Hess et al. 2013). Similarly, greenlip abalone (*Haliotis laevigata*) showed very low differentiation using 8,786 putatively neutral loci but 5 divergent population clusters using 323 candidate adaptive loci (Sandoval-Castillo et al. 2018). These studies demonstrate that genomic approaches can identify population structure that is not apparent when using a few neutral markers.

3.2 Genomic Impacts on Ocean Fisheries

Understanding population structure is important for managing fisheries stocks because independent fisheries stocks are likely to have independent population dynamics and respond differently to changing environmental conditions and fishing pressures. An important fish stock in North Atlantic waters is Atlantic herring (*Clupea harengus*); previous studies using a limited number of genetic markers found no genetic differentiation between Atlantic herring sampled from different regions. Atlantic herring is a pelagic fish in North Atlantic waters. It is a major food source for many marine animals and is widely used for producing fish feed for aquaculture (Lamichhaney et al. 2012). Thus, understanding the genetic differentiation between herring stocks is critical for sustainably managing this species. Using transcriptome sequencing, more than 440,000 SNPs were identified across herring from a wide geographic range, and most showed no allele frequency differences among populations. However, in contrast to this lack of genetic differentiation for most loci, several thousand SNPs (2–3%) showed strong allele frequency differences (Lamichhaney et al. 2012) and define a number of genetically distinct herring populations in the North Atlantic. Many of the differentiated loci are correlated with salinity and associated with osmoregulation in other species, suggesting that salinity differences across geographic regions might be driving the genetic differentiation.

Population genomic approaches have been used with a number of other commercially important marine species, and there has been an exponential increase in fisheries-related population genomic studies from 2009 to 2014 (Valenzuela-Quinonez 2016). For example, another important fish stock in North Atlantic waters is turbot (*Scophthalmus maximus*), which inhabits the European continental shelf. Among 20 turbot populations collected from across its range, genotyping via double-digest RAD sequencing showed that this flatfish species is structured into four main regions: Baltic Sea, Atlantic Ocean, Adriatic Sea, and Black Sea. Genetic variation correlates with temperature and salinity, suggesting that these two parameters are driving the genetic differentiation (Prado et al. 2018). In another study, a targeted genome scan was used specifically to determine whether Atlantic cod (*Gadus morhua*) populations are adapted to local environmental conditions (Nielsen et al. 2009a) and showed stable interpopulation variation over a 24-year time period. This interpopulation variation was better correlated with spawning ground temperature and/or salinity conditions during spawning than with geographic distance. While the mechanisms maintaining local adaptation despite high gene flow are still poorly understood, a subsequent study hints at the importance of genomic architecture: cod populations locally adapted to low salinity fjord environments have a significant overrepresentation of a large (~5 Mb) chromosomal rearrangement (Barth et al. 2017).

Additional population genomic studies show limited effective dispersal that structures sea scallop (*Placopecten magellanicus*) populations along eastern North America (Van Wyngaarden et al. 2016), two well-defined anchovy ecotypes for the European anchovy (*Engraulis encrasicolus*) collected from Atlantic and Mediterranean locations that correlate with habitat (Montes et al. 2013), and spatially varying selection acting on glass eels (an intermediary stage in the eel's complex life history between the leptocephalus stage and the juvenile [elver] stage) in the otherwise panmictic European eel, *Anguilla anguilla* (Pujolar et al. 2014), and American eel (*Anguilla rostrata*) (Gagnaire et al. 2012). Studies in two different lobster species, the southern rock lobster (*Jasus edwardsii*) and American lobster (*Homarus americanus*), identified genetic markers that can be used for assignment tests to the original population (Benestan et al. 2015; Villacorta-Rath et al. 2016). These findings show that using the large number of genetic markers available through population genomic approaches can improve the identification of fine-scale structure and be used to better define appropriate stock management scales and conservation units in these commercially valuable species. Additionally, these approaches can be used to identify population origins, which is critical for enforcing management policies.

However, not all population genomic approaches reveal previously unknown population structure. For instance, in a commercially harvested abalone species (*Haliotis rubra*) from southeastern Australia, genotyping by sequencing results using up to 1,700 SNPs indicate high levels of gene flow and no significant genetic structure within or between benthic reef habitats across 1,400 km of coastline (Miller et al. 2016). Given that abalone along this coast inhabit reef patches up to at least 6,600 m apart, this suggests that recruitment success along this coast does not predominantly depend on local reef sources.

3.3 Local Adaptation

Overall, most marine population genomic studies comparing between groups or populations consistently identify a few percentages of SNPs with elevated divergence, which exceeds neutral expectations. Thus, in addition to resolving previously unresolved population structure, marine genomic studies also are revealing a plethora of potentially adaptive loci. Perhaps the most well-known example of a marine organism adaptation using genomic approaches is repeated stickleback (*Gasterosteus aculeatus*) adaptation from oceanic to freshwater habitats. A genome scan using over 45,000 SNPs identified parallel genetic divergence across independent populations in both previously characterized and novel genomic regions (Hohenlohe et al. 2010). How quickly stickleback freshwater adaptation occurs was addressed in a subsequent population genomic study that examined sticklebacks from freshwater habitats that were only recently colonized by sticklebacks from ocean populations. These freshwater habitats were formed on earthquake-uplifted islands in Alaska in 1964. Fifty years later, these populations have phenotypically diverged from the oceanic phenotypes to nearly the same extent as much older freshwater stickleback populations and also show genetic divergence between oceanic and freshwater populations (Messer et al. 2016). The lower genetic divergence between oceanic and freshwater stickleback populations compared to the divergence among the freshwater populations suggests independent invasions of the freshwater habitats and further differences among freshwater habitats that have occurred within the last 50 years, despite likely recurrent gene flow between oceanic and freshwater populations.

The stickleback study suggests that freshwater adaptation occurs quickly, within the first few decades of freshwater invasion, and raises the question of how rapidly adaptation occurs in nature. With strong selection, directional selection is often rapid. For example, four independent *F. heteroclitus* populations have adapted to strong pollution clines within 50 generations (Reid et al. 2016). Similarly, introduced Chinook salmon (*Oncorhynchus tshawytscha*) show rapid trait divergence between populations within at most 30 generations (Quinn et al. 2001), and the Atlantic silverside (*Menidia menidia*) showed selection for slower or faster growth rates in response to size-selected harvest in just 4 generations (Conover and Munch 2002). While this last example is due to artificial selection in the laboratory, there is a growing body of evidence that rapid phenotypic evolution is common in nature (Messer et al. 2016).

While geographically varying selection is widely accepted as an important factor for maintaining genetic variation, less attention has been paid to temporally fluctuating selection (Messer et al. 2016). Temporal fluctuations as exemplified by cold years in the North Atlantic with a general warming trend associated with global warming have affected species distributions (Wethey et al. 2011). What is less well understood is whether these types of temporal variations affect genetic diversity within a species or divergence among populations. Since global warming is associated with higher variation in climatic conditions, understanding the effect of temporal variations is an important avenue for future population genomics research.

Population level genomics will allow us to better understand genetic variation over time as well as space. This may be particularly relevant for marine populations with large, well-connected populations harboring lots of genetic variation. Importantly, shifting habitats and environmental conditions as might be influenced by seasonal current shifts, large-scale ocean-atmosphere oscillations (e.g., El Niño/Southern Oscillation or ENSO, Antarctic, Arctic, and North Atlantic Oscillations among others), as well as environmental disturbances due to local (e.g., pollution and eutrophication) and global (e.g., global warming and ocean acidification (Sunday et al. 2014)) climate change can cause strong directional selection and require a rapid evolutionary response. We now have the resolution to analyze very recently diverged populations at the genomic level, even due to seasonal changes (Garud et al. 2015). Understanding the relevant time scales and ecological factors affecting rapidly fluctuating selection will require extensive sampling of both populations and relevant environmental parameters (Messer et al. 2016) and will have important implications for how we protect and manage marine populations in today's changing environments.

3.4 Epigenomic Adaptation

In contrast to local adaptations, which become hardwired into organisms' genomes, epigenetic changes provide organisms with alternative ways to deal with changing environments. Epigenetic changes are heritable changes in the genome that do not alter the DNA sequence (Deans and Maggert 2015), and the best studied epigenetic modification is DNA methylation. DNA methylation studies across 17 eukaryotic genomes, including marine species genomes, suggest that gene body methylation is conserved between plants and animals (Zemach et al. 2010). Other major epigenetic modifications include chromatin remodeling, histone modifications, and noncoding RNA mechanisms. These epigenetic mechanisms are shared across most taxa.

At the population level, environmental epigenomic studies of marine organisms are just beginning and mostly focus on DNA methylation. For marine populations impacted by rapid environmental change, epigenetic mechanisms may give impacted populations enough time to genetically adapt. This may be especially important for sessile marine invertebrates that have no choice but to cope with the environment they inhabit. Interestingly, a study examining the role of genome-wide DNA methylation in the adaptation of a marine stickleback population to freshwater conditions found that the genes that harbor genetic and epigenetic changes were not the same, suggesting that epigenetic adaptation complements but does not replace natural selection (Artemov et al. 2017).

Examples concerning epigenetic effects due to environmental change include studies examining pollution, temperature, and pCO2 effects. Thus, environmental pollutants have been shown to affect genomic methylation levels in three-spined stickleback (Aniagu et al. 2008), flatfish dab liver tumors (Mirbahai et al. 2011), and European eels (Pierron et al. 2014). Across fish species living at different

temperatures, polar fishes exhibit higher DNA methylation levels than tropical and temperate fishes (Varriale and Bernardi 2006). In the context of global environmental change, a recent study in an Antarctic marine polychaete showed both physiological and epigenetic responses to increased temperatures. When cultured in the laboratory, the Antarctic polychaete *Spiophanes tcherniai* rapidly responded to increased temperatures: within 4 weeks of a high temperature stress (from −1.4°C to +4°C), metabolic rates return to normal. Additionally, these worms showed an 11% increase in CpG methylation state genome wide, with 85% of changes showing a net increase in methylation (Marsh and Pasqualone 2014). Similarly, larval European sea bass exposed to just 2°C warmer temperatures, the temperature increase predicted by recent global warming models, changed both global DNA methylation and the expression of ecologically relevant genes related to DNA methylation, stress response, and muscle and organ formation (Anastasiadi et al. 2017). Another example relevant to global environmental change occurs in corals exposed to increased CO_2 levels to simulate ocean acidification. Two different reef-building coral species, *Pocillopora damicornis* and *Montipora capitata*, were exposed to ambient and ocean acidification conditions in common garden tanks for ~6 weeks. *Pocillopora damicornis* showed an epigenetic response, while *Montipora capitata* did not (Putnam et al. 2016). Not surprisingly, inducible DNA methylation varies by taxa.

Environmental change also occurs for invasive species and invasive species dynamics, which can provide insight into how populations might adapt to rapid environmental change. During the expansive phase of a recent invasion (within 2 years), pygmy mussel (*Xenostrobus securis*) showed significantly reduced global methylation levels. In older introductions such epigenetic signatures of invasion were progressively reduced. Decreased methylation was interpreted as a rapid way of increasing phenotypic plasticity that would help invasive populations to thrive. As reported for introduced plants and vertebrates, epigenetic variation could compensate for relatively lower genetic variation caused by founder effects (Ardura et al. 2017). Overall, epigenetic changes may be a rapid and powerful way in which marine organisms can respond to rapid environmental change.

4 Conservation and Management Considerations

Both population connectivity and how organisms are able to and do adapt to changing environments have significant implications for how marine systems are conserved and managed. Population connectivity is critically important when considering how best to manage valuable resources such as ecosystem diversity and is unknown for the vast majority of marine species. If source populations are not protected, marine protected areas will be ineffective. Thus, the ability of a marine protected area to sustain locally endangered populations depends on its connectivity to other protected areas or other non-endangered populations and requires an understanding of larval and adult exchange between locations (Palumbi 2003).

Population connectivity is unknown for many marine species and with the many potential physical and biological barriers to dispersion cannot be assumed to be boundless. Furthermore, with today's changing environments, population connectivity will change. This is evident with man-made habitat fragmentation but also with more subtle changes such as increasing mean annual ocean temperatures. Increased temperatures are causing range expansion for many species by allowing adults to survive and reproduce in the higher latitudes (Sorte et al. 2010; Chen et al. 2011). However, increased temperatures also will shorten pelagic larval duration for many species due to increased metabolism (O'Connor et al. 2007). Thus, this shortened pelagic larval duration will limit connectivity even as adult ranges increase. Understanding the biological and environmental interactions and how they affect marine connectivity will remain an important factor for successfully protecting marine diversity.

The identification of potentially adaptive loci in marine populations also has implications for marine management and conservation. In the oceans, adaptive population differentiation occurs across different spatial scales and for species with different life histories. Understanding local adaptation provides insights into how organisms will deal with climate change and thus how best to manage and conserve marine species with climate change. Studies of domestication and experimental selection in yeast are making it clear that local adaptation over ecological time scales selects from standing genetic variation (Burke et al. 2014; Boitard et al. 2016). This highlights the need to protect genetic diversity in marine populations if these populations are to retain the ability to respond to a changing environment. However, an open question is whether or how much of all the potentially adaptive genetic differences recently revealed by population genomic studies are relevant for conservation and species management. Given the complexity of adaptation in the marine environment with fluctuating selective pressures, likely polygenic adaptation where many genes have small, nonmeasurable effects (Rockman 2012), and the fact that neutral and adaptive markers provide different types of information (Funk et al. 2012), indeed, the best conservation approach may simply be to preserve as much genetic variation as possible so that species can maintain the full extent of their evolutionary potential (Pearse 2016).

Regardless of whether or not newly discovered adaptive loci will or should impact management decisions, these potentially adaptive loci do have an important role in marine conservation with respect to identifying population origin. This is important for exploited and endangered species because illegal, unregulated, and unreported fishing significantly contributes to fish population overexploitation and negatively affects population and ecosystem recovery. Illegal, unregulated, and unreported fishing in high seas causes economic losses between $10 and $23.5 billion annually and is highly correlated with governance (Agnew et al. 2009). The ability to identify and keep track of the origin of fishery products along the supply chain will make controlling and enforcing regulations easier (Ogden 2008), and genetic markers identified with population genomics approaches provide this ability. For example, genome scans were used with four economically important fish species (Atlantic cod [*Gadus morhua*], Atlantic herring [*Clupea harengus*],

sole [*Solea solea*], and European hake [*Merluccius merluccius*]), all threatened by overfishing and illegal, unregulated, and unreported fishing activities, to identify genetic markers with high genetic differentiation. These markers correctly assigned 93–100% of individuals to correct population origin. Thus, this marine population genomics approach provides a powerful, readily developed, and standardized means to identify population origin and thus enhance fishing governance (Nielsen et al. 2012).

5 Conclusions and Future Prospects

Marine population genomics has given us the unprecedented ability to resolve population structure, identify genetic divergence among populations, and detect selectively important genes. These data are important because they inform us about the conservation genetics of isolated populations, the genes affecting important phenotypes (e.g., reproductive schedules) and the frequency and effectiveness of adaptive change in a changing environment. An ever-expanding number of genomic studies suggest that marine species have greater population structure than previously appreciated. Additionally, many of these studies identify lots of loci apparently evolving by natural selection over both long and short evolutionary time scales. A growing challenge will be to determine the functional effects of these loci evolving by natural selection and predict which of the genetic differences revealed by population genomic approaches are relevant for conservation and species management. Regardless, these loci allow high-resolution stock identification and have important implications for regulating illegal fishing. The number and frequency of loci apparently evolving by natural selection suggests that natural selection is more effective than currently appreciated, resulting in marine populations adapted to local environmental conditions. If it is true that natural selection is more effectively shaping population-specific genotypes, it suggests that current climate changes will be mitigated by adaptive change in many marine organisms with sufficient genetic variation (Crawford and Oleksiak 2016). This optimism is tempered by the realization that while one or many species may adapt to climate change, the spatial and temporal interactions among species could alter and have negative effects on ecosystems.

The prospects and challenges for marine population genomics are similar to those for any natural population, marine or terrestrial. In addition to linking genotype to phenotype to determine the functional effects of loci evolving by natural selection, which is especially difficult when life histories are unknown and the species themselves cannot be cultured as is still true for many marine species, further challenges include understanding the genetic architecture underlying adaptive phenotypes in the presence of gene flow in large marine populations without strict boundaries and assessing changing population dynamics in today's fast-changing environments. Whole genome sequencing of marine organisms (along with the bioinformatic tools to analyze these sequences and genomes) is accelerating. This in turn will accelerate whole genome sequencing of marine populations, allowing us to study the

genomic landscapes and allelic diversity variance within and between populations truly at the genome-wide level (Ellegren 2014). Consequently, by further developing marine population genomics, it will be possible to better understand how populations respond to and are affected by their environment and eventually gain insight into how population dynamics affect ecosystem functioning as a whole.

References

Able KW, Hagan SM, Brown SA. Habitat use, movement, and growth of young-of-the-year Fundulus spp. in southern New Jersey salt marshes: comparisons based on tag/recapture. J Exp Mar Biol Ecol. 2006;335(2):177–87.

Able KW, Vivian DN, Petruzzelli G, Hagan SM. Connectivity among salt marsh subhabitats: residency and movements of the mummichog (Fundulus heteroclitus). Estuar Coasts. 2012;35(3):743–53.

Agnew DJ, Pearce J, Pramod G, Peatman T, Watson R, Beddington JR, Pitcher TJ. Estimating the worldwide extent of illegal fishing. PLoS One. 2009;4(2):e4570.

Allendorf FW, Hohenlohe PA, Luikart G. Genomics and the future of conservation genetics. Nat Rev Genet. 2010;11(10):697–709.

Almany GR, Berumen ML, Thorrold SR, Planes S, Jones GP. Local replenishment of coral reef fish populations in a marine reserve. Science. 2007;316(5825):742–4.

Anastasiadi D, Diaz N, Piferrer F. Small ocean temperature increases elicit stage-dependent changes in DNA methylation and gene expression in a fish, the European sea bass. Sci Rep. 2017;7(1):12401.

Aniagu SO, Williams TD, Allen Y, Katsiadaki I, Chipman JK. Global genomic methylation levels in the liver and gonads of the three-spine stickleback (Gasterosteus aculeatus) after exposure to hexabromocyclododecane and 17-beta oestradiol. Environ Int. 2008;34(3):310–7.

Apodaca JJ, Trexler JC, Jue NK, Schrader M, Travis J. Large-scale natural disturbance alters genetic population structure of the Sailfin Molly, Poecilia latipinna. Am Nat. 2013;181(2):254–63.

Ardura A, Zaiko A, Moran P, Planes S, Garcia-Vazquez E. Epigenetic signatures of invasive status in populations of marine invertebrates. Sci Rep. 2017;7:42193.

Artemov AV, Mugue NS, Rastorguev SM, Zhenilo S, Mazur AM, Tsygankova SV, Boulygina ES, Kaplun D, Nedoluzhko AV, Medvedeva YA, Prokhortchouk EB. Genome-wide DNA methylation profiling reveals epigenetic adaptation of stickleback to marine and freshwater conditions. Mol Biol Evol. 2017;34(9):2203–13.

Avise JC. Molecular population structure and the biogeographic history of a regional fauna: a case history with lessons for conservation biology. Oikos. 1992;63(1):62–76.

Baird N, Etter P, Atwood T, Currey M, Shiver A, Lewis Z, Selker E, Cresko W, Johnson E. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 2008;3:e3376.

Barabas G, D'Andrea R. The effect of intraspecific variation and heritability on community pattern and robustness. Ecol Lett. 2016;19(8):977–86.

Barber PH, Palumbi SR, Erdmann MV, Moosa MK. Biogeography. A marine Wallace's line? Nature. 2000;406(6797):692–3.

Barth JMI, Berg PR, Jonsson PR, Bonanomi S, Corell H, Hemmer-Hansen J, Jakobsen KS, Johannesson K, Jorde PE, Knutsen H, Moksnes PO, Star B, Stenseth NC, Svedang H, Jentoft S, Andre C. Genome architecture enables local adaptation of Atlantic cod despite high connectivity. Mol Ecol. 2017;26(17):4452–66.

Baums I, Johnson M, Devlin-Durante M, Miller M. Host population genetic structure and zooxanthellae diversity of two reef-building coral species along the Florida Reef Tract and wider Caribbean. Coral Reefs. 2010;29(4):835–42.

Baums IB, Boulay JN, Polato NR, Hellberg ME. No gene flow across the Eastern Pacific Barrier in the reef-building coral Porites lobata. Mol Ecol. 2012;21(22):5418–33.

Belanger CL, Jablonski D, Roy K, Berke SK, Krug AZ, Valentine JW. Global environmental predictors of benthic marine biogeographic structure. Proc Natl Acad Sci U S A. 2012;109(35):14046–51.

Benestan L, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L. RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (Homarus americanus). Mol Ecol. 2015;24(13):3299–315.

Boitard S, Boussaha M, Capitan A, Rocha D, Servin B. Uncovering adaptation from sequence data: lessons from genome resequencing of four cattle breeds. Genetics. 2016;203(1):433–50.

Bozinovic G, Oleksiak MF. Genomic approaches with natural fish populations from polluted environments. Environ Toxicol Chem. 2011;30(2):283–9.

Breitburg DL, Hondorp DW, Davias LA, Diaz RJ. Hypoxia, nitrogen, and fisheries: integrating effects across local and global landscapes. Annu Rev Mar Sci. 2009;1:329–49.

Burke MK, Liti G, Long AD. Standing genetic variation drives repeatable experimental evolution in outcrossing populations of Saccharomyces cerevisiae. Mol Biol Evol. 2014;31(12):3228–39.

Burton RS. Intraspecific phylogeography across the point conception biogeographic boundary. Evolution. 1998;52(3):734–45.

Burton RS, Feldman MW, Curtsinger JW. Population-genetics of Tigriopus-californicus (Copepoda, Harpacticoida). 1. Population-structure along the Central California coast. Mar Ecol Prog Ser. 1979;1(1):29–39.

Chen I-C, Hill JK, Ohlemüller R, Roy DB, Thomas CD. Rapid range shifts of species associated with high levels of climate warming. Science. 2011;333(6045):1024–6.

Clarke A, Barnes DK, Hodgson DA. How isolated is Antarctica? Trends Ecol Evol. 2005;20(1):1–3.

Conover DO, Munch SB. Sustaining fisheries yields over evolutionary time scales. Science. 2002;297(5578):94–6.

Conover DO, Clarke LM, Munch SB, Wagner GN. Spatial and temporal scales of adaptive divergence in marine fishes and the implications for conservation. J Fish Biol. 2006;69:21–47.

Crawford DL, Oleksiak MF. Ecological population genomics in the marine environment. Brief Funct Genomics. 2016;15(5):342–51.

Crawford DL, Powers DA. Molecular-basis of evolutionary adaptation at the lactate dehydrogenase-B locus in the fish Fundulus-heteroclitus. Proc Natl Acad Sci U S A. 1989;86(23):9365–9.

Dana JD. On an isothermal oceanic chart, illustrating the geographical distribution of marine animals. Am J Sci Arts. 1853;16:153–67. 314–327

Deans C, Maggert KA. What do you mean, "epigenetic"? Genetics. 2015;199(4):887–96.

DiMichele L, Powers DA. LDH-B genotype-specific hatching times of *Fundulus heteroclitus* embryos. Nature. 1982a;296(5857):563–4.

DiMichele L, Powers DA. Physiological basis for swimming endurance differences between LDH-B genotypes of *Fundulus heteroclitus*. Science. 1982b;216(4549):1014–6.

DiMichele L, Paynter KT, Powers DA. Evidence of lactate dehydrogenase-B allozyme effects in the teleost, *Fundulus heteroclitus*. Science. 1991;253(5022):898–900.

Drury C, Dale KE, Panlilio JM, Miller SV, Lirman D, Larson EA, Bartels E, Crawford DL, Oleksiak MF. Genomic variation among populations of threatened coral: Acropora cervicornis. BMC Genomics. 2016;17:286.

Eastman J. Antarctic fish biology: evolution in a unique environment. San Diego: Academic Press; 1993.

Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 2014;29(1):51–63.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6(5):e19379.

Etter P, Bassham S, Hohenlohe P, Johnson E, Cresko W. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Methods Mol Biol. 2011;772:157–78.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW. Harnessing genomics for delineating conservation units. Trends Ecol Evol. 2012;27(9):489–96.

Gagnaire PA, Normandeau E, Cote C, Moller Hansen M, Bernatchez L. The genetic consequences of spatially varying selection in the panmictic American eel (Anguilla rostrata). Genetics. 2012;190(2):725–36.

Gagnaire PA, Broquet T, Aurelle D, Viard F, Souissi A, Bonhomme F, Arnaud-Haond S, Bierne N. Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. Evol Appl. 2015;8(8):769–86.

Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS Genet. 2015;11(2):e1005004.

Grigg RW, Hey R. Paleoceanography of the tropical eastern Pacific Ocean. Science. 1992;255(5041):172.

Hamilton PB, Cowx IG, Oleksiak MF, Griffiths AM, Grahn M, Stevens JR, Carvalho GR, Nicol E, Tyler CR. Population-level consequences for wild fish exposed to sublethal concentrations of chemicals - a critical review. Fish Fish. 2016;17(3):545–66.

Hand BK, Lowe WH, Kovach RP, Muhlfeld CC, Luikart G. Landscape community genomics: understanding eco-evolutionary processes in complex environments. Trends Ecol Evol. 2015;30(3):161–8.

Hedrick PW. Perspective: highly variable loci and their interpretation in evolution and conservation. Evolution. 1999;53:313–8.

Hess JE, Campbell NR, Close DA, Docker MF, Narum SR. Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. Mol Ecol. 2013;22(11):2898–916.

Hilbish TJ, Koehn RK. The physiological basis of natural selection at the LAP locus. Evolution. 1985;39(6):1302–17.

Hoffmann RJ. Evolutionary genetics of Metridium senile. II. Geographic patterns of allozyme variation. Biochem Genet. 1981;19(1):145–54.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 2010;6(2):e1000862.

Holland LZ, McFall-Ngai M, Somero GN. Evolution of lactate dehydrogenase-A homologs of barracuda fishes (genus Sphyraena) from different thermal environments: differences in kinetic properties and thermal stability are due to amino acid substitutions outside the active site. Biochemistry. 1997;36(11):3207–15.

Hunter RL, Halanych KM. Evaluating connectivity in the brooding brittle star Astrotoma agassizii across the Drake Passage in the Southern Ocean. J Hered. 2008;99(2):137–48.

Hutchins LW. The bases for temperature zonation in geographical distribution. Ecol Monogr. 1947;17(3):325–35.

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337(6096):816–21.

Jones GP, Planes S, Thorrold SR. Coral reef fish larvae settle close to home. Curr Biol. 2005;15(14):1314–8.

Kinlan BP, Gaines SD. Propagule dispersal in marine and terrestrial environments: a community perspective. Ecology. 2003;84(8):2007–20.

Kliman R, Sheehy B, Schultz J. Genetic drift and effective population size. Nat Educ. 2008;1(3):3.

Lamichhaney S, Martinez Barrio A, Rafati N, Sundstrom G, Rubin CJ, Gilbert ER, Berglund J, Wetterbom A, Laikre L, Webster MT, Grabherr M, Ryman N, Andersson L. Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proc Natl Acad Sci U S A. 2012;109(47):19345–50.

Lessios HA, Robertson DR. Crossing the impassable: genetic connections in 20 reef fishes across the eastern Pacific barrier. Proc Biol Sci. 2006;273(1598):2201–8.

Lessios HA, Kessing BD, Robertson DR. Massive gene flow across the world's most potent marine biogeographic barrier. Proc R Soc B Biol Sci. 1998;265(1396):583–8.

Lessios HA, Kane J, Robertson DR. Phylogeography of the pantropical sea urchin Tripneustes: contrasting patterns of population structure between oceans. Evolution. 2003;57(9):2026–36.

Levinton J, Koehn R. Population genetics of mussels. In: Marine mussels, their ecology and physiology. Cambridge: Cambridge University Press; 1976. p. 357–84.

Lotrich VA. Summer home range and movements of *Fundulus heteroclitus* (Pisces: Cyprinodontidae) in tidal creek. Ecology. 1975;56:191–8.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet. 2003;4(12):981–94.

Marsh AG, Pasqualone AA. DNA methylation and temperature stress in an Antarctic polychaete, Spiophanes tcherniai. Front Physiol. 2014;5:173.

Mcmillan WO, Palumbi SR. Concordant evolutionary patterns among Indo-West Pacific butterflyfishes. Proc R Soc Lond B Biol Sci. 1995;260(1358):229–36.

Medlin L, Lange M, Baumann M. Genetic differentiation among three colony-forming species of Phaeocystis: further evidence for the phylogeny of the Prymnesiophyta. Phycologia. 1994;33(3):199–212.

Messer PW, Ellner SP, Hairston NG. Can population genetics adapt to rapid evolution? Trends Genet. 2016;32(7):408–18.

Miller AD, van Rooyen A, Rasic G, Ierodiaconou DA, Gorfine HK, Day R, Wong C, Hoffmann AA, Weeks AR. Contrasting patterns of population connectivity between regions in a commercially important mollusc Haliotis rubra: integrating population genetics, genomics and marine LiDAR data. Mol Ecol. 2016;25(16):3845–64.

Mirbahai L, Yin G, Bignell JP, Li N, Williams TD, Chipman JK. DNA methylation in liver tumorigenesis in fish from the environment. Epigenetics. 2011;6(11):1319–33.

Montes I, Conklin D, Albaina A, Creer S, Carvalho GR, Santos M, Estonba A. SNP discovery in European anchovy (Engraulis encrasicolus, L) by high-throughput transcriptome and genome sequencing. PLoS One. 2013;8(8):e70051.

Mora C, Tittensor DP, Adl S, Simpson AG, Worm B. How many species are there on earth and in the ocean? PLoS Biol. 2011;9(8):e1001127.

Newell RIE. Species profiles: life histories and environmental requirements of coastal fishes and invertebrates (North and Mid-Atlantic) – blue mussel. US Fish and Wildlife Service Biology Report 82(11. 102), US Army Corps of Engineers, TR El-82-4; 1989. p. 25.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461(7261):272–6.

Nielsen EE, Hemmer-Hansen J, Poulsen NA, Loeschcke V, Moen T, Johansen T, Mittelholzer C, Taranger GL, Ogden R, Carvalho GR. Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (Gadus morhua). BMC Evol Biol. 2009a;9(1):276.

Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D. Population genomics of marine fishes: identifying adaptive variation in space and time. Mol Ecol. 2009b;18(15):3128–50.

Nielsen EE, Cariani A, Mac Aoidh E, Maes GE, Milano I, Ogden R, Taylor M, Hemmer-Hansen J, Babbucci M, Bargelloni L, Bekkevold D, Diopere E, Grenfell L, Helyar S, Limborg MT, Martinsohn JT, McEwing R, Panitz F, Patarnello T, Tinti F, Van Houdt JK, Volckaert FA, Waples RS, FishPopTrace Consortium, Albin JE, Vieites Baptista JM, Barmintsev V, Bautista JM, Bendixen C, Berge JP, Blohm D, Cardazzo B, Diez A, Espineira M, Geffen AJ, Gonzalez E, Gonzalez-Lavin N, Guarniero I, Jerame M, Kochzius M, Krey G, Mouchel O, Negrisolo E, Piccinetti C, Puyet A, Rastorguev S, Smith JP, Trentini M, Verrez-Bagnis V, Volkov A, Zanzi A, Carvalho GR. Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. Nat Commun. 2012;3:851.

Nishida M, Lucas JS. Genetic differences between geographic populations of the Crown-of-thorns starfish throughout the Pacific region. Mar Biol. 1988;98(3):359–68.
Nunez JC, Seale TP, Fraser MA, Burton TL, Fortson TN, Hoover D, Travis J, Oleksiak MF, Crawford DL. Population genomics of the euryhaline teleost Poecilia latipinna. PLoS One. 2015;10(9):e0137077.
O'Connor MI, Bruno JF, Gaines SD, Halpern BS, Lester SE, Kinlan BP, Weiss JM. Temperature control of larval dispersal and the implications for marine ecology, evolution, and conservation. Proc Natl Acad Sci U S A. 2007;104(4):1266–71.
Ogden R. Fisheries forensics: the use of DNA tools for improving compliance, traceability and enforcement in the fishing industry. Fish Fish. 2008;9(4):462–72.
Oleksiak MF. Genomic approaches with natural fish populations. J Fish Biol. 2010;76(5):1067–93.
Oleksiak MF. Marine genomics: insights and challenges. Brief Funct Genomics. 2016;15(5):331–2.
Oleksiak MF, Churchill GA, Crawford DL. Variation in gene expression within and among natural populations. Nat Genet. 2002;32(2):261–6.
Page TJ, Linse K. More evidence of speciation and dispersal across the Antarctic polar front through molecular systematics of Southern Ocean Limatula (Bivalvia: Limidae). Polar Biol. 2002;25(11):818–26.
Palumbi SR. Population genetics, demographic connectivity, and the design of marine reserves. Ecol Appl. 2003;13(1):S146–58.
Patarnello T, Bargelloni L, Varotto V, Battaglia B. Krill evolution and the Antarctic Ocean currents: evidence of vicariant speciation as inferred by molecular data. Mar Biol. 1996;126(4):603–8.
Pearse DE. Saving the spandrels? Adaptive genomic variation in conservation and fisheries management. J Fish Biol. 2016;89(6):2697–716.
Pierce VA, Crawford DL. Phylogenetic analysis of glycolytic enzyme expression. Science. 1997;276(5310):256–9.
Pierron F, Baillon L, Sow M, Gotreau S, Gonzalez P. Effect of low-dose cadmium exposure on DNA methylation in the endangered European eel. Environ Sci Technol. 2014;48(1):797–803.
Place AR, Powers DA. Genetic bases for protein polymorphism in Fundulus heteroclitus (L.). I. Lactate dehydrogenase (Ldh-B), malate dehydrogenase (Mdh-A), glucosephosphate isomerase (Gpi-B), and phosphoglucomutase (Pgm-A). Biochem Genet. 1978;16(5–6):577–91.
Place AR, Powers DA. Genetic variation and relative catalytic efficiencies: lactate dehydrogenase B allozymes of *Fundulus heteroclitus*. Proc Natl Acad Sci U S A. 1979;76(5):2354–8.
Place AR, Powers DA. Kinetic characterization of the lactate dehydrogenase (LDH-B4) allozymes of *Fundulus heteroclitus*. J Biol Chem. 1984;259(2):1309–18.
Podrabsky JE, Javillonar C, Hand SC, Crawford DL. Intraspecific variation in aerobic metabolism and glycolytic enzyme expression in heart ventricles. Am J Phys Regul Integr Comp Phys. 2000;279(6):R2344–8.
Powers DA, Place AR. Biochemical genetics of Fundulus heteroclitus (L.). I. Temporal and spatial variation in gene frequencies of Ldh-B, Mdh-A, Gpi-B, and Pgm-A. Biochem Genet. 1978;16(5–6):593–607.
Powers DA, Lauerman T, Crawford D, DiMichele L. Genetic mechanisms for adapting to a changing environment. Annu Rev Genet. 1991;25:629–59.
Prado FD, Vera M, Hermida M, Bouza C, Pardo BG, Vilas R, Blanco A, Fernández C, Maroso F, Maes GE, Turan C, Volckaert FAM, Taggart JB, Carr A, Ogden R, Nielsen EE, The Aquatrace Consortium, Martínez P. Parallel evolution and adaptation to environmental factors in a marine flatfish: implications for fisheries and aquaculture management of the turbot (Scophthalmus maximus). Evol Appl. 2018. https://doi.org/10.1111/eva.12628.
Pujolar JM, Jacobsen MW, Als TD, Frydenberg J, Munch K, Jonsson B, Jian JB, Cheng L, Maes GE, Bernatchez L, Hansen MM. Genome-wide single-generation signatures of local selection in the panmictic European eel. Mol Ecol. 2014;23(10):2514–28.
Putnam HM, Davidson JM, Gates RD. Ocean acidification influences host DNA methylation and phenotypic plasticity in environmentally susceptible corals. Evol Appl. 2016;9(9):1165–78.

Quinn TP, Kinnison MT, Unwin MJ. Evolution of Chinook salmon (Oncorhynchus tshawytscha) populations in New Zealand: pattern, rate, and process. Genetica. 2001;112–113:493–513.
Reid NM, Proestou DA, Clark BW, Warren WC, Colbourne JK, Shaw JR, Karchner SI, Hahn ME, Nacci D, Oleksiak MF, Crawford DL, Whitehead A. The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. Science. 2016;354(6317):1305–8.
Reid NM, Jackson CE, Gilbert D, Minx P, Montague MJ, Hampton TH, Helfrich LW, King BL, Nacci DE, Aluru N, Karchner SI, Colbourne JK, Hahn ME, Shaw JR, Oleksiak MF, Crawford DL, Warren WC, Whitehead A. The landscape of extreme genomic variation in the highly adaptable Atlantic killifish. Genome Biol Evol. 2017. https://doi.org/10.1093/gbe/evx023.
Reitzel A, Herrera S, Layden M, Martindale M, Shank T. Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. Mol Ecol. 2013;22(11):2953–70.
Ricklefs RE. Ecology. Asheville: Chiron Press; 1979.
Riginos C, Nachman MW. Population subdivision in marine environments: the contributions of biogeography, geographical distance and discontinuous habitat to genetic differentiation in a blennioid fish, Axoclinus nigricaudus. Mol Ecol. 2001;10(6):1439–53.
Rockman MV. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution. 2012;66(1):1–17.
Rosenblatt RH, Waples RS. A genetic comparison of allopatric populations of shore fish species from the eastern and Central Pacific Ocean: dispersal or vicariance? Copeia. 1986;1986(2):275–84.
Rudman SM, Rodriguez-Cabal MA, Stier A, Sato T, Heavyside J, El-Sabaawi RW, Crutsinger GM. Adaptive genetic variation mediates bottom-up and top-down control in an aquatic ecosystem. Proc R Soc B Biol Sci. 2015;282(1812):20151234.
Sandoval-Castillo J, Robinson NA, Hart AM, Strain LWS, Beheregaray LB. Seascape genomics reveals adaptive divergence in a connected and commercially important mollusc, the greenlip abalone (Haliotis laevigata), along a longitudinal environmental gradient. Mol Ecol. 2018;27(7):1603–20.
Saunders NC, Kessler LG, Avise JC. Genetic variation and geographic differentiation in mitochondrial DNA of the horseshoe crab, Limulus polyphemus. Genetics. 1986;112(3):613–27.
Shaw PW, Arkhipkin AI, Al-Khairulla H. Genetic structuring of Patagonian toothfish populations in the Southwest Atlantic Ocean: the effect of the Antarctic polar front and deep-water troughs as barriers to genetic exchange. Mol Ecol. 2004;13(11):3293–303.
Shulzitski K, Sponaugle S, Hauff M, Walter KD, Cowen RK. Encounter with mesoscale eddies enhances survival to settlement in larval coral reef fishes. Proc Natl Acad Sci U S A. 2016;113(25):6928–33.
Sorte CJ, Williams SL, Carlton JT. Marine range shifts and species introductions: comparative spread rates and community impacts. Glob Ecol Biogeogr. 2010;19(3):303–16.
Spalding MD, Agostini VN, Rice J, Grant SM. Pelagic provinces of the world: a biogeographic classification of the world's surface pelagic waters. Ocean Coast Manag. 2012;60:19–30.
Stillman JH, Armstrong E. Genomics are transforming our understanding of responses to climate change. Bioscience. 2015;65(3):237–46.
Strand AE, Williams LM, Oleksiak MF, Sotka EE. Can diversifying selection be distinguished from history in geographic clines? A population genomic study of killifish (Fundulus heteroclitus). PLoS One. 2012;7(9):e45138.
Sunday JM, Calosi P, Dupont S, Munday PL, Stillman JH, Reusch TB. Evolution in an acidifying ocean. Trends Ecol Evol. 2014;29(2):117–25.
Swearer SE, Shima JS, Hellberg ME, Thorrold SR, Jones GP, Robertson DR, Morgan SG, Selkoe KA, Ruiz GM, Warner RR. Evidence of self-recruitment in demersal marine populations. Bull Mar Sci. 2002;70(1):251–71.
Therkildsen NO, Palumbi SR. Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Mol Ecol Resour. 2016;17(2):194–208.

Thornhill DJ, Mahon AR, Norenburg JL, Halanych KM. Open-ocean barriers to dispersal: a test case with the Antarctic polar front and the ribbon worm Parborlasia corrugatus (Nemertea: Lineidae). Mol Ecol. 2008;17(23):5104–17.

Trexler JC, Travis J, Dinep A. Variation among populations of the sailfin molly in the rate of concurrent multiple paternity and its implications for mating-system evolution. Behav Ecol Sociobiol. 1997;40(5):297–305.

Valenzuela-Quinonez F. How fisheries management can benefit from genomics? Brief Funct Genomics. 2016;15(5):352–7.

Van Wyngaarden M, Snelgrove PV, DiBacco C, Hamilton LC, Rodríguez-Ezpeleta N, Jeffery NW, Stanley RR, Bradbury IR. Identifying patterns of dispersal, connectivity, and selection in the sea scallop, Placopecten magellanicus, using RAD-seq derived SNPs. Evol Appl. 2016;10(1):102–17.

Varriale A, Bernardi G. DNA methylation and body temperature in fishes. Gene. 2006;385:111–21.

Vasemagi A. The adaptive hypothesis of clinal variation revisited: single-locus clines as a result of spatially restricted gene flow. Genetics. 2006;173(4):2411–4.

Villacorta-Rath C, Ilyushkina I, Strugnell JM, Green BS, Murphy NP, Doyle SR, Hall NE, Robinson AJ, Bell JJ. Outlier SNPs enable food traceability of the southern rock lobster, Jasus edwardsii. Mar Biol. 2016;163(11):223.

Vollmer SV, Palumbi SR. Restricted gene flow in the Caribbean staghorn coral Acropora cervicornis: implications for the recovery of endangered reefs. J Hered. 2007;98(1):40–50.

Waples RS. Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. J Hered. 1998;89(5):438–50.

Waples RS, Gaggiotti O. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol Ecol. 2006;15(6):1419–39.

Ward RD, Woodwark M, Skibinski DOF. A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes. J Fish Biol. 1994;44(2):213–32.

Wethey DS, Woodin SA, Hilbish TJ, Jones SJ, Lima FP, Brannock PM. Response of intertidal populations to climate: effects of extreme events versus long term change. J Exp Mar Biol Ecol. 2011;400(1):132–44.

Willette DA, Allendorf FW, Barber PH, Barshis DJ, Carpenter KE, Crandall ED, Cresko WA, Fernandez-Silva I, Matz MV, Meyer E, Santos MD, Seeb LW, Seeb JE. So, you want to use next-generation sequencing in marine systems? Insight from the Pan-Pacific Advanced Studies Institute. Bull Mar Sci. 2014;90(1):79–122.

Williams S, Benzie J. Evidence of a biogeographic break between populations of a high dispersal starfish: congruent regions within the Indo-West Pacific defined by color morphs, mtDNA, and allozyme data. Evolution. 1998;52(1):87–99.

Wood CW, Brodie ED 3rd. Evolutionary response when selection and genetic variation covary across environments. Ecol Lett. 2016;19(10):1189–200.

Wright S. The genetical structure of populations. Ann Eugenics. 1949;15(1):323–54.

Zamer WE, Hoffmann RJ. Allozymes of glucose-6-phosphate isomerase differentially modulate pentose-shunt metabolism in the sea anemone Metridium senile. Proc Natl Acad Sci U S A. 1989;86(8):2737–41.

Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science. 2010;328(5980):916.

Population Genomics of Speciation and Admixture

Nicola J. Nadeau and Takeshi Kawakami

Abstract The application of population genomics to the understanding of speciation has led to the emerging field of speciation genomics. This has brought new insight into how divergence builds up within the genome during speciation and is also revealing the extent to which species can continue to exchange genetic material despite reproductive barriers. It is also providing powerful new approaches for linking genotype to phenotype in admixed populations. In this chapter, we give an overview of some of the methods that have been used and some of the novel insights gained. We also outline some of the pitfalls of the most commonly used methods and possible problems with interpretation of the results.

Keywords Admixture · Divergence · Hybrid zone · Hybridisation · Introgression · Population genomics · Speciation

1 Introduction

Speciation is a fundamental process in evolution, giving rise to biological diversity (Box 1). It involves the divergence of populations, with the establishment of reproductive isolation (RI) being an essential feature for maintaining distinctive characteristics of the incipient species (Coyne and Orr 2004). The emerging field of speciation genomics makes use of dense genome-wide markers to understand how genetic differences build up within the genome and to identify genetic loci that contribute to speciation (Butlin 2008; Nosil and Feder 2012; Seehausen et al. 2014).

N. J. Nadeau (✉)
Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK
e-mail: n.nadeau@sheffield.ac.uk

T. Kawakami
Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK

Department of Evolutionary Biology, Evolutionary Biology Centre (EBC), Uppsala University, Uppsala, Sweden

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_24,

Gene flow between diverging populations slows down genome divergence by homogenising genetic variation. Establishment of barrier loci involved in RI is also difficult in the face of gene flow because such loci may be quickly eliminated by selection, and therefore, key questions in speciation research are how differences accumulate and how RI mechanisms are established (Coyne and Orr 2004). Speciation genomics studies have shown that divergence can persist in the face of gene flow, with sharing of alleles being detected to a greater or lesser extent between a wide range of taxa, including those that are considered good species (Payseur and Rieseberg 2016). By studying species at different points along the 'speciation' or 'divergence continuum', from those that have diverged little to species that may not even be sister to one another, we are starting to understand how genetic differentiation has accumulated within the genome (Feulner et al. 2015; Nadeau et al. 2013; Riesch et al. 2017; Seehausen et al. 2014). However, species may currently maintain distinctive features despite some ongoing gene flow, but this does not mean that these differences were accumulated initially in the face of gene flow (i.e. in sympatry, Box 1). Hybridisation can happen in different spatial contexts, from narrow hybrid zones where parapatric populations meet, to complete sympatry (Abbott et al. 2013). Similarly, there can be a diversity of different temporal contexts, ranging from brief periods of secondary contact to continuous contact with divergent selection. Distinguishing these different scenarios from genomic data is not straightforward (Payseur and Rieseberg 2016). Therefore, empirical studies need to be interpreted in the light of a sound theoretical understanding of how differences accumulate in the genome under different scenarios (Nosil and Feder 2012; Payseur and Rieseberg 2016).

Box 1 Definitions and Concepts

Speciation – the separation of populations, originally able to interbreed, into distinct species, no longer able to interbreed freely. This definition depends on the species definition being used. The strict biological species concept demands complete reproductive isolation between species, while other definitions may relax this (Coyne and Orr 2004). Speciation genomics studies often consider taxa below the species level, with the idea being that these may be in the early stages of speciation (Seehausen et al. 2014) (see 'The speciation continuum').

Sympatry – occurring together or with overlapping geographic areas. Sympatric speciation occurs without any physical barriers to gene flow.

Allopatry – occurring in separate, nonoverlapping geographic areas. Allopatric speciation occurs when populations are physically isolated and so unable to exchange genetic material.

Parapatry – occurring in partially overlapping geographic areas or areas with a partial barrier between them.

(continued)

Box 1 (continued)

Reproductive isolation (RI) – a reduction or absence of gene flow between populations beyond that caused by geographic barriers, usually due to incompatibilities in the reproductive systems of the organisms, either before fertilisation (prezygotic, e.g. timing of reproduction, courtship, mate choice or physical incompatibilities) or after fertilisation (post-zygotic, e.g. inviable or infertile offspring, offspring with reduced fitness).

The speciation continuum – the idea that speciation proceeds gradually and so it should be possible to observe populations with different levels of divergence that are at different points long the continuum. By studying these populations, we can understand how speciation proceeds. One possible problem with this paradigm is that some of the populations with low levels of divergence may be at a stable point and not in fact proceeding towards becoming full species.

Genome scan – an analysis of genome-wide genetic markers to detect loci with elevated genetic differentiation. In this chapter we are mostly referring to scans of F_{ST} between two populations in order to detect loci that are under divergent selection or exhibit reduced gene flow between populations.

Speciation/differentiation islands – regions of the genome showing increased levels of differentiation between two populations. These are usually inferred to contain genetic loci responsible for maintaining differences between the populations.

Admixture – mixing of genetically distinct populations through interbreeding.

Hybridisation – mating between individuals of different species or distinct populations.

Introgression – the transfer of genetic loci from one species to another following hybridisation and repeated backcrossing.

Gene flow – the movement of genetic material between populations, usually by migration and interbreeding.

Hybrid zone – a restricted geographic region where phenotypically or genetically distinct populations or species meet and interbreed, forming hybrids.

Cline – a spatial transition from one genotypic or phenotypic form to another, or a change in allele frequency across a geographical region.

F_{ST} (also known as Wright's fixation index) – a measure of genetic differentiation between populations varying between zero (no difference) and one (a fixed genetic difference). It involves comparing how similar two individuals from the same subpopulation are as compared to the total population, so giving a measure of the amount of genetic variance that can be explained by population structure. The formula normally used for DNA sequence data is: $F_{ST} = \frac{\pi_{Between} - \pi_{Within}}{\pi_{Between}}$ where $\pi_{Between}$ and π_{Within} are the pairwise genetic differences between individuals sampled from within a (sub)population (π_{Within}) or from different populations ($\pi_{Between}$).

(continued)

Box 1 (continued)

Hard sweep – a selective sweep by positive selection acting on a new mutation. This results in advantageous variants reaching fixation in a population. Genetic variation at sites that are tightly linked to the selected sites is eliminated by genetic hitchhiking.

Soft sweep – selection acting on variants that segregate in a population as standing genetic variation. These variants may not confer a selective advantage in one population or under one set of conditions but do so in another population under different conditions. Because the selected variants are present in a variety of different genetic backgrounds, variation at linked sites is not reduced to the same extent as in a hard sweep.

Linkage disequilibrium (LD) – the non-random association of alleles at different loci within a population. This is most often due to physical linkage between loci but can also be found between unlinked loci. For example, unlinked loci under divergent selection between two populations will tend to be in LD. LD is also elevated in admixed populations because of associations between loci coming from the same parental population.

Barrier loci – positions in the genome that contribute to restriction of gene flow between diverging populations. These loci may be involved in various types of reproductive isolation, including divergent ecological selection (extrinsic reproductive isolation), mate choice (pre-mating reproductive isolation), egg-sperm incompatibility (post-mating-prezygotic reproductive isolation) and hybrid sterility/inviability (post-zygotic reproductive isolation).

Studies of speciation have long made use of hybrid zones (Box 1), where distinct populations or species come into contact and interbreed (Kawakami and Butlin 2012). When high-resolution genomic tools were not available, studying phenotypic variation and few loci within and across hybrid zones provided useful insight into the nature of barriers to gene exchange and the selective forces at play in keeping distinct populations from fully mixing (Barton and Hewitt 1985). Building on this solid foundation, population genomic analyses of hybrid zones can bring new insights at a much finer scale, for example, determining the extent and nature of barriers to gene flow by characterising how much of the genome is being exchanged (Gompert et al. 2017; Harrison and Larson 2016). In this chapter, we explore the new insights that population genomics approaches are bringing to the field of speciation research, as well as how population genomics of admixed populations and hybrid zones can help to identify the genetic basis of phenotypic differences more broadly. Key systems in the speciation genomics literature are summarised in Table 1.

Table 1 Key systems in population genomic studies of speciation and admixture

System	Key studies
Humans *Homo*	Admixture mapping within modern human populations (Shriver et al. 2003). Analysis of admixture with ancient *Homo* species (Patterson et al. 2012; Sankararaman et al. 2012)
Bears *Ursus*	Analysis of admixture and divergent selection between brown and polar bears (Liu et al. 2014b)
Rabbits *Oryctolagus*	Divergence scan and admixture analysis between parapatric species (Carneiro et al. 2014)
House mice *Mus*	Cline analyses of a hybrid zone between species (Gompert and Buerkle 2009; Janoušek et al. 2012; Teeter et al. 2008). Modelling of population history and gene flow (Duvaux et al. 2011)
Flycatchers *Ficedula*	Divergence scan of parapatric species (Ellegren et al. 2012). Effect of linked selection and recombination rate on divergence patterns (Burri et al. 2015)
Crows *Corvus*	Divergence scan of parapatric species and identification of genes controlling colour differences (Poelstra et al. 2014)
Sparrows *Passer*	Divergence scan and admixture analyses of parapatric species (Elgvin et al. 2017)
Darwin's finches *Geospiza*	Within and between species divergence scans, admixture analyses and identification of a genes controlling beak shape (Han et al. 2017; Lamichhaney et al. 2015, 2017)
Great tits *Parus*	Divergence scans between populations, leading to identification of a divergently selected beak shape gene (Bosse et al. 2017)
Warblers *Vermivora*	Divergence scans and admixture modelling in hybridising species leading to identification of genes controlling colour differences (Toews et al. 2016)
Sticklebacks *Gasterosteus*	Divergence scans in parallel populations and at different levels of divergence, identifying repeated selection of the same alleles (Feulner et al. 2015; Jones et al. 2012b). Identification of genes controlling phenotypic differences (Chan et al. 2010; Colosimo et al. 2005)
Whitefish *Coregonus*	Divergence scans and admixture analyses between parallel population pairs (Gagnaire et al. 2013; Renaut et al. 2012; Rogers et al. 2001)
Reef fish *Hypoplectrus*	Divergence scans within and between species (Picq et al. 2016; Puebla et al. 2014)
Fruit flies *Drosophila*	Divergence scans at multiple levels of divergence (Begun et al. 2007; McGaugh and Noor 2012). Cline analysis within species, identifying genes controlling phenotypic differences (McKechnie et al. 2010; Turner et al. 2008). Admixture analyses (Pool et al. 2012)
Neotropical butterflies *Heliconius*	Divergence scans and association mapping across hybrid zones (Nadeau et al. 2014). Divergence scans and admixture analyses at multiple levels of divergence (Martin et al. 2013; Nadeau et al. 2013)
Stick insects *Timema*	Divergence scans between parallel population pairs and allele frequency changes in relocation experiments (Soria-Carrasco et al. 2014). Divergence scans at multiple levels of divergence (Riesch et al. 2017). Identification of loci controlling colour variation (Comeault et al. 2015)

(continued)

Table 1 (continued)

System	Key studies
Fruit fly *Rhagoletis*	Divergence scans, cline analysis and experimental evolution indicating many divergent loci between host races (Egan et al. 2015; Michel et al. 2010). Divergence scans at different levels of divergence (Powell et al. 2013)
Mosquitos *Anopheles*	Divergence scans, identifying inversions between forms (Turner et al. 2005). Divergence scans between species (Caputo et al. 2016). Divergence scans and admixture analyses at multiple levels of divergence (Crawford et al. 2015)
Periwinkles *Littorina*	Divergence scans between multiple parallel ecotype pairs, showing low levels of shared divergence outliers (Ravinet et al. 2016)
Poplar trees *Populus*	Admixture mapping of quantitative trait differences between hybridising species (Lindtke et al. 2013). Admixture analysis of several parallel hybrid zones to identify RI loci (Lindtke et al. 2012)
Monkey-flowers *Mimulus*	Cline analysis across an ecotype hybrid zone (Stankowski et al. 2017). Studies of divergence and admixture between species (Brandvain et al. 2014; Vallejo-Marín et al. 2015)
Sunflowers *Helianthus*	Divergence scans at multiple levels of divergence (Andrew and Rieseberg 2013). Genomic cline analysis across hybrid zones (Gompert and Buerkle 2009)

2 Genomic Signatures of Speciation and Reproductive Isolation

High-throughput sequencing technologies allow biologists to investigate genome-wide patterns of genetic differentiation between diverging populations (Seehausen et al. 2014; Wolf and Ellegren 2016). Speciation can be driven by extrinsic (environmental) factors where divergent selection reduces gene flow between populations or by intrinsic factors where genes incompatible in a foreign genomic background result in reduced fitness in hybrids (Coyne and Orr 2004). The former is known as ecological speciation, where locally adapted populations are exposed to divergent ecological selection in different environments, leading to the establishment of barriers to gene flow (Nosil 2012). The latter cases are formulated by theoretical models where epistatic interactions of incompatible alleles at multiple loci have evolved in diverging populations, resulting in a reduction in hybrid fitness (Dobzhansky-Muller incompatibility) (Dobzhansky 1936; Muller 1940; Orr 1995). However, this binary classification does not fit all situations; for example, local adaptation may be mediated by epistatically interacting alleles that could also give rise to reduced hybrid fitness. Regardless of the types of reproductive barriers, the genic model of speciation predicts that genetic differentiation is initially accumulated at a small number of genomic regions that are under selection associated with RI. These barrier loci are resistant to gene flow, either by ecological divergent selection or intrinsic incompatibility, while the rest of the genome is homogenised by gene flow (Wu 2001).

Barrier loci could be established in the face of gene flow or in geographically isolated populations. In the case of gene flow, the level of genetic differentiation is kept low in regions unlinked to the barrier loci, whereas under geographic isolation, genetic differentiation accumulated during the allopatric period may be eroded by gene flow following secondary contact occurring across the genome except in regions containing barrier loci. In both cases, at the very onset of speciation, the genomes of two diverging populations may be characterised as a small number of regions with elevated differentiation surrounded by regions of low differentiation (hence often referred to as 'differentiation islands' or 'genomic islands of divergence' as an oceanic island metaphor) (Nosil and Feder 2012; Turner et al. 2005). In addition, it has been proposed that restricted gene flow near differentiation islands can allow for the sequential accumulation of additional barrier loci at neighbouring sites, and as a result, these differentiation islands can increase in height and width as speciation proceeds (Via 2012). As additional barrier loci accumulate in a genome, either at proximal or distal regions of existing differentiation islands, the strength of RI increases and genetic differentiation would increase across the whole genome (Fig. 1). There is a suggestion from both theory (Feder et al. 2012; Flaxman et al. 2014) and empirical evidence (Riesch et al. 2017) that this increase does not occur linearly and that there may be a 'tipping point' in either the strength of RI or the number of differentiated regions, at which point populations transition from having a small number of differentiation islands to effectively genome-wide differentiation (Nosil et al. 2017). Nevertheless, the idea of differentiation islands has motivated a number of researchers to characterise genome-wide patterns of genetic differentiation between closely related species and between diverging lineages, aiming to characterise underlying genetic mechanisms of RI.

2.1 Genome Divergence Scans to Identify Barrier Loci

There are an increasing number of studies reporting heterogeneous patterns of genomic differentiation (Ellegren et al. 2012; Nadeau et al. 2012, 2014; Renaut et al. 2013; Turner et al. 2005; Via et al. 2012), but interpretation of these differentiation islands is not as straightforward as one might think based on the genic model of speciation. Specifically, it remains challenging to determine whether the differentiation islands evolved as a result of speciation (i.e. 'speciation islands') or by other processes independent of the evolution of RI mechanisms (i.e. 'incidental islands') (Cruickshank and Hahn 2014). Under the genic model of speciation, gene flow plays a critical role in the formation of differentiation islands by homogenising genetic diversity between species at the vast majority of genomic regions that do not harbour loci involved in RI. However, there are several studies reporting similar patterns of heterogeneous differentiation between geographically isolated populations, which have no apparent contemporary gene flow between them (Martin et al. 2013; Renaut et al. 2013; Vijay et al. 2016). Incomplete lineage sorting of ancestral polymorphisms and stochasticity in allele frequency changes can result

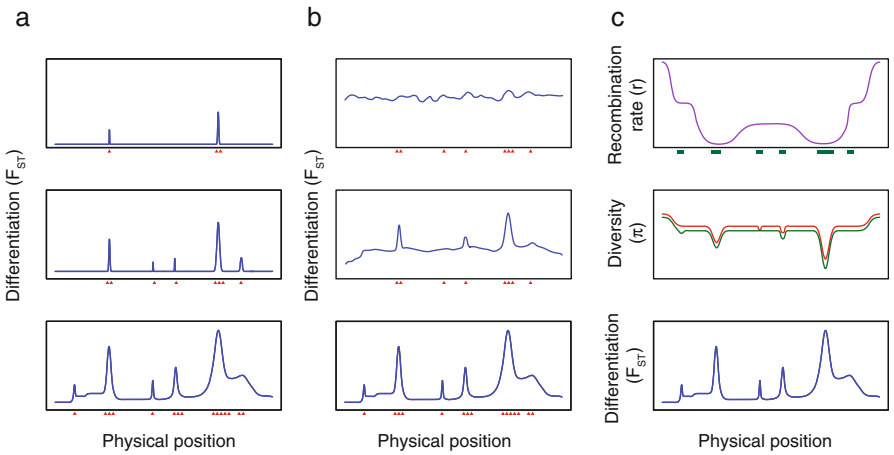

Fig. 1 Three models for the evolution of differentiation islands. (**a**) Speciation island model without geographic isolation. Gene flow maintains low genetic differentiation throughout the speciation continuum by homogenising genetic materials elsewhere in genomes of diverging populations except loci involved in divergent selection and reproductive isolation (red triangles). As speciation progresses (from the top to the bottom panels), new reproductive isolation loci are accumulated either at proximal region to the existing differentiation islands or at novel regions. This makes the differentiation islands higher and wider. (**b**) Speciation island model with geographic isolation. Genetic differentiation is initially accumulated in geographically isolated populations (top panel). Reproductive isolation loci are also accumulated during this period. Upon secondary contact (middle panel), gene flow erodes genetic differentiation elsewhere in a genome except reproductive isolation loci. Additional reproductive isolation loci may be accumulated, further strengthening the barrier to gene flow. (**c**) Incidental island model. Because of the heterogeneity in recombination rate and gene density (purple line and green rectangles, respectively, in the top panel), shared ancestral polymorphisms between diverging populations are removed more extensively at regions of low recombination rate and high gene density. This results in non-uniform reduction of nucleotide diversity (π) in each population (middle panel) and heterogeneous differentiation landscape (bottom panel)

in heterogeneity in genetic variation between closely related species even in the absence of current gene flow, especially when selection acts on linked neighbouring sites (Cruickshank and Hahn 2014; Nachman and Payseur 2012; Noor and Bennett 2009). Selection in this case can be either positive or negative (background/purifying selection) and does not have to be directly associated with RI mechanisms. The rationale is that recurrent positive and negative selection removes polymorphisms not only at target sites of selection but also at neighbouring sites in linkage disequilibrium (LD). This process, collectively referred to as 'linked selection' (Cutter and Payseur 2013), can create regions with locally reduced effective population size (N_e), which in turn accelerates lineage sorting, decreases genetic diversity (π), and increases differentiation (F_{ST}). Because the effect of linked selection is a function of the density of target sites for selection (e.g. gene density) and local recombination rate, the magnitude of lineage sorting and, hence, genetic differentiation is inherently heterogeneous across a genome. Therefore, it is necessary to understand the

underlying genetic mechanisms and evolutionary processes in the formation of differentiation islands.

One way of distinguishing these two scenarios is to compare patterns of genetic differentiation along a genome between multiple pairs of closely related species with different divergence times. An important assumption is that conserved recombination rate and genome structure among closely related species, such as gene density and distribution, result in shared patterns of genetic differentiation by linked selection. There are several studies showing that differentiation islands observed both in very early stages of speciation (i.e. less divergent pairs of species) and more advanced stages of speciation (i.e. more divergent pairs of species) likely represent 'incidental islands', while population-specific differentiation islands represent candidate 'speciation islands' (Andrew and Rieseberg 2013; Burri et al. 2015; Poelstra et al. 2014; Renaut et al. 2013; Vijay et al. 2016). In addition, at the early stage of speciation, 'speciation islands' can be distinguished from heterogeneous genomic differentiation shaped by linked selection unrelated to RI, because strong divergent selection can create a small number of extremely differentiated regions with long haplotype blocks (Andrew and Rieseberg 2013; Poelstra et al. 2014).

Due to the increasing accessibility of genome-wide polymorphism data in various organisms, the genome scan approach is a tractable first step towards the understanding of the genetic basis of reproductive isolation by characterising genetic differentiation along a genome. One advantage of this approach is that phenotypic differences do not need to have been previously characterised, meaning that it has the potential to identify loci underlying novel divergently selected traits. In addition, it can be a powerful tool for detecting divergently selected regions between readily interbreeding taxa, because it makes use of the genomic signatures left by both selection and gene flow. However, deciphering underlying mechanisms for the formation of differentiation islands (i.e. divergent selection related to reproductive isolation vs. linked selection) remains a challenge, not only because these two processes can take place simultaneously but also because these two processes would leave very similar signals (Yeaman et al. 2016). One way forward is to combine trait information with genome scan analysis, by which one can further narrow down the candidate genomic regions from numerous differentiation islands identified by the genome scan. In practice this has rarely been done for traits that were not already well characterised or genetically mapped. A study by Toews et al. (2016) on warblers is one of the few examples to use an outlier approach to identify anonymous outlier loci and to then link these to phenotypic differences between populations (Fig. 2), although, even in this case, the phenotypes were well-characterised differences in colouration. The process of linking anonymous loci to phenotypes necessarily starts with an informed guess, which makes it difficult for the genome scan approach to identify really novel or unexpected divergently selected traits (but see Bosse et al. 2017). Although undoubtedly useful for characterising the patterns of divergence across the genome, genome scan analysis alone may have a limited power to identify causal genes for reproductive isolation. In section 3, we describe approaches that gain additional power from the information present in admixed populations to identify barrier loci and those loci underlying divergent traits more broadly.

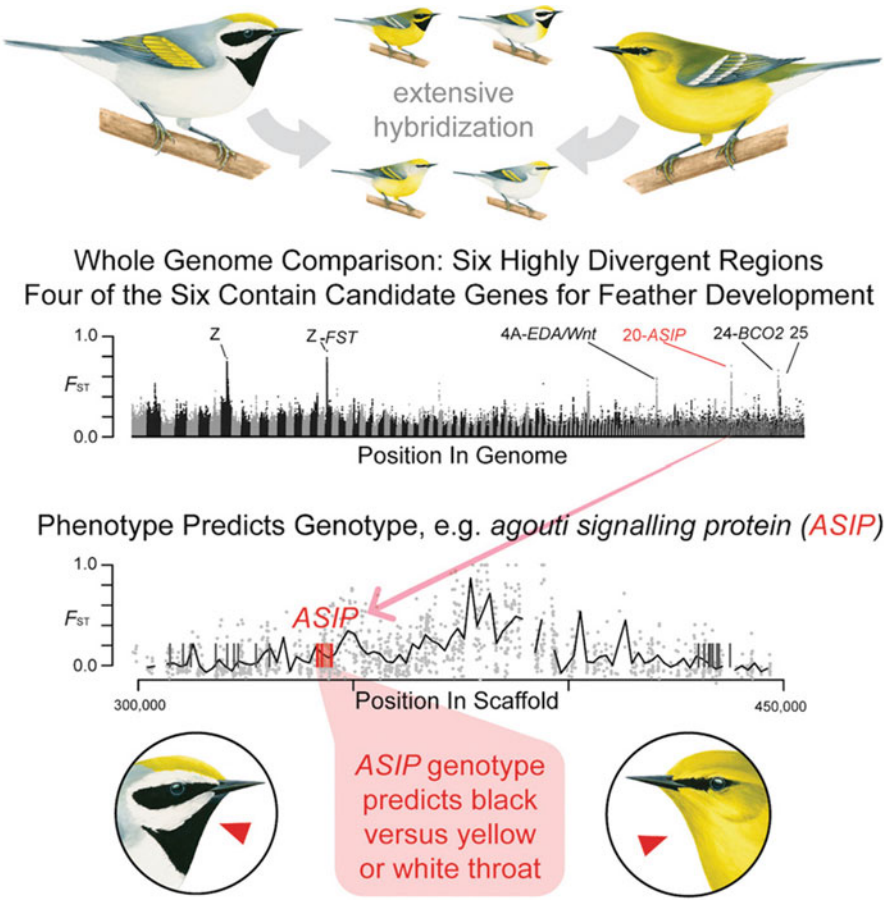

Fig. 2 F_{ST} outlier scan between golden-winged and blue-winged warblers (*Vermivora chrysoptera* and *V. cyanoptera*) identified six divergent regions between species, four of which contained candidate plumage colour genes. Associations between these loci and particular plumage colour elements were then confirmed by characterising particular SNPs in a larger number of individuals, including hybrids. Reprinted from Toews et al. (2016), with permission from Elsevier

2.2 *Key Examples of Applications of Genome Divergence Scans in Speciation Population Genomics*

Here we outline 'genome scans' performed on the three-spine stickleback (*Gasterosteus aculeatus*) to show how genetic differentiation accumulates along a genome at different stages of the 'speciation continuum'. Two avian examples, highlighting some of the issues with interpretation of divergence scans, are also presented. Additional examples are summarised in Table 1 and have been reviewed elsewhere (Haasl and Payseur 2016; Ravinet et al. 2017; Wolf and Ellegren 2016).

2.2.1 Three-Spine Stickleback

The three-spine stickleback provides a powerful model system for studying the genetic basis of adaptation and ecological speciation. This small fish is widely distributed in the Northern hemisphere and shows a remarkable history of independent colonisation from the marine environment to freshwater ecosystems after the glacial retreat (ca. 12,000 years ago) (Bell and Foster 1994). Freshwater and marine ecotypes show marked differences in body size and shape, colouration, courtship behaviour, trophic specialisation, the number of skeletal armour plates, and spine length (Fig. 1a) (McKinnon and Rundle 2002). The repeated observation of these morphological and behavioural shifts at multiple locations in North America and Europe suggests that the selection pressures associated with the colonisation of freshwaters have been instrumental in driving recurrent/parallel evolution. In fact, parallel evolution of freshwater-adapted phenotypes has likely been facilitated through repeated selection of rare genetic variants segregating in the marine ancestor (Colosimo et al. 2005; Jones et al. 2012b; Roesti et al. 2015). After the colonisation of freshwaters, populations have further diversified into several distinctive ecotypes. For example, populations in open water lake habitat show ecologically distinctive life history traits by having pelagic lifestyle feeding on zooplankton ('lake ecotypes' or 'limnetic ecotypes'), whereas populations in rivers and small stream habitat show a benthic lifestyle by feeding on macroinvertebrates ('stream ecotypes' or 'benthic ecotypes') (Berner et al. 2010; Moser et al. 2015). In both cases, increases in allele frequency of adaptive variants in newly colonised habitat may leave a specific signature in their genomes, and genome scan analysis, in theory, can detect such a signature as an elevated differentiation relative to the surrounding genomic regions. Moreover, repeated occurrence of differentiation islands at the same genomic location between multiple, independent pairs of ecotypes is commonly taken as evidence of parallel evolution at the molecular level (Hohenlohe et al. 2010). However, linked selection unrelated to adaptive divergence could also contribute to the parallel evolution of differentiation islands because these ecotypes likely share common genomic features important to the magnitude of linked selection across a genome, such as variation in gene density and recombination rate, which would then result in positive correlation in the magnitude of differentiation between ecotype comparisons.

Several studies have identified key genes associated with phenotypic traits that confer adaptation to the newly colonised habitat in sticklebacks (Chan et al. 2010; Colosimo et al. 2005). For example, higher predation pressure in open-water habitat (either in marine populations or lake populations) than small stream populations, has resulted in more complete lateral armour plates (Bell and Foster 1994; Berner et al. 2010; Roesti et al. 2015). Allelic variation at the *Ectodysplasin* (*Eda*) gene on chromosome 4 is strongly associated with phenotypic variation in this trait (Berner et al. 2014; Colosimo et al. 2005), representing a prime candidate for selection. Another well-studied candidate gene for adaptive evolution is *Pituitary homeobox transcription factor 1* (*Pitx1*) gene, whose regulatory mutations resulted in partial or

complete loss of pelvic spines in freshwater ecotypes (Chan et al. 2010). While *EDA* represents a classic case for adaptation from standing genetic variation, the evolution of *Pitx1* has involved repeated de novo mutations in multiple populations. Therefore, these loci offer an opportunity to test a predicted genomic pattern, in which divergent selective sweeps increase genetic differentiation at these loci, while ongoing gene flow maintains low differentiation at the genomic background.

To identify putative genomic regions under divergent selection, several studies took a population genomics approach to characterise genome-wide patterns of genetic differentiation between marine and freshwater ecotypes and between lake and stream ecotypes by using RAD-sequencing approaches and whole-genome re-sequencing approaches (Deagle et al. 2011; Feulner et al. 2015; Hohenlohe et al. 2010, 2012; Jones et al. 2012a, b; Roesti et al. 2012, 2014, 2015). The first genome-wide survey of genetic differentiation identified nine differentiation islands in three comparisons of ancestral oceanic populations versus derived freshwater populations in Alaska by using over 45,000 RAD sequencing markers (Hohenlohe et al. 2010). Jones et al. (2012b) further expanded sampling populations (21 locations across Northern hemisphere) as well as the number of SNPs covering the entire assembled genome and identified 174 regions of elevated differentiation between marine and freshwater ecotypes with median size of 3 kb with 5% false discovery rate (FDR). Consistent with the a priori expectation, the *EDA* locus showed elevated differentiation in both studies, confirming that genome scan analysis can recover signatures of divergent selection; however, *Ptx1* was not located at differentiation islands despite the apparent phenotypic differences between the ecotypes (Hohenlohe et al. 2012; Jones et al. 2012b). The insignificant genetic difference at *Ptx1* could be interpreted as (1) weak or no selection on *Ptx1* or (2) difficulties in detecting a selective signature by this approach if adaptive causal variants are found in multiple haplotype backgrounds (i.e. 'soft sweeps', see Box 1) (Hohenlohe et al. 2010). In addition to these a priori candidate genes, other differentiation islands contained a number of genes with functions related to skeletal traits, response to osmotic stress, signal response, behavioural interaction between organisms, amine and fatty acid metabolism, cell–cell junctions, WNT developmental signalling, epithelial barrier, and immune functions (Jones et al. 2012a, b), which represent candidate genes for functional analysis. Importantly, there are many other differentiation islands distributed in intergenic regions, implying that adaptive divergence can involve changes in both protein coding genes and non-coding regulatory regions (Jones et al. 2012b).

Lake and stream population pairs also provide a useful system for parallel evolution of differentiation. For instance, after colonisation to Lake Constance in Central Europe, small creeks and streams connected to the lake were subsequently colonised by stickleback populations (Roesti et al. 2015), thus possibly representing much more recent divergence than the marine-freshwater comparison. Despite the short evolutionary time window, Marques et al. (2016) identified 37 differentiation islands that consisted of 1–26 SNPs. Importantly, 19 out of these 37 differentiation islands were consistently identified in two pairs of stream and lake ecotypes, indicating potential parallel change in allele frequency driven by ecological

adaptation. Three other tributaries of Lake Constance also showed heterogeneous genetic differentiation with 2–25 highly differentiated SNPs scattered across the genome in at least one of the three comparisons of lake versus stream populations (Roesti et al. 2015). All three comparisons showed a similar shift in allele frequency at these loci, supporting the parallel action of similar ecological pressure at the genomic level.

An important difference from the marine-freshwater comparison is that signature of selection at the *Eda* locus is much weaker in the lake-stream comparison as represented by the inconsistent elevation of genetic differentiation at this locus among population pairs (Roesti et al. 2015). This may possibly be because of recent recolonisation history in the lake-stream system where the selective sweep is likely incomplete. An additional complication with regard to the genetic differentiation at the *Eda* locus is that adaptive alleles can be unconditionally favoured in both stream and lake populations on Vancouver Island in Canada, which generates a peculiar pattern where genetic differentiation is reduced at *Eda* locus due to the fixation of shared ancestral haplotypes, while the surrounding neutral regions of the *Eda* locus are characterised as elevated differentiation (Roesti et al. 2012, 2014). Unconditional selection on the adaptive alleles at the *Eda* locus, if any, could be due to similar ecological selective pressure in lakes and streams on Vancouver Island, whereas selective pressure may be more contrasting in the Lake Constance system in Central Europe because of its larger size. Altogether, this highlights difficulties and challenges in using genome scans to detect signatures of selective sweeps, even at genomic regions with strong candidate genes under ecological selection.

Since stickleback recolonisation has likely taken place independently at different times, multiple pairs of lake-stream populations can also provide an opportunity to test how differentiation islands emerge and increase in number and size along the speciation continuum (Fig. 1). If an increase in differentiation in the background genomic region is accompanied by increase in the number and size of differentiation islands as predicted by the genic model of speciation, then population pairs with higher genome-wide differentiation should have more and wider differentiation islands than population pairs with lower genome-wide differentiation. Feulner et al. (2015) compared genetic differentiation (measured as F_{ST}) among five pairs of lake-stream populations in the Northern Hemisphere, with varying degrees of genome-wide F_{ST}, ranging from 0.10 to 0.28 (Fig. 3b). They found no apparent growth of differentiation islands despite the significant difference in the background F_{ST}, which may partly be due to population-specific selection for each locality and/or differences in the extent of divergent selection. Similar patterns were also found in *Timema* stick insects (Riesch et al. 2017), although evidence for the growth of differentiation islands has been reported in *Heliconius* butterflies (Nadeau et al. 2013). Theoretical studies have suggested that the differentiation islands could grow in size by accumulating additional RI loci in the presence of gene flow, but their growth may require specific conditions composed of rather narrow parameter space, such as low migration, strong selection, low level of differentiation in background regions, and locally reduced recombination rate (Feder and Nosil 2010; Yeaman et al. 2016). In addition, since a transition from an early stage of

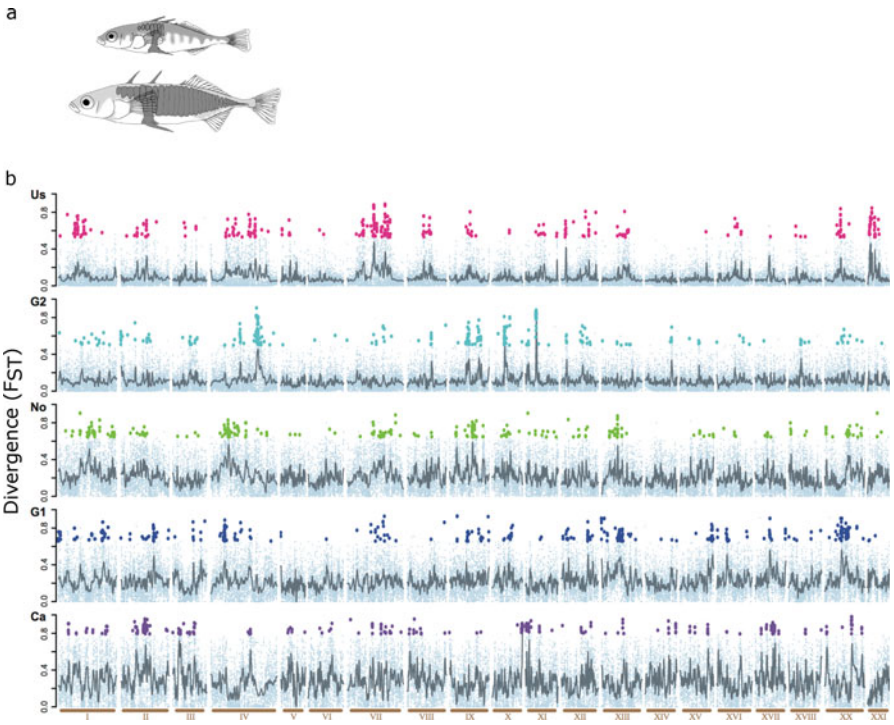

Fig. 3 Divergent phenotypes of three-spine stickleback (*G. aculeatus*) and genome-wide patterns of genetic differentiation between ecotypes. (**a**) Freshwater (top) and marine (bottom) ecotypes. Reprinted by permission from Macmillan Publishers Ltd: Nature (Jones et al. 2012b), copyright 2012 (**b**) Distribution of F_{ST} along a genome in five pairs of stream and lake ecotypes with different levels of genome average F_{ST} (smallest at the top and biggest at the bottom panels). Note that location of loci that are exceptionally different (i.e. elevated F_{ST}, coloured dots) is not always conserved between population pairs, and the number and intensity of these high differentiation regions are not correlated with background level of F_{ST}. Reprinted from Feulner et al. (2015) under the Creative Commons Attribution License

speciation with detectable differentiation islands to an advanced stage with genome-wide differentiation (Fig. 1) may happen rapidly, detecting signals for the growth of differentiation islands may be challenging (Feder and Nosil 2010). Additional empirical studies may refine these theoretical models to predict necessary conditions for broadening the regions of differentiation under various demographic scenarios.

Linked selection also plays a critical role in the formation of heterogeneous differentiation landscape along the stickleback genome by removing genetic variation, particularly at low recombination regions. Like many other species (Auton et al. 2012; Kawakami et al. 2014), recombination rate is highly variable along the stickleback genome with the rate generally increasing towards the ends of chromosomes (Roesti et al. 2013). This 'U-shape' distribution of recombination events along a chromosome is inversely correlated with genetic differentiation at a global

genomic scale (Roesti et al. 2012, 2013), supporting an action of linked selection where lineage sorting takes place much more extensively at low recombination regions by the removal of shared ancestral genetic variation by positive selection or negative (background) selection (Fig. 1). The strong influence of linked selection at low recombination regions is consistent with the pattern reported in a wide variety of species (Burri et al. 2015; Martin et al. 2016; Vijay et al. 2016; Wang et al. 2016). These studies also show that the effect of linked selection is stronger at gene dense regions because the extent of the removal of genetic variation at physically linked sites is proportional to gene density. Given the significant correlation between genetic diversity and recombination rate, it is important to take into account the variation in recombination rate between diverging populations, which can potentially create population-specific patterns of diversity landscape along a genome (Kawakami et al. 2017; Smukowski and Noor 2011).

2.2.2 Flycatchers

Collared flycatcher (*Ficedula albicollis*) and pied flycatcher (*F. hypoleuca*) have been intensively studied to identify divergence in life history traits, traits under sexual selection (e.g. plumage colour and song), and hybrid fitness reduction (Qvarnström et al. 2010). Both species are small migratory passerine birds that overwinter in sub-Saharan Africa but return to their breeding ranges in summer in Europe. Occasional hybridisation has been reported at regions where two species overlap in central Europe (Svedin et al. 2008), but reproductive isolation is near complete despite their relatively recent divergence (<1 million years) (Nadachowska-Brzyska et al. 2013). By sequencing whole-genomes of 20 individuals (10 individuals/species), Ellegren et al. (2012) discovered that the pattern of genetic differentiation was highly heterogeneous along the genome with about 50 regions with elevated differentiation between species (measured by F_{ST}). To further investigate underlying mechanisms for the formation of these 'differentiation islands', Burri et al. (2015) expanded the samples to other flycatcher species. These multi-population and multi-species comparisons revealed similar patterns of genetic differentiation both within species and between species, regardless of their divergence time (Fig. 4), indicating that shared genomic features among these *Ficedula* species are likely responsible for the emergence of differentiation islands. In fact, recombination rate estimated based on the linkage map in collared flycatcher (Kawakami et al. 2014) and the density of coding sequence were significantly correlated with genetic diversity (π) and genetic differentiation (F_{ST} and d_{XY}), suggesting that 'linked selection' plays a more predominant role than gene flow in the formation of differentiation islands in flycatcher.

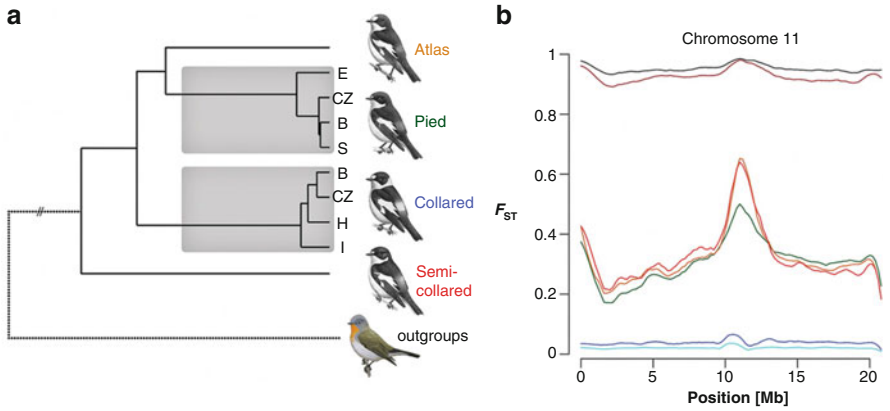

Fig. 4 (**a**) About 20 genomes per population were sequenced (collared flycatcher [*F. albicollis*], pied flycatcher [*F. hypoleuca*], atlas flycatcher [*F. speculigera*], and semicollared flycatcher [*F. semitorquata*]). Outgroup species were red-breasted flycatcher (*F. parva*) and snowy-browed flycatcher (*F. hyperythra*) (not shown). Four populations each of collared flycatcher and pied flycatcher were sampled across Europe (*E* Spain, *CZ* Czech Republic, *B* Baltic, *S* Sweden, *H* Hungary, *I* Italy), which allowed within-species comparisons. (**b**) Genetic difference (F_{ST}) along an example chromosome (chromosome 11). Differentiation islands observed in collared-pied comparison (green) were also observed in collared-atlas comparison (orange), collared-semicollared comparison (red), collared-red-breasted comparison (dark red), and collared-snowy-browed comparison (black). Importantly, the differentiation island starts emerging within species comparisons (I-H collared flycatcher populations [dark blue] and I-B collared flycatcher populations [light blue]). Modified from Burri et al. (2015) with permission

2.2.3 Crows

The *Corvus* crow species complex in Eurasia (*Corvus [corone] corone*, *C. [c]. cornix*, *C. [c]. orientalis* and *C. [c]. pectoralis*) represents another classic example of speciation in birds (Mayr 1942). This species complex has been extensively studied to understand genetic mechanisms of the traits under divergent selection, which are the key in the maintenance of stable hybrid zones (Fig. 5) (Randler 2007). Because RI between carrion crow (*C. [c]. corone*) and hooded crow (*C. [c]. cornix*) is incomplete with frequent backcrossing of hybrids, this pair of taxa may be at an earlier stage of the speciation continuum than the flycatcher species pair. By using the whole genome sequencing approach, Poelstra et al. (2014) identified five 'differentiation islands' based on F_{ST} outlier analysis. The largest differentiation island, identified on chromosome 18, harboured genes associated with colour pigmentation and visual perception, which are likely responsible for differences in plumage colour and assortative mating. In addition, long-range sequencing analysis using PacBio and Nanopore optical mapping revealed that this region coincided with putative centromeric region, suggesting that the combined effect of low recombination and positive selection resulted in the elevated genetic differentiation (Weissensteiner et al. 2017). In addition, Vijay et al. (2016) identified several differentiation islands in the other species pairs (Siberian hybrid zone

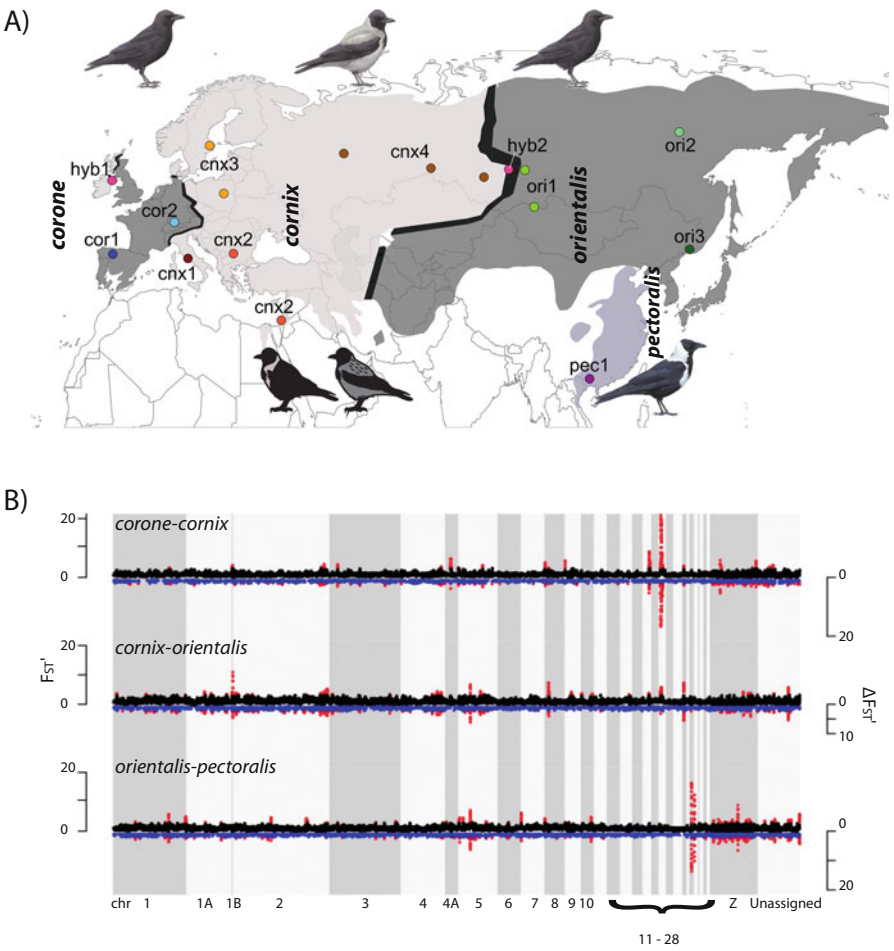

Fig. 5 (**a**) Distribution of *Corvus* crow species complex. (**b**) The *corone-cornix* hybrid zone in central Europe was used in Poelstra et al. (2014), revealing a strong genetic difference on chromosome 18 (top panel). The *cornix-orientalis* comparison (middle panel) and the *orientalis-pectoralis* comparison (bottom panel) showed differentiation islands that are at different genomic regions. Standardised genetic differentiation F_{ST}' (black, positive axis) and net genetic differentiation $\Delta F_{ST}'$ (blue, mirrored to the negative axis) in 50 kb windows across the genome. Genomic regions of extreme differentiation (499th percentile) are shown in red for both F_{ST}' and $\Delta F_{ST}'$. Modified from Vijay et al. (2016) with permission

between *C. [c] cornix* and *C. [c] orientaris* and Asian hybrid zone between *C. [c] orientaris* and *C. [c] pectolaris*) (Fig. 5). Importantly, the locations of these islands were mostly different from the ones identified in the *corone-cornix* hybrid zone, and consequently, genes identified on the differentiation islands hardly overlapped between three species pairs. Nevertheless, these differentiation islands also contained genes involved in pigmentation and melanogenesis, suggesting that

parallel divergent selection acts on plumage colour at multiple independent hybrid zones but on different genes in the same melanogenesis pathways. The pattern found in the crow species complex is quite contrasting to that found in *Heliconius* butterflies, in which the parallel patterns of phenotypic divergence are largely based on selection acting on the same genomic regions (Nadeau et al. 2014).

3 Using Admixture to Identify Genes Underlying Divergent Traits

Hybrid zones or other situations in which phenotypically distinct populations meet and interbreed provide a valuable opportunity to study the genetic basis of divergent traits. It has long been recognised that hybrid zones can act as natural laboratories in which many generations of crossing generate novel genetic combinations and the potential to identify loci contributing to adaptive phenotypic differences (Barton and Hewitt 1985). However, it is only relatively recently, with the advent of population genomics approaches, that this potential has begun to be realised.

3.1 Clines

The rate of change in allele frequency across a cline can be used to infer the strength of divergent selection acting on that locus if the average dispersal distance for the organism is also known (Barton and Hewitt 1985). This approach has been used extensively for single loci or phenotypes but has rarely been used with population genomic data. Nevertheless, it does have the potential to narrow down lists of candidates identified through outlier scan approaches by identifying the loci with the steepest allele frequency changes and with cline centres corresponding to the centre of the hybrid zone or phenotypic transition. Stankowski et al. (2017) applied this approach to a hybrid zone between monkey-flowers (*Mimulus aurantiacus*) with different floral traits and found that just 130 out of the 426 most differentiated loci had clines similar to that of the phenotypic trait (Fig. 6).

It is also possible to use sets of hybrid individuals to infer ‘genomic clines’ that can be independent of geographical clines. The method, developed by Gompert and Buerkle (2009) and Gompert and Alex Buerkle (2010), uses multiple loci to estimate a genomic background level of admixture for each individual and then detects loci that deviate significantly from this neutral background rate across the population. These loci can either show increased rates of introgression, indicating that they are under positive selection and sweeping through both populations (or spreading from one to the other), or reduced introgression, indicating that they are under divergent selection and not spreading between the populations (Fig. 7). The admixture proportions generated by this method can also be useful for inferring the age of the

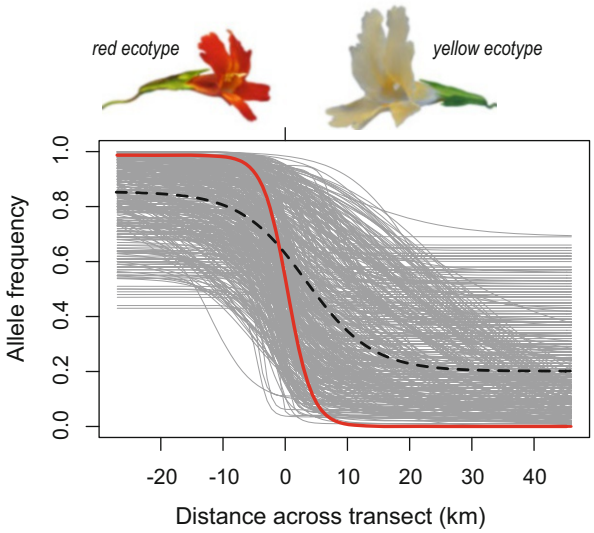

Fig. 6 Geographic clines across a hybrid zone between yellow and red forms of monkey-flowers (*Mimulus aurantiacus*) for 426 highly differentiated loci (top 1% of the F_{CT} distribution). The red line shows the cline at the colour controlling locus, *MaMyb2*. The dashed line is the average cline across all 426 loci. Most markers have cline slopes shallower than those seen at the known divergently selected locus, suggesting that only a subset are under divergent selection, despite all showing high differentiation. Reproduced from Stankowski et al. (2017), with permission

hybrid zone and the strength of barriers to gene flow, by establishing the proportions of early versus late generation hybrids that are present (Gompert et al. 2014; Nadeau 2014). However, unlike the geographic cline approach, where populations may differ at only a small number of loci, the genomic cline approach requires the parental populations to have marked allele frequency differences in order to reconstruct a background genomic cline.

3.2 Admixture Mapping

The most widely used approach for identifying genetic loci underlying a particular trait is to perform controlled laboratory crosses. Offspring from F2 or backcross generations can be genotyped with a relatively small number of parentally informative markers to identify the inheritance of large chromosomal blocks and to characterise where recombination breaks have occurred. This is then used to generate a dense linkage map and identify the genomic location of either Mendelian loci or quantitative trait loci (QTL) (many descriptions of these methods have been published previously, e.g. Liu 1997). This approach has been extensively and successfully used but is limited to taxa that can be reared in captivity and can usually

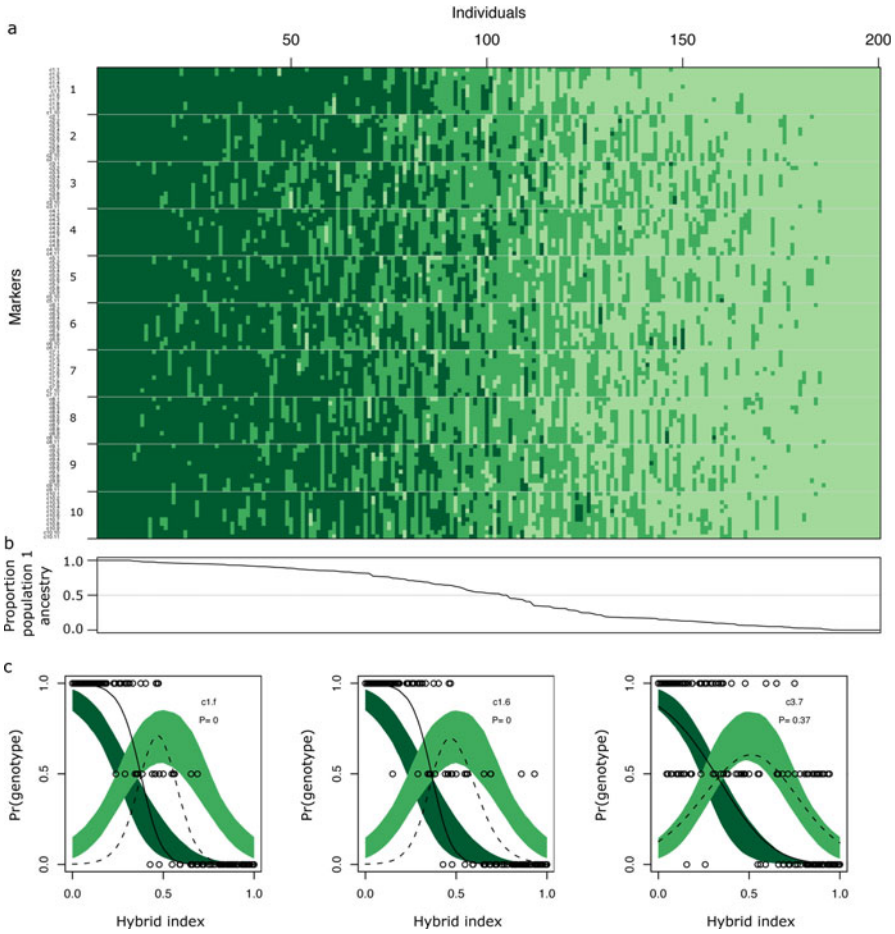

Fig. 7 Genomic clines from a simulated data set. (**a**) Across all loci and individuals. Loci are ordered based on map locations, and individuals are ordered based on their hybrid index (fraction of alleles coming from population 1). Each block in the plot denotes an individual's genotype at that locus (dark green, homozygous population 1; green, heterozygous population 1/population 2; light green, homozygous population 2). (**b**) Hybrid index of each individual. (**c**) Clines at three individual loci (black lines, proportion homozygous population 1; dashed lines, proportion heterozygous) compared to the 95% confidence intervals for the genomic background (dark green and light green). The left plot shows a locus under selection, the middle plot is a locus linked to this, and the right plot is a locus not under selection. The circles show the raw genotype data for each individual. Reproduced from Gompert and Alex Buerkle (2010), with permission

only identify fairly large genomic intervals because of the limited number of recombination events that occur within a limited number of offspring and a few generations.

Admixture mapping has the same underlying rational as linkage mapping using crosses, but instead uses naturally admixed populations (Winkler et al. 2010). This

also relies on the mixing populations being sufficiently genetically distinct that they will consistently differ at many positions across the genome, allowing blocks of the genome in admixed individuals to be assigned to one or other of the parental populations (Fig. 8). It then looks for statistical associations between inheriting a particular chromosomal block from one parental population and a trait found in that population. The main applications of admixture mapping have been to map phenotypic and disease traits in admixed human populations, for example, African Americans who can trace their ancestry to both African and European populations (Shriver et al. 2003). However, it can also be applied to other species, particularly where genetically distinct populations meet and mix in hybrid zones. For example,

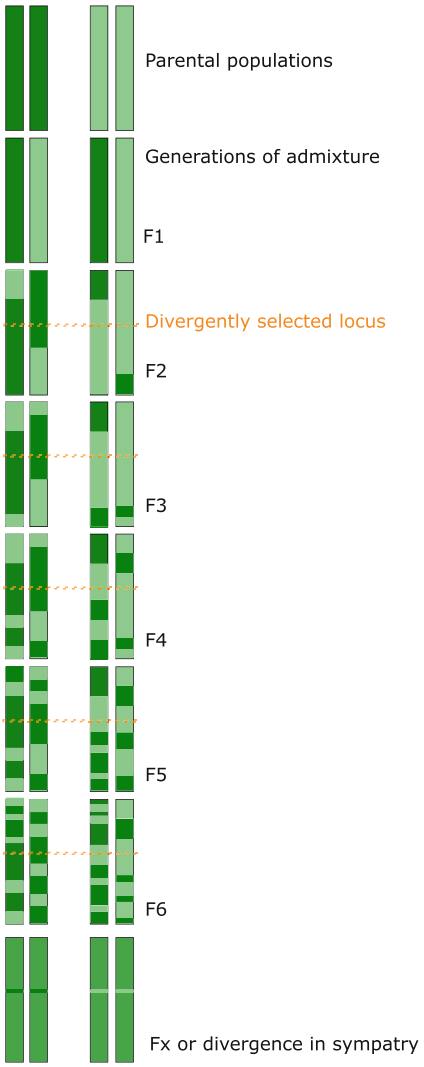

Fig. 8 Schematic representation of admixture between two starting (parental) populations, on one pair of chromosomes, over a limited number of generations (F1–F6). After many generations (Fx), the genotypes of the two populations have become homogenised, except for regions tightly linked to those under divergent selection, which resembles the situation of divergence in sympatry. Populations in which distinct genomic blocks can be assigned to one or other parental population are suitable for admixture mapping, while those that are more genetically homogenous are more suitable for genome-wide association mapping

QTL for leaf morphological traits have been mapped in naturally occurring hybrids of white poplar (*Populus alba*) and European aspen (*Populus tremula*) tree species (Lindtke et al. 2013). The *Populus* system is ideally suited to admixture mapping because the parental populations (species in this case) show marked allele frequency differences and natural hybrid zones occur at the boundaries of the preferred habitat (flood plain vs. upland) of each species.

A major advantage of admixture mapping over traditional QTL mapping using crosses is that there are likely to have been many generations of hybridisation and recombination, leading to small ancestry blocks, giving the potential to map loci to narrow genomic intervals. However, in reality the power of admixture mapping to identify QTL decreases with the number of generations of admixture (Lindtke et al. 2013), because the genomic blocks inherited from each parental population become too small to be identified. Ultimately this comes down to the same issue as low genetic differentiation between the parental populations; many generations of hybridisation will erode the genetic differentiation between the parental populations, leading to an inability to assign genetic markers to a population of origin (Fig. 8).

3.3 Genome-Wide Association Mapping

In situations in which genetic differentiation between parental populations is too low to allow admixture mapping, a suitable alternative approach can be to use genome-wide association (GWA) mapping. This method is dealt with in detail in another chapter, but it is worth highlighting some of the considerations when applying this technique to hybrid zone populations. Like admixture mapping, GWA mapping was first developed for human populations, with the idea of being able to map loci linked to disease susceptibility. Although this approach has been reasonably successful, a major limitation has been that the traits being mapped are usually due to rare alleles, after all, alleles causing disease will tend to be removed by purifying selection. This is compounded in cases of complex phenotypes, where individual loci often have small effect sizes (Fig. 9). Together, these factors mean that extremely large sample sizes are needed in order to have the power to detect loci (Bush and Moore 2012; Kardos et al. 2016).

In contrast, loci controlling traits that differ across hybrid zones will usually have alleles at high frequency on either side of the hybrid zone. Therefore, sampling evenly from across the hybrid zone will tend to sample each allele at around 50%, making these potentially extremely powerful situations in which to use GWA mapping (Fig. 9). In addition, many traits that differ across hybrid zones have been found to be controlled by major-effect loci (Nadeau et al. 2014; Scordato and Safran 2016). In these situations, relatively small numbers of individuals (less than 100) can be sufficient to identify loci underlying phenotypic differences using a GWA framework. For example, just 30 individuals sampled from across a natural hybrid zone were successfully used to map major effect loci controlling colour pattern differences in the butterfly *Heliconius melpomene* (Nadeau et al. 2014)

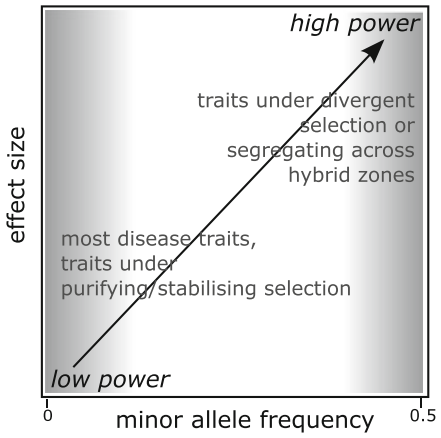

Fig. 9 The power of genome-wide association studies increases with both the minor allele frequency and the effect size of the underlying loci. Traits that segregate across hybrid zones will tend to have a balanced allele frequency and in many cases are also controlled by large-effect loci

(Fig. 10). Hybridising populations will also tend to have relatively high linkage disequilibrium (LD) between loci within the genome (Box 1), with the extent depending on the level of genetic differentiation between the parental populations (as well as general factors such as population size and recombination rate). High LD will also tend to increase the power of GWA studies, particularly when reduced representation sequencing is used, due to larger numbers of loci being in LD with the causative site (Kardos et al. 2016). However, high LD between loci, especially the long-range LD that can occur in contact zones, will also increase the false positive rate and make fine mapping of functional variants more difficult.

Population structure, causing genome-wide LD between unlinked loci, is a consideration for all GWA studies (Segura et al. 2012). However, it can be particularly problematic for hybrid zones. Even if there is little genetic differentiation between parental populations, a trait that changes along a linear transect will tend to be correlated with genome-wide genetic differences due to isolation by distance. For traits controlled by large-effect loci that change rapidly over short geographical distances, such as wing pattern in *H. melpomene* (Fig. 10), the problem is reduced because loci tightly linked to those controlling the traits will tend to show much stronger associations with phenotype than other loci in the genome. Similarly, a GWA study identified a major-effect locus controlling colour pattern in the stick insect *Timema cristinae* (Comeault et al. 2015). In this case the colour pattern morphs are cryptic on different host plants, which occur in mosaic patches within the landscape. In mosaic hybrid zones with high gene flow, such as this, background genetic structure is more likely to be decoupled from divergently selected loci (Nosil et al. 2002), making GWA mapping a potentially powerful tool. However, for traits that change gradually and linearly with distance and have a polygenic architecture, disentangling real versus correlated genetic associations is likely to be difficult. Although many ecologically relevant traits are likely to follow this pattern, there have been few attempts so far to apply GWA mapping to polygenic traits with broad geographic clines, perhaps because of the inherent challenge this poses. However, efficient mixed model approaches have been successfully used to control for

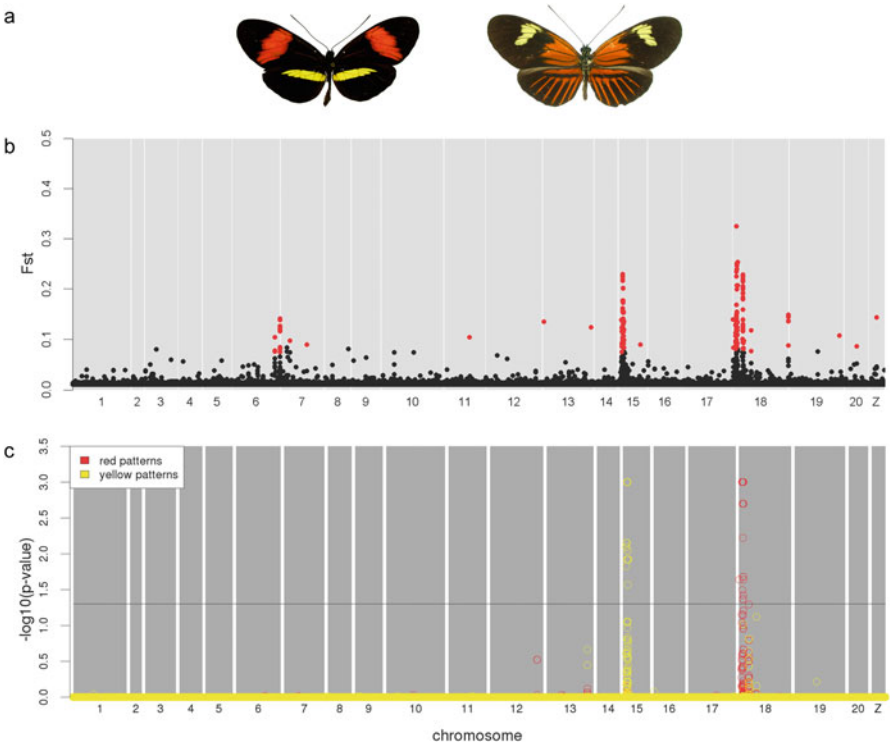

Fig. 10 Identifying loci underlying divergently selected traits in the butterfly *Heliconius melpomene*: a comparison of genomic differentiation and phenotypic association methods. (**a**) Butterflies from high elevation (left) and low elevation (right) near Tarapoto in Peru have very different wing colour patterns (photographs courtesy of Mathieu Joron). A narrow hybrid zone exists between these populations, which are maintained by strong positive frequency dependent selection, due to predator recognition of particular warning colour patterns (Mallet and Barton 1989). (**b**) Genome-wide differentiation (F_{ST}) between high and low elevation populations shows little background differentiation and few regions of high differentiation predicted to be under divergent selection (red points). Each point represents one SNP. (**c**) Genome-wide association mapping of red colour pattern elements (red points) and yellow colour pattern elements (yellow points) from 30 individuals, including 10 with hybrid phenotypes, from across the hybrid zone. This clearly identifies distinct loci for each trait, which correspond to the two most prominent divergently selected loci. The patterns of phenotypic association are less noisy than the patterns of genomic differentiation. Produced with data from Nadeau et al. (2014)

complex population structure in GWA studies of humans and plants (Berg and Coop 2014; Segura et al. 2012; Zhou and Stephens 2012), demonstrating the potential of GWA mapping to identify the genetic basis of complex traits that show clinal variation.

In summary, the presence of admixture between populations or species provides a valuable opportunity to identify loci that are divergently selected or control particular phenotypes. There are several methods for detecting these loci, and their applicability depends partly on the extent of gene flow between species. Methods

that test for associations between genotype and phenotype are the most powerful (Crawford and Nielsen 2013) and arguably also the most informative in terms of understanding the underlying selective pressures.

4 Detecting Hybridisation and Gene Flow Between Species

A major insight from population genomics studies has been the extent and prevalence of gene flow between species at multiple levels of divergence. Genome-wide markers allow introgressed variation to be identified, quantified and the history of hybridisation modelled through time, as never before. A large number of methods have emerged for identifying, quantifying and/or characterising gene flow between species, which are summarised in Table 2 taken from a thorough review of the topic by Payseur and Rieseberg (2016). Some of these methods overlap with those described in the previous sections for characterising divergently selected loci. Detecting gene flow is in some regards the inverse of this, and for populations where gene flow is high, these methods can be appropriate. However, for situations in which gene flow is rare or more ancient, more sensitive methods are needed.

Studies of humans have again largely paved the way in these approaches, motivated by the question of whether modern humans hybridised with Neanderthals during their colonisation of Europe. Sequencing of Neanderthal mitochondrial DNA did not reveal any evidence for hybridisation, with all Neanderthal sequences forming a cluster distinct from that of modern humans (Caramelli et al. 2006). However, genome-wide sequencing revealed an excess of genetic variants shared between Neanderthals and present-day Eurasian populations as compared to present-day African populations, suggesting gene flow may have occurred between Neanderthals and Eurasian modern humans (Green et al. 2010). This analysis was formalised as the D-statistic (or ABBA-BABA test), which uses an outgroup to test for an excess of shared derived SNPs between two putatively hybridising taxa (Fig. 8). Unfortunately this analysis has some problems, the most significant being that similar patterns of shared derived SNPs can be found if spatial population structure is present in the ancestral populations that both species diverged from, which is likely to have been the case in these archaic hominins (Durand et al. 2011; Eriksson and Manica 2012).

Nevertheless, subsequent studies using other approaches have also found evidence for gene flow between Neanderthals and modern humans. Sankararaman et al. (2012) used the extent of LD within the genomes of present-day Europeans to confirm and date the periods of gene flow with Neanderthals. LD is expected to break down with time, so if shared genetic variants were due to ancient population structure then blocks of LD would be shorter than if these were due to introgression events. Based on the size of the LD blocks containing variants shared between Europeans and Neanderthals, they concluded that introgression occurred between 37,000 and 86,000 years ago, long after the split between modern humans and Neanderthals. Subsequently, Lohse and Frantz (2014), estimated the maximum

Table 2 Genomic methods for detecting and characterising gene flow

Method	Characterisation of hybridisation						Focal pattern of variation	References
	Rate of gene flow	Timing of gene flow	Variable gene flow across genome	Variable gene flow across time	Individual ancestry proportions (genome-wide)	Individual locus-specific ancestries		
Geographic clines	Yes	No	Yes	No	No	No	Geographic gradient of allele frequencies across populations (hybrid zone)	Barton and Hewitt (1985), Porter et al. (1997), and Szymura and Barton (1986)
Genomic clines	No	No	Yes	No	No	No	Individual genotypes (hybrid zone)	Fitzpatrick (2013), Gompert and Buerkle (2009, 2011), Rieseberg et al. (1999), and Szymura and Barton (1986)
Structure/Structurama/ Frappe/Admixture/ FastStructure	No	No	No	No	Yes	No	Individual genotypes	Alexander et al. (2009), Falush et al. (2003), Hubisz et al. (2009), Huelsenbeck and Andolfatto (2007), Pritchard et al. (2000), Raj et al. (2014), and Tang et al. (2005)
Principal components analysis (PCA)	No	No	No	No	No	No	Individual genotypes	Patterson et al. (2006) and Price et al. (2006)
HapMix/recombination via ancestry switch probabilities (RASPberry)	No	No	Yes	No	Yes	Yes	Individual haplotypes	Price et al. (2009) and Wegmann et al. (2011)
Ancestry tract lengths/ shared haplotype lengths	Yes	Yes	No	Yes	No	No	Individual haplotypes	Gravel (2012), Harris and Nielsen (2013), Patterson et al. (2012), and Pool and Nielsen (2009)

Markov chain Monte Carlo – fit to isolation with migration model	Yes	No	No	No	No	No	Numbers of unique, shared and divergent polymorphisms	Becquet and Przeworski (2007), Hey and Nielsen (2004, 2007), Sethuraman and Hey (2016), Sousa et al. (2013), and Wang and Hey (2010)
Coalescent hidden Markov model (CoalHMM) – fit to isolation with migration model	Yes	Yes	No	No	No	No	Individual haplotypes	Mailund et al. (2012)
Blockwise likelihood – fit to isolation with migration model	Yes	Yes	No	No	No	No	Numbers of unique, shared and divergent polymorphisms	Lohse et al. (2011, 2016) and Lohse and Frantz (2014)
Diffusion approximations for demographic inference (dadi) – fit to isolation with migration model	Yes	No	No	No	No	No	Joint site frequency spectrum	Gutenkunst et al. (2009)
Sequentially Markov conditional sampling distribution – fit to isolation with migration model	Yes	Yes	No	Yes	No	No	Individual haplotypes	Steinrücken et al. (2015)
fineSTRUCTURE	No	No	No	No	Yes	No	Individual haplotypes	Lawson et al. (2012)
TreeMix	No	No	No	No	No	No	Joint site frequency spectrum	Pickrell and Pritchard (2012)
Divergence time heterogeneity	No	No	Yes	No	No	No	Numbers of unique, shared and divergent polymorphisms	Garrigan et al. (2012) and Yang (2010)
Approximate Bayesian computation	Yes	Yes	Yes	Yes	No	No	Array of summary statistics	Beaumont et al. (2002) and Robinson et al. (2014)

(continued)

Table 2 (continued)

Method	Characterisation of hybridisation						Focal pattern of variation	References
	Rate of gene flow	Timing of gene flow	Variable gene flow across genome	Variable gene flow across time	Individual ancestry proportions (genome-wide)	Individual locus-specific ancestries		
Genomic outliers (summary statistics, e.g. F_{ST})	No	No	Yes	No	No	No	Summaries of population differentiation	Beaumont and Balding (2004)
ABBA-BABA/D-statistics	No	No	No	No	No	No	Pattern of shared-derived changes	Durand et al. (2011), Green et al. (2010), Martin et al. (2014), Patterson et al. (2012), and Pease and Hahn (2015)
Phylogenetic discordance	No	No	No	No	No	No	Pattern of shared-derived changes	Ané et al. (2007) and Meng and Kubatko (2009)
Phylogenetic networks	No	No	No	No	No	No	Pattern of shared-derived changes	Liu et al. (2014a) and Yu et al. (2012, 2013, 2014)

Reproduced from Payseur and Rieseberg (2016) with permission

likelihood fit to models of admixture or ancestral population structure, using small non-recombining blocks of the genomes of two modern humans and a Neanderthal. They found strong support for Neanderthal admixture and obtained higher estimates of the rate of admixture (3.4–7.3%) than previous methods.

The *Heliconius* butterflies are another system in which population genomics has been used to characterise the extent and timing of gene flow between species. It had long been suspected that species within this genus did hybridise occasionally in the wild, with hybrids even found between fairly distantly related species (Dasmahapatra et al. 2007). The first population genomic evidence for gene flow between species again used ABBA-BABA D-statistics to show an excess of shared derived variants between sympatric sister species as opposed to allopatric populations of these species (Heliconius Genome Consortium 2012) (Fig. 11). In this case it was hard to envisage a scenario under which ancestral population structure could have given rise to these shared variants, because increased levels of shared variants were found in multiple sympatric population pairs in different geographic locations (Martin et al. 2013; Nadeau et al. 2013). However, other problems with the D-statistic were revealed. In particular, D-statistics do not reliably give the location of introgressed variants in the genome because genome-wide patterns are strongly correlated with nucleotide diversity, and simulations revealed that they could not reliably be used to compare the extent of gene flow between loci (Martin et al. 2014). Instead a different statistic, *f*, has been proposed, which also makes use of ABBA-BABA patterns but was found to be more robust to variation in nucleotide diversity and a better estimator of localised gene flow within the genome (Fig. 11).

A wide range of population genomic methods for inferring gene flow between species now exist. While some of these, such as ABBA-BABA and F_{ST}, are attractive because of their intuitive simplicity, they can be influenced by factors other than migration and do not provide estimates of the rate or timing of gene flow. Undoubtedly better are methods that test the fit of population genomic models, which can include varying amounts and timings of gene flow and can also incorporate factors such as population structure and varying population size to either patterns of nucleotide variation (Lohse and Frantz 2014) or the frequency spectrum of genetic variants (Gutenkunst et al. 2009). Roux et al. (2016) used an approximate Bayesian computation (ABC) framework to assess the extent of gene flow between 61 pairs of diverse animal species/populations from across the divergence continuum. They found a strong relationship between a simple divergence metric, Da (relative divergence, corrected for within-species diversity, which is strongly correlated with F_{ST}) and the extent of gene flow. However, both distinct species with virtually no gene flow and populations with high gene flow were present within a 'grey zone' between 0.5 and 2% net synonymous divergence, demonstrating the increased power of model-based approaches to detect and quantify gene flow.

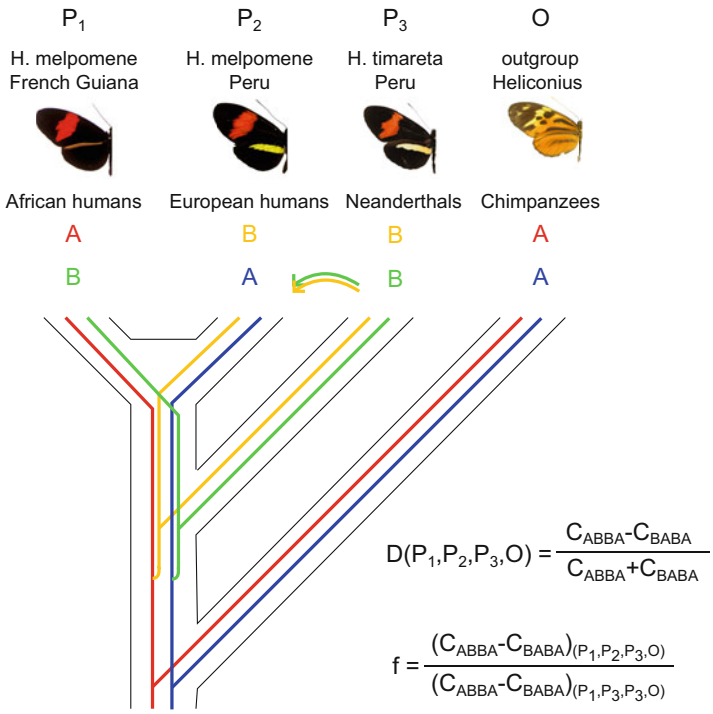

Fig. 11 ABBA-BABA methods for detecting gene flow between species. Initially formulated to test for introgression from Neanderthals (P_3) into modern European humans (P_2), by comparison to an outgroup (O) and an ingroup that would not have experienced gene flow (P_1, Africans in this case) (Green et al. 2010). The coloured lines show the situation of incomplete lineage sorting, where ABBA and BABA patterns can arise due to polymorphism in the ancestor of P_1, P_2 and P_3, which is sorted between the species. Without gene flow, an equal number of ABBA and BABA sites should be present, while gene flow will increase the number of ABBA sites. The D-statistic measures the relative proportion of ABBA to BABA sites, with C_{ABBA} and C_{BABA} being counts of the number of sites showing ABBA and BABA patterns, respectively. The *f* statistic was initially proposed to quantify the fraction of the genome shared through introgression, by comparing the difference between C_{ABBA} and C_{BABA} to the maximum difference possible by substituting P_2 for P_3 (Green et al. 2010). This statistic and variations thereof were also proposed to be more suitable for identifying introgressed regions of the genome, for example, to test if colour pattern controlling loci had introgressed between sympatric populations of the butterflies *Helcionius melpomene* and *H. timareta* (Martin et al. 2014). Butterfly photographs courtesy of Mathieu Joron

5 Future Perspectives

As with all areas of population genomics, the field is moving quickly, with new methods and approaches continually being developed. The field of speciation genomics essentially started with genome scans for divergence or differentiation outliers, but the challenges in this approach are now widely appreciated (Ravinet et al. 2017). Comparative genome scan approaches using multiple pairs of species can provide a

powerful framework to distinguish differentiation islands containing barrier loci from high divergence regions not directly associated with barriers to gene flow (incidental islands) (Burri 2017). Nevertheless, we also need an explicit null model, to understand how baseline genetic diversity varies under background selection at linked sites (Comeron 2017; Ravinet et al. 2017). Current differentiation outlier analyses implicitly assume uniform *Ne* along a genome and stable *Ne* over evolutionary time (i.e. uniform and stable recombination rate and gene density), but we are increasingly aware of the heterogeneity of these parameters associated with variation in the effect of background selection. Signatures of selection at barrier loci can be detected by comparing the observed patterns of genetic diversity with those expected under a null model with background selection. Ideally, detailed recombination maps for the organism in question would be used to simulate baseline genetic diversity, but such maps are rarely available. However, a broad 'U-shape' recombination landscape (i.e. higher recombination rate at the both ends of chromosomes) appears to be a general pattern in various species (Berner and Roesti 2017) and can be used as a proxy for species without detailed recombination maps.

The problems raised by variable *Ne* across the genome are not unique to divergence measures and will also influence other metrics such as cline shape and ABBA-BABA D-statistics (Gompert et al. 2017; Martin et al. 2014). Neutral processes (drift) and background selection combined with variation in recombination and mutation rates across the genome will produce variation in cline shapes. Therefore, to reliably detect either barrier loci or adaptively introgressed loci between species, null distributions for genomic clines and admixture proportions are needed. These should again ideally take into account recombination rate variation across the genome (Payseur and Rieseberg 2016). In addition, if the ultimate goal is to understand the role of natural selection in speciation, outlier loci detected by any method need to be linked to the phenotypes they control. Therefore, studies of genomes alone can only take us so far and need to be partnered with a detailed understanding of the phenotypes and ecology of the organisms in question.

The recent advances in long-read sequencing (e.g. PacBio and Oxford Nanopore), linked read sequencing (e.g. 10× Genomics) and long-range scaffolding technologies (e.g. optical mapping and Hi-C chromosome conformation capture) are beginning to substantially improve the contiguity of reference genomes. For instance, recently published reference genomes of mosquito (*Aedes aegypti*), grey mouse lemur (*Microcebus murinus*) and hooded crow (*Corvus* [*corone*] *cornix*) cover almost entire chromosomes, including highly repetitive regions, such as centromeres and pericentromeric regions (Dudchenko et al. 2017; Larsen et al. 2017; Weissensteiner et al. 2017). These regions tend to have low recombination rate due to their heterochromatic nature and likely coincide with elevated differentiation (Ellegren et al. 2012), possibly due to the effect of selection at linked sites. Importantly, one of these low recombination regions in crows contained several genes associated with plumage colour difference, which are likely to be under divergent selection (Fig. 5) (Poelstra et al. 2014).

Long-read and long-range sequencing technologies are also key tools for the identification of large structural variants, such as inversions and translocations

(Peichel et al. 2017). Chromosomal rearrangements have been suggested to play a key role in speciation by suppressing recombination and extending the effects of linked barrier loci (Kirkpatrick and Barton 2006; Navarro and Barton 2003; Noor et al. 2001; Rieseberg 2001). These models suggest that selection can facilitate the establishment and spread of new chromosomal mutations that harbour combinations of alleles contributing to local adaptation, or that rearrangements protect combinations of alleles that contribute to reproductive isolation from being disrupted by recombination. There are a growing number of examples showing an association between inversions and segregating phenotypes under divergent selection (Feder et al. 2003; Lowry and Willis 2010; McGaugh and Noor 2012; Turner et al. 2005). However, in other systems, such as *Heliconius*, divergence at many loci can be maintained in the absence of major structural variants or suppression of recombination (Davey et al. 2017). It is therefore not clear whether recombination modifiers, such as inversions, or more generally regions of low recombination (as found in crows), are necessary for the process of divergence with gene flow. New sequencing technologies will provide new insights into the frequencies of structural polymorphisms and their potential roles in speciation.

6 Conclusions

As high-throughput sequencing technologies have become accessible to many evolutionary biologists, there are a number of empirical studies published every year, describing genetic differences between genomes of diverging species and quantifying the level of gene flow between hybridising taxa. Nevertheless, despite the prediction based on the genic model of speciation (Wu 2001), genomic regions of elevated differentiation do not always harbour genes involved in RI or divergent selection. This does not necessarily mean that the model or analytical approaches are incorrect, but we need to develop an analytically tractable null model to predict genome-wide pattern of genetic diversity. Hybrid zones and admixed populations have been known as powerful model systems in speciation research for decades, but the advent of big population genomic data allows to fully exploit the power of these research systems by applying both traditional cline analysis and GWA. A combination of emerging new sequencing technologies and the development of analytical models will further provide a clearer picture of species divergence in the face of gene flow and identify barrier loci and their relative roles in the process of speciation.

Acknowledgements We would like to thank Marius Roesti and Roger Butlin for their helpful comments on this chapter.

References

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, et al. Hybridization and speciation. J Evol Biol. 2013;26(2):229–46.

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64.

Andrew RL, Rieseberg LH. Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes. Evolution. 2013;67:2468–82.

Ané C, Larget B, Baum DA, Smith SD, Rokas A. Bayesian estimation of concordance among gene trees. Mol Biol Evol. 2007;24(2):412–26.

Auton A, Fledel-Alon A, Pfeifer S, Venn O, Segurel L, Street T, et al. A fine-scale chimpanzee genetic map from population sequencing. Science. 2012;336:193–8.

Barton NH, Hewitt GM. Analysis of hybrid zones. Annu Rev Ecol Syst. 1985;16(1):113–48.

Beaumont MA, Balding DJ. Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol. 2004;13(4):969–80.

Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002;162(4):2025–35.

Becquet C, Przeworski M. A new approach to estimate parameters of speciation models with application to apes. Genome Res. 2007;17(10):1505–19.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, et al. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol. 2007;5(11):2534–59.

Bell MA, Foster SA. The evolutionary biology of the threespine stickleback. Oxford: Oxford University Press; 1994.

Berg JJ, Coop G. A population genetic signal of polygenic adaptation. PLoS Genet. 2014;10(8): e1004412.

Berner D, Roesti M. Genomics of adaptive divergence with chromosome-scale heterogeneity in crossover rate. Mol Ecol. 2017;26(22):6351–69.

Berner D, Roesti M, Hendry AP, Salzburger W. Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. Mol Ecol. 2010;19:4963–78.

Berner D, Moser D, Roesti M, Buescher H, Salzburger W. Genetic architecture of skeletal evolution in European lake and stream stickleback. Evolution. 2014;68:1792–805.

Bosse M, Spurgin LG, Laine VN, Cole EF, Firth JA, Gienapp P, et al. Recent natural selection causes adaptive evolution of an avian polygenic trait. Science. 2017;358(6361):365–8.

Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. Speciation and introgression between Mimulus nasutus and Mimulus guttatus. PLoS Genet. 2014;10(6):e1004410.

Burri R. Interpreting differentiation landscapes in the light of long-term linked selection. Evol Lett. 2017;1(3):118–31.

Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, et al. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula flycatchers*. Genome Res. 2015;25:1656–65.

Bush WS, Moore JH. Genome-wide association studies. PLoS Comput Biol. 2012;8(12):e1002822.

Butlin RK. Population genomics and speciation. Genetica. 2008;138(4):409–18.

Caputo B, Pichler V, Mancini E, Pombi M, Vicente JL, Dinis J, et al. The last bastion? X chromosome genotyping of Anopheles gambiae species pair males from a hybrid zone reveals complex recombination within the major candidate 'genomic island of speciation'. Mol Ecol. 2016;25(22):5719–31.

Caramelli D, Lalueza-Fox C, Condemi S, Longo L, Milani L, Manfredini A, et al. A highly divergent mtDNA sequence in a Neandertal individual from Italy. Curr Biol. 2006;16(16): R630–2.

Carneiro M, Albert FW, Afonso S, Pereira RJ, Burbano H, Campos R, et al. The genomic architecture of population divergence between subspecies of the European rabbit. PLoS Genet. 2014;10(8):e1003519.

Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, et al. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science. 2010;327 (5963):302–5.

Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Dickson M, Grimwood J, et al. Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science. 2005;307(5717):1928–33.

Comeault AA, Flaxman SM, Riesch R, Curran E, Soria-Carrasco V, Gompert Z, et al. Selection on a genetic polymorphism counteracts ecological speciation in a stick insect. Curr Biol. 2015;25 (15):1975–81.

Comeron JM. Background selection as null hypothesis in population genomics: insights and challenges from Drosophila studies. Philos Trans R Soc Lond B Biol Sci. 2017;372(1736).

Coyne JA, Orr HA. Speciation. New York: W. H. Freeman; 2004.

Crawford JE, Nielsen R. Detecting adaptive trait loci in nonmodel systems: divergence or admixture mapping? Mol Ecol. 2013;22(24):6131–48.

Crawford JE, Riehle MM, Guelbeogo WM, Gneme A, Sagnon N, Vernick KD, et al. Reticulate speciation and barriers to introgression in the Anopheles gambiae species complex. Genome Biol Evol. 2015;7(11):3116–31.

Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 2014;23(13):3133–57.

Cutter AD, Payseur BA. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet. 2013;14(4):262–74.

Dasmahapatra KK, Silva-Vásquez A, Chung J-W, Mallet J. Genetic analysis of a wild-caught hybrid between non-sister Heliconius butterfly species. Biol Lett. 2007;3(6):660–3.

Davey JW, Barker SL, Rastas PM, Pinharanda A, Martin SH, Durbin R, et al. No evidence for maintenance of a sympatric Heliconius species barrier by chromosomal inversions. Evol Lett. 2017;1(3):138–54.

Deagle BE, Jones FC, Chan YF, Absher DM, Kingsley DM, Reimchen TE. Population genomics of parallel phenotypic evolution in stickleback across stream–lake ecological transitions. Proc R Soc B Biol Sci. 2011. http://rspb.royalsocietypublishing.org/content/early/2011/10/04/rspb.2011.1552.abstract. Cited 13 Dec 2011.

Dobzhansky T. Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. Genetics. 1936;21:113–35.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356 (6333):92–5.

Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Mol Biol Evol. 2011;28(8):2239–52.

Duvaux L, Belkhir K, Boulesteix M, Boursot P. Isolation and gene flow: inferring the speciation history of European house mice. Mol Ecol. 2011;20(24):5248–64.

Egan SP, Ragland GJ, Assour L, Powell THQ, Hood GR, Emrich S, et al. Experimental evidence of genome-wide impact of ecological selection during early stages of speciation-with-gene-flow. Ecol Lett. 2015;18(8):817–25.

Elgvin TO, Trier CN, Tørresen OK, Hagen IJ, Lien S, Nederbragt AJ, et al. The genomic mosaicism of hybrid speciation. Sci Adv. 2017;3(6):e1602996.

Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, et al. The genomic landscape of species divergence in Ficedula flycatchers. Nature. 2012;491(7426):756–60.

Eriksson A, Manica A. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc Natl Acad Sci. 2012;109 (35):13956–60.

Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164(4):1567–87.

Feder JL, Nosil P. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. Evolution. 2010;64(6):1729–47.

Feder JL, Roethele JB, Filchak K, Niedbalski J, Romero-Severson J. Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, Rhagoletis pomonella. Genetics. 2003;163(3):939–53.
Feder JL, Gejji R, Yeaman S, Nosil P. Establishment of new mutations under divergence and genome hitchhiking. Philos Trans R Soc Lond B Biol Sci. 2012;367(1587):461–74.
Feulner PGD, Chain FJJ, Panchal M, Huang Y, Eizaguirre C, Kalbe M, et al. Genomics of divergence along a continuum of parapatric population differentiation. PLoS Genet. 2015;11: e1004966.
Fitzpatrick BM. Alternative forms for genomic clines. Ecol Evol. 2013;3(7):1951–66.
Flaxman SM, Wacholder AC, Feder JL, Nosil P. Theoretical models of the influence of genomic architecture on the dynamics of speciation. Mol Ecol. 2014;23(16):4074–88.
Gagnaire P-A, Pavey SA, Normandeau E, Bernatchez L. The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by rad sequencing. Evolution. 2013;67(9):2483–97.
Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR, et al. Genome sequencing reveals complex speciation in the Drosophila simulans clade. Genome Res. 2012;22(8):1499–511.
Gompert Z, Alex Buerkle C. Introgress: a software package for mapping components of isolation in hybrids. Mol Ecol Resour. 2010;10(2):378–84.
Gompert Z, Buerkle CA. A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. Mol Ecol. 2009;18(6):1207–24.
Gompert Z, Buerkle CA. Bayesian estimation of genomic clines. Mol Ecol. 2011;20(10):2111–27.
Gompert Z, Lucas LK, Buerkle CA, Forister ML, Fordyce JA, Nice CC. Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. Mol Ecol. 2014;23(18):4555–73.
Gompert Z, Mandeville EG, Buerkle CA. Analysis of population genomic data from hybrid zones. Annu Rev Ecol Evol Syst. 2017;48(1):207–29.
Gravel S. Population genetics models of local ancestry. Genetics. 2012;191(2):607–19.
Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the neandertal genome. Science. 2010;328(5979):710–22.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5(10):e1000695.
Haasl RJ, Payseur BA. Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. Mol Ecol. 2016;25(1):5–23.
Han F, Lamichhaney S, Grant BR, Grant PR, Andersson L, Webster MT. Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. Genome Res. 2017. http://genome.cshlp.org/content/early/2017/04/25/gr.212522.116. Cited 10 May 2017.
Harris K, Nielsen R. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 2013;9(6):e1003521.
Harrison RG, Larson EL. Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. Mol Ecol. 2016;25(11):2454–66.
Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 2012;487(7405):94–8.
Hey J, Nielsen R. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics. 2004;167(2):747–60.
Hey J, Nielsen R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci. 2007;104(8):2785–90.
Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 2010;6 (2):e1000862.

Hohenlohe PA, Bassham S, Currey M, Cresko WA. Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. Philos Trans R Soc Lond B Biol Sci. 2012;367:395–408.

Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour. 2009;9(5):1322–32.

Huelsenbeck JP, Andolfatto P. Inference of population structure under a Dirichlet process model. Genetics. 2007;175(4):1787–802.

Janoušek V, Wang L, Luzynski K, Dufková P, Vyskočilová MM, Nachman MW, et al. Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between Mus musculus musculus and M. m. domesticus. Mol Ecol. 2012;21(12):3032–47.

Jones FC, Chan YF, Schmutz J, Grimwood J, Brady SD, Southwick AM, et al. A genome-wide SNP genotyping array reveals patterns of global and repeated species-pair divergence in sticklebacks. Curr Biol. 2012a;22:83–90.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012b;484(7392):55–61.

Kardos M, Husby A, McFarlane SE, Qvarnström A, Ellegren H. Whole-genome resequencing of extreme phenotypes in collared flycatchers highlights the difficulty of detecting quantitative trait loci in natural populations. Mol Ecol Resour. 2016;16(3):727–41.

Kawakami T, Butlin RK. Hybrid zones. eLS. 2012. http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0001752.pub2/abstract. Cited 20 Jul 2012.

Kawakami T, Smeds L, Backstrom N, Husby A, Qvarnstrom A, Mugal CF, et al. A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. Mol Ecol. 2014;23:4035–58.

Kawakami T, Mugal CF, Suh A, Nater A, Burri R, Smeds L, et al. Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. Mol Ecol. 2017;26:4158–72.

Kirkpatrick M, Barton N. Chromosome inversions, local adaptation and speciation. Genetics. 2006;173(1):419–34.

Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A, et al. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature. 2015;518(7539):371.

Lamichhaney S, Han F, Webster MT, Andersson L, Grant BR, Grant PR. Rapid hybrid speciation in Darwin's finches. Science. 2017:eaao4593. https://doi.org/10.1126/science.aao4593.

Larsen PA, Harris RA, Liu Y, Murali SC, Campbell CR, Brown AD, et al. Hybrid de novo genome assembly and centromere characterization of the gray mouse lemur (Microcebus murinus). BMC Biol. 2017;15:110.

Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8(1):e1002453.

Lindtke D, Buerkle CA, Barbará T, Heinze B, Castiglione S, Bartha D, et al. Recombinant hybrids retain heterozygosity at many loci: new insights into the genomics of reproductive isolation in Populus. Mol Ecol. 2012;21(20):5042–58.

Lindtke D, González-Martínez SC, Macaya-Sanz D, Lexer C. Admixture mapping of quantitative traits in Populus hybrid zones: power and limitations. Heredity. 2013;111(6):474–85.

Liu BH. Statistical genomics: linkage, mapping, and QTL analysis. Boca Raton: CRC Press; 1997.

Liu KJ, Dai J, Truong K, Song Y, Kohn MH, Nakhleh L. An HMM-based comparative genomic framework for detecting introgression in eukaryotes. PLoS Comput Biol. 2014a;10(6):e1003649.

Liu S, Lorenzen ED, Fumagalli M, Li B, Harris K, Xiong Z, et al. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell. 2014b;157(4):785–94.

Lohse K, Frantz LAF. Neandertal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. Genetics. 2014;196(4):1241–51.

Lohse K, Harrison RJ, Barton NH. A general method for calculating likelihoods under the coalescent process. Genetics. 2011;189(3):977–87.
Lohse K, Chmelik M, Martin SH, Barton NH. Efficient strategies for calculating blockwise likelihoods under the coalescent. Genetics. 2016;202(2):775–86.
Lowry DB, Willis JH. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. PLoS Biol. 2010;8(9): e1000500.
Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, Andersen LN, et al. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genet. 2012;8(12):e1003125.
Mallet J, Barton NH. Strong natural selection in a warning-color hybrid zone. Evolution. 1989;43 (2):421–31.
Marques DA, Lucek K, Meier JI, Mwaiko S, Wagner CE, Excoffier L, et al. Genomics of rapid incipient speciation in sympatric threespine stickleback. PLoS Genet. 2016;12:e1005887.
Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, et al. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 2013;23 (11):1817–28.
Martin SH, Davey JW, Jiggins CD. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. Mol Biol Evol. 2014;32(1):244–57.
Martin SH, Möst M, Palmer WJ, Salazar C, McMillan WO, Jiggins FM, et al. Natural selection and genetic diversity in the butterfly Heliconius melpomene. Genetics. 2016;203:525–41.
Mayr E. Systematics and the origin of species, from the viewpoint of a zoologist. London: Columbia University Press; 1942.
McGaugh SE, Noor MAF. Genomic impacts of chromosomal inversions in parapatric Drosophila species. Philos Trans R Soc Lond B Biol Sci. 2012;367(1587):422–9.
McKechnie SW, Blacket MJ, Song SV, Rako L, Carroll X, Johnson TK, et al. A clinally varying promoter polymorphism associated with adaptive variation in wing size in Drosophila. Mol Ecol. 2010;19(4):775–84.
McKinnon JS, Rundle HD. Speciation in nature: the threespine stickleback model systems. Trends Ecol Evol. 2002;17:480–8.
Meng C, Kubatko LS. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor Popul Biol. 2009;75(1):35–45.
Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL. Widespread genomic divergence during sympatric speciation. Proc Natl Acad Sci. 2010;107(21):9724–9.
Moser D, Kueng B, Berner D. Lake-stream divergence in stickleback life history: a plastic response to trophic niche differentiation? Evol Biol. 2015;42:328–38.
Muller HJ. Bearing of the Drosophila work on systematics. In: Huxley J, editor. New systematics. Oxford: Oxford University Press; 1940. p. 185–268.
Nachman MW, Payseur BA. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Philos Trans R Soc Lond B Biol Sci. 2012;367:409–21.
Nadachowska-Brzyska K, Burri R, Olason PI, Kawakami T, Smeds L, Ellegren H. Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. PLoS Genet. 2013;9(11):e1003942.
Nadeau N. Butterfly genomics sheds light on the process of hybrid speciation. Mol Ecol. 2014;23 (18):4441–3.
Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, et al. Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. Philos Trans R Soc Lond B Biol Sci. 2012;367(1587):343–53.
Nadeau NJ, Martin SH, Kozak KM, Salazar C, Dasmahapatra KK, Davey JW, et al. Genome-wide patterns of divergence and gene flow across a butterfly radiation. Mol Ecol. 2013;22(3):814–26.

Nadeau NJ, Ruiz M, Salazar P, Counterman B, Medina JA, Ortiz-Zuazaga H, et al. Population genomics of parallel hybrid zones in the mimetic butterflies, H. melpomene and H. erato. Genome Res. 2014;24(8):1316–33.
Navarro A, Barton NH. Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. Evol Int J Org Evol. 2003;57(3):447–59.
Noor MAF, Bennett SM. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity. 2009;103:439–44.
Noor MAF, Grams KL, Bertucci LA, Reiland J. Chromosomal inversions and the reproductive isolation of species. Proc Natl Acad Sci. 2001;98(21):12084–8.
Nosil P. Ecological speciation. Oxford: Oxford University Press; 2012.
Nosil P, Feder JL. Genomic divergence during speciation: causes and consequences. Philos Trans R Soc Lond B Biol Sci. 2012;367:332–42.
Nosil P, Crespi BJ, Sandoval CP. Host-plant adaptation drives the parallel evolution of reproductive isolation. Nature. 2002;417(6887):440–3.
Nosil P, Feder JL, Flaxman SM, Gompert Z. Tipping points in the dynamics of speciation. Nat Ecol Evol. 2017;1(2):0001.
Orr HA. The population-genetics of speciation – the evolution of hybrid incompatibilities. Genetics. 1995;139:1805–13.
Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12): e190.
Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192(3):1065–93.
Payseur BA, Rieseberg LH. A genomic perspective on hybridization and speciation. Mol Ecol. 2016;25(11):2337–60.
Pease JB, Hahn MW. Detection and polarization of introgression in a five-taxon phylogeny. Syst Biol. 2015;64(4):651–62.
Peichel CL, Sullivan ST, Liachko I, White MA. Improvement of the threespine stickleback genome using a Hi-C-based proximity-guided assembly. J Hered. 2017;108(6):693–700.
Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8(11):e1002967.
Picq S, McMillan WO, Puebla O. Population genomics of local adaptation versus speciation in coral reef fishes (Hypoplectrus spp, Serranidae). Ecol Evol. 2016;6(7):2109–24.
Poelstra JW, Vijay N, Bossu CM, Lantz H, Ryll B, Müller I, et al. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science. 2014;344:1410–4.
Pool JE, Nielsen R. Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics. 2009;181(2):711–9.
Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, et al. Population genomics of sub-Saharan Drosophila melanogaster: African diversity and non-African admixture. PLoS Genet. 2012;8(12):e1003080.
Porter AH, Wenger R, Geiger H, Scholl A, Shapiro AM. The pontia daplidice-edusa hybrid zone in northwestern Italy. Evolution. 1997;51(5):1561–73.
Powell THQ, Hood GR, Murphy MO, Heilveil JS, Berlocher SH, Nosil P, et al. Genetic divergence along the speciation continuum: the transition from host race to species in Rhagoletis (Diptera: Tephritidae). Evolution. 2013;67(9):2561–76.
Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38 (8):904–9.
Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 2009;5(6): e1000519.
Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59.

Puebla O, Bermingham E, McMillan WO. Genomic atolls of differentiation in coral reef fishes (Hypoplectrus spp., Serranidae). Mol Ecol. 2014;23(21):5291–303.
Qvarnström A, Rice AM, Ellegren H. Speciation in Ficedula flycatchers. Philos Trans R Soc Lond B Biol Sci. 2010;365(1547):1841–52.
Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 2014;197(2):573–89.
Randler C. Assortative mating of Carrion Corvus corone and Hooded Crows C. cornix in the hybrid zone in eastern Germany. Ardea. 2007;95(1):143–9.
Ravinet M, Westram A, Johannesson K, Butlin R, André C, Panova M. Shared and nonshared genomic divergence in parallel ecotypes of Littorina saxatilis at a local scale. Mol Ecol. 2016;25 (1):287–305.
Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, et al. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J Evol Biol. 2017;30 (8):1450–77.
Renaut S, Maillet N, Normandeau E, Sauvage C, Derome N, Rogers SM, et al. Genome-wide patterns of divergence during speciation: the lake whitefish case study. Philos Trans R Soc Lond B Biol Sci. 2012;367(1587):354–63.
Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC, et al. Genomic islands of divergence are not affected by geography of speciation in sunflowers. Nat Commun. 2013;4:1827.
Riesch R, Muschick M, Lindtke D, Villoutreix R, Comeault AA, Farkas TE, et al. Transitions between phases of genomic differentiation during stick-insect speciation. Nat Ecol Evol. 2017;1:0082.
Rieseberg LH. Chromosomal rearrangements and speciation. Trends Ecol Evol. 2001;16(7):351–8.
Rieseberg LH, Whitton J, Gardner K. Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. Genetics. 1999;152(2):713–27.
Robinson JD, Bunnefeld L, Hearn J, Stone GN, Hickerson MJ. ABC inference of multi-population divergence with admixture from unphased population genomic data. Mol Ecol. 2014;23 (18):4458–71.
Roesti M, Hendry AP, Salzburger W, Berner D. Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. Mol Ecol. 2012;21 (12):2852–62.
Roesti M, Moser D, Berner D. Recombination in the threespine stickleback genome – patterns and consequences. Mol Ecol. 2013;22:3014–27.
Roesti M, Gavrilets S, Hendry AP, Salzburger W, Berner D. The genomic signature of parallel adaptation from shared genetic variation. Mol Ecol. 2014;23:3944–56.
Roesti M, Kueng B, Moser D, Berner D. The genomics of ecological vicariance in threespine stickleback fish. Nat Commun. 2015;6:8767.
Rogers SM, Campbell D, Baird SJE, Danzmann RG, Bernatchez L. Combining the analyses of introgressive hybridisation and linkage mapping to investigate the genetic architecture of population divergence in the lake whitefish (Coregonus clupeaformis, Mitchill). Genetica. 2001;111(1–3):25–41.
Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. Shedding light on the grey zone of speciation along a continuum of genomic divergence. PLoS Biol. 2016;14(12):e2000234.
Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. The date of interbreeding between Neandertals and modern humans. PLoS Genet. 2012;8(10):e1002947.
Scordato ESC, Safran RJ. Evolutionary genetics: small genomic regions make a big impact. Curr Biol. 2016;26(21):R1155–7.
Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, et al. Genomics and the origin of species. Nat Rev Genet. 2014;15:176–92.
Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet. 2012;44(7):825–30.

Sethuraman A, Hey J. IMa2p – parallel MCMC and inference of ancient demography under the isolation with migration (IM) model. Mol Ecol Resour. 2016;16(1):206–15.
Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, et al. Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet. 2003;112(4):387–99.
Smukowski CS, Noor MAF. Recombination rate variation in closely related species. Heredity. 2011;107(6):496–508.
Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, Johnston JS, et al. Stick insect genomes reveal natural selection's role in parallel speciation. Science. 2014;344(6185):738–42.
Sousa VC, Carneiro M, Ferrand N, Hey J. Identifying loci under selection against gene flow in isolation-with-migration models. Genetics. 2013;194(1):211–33.
Stankowski S, Sobel JM, Streisfeld MA. Geographic cline analysis as a tool for studying genome-wide variation: a case study of pollinator-mediated divergence in a monkeyflower. Mol Ecol. 2017;26(1):107–22.
Steinrücken M, Kamm JA, Song YS. Inference of complex population histories using whole-genome sequences from multiple populations. bioRxiv. 2015:026591. https://doi.org/10.1101/026591.
Svedin N, Wiley C, Veen T, Gustafsson L, Qvarnström A. Natural and sexual selection against hybrid flycatchers. Proc R Soc Lond B Biol Sci. 2008;275(1635):735–44.
Szymura JM, Barton NH. Genetic analysis of a hybrid zone between the fire-bellied toads, Bombina bombina and B. variegata, near Cracow in southern Poland. Evolution. 1986;40(6):1141–59.
Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol. 2005;28(4):289–301.
Teeter KC, Payseur BA, Harris LW, Bakewell MA, Thibodeau LM, O'Brien JE, et al. Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Res. 2008;18(1):67–76.
Toews DPL, Taylor SA, Vallender R, Brelsford A, Butcher BG, Messer PW, et al. Plumage genes and little else distinguish the genomes of hybridizing warblers. Curr Biol. 2016;26(17):2313–8.
Turner TL, Hahn MW, Nuzhdin SV. Genomic islands of speciation in Anopheles gambiae. PLoS Biol. 2005;3:e285.
Turner TL, Levine MT, Eckert ML, Begun DJ. Genomic analysis of adaptive differentiation in Drosophila melanogaster. Genetics. 2008;179(1):455–73.
Vallejo-Marín M, Buggs RJA, Cooley AM, Puzey JR. Speciation by genome duplication: repeated origins and genomic composition of the recently formed allopolyploid species Mimulus peregrinus. Evolution. 2015;69(6):1487–500.
Via S. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. Philos Trans R Soc Lond B Biol Sci. 2012;367(1587):451–60.
Via S, Conte G, Mason-Foley C, Mills K. Localizing F(ST) outliers on a QTL map reveals evidence for large genomic regions of reduced gene exchange during speciation-with-gene-flow. Mol Ecol. 2012;21:5546–60.
Vijay N, Bossu CM, Poelstra JW, Weissensteiner MH, Suh A, Kryukov AP, et al. Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. Nat Commun. 2016;7:13195.
Wang Y, Hey J. Estimating divergence parameters with small samples from a large number of loci. Genetics. 2010;184(2):363–79.
Wang J, Street NR, Scofield DG, Ingvarsson PK. Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American Aspens. Mol Biol Evol. 2016;33:1754–67.
Wegmann D, Kessner DE, Veeramah KR, Mathias RA, Nicolae DL, Yanek LR, et al. Recombination rates in admixed individuals identified by ancestry-based inference. Nat Genet. 2011;43(9):847–53.
Weissensteiner MH, Pang AWC, Bunikis I, Höijer I, Vinnere-Pettersson O, Suh A, et al. Combination of short-read, long-read and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. Genome Res. 2017;27(5):697–708.

Winkler CA, Nelson GW, Smith MW. Admixture mapping comes of age. Annu Rev Genomics Hum Genet. 2010;11(1):65–89.
Wolf JBW, Ellegren H. Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet. 2016;18:87–100.
Wu CI. The genic view of the process of speciation. J Evol Biol. 2001;14:851–65.
Yang Z. A likelihood ratio test of speciation with gene flow using genomic sequence data. Genome Biol Evol. 2010;2:200–11.
Yeaman S, Aeschbacher S, Bürger R. The evolution of genomic islands by increased establishment probability of linked alleles. Mol Ecol. 2016;25(11):2542–58.
Yu Y, Degnan JH, Nakhleh L. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet. 2012;8(4):e1002660.
Yu Y, Barnett RM, Nakhleh L. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. Syst Biol. 2013;62(5):738–51.
Yu Y, Dong J, Liu KJ, Nakhleh L. Maximum likelihood inference of reticulate evolutionary histories. Proc Natl Acad Sci. 2014;111(46):16448–53.
Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44(7):821–4.

Population Genomics of Colonization and Invasion

Shana R. Welles and Katrina M. Dlugosch

Abstract Population genomic analyses can reveal the mechanisms shaping the evolution of colonizing and invasive taxa, as for any species, including the fundamental processes of mutation, genetic drift, gene flow, and selection. Colonization events associated with species introductions, range shifts, and invasions pose a number of unique evolutionary questions, however, for which population genomic approaches are especially well-equipped to answer. These include quantifying the extent of founder effects, genetic bottlenecks, gene flow, and admixture that give rise to successful colonizing populations, identifying the nature and architecture of adaptive variation that resides in these populations (including types of mutations, their effect sizes, and their standing levels of variation), disentangling signatures of adaptation from other mechanisms of evolution, and identifying the ecological traits that have been the targets of natural selection and might be directly involved in the evolution of colonizing ability itself. We address each of these topics in this chapter, highlighting examples of recent research and the potential for population genomics to provide answers to some of the most pressing questions in the biology of colonizing and invasive species.

Keywords Adaptation · Admixture · Colonizers · Gene flow · Genetic drift · Invasive species · Mutations · Phylogeography · Population genetics

1 Introduction

Throughout the evolutionary and ecological history of life, lineages have colonized new locations through their natural modes of dispersal and, more recently, through the aid of humans (Baker 1955; Elton 1958; Parmesan and Yohe 2003; Jeschke and Strayer 2005; Vermeij 2005; Ellis et al. 2012a). The propagules that establish new

S. R. Welles · K. M. Dlugosch (✉)
Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA
e-mail: welles@chapman.edu; kdlugosch@email.arizona.edu

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_22,

populations are necessarily samples of genotypes from one or more source populations. Population genetics predicts that such sampling and establishment of genotypes in new environments should often be associated with significant genetic changes, including founder effects, genetic bottlenecks, and adaptation to the new environment (Baker and Stebbins 1965; Nei et al. 1975; Carroll et al. 2003; Excoffier et al. 2009a; Barrett et al. 2017). Human-mediated species introductions are providing many examples of colonization events that are observable on contemporary timescales, and indeed there are now many documented cases of evolutionary changes in introduced species (Lee 2002; Cox 2004; Colautti and Barrett 2013; Bock et al. 2015; Colautti and Lau 2015). Moreover, evidence is accumulating that evolutionary changes are not only common during colonization but also that they might contribute directly to colonization success and the evolution of highly invasive species (Ellstrand and Schierenbeck 2000; Facon et al. 2011; Williams et al. 2016; Ochocki and Miller 2017; Wagner et al. 2017).

Population genomic approaches offer the opportunity to identify the mechanisms underlying evolutionary changes during colonization. A mechanistic understanding of colonization genetics and evolution is central to addressing many open questions in the biology of colonizing and invasive species, particularly how evolutionary changes may enhance or impede the success of colonizers (Fig. 1). These questions generally revolve around understanding the role that each of the major evolutionary forces (selection, gene flow, genetic drift, and mutation) play in shaping phenotypes in a novel habitat. In this chapter we address the following areas:

- Phylogeography and historical demography. We start by considering the identification of the source(s) and demographic history of colonizing populations. Historical context is fundamental to uncovering the evolutionary changes that have occurred during colonization and inferring the action of specific evolutionary forces. Population genomic inference of historical demography has revolutionized these efforts.
- Genetic drift. During both founding events and expansion into new regions, small populations and/or inbreeding could lead to strong genetic drift. An active area of theoretical, genomic, and ecological research is considering whether these effects are strong enough to create barriers to successful colonization and/or adaptation in novel environments.
- Gene flow. Gene flow among colonizing populations, between introductions from different sources (admixture), and between species (hybridization) could ameliorate the effects of genetic drift or provide particularly favorable genetic combinations for colonizers. Population genomic approaches are especially powerful for uncovering what impact these different types of gene flow have on genetic variation, adaptive variation, and fitness.
- Selection. Colonized environments almost certainly present novel patterns of selection for founding populations, and there is mounting evidence that adaptive evolution is a common feature of colonization. Population genomics offers

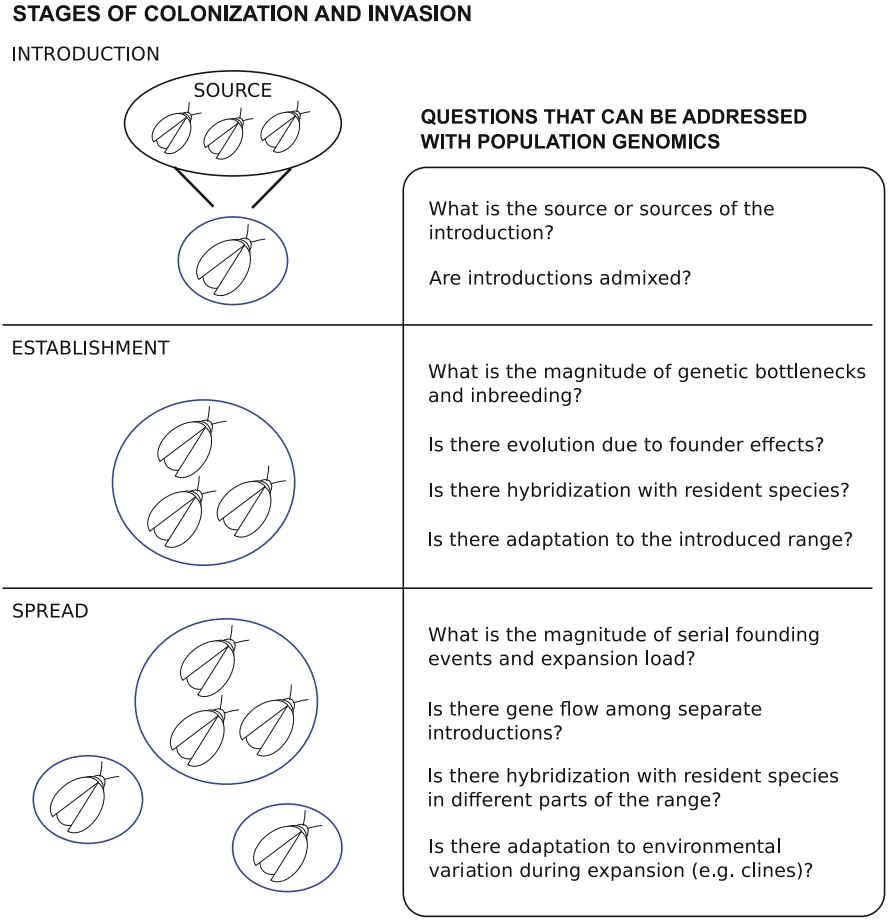

Fig. 1 Major open questions in the study of colonization and invasion, at each stage of colonization

opportunities to disentangle responses to selection from other forms of evolution and can provide insights into the genetic changes that might have had the most impact on ecological success in a new environment.

- Mutation. Finally, we consider how population genomic approaches can give insights into the mutations underlying variation in colonizer phenotypes. The types of mutation, their frequency of de novo formation during colonization, and their propensity to form standing genetic variation in source populations will interact with genetic drift, gene flow, and selection to influence ecological variation and colonization success.

2 Inferring the History of Colonization and Invasion

The first step in any study of colonization genetics is the reconstruction of colonization history. It is crucial to know the source(s) of colonizers in order to set up appropriate comparisons for studies of genomic and trait evolution, and the accuracy with which sources are identified will affect the power and quality of any subsequent analyses (Dlugosch and Parker 2008; Estoup and Guillemaud 2010; Tiffin and Ross-Ibarra 2014; Cristescu 2015). Source populations can vary greatly in their genetic structure, composition, and resulting phenotypes (Colautti et al. 2009). Misidentification or insufficient characterization of sources can lead to erroneous conclusions about colonization events, even inferences of genetic changes that are opposite of the true history of evolution in colonizing populations (Dlugosch and Parker 2008; Colautti and Lau 2015).

Genetic, especially population genetic, tools have long been used to identify the sources of colonizing populations, both ancient and contemporary (Avise 1994). While historical records can provide important information about introduction pathways for human-mediated introductions, the specific source(s) of colonists that contributed genetic material to founding populations is still often complex or unclear, and historical information for many (if not most) colonization events is incomplete or lacking (Estoup and Guillemaud 2010; Cristescu 2015). Genetic determination of source populations relies on identifying unique genetic features (i.e., private alleles, allele frequency patterns) of different native populations and associating these with the genetic makeup of the colonizing populations, using methods that include phylogenetic trees or networks, genetic distances among populations, and/or inference of the probability of observing the colonizing population given different possible sources (e.g., through comparisons of model likelihood) (Gutenkunst et al. 2009; Knowles 2009; Estoup and Guillemaud 2010). Analyses that reconstruct colonizing population genetics from source genotypes can then be used to infer additional historical features, such as the number of independent introductions, effective population sizes, and the amount and direction of gene flow between populations (Emerson et al. 2001; Knowles and Maddison 2002; Knowles 2009).

Genome-scale population genetic information can be particularly powerful for historical reconstruction because of the high resolution afforded by a large dataset of polymorphisms, which can provide many opportunities to observe private alleles and unique multi-locus allele frequency signatures (Brumfield et al. 2003; Emerson et al. 2001). A variety of genome-scale population genetic methods have been explored for this purpose (e.g., AFLPs; Meudt and Clarke 2007), but the field is now rapidly expanding through the use of high-throughput reduced-representation nuclear genomic sequencing methods (e.g., GBS, RADseq), as well as whole genome sequencing and resequencing. These methods typically identify many thousands of nucleotide polymorphisms across the genome, maximizing the information available for historical inference as well as our ability to analyze the datasets with powerful models of nuclear sequence evolution (Narum et al. 2013).

One of the most common approaches for understanding colonization history through population genomic datasets is the use of a Bayesian clustering algorithm in the program Structure (Pritchard et al. 2000; Raj et al. 2014) with a staggering ~20,000 citations in Google Scholar at the time of this writing. Structure performs Bayesian assignment of individuals to a given number of population clusters, maximizing Hardy-Weinberg and linkage equilibrium within clusters. Colonizing populations can be assigned to a defined set of source clusters, or clustering can be done simultaneously on colonizer and source individuals (to evaluate whether colonists naturally cluster with source individuals). This method is operationally easy to carry out and was a breakthrough for the study of colonization, particularly because the approach allowed the quantitative inference of mixed ancestry and because it allowed the inference of geographic structure (i.e., genetically distinct source regions and unique founding populations) from data without a priori knowledge of population boundaries.

There are several limitations to using a clustering approach to make conclusions about colonizing population ancestry, however, and the authors of Structure in particular have long recommended it as primarily an exploratory tool (Pritchard et al. 2000; Falush et al. 2016; Novembre 2016). Colonization histories can be complex in terms of the number of contributing source populations, evolution post-colonization, and gene flow among different founding lineages. Clustering methods have been shown to suffer from poor accuracy under complex evolutionary histories (Estoup and Guillemaud 2010). Particularly where there is genetic admixture or incomplete lineage sorting in divergent populations, the clustering algorithms can produce incorrect assignment of admixed ancestry (Kalinowski 2011; Falush et al. 2016; Novembre 2016; Wang 2016). Additionally, the full range of a species can be large, including geographic areas that may be logistically difficult or impossible to sample. Unsampled source populations are likely to lead to incorrect conclusions regarding colonization history when using a clustering approach; however incomplete sampling can limit the ability to reconstruct invasion history under any of the available methods to date (Guillemaud et al. 2010; Cristescu 2015).

More recently, demographic inference methods have been used to explicitly evaluate support for alternative evolutionary scenarios. For example, competing models may include the presence or absence of bottlenecks, different source populations, unsampled source populations, and gene flow (Fig. 2). Demographic inference methods compare the observed dataset with multiple simulated datasets created under different demographic and genetic histories (Bertorelle et al. 2010; Csilléry et al. 2010). The probability of observing the data under alternative evolutionary histories is calculated, and the best fitting model can be identified. Historical information and exploratory analyses (such as Structure) can be used to inform decisions about how to delineate populations and which competing scenarios to select for alternative models.

Demographic inference is computationally intensive, and multiple programs are currently available for the data simulation, calculation of summary statistics for each simulation, and likelihood of each scenario. Particularly popular are coalescent simulations in the approximate Bayesian computation (ABC) framework. Available

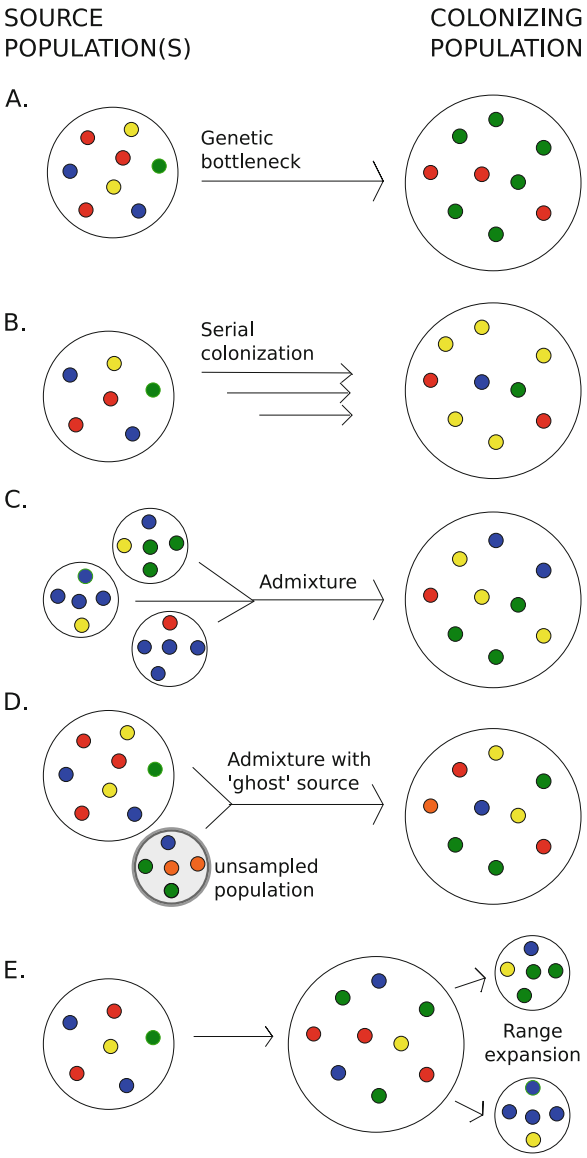

Fig. 2 Examples of colonization scenarios that can be tested under competing models of historical demography, including (**a**) genetic bottlenecks during founding, (**b**) serial colonization events into the same region, (**c**) admixture among multiple source populations, (**d**) admixture with unsampled source populations, and (**e**) sources of new colonization events within an expanding range

programs for this approach differ in the scenarios that they can model, ease of use, and speed (often with trade-offs among these features). DIYABC (Cornuet et al. 2008) and PopABC (Lopes et al. 2009) are the most user-friendly and commonly used programs. ABCtoolbox (Wegmann et al. 2010) is a powerful command-line program that can model many introduction scenarios and can be easily integrated with other programs. Additional implementations are available in the R statistical environment [abc (Csilléry et al. 2010), EasyABC (Jabot et al. 2013), abcrf

(Pudlo et al. 2016)]. Coalescent methods are also not the only approaches available for historical demography. For example, δaδi uses a diffusion approximation to model the evolution of the joint allele frequency spectrum of populations under alternative models, with very fast computation times relative to ABC simulations (Gutenkunst et al. 2009; Huber et al. 2014; Qi et al. 2017).

The power of these genomic approaches can be demonstrated by a recent phylogeographic study of the plant yellow starthistle (*Centaurea solstitialis*). Native to Eurasia, the plant is highly invasive in its introductions to the Americas (Gerlach 1997). A previous microsatellite-based study had found evidence for admixture in the major invasions of California, USA, largely based on clustering assignments (Eriksen et al. 2014). Barker et al. (2017a) generated a large double-digest RADseq dataset for the species and used an ABC approach to test several alternative models of both the recent introduction history and the history of ancient colonization among genetically differentiated regions of the native range. Using this approach, Barker et al. (2017a) found strong support for a model with no admixture in the recent introduction to California. Instead, ancient admixture was identified within the native source region for the introductions. This case illustrates both the difficulty of assigning admixture under clustering methods and the power of current population genomic techniques to resolve complex colonization histories across multiple time scales.

Although demographic inference models are a powerful approach to reconstructing invasion histories, they do have limitations (Estoup and Guillemaud 2010). It can be difficult to distinguish between some scenarios; for example, it is typically difficult to distinguish between a single large founding population and serial colonization from the same source population. The quality of the results is limited by the thoroughness of the sampling and the number of genetic markers used. Alternative models must be chosen by the user and may not include the true history. Despite these limitations, the combination of high-throughput population genomic sequencing and powerful models of historical demography maximizes our ability to differentiate between colonization scenarios that produce only subtle differences in genetic patterns.

3 Genetic Drift: Founder Effects, Bottlenecks, Inbreeding, and Allele Surfing

Genetic drift has the potential to be one of the most important evolutionary forces operating during colonization, due to genetic sampling effects inherent in the establishment of a limited number of dispersing propagules, as well as the small effective population sizes experienced as populations establish and spread (Nei et al. 1975; Frankham 2005). These effects are relevant for understanding patterns of genetic variation and differentiation in colonizing populations, but they may also have important impacts on the establishment, spread, and persistence (i.e., the

ecology) of founding events. There is evidence that some founding populations might face negative fitness consequences of genetic drift due to fixation of deleterious alleles (Briskie and Mackintosh 2004), reduced genetic diversity (Crawford and Whitney 2010; Szűcs et al. 2014, 2017), and reduced efficacy of selection (Peischl et al. 2015). These challenges to colonization have led invasion biologists to suggest a "genetic paradox of invasion," in which introduced species rapidly expand their range in novel environments despite being expected to suffer the deleterious effects of genetic drift during founding (Allendorf and Lundquist 2003; Frankham 2005).

The prevalence of genetic bottlenecks during colonization is an especially active area of current research. Although theory suggests that reductions in genetic diversity associated with small population sizes during colonization should be common, empirical observations from contemporary species introductions indicate otherwise (Gray et al. 1986; Barrett et al. 1990; Kolbe et al. 2004; Dlugosch et al. 2015a). Surveys of the literature have suggested that human-mediated colonization events are associated with only very weak genetic bottlenecks in most (though not all) cases (Dlugosch and Parker 2008; Uller and Leimu 2011), and a recent analysis of transcriptome evolution across several related taxa found no evidence of excess fixation of deleterious alleles in colonists (Hodgins et al. 2015). On the other hand, persistent signatures of genetic bottlenecks can be seen in historical cases (Ramachandran et al. 2005; Moreau et al. 2011; Domingues et al. 2012), and long-distance colonization (such as of remote islands) appears to be associated with evolutionary changes consistent with very small founder sizes (Baker 1955; Pannell 2015).

Of particular recent interest is the potential for strong effects of genetic drift as colonizing populations expand across space. If dispersal tends to originate from the leading edge, chance increases in allele frequency at the wave front can quickly lead to stochastic fixation of those alleles, a pattern known as "allele surfing" (Edmonds et al. 2004; Excoffier et al. 2009a). Allele surfing is expected to lead to increased homozygosity at the leading edge and increased differentiation among different regions of the leading edge (Hallatschek et al. 2007; Excoffier and Ray 2008; Peischl and Excoffier 2015; Peischl et al. 2015). Both of these patterns have been observed along expansion trajectories in recently colonizing populations (Ramakrishnan et al. 2010; Graciá et al. 2013; White et al. 2013; Pierce et al. 2014). In general, the effects of allele surfing are expected to be weaker under high levels of gene flow, which will tend to restore diversity to the wave front (Pierce et al. 2014; Peischl et al. 2015). The accumulation of empirical studies will be valuable for revealing both the extent to which allele surfing affects colonization events and the persistence of these effects over evolutionary time.

A surfing allele can have the appearance of responding to selection during range expansion, rapidly increasing in frequency relative to the genetic background. Allele surfing can be random, however, and should often lead to reduced response to selection and the fixation of deleterious alleles ("expansion load") at the wave front (Peischl and Excoffier 2015; Peischl et al. 2015). Signatures of surfing effects on adaptive variation have been found in humans as well as experimental studies of

microbial populations, though it is notable that in some cases the surfing of rare beneficial mutations to high frequency can lead to increased response to selection and overall fitness gains (Hallatschek et al. 2007; Moreau et al. 2011; Gralka et al. 2016).

Population genomic datasets offer opportunities to shed light on the influence of genetic drift during colonization by quantifying the magnitude of genetic bottlenecks and effective population size changes experienced during colonization. As discussed above, historical demography provides inferences of past effective population sizes and testing of alternative models with and without genetic bottleneck events (Knowles 2009). Genomic datasets are also powerful for identifying signatures of genetic drift through serial founding events and allele surfing during range expansion (Moreau et al. 2011; White et al. 2013). Such analyses are increasingly recognized as critical for distinguishing evolution due to genetic drift from evolution due to adaptation during colonization (Keller and Taylor 2008; Colautti and Lau 2015), and they can help to identify whether there is likely to be sufficient statistical power available for distinguishing the action of selection in colonization events that have experienced strong genetic drift (Poh et al. 2014).

For example, White et al. (2013) studied the population genomics of the introduced bank vole (*Myodes glareolus*) invading Ireland. Using multiple transects from the introduction core to the leading edge of the invasion, they identified declines in genetic diversity at 5979 SNPs (using GBS) along each transect, consistent with predictions from allele surfing dynamics. They did not find increases in putative deleterious mutations, however, suggesting that any expansion load is weak in this case. The authors were able to use consistent changes in allele frequency along all transects to identify candidates for likely responses to selection, because allele surfing predicts the fixation of different random variants at different locations along the wave front. In this way, they identified several outliers with strong potential to affect ecologically important phenotypes.

For studies of adaptive variation during colonization, population genomics will be valuable for helping to clarify whether species typically experience losses of ecologically relevant trait variation during colonization. Stronger genetic bottlenecks are expected to reduce opportunities for adaptation, but this outcome depends on the genetic architecture of traits, and the relationship between relevant trait variation and colonization dynamics in natural populations is still poorly understood (Dlugosch et al. 2015a). Studies that identify adaptive variation in the source region and follow this variation through population genetic analyses of colonizing populations would be particularly useful for illuminating these relationships.

4 Gene Flow, Admixture, and Hybridization

Gene flow between colonizing populations derived from different colonization events, different source populations, and different species has the potential to strongly influence the genetics and success of colonizers (Ellstrand and

Schierenbeck 2000; Sakai et al. 2001; Lee 2002; Kolbe et al. 2004; Lavergne and Molofsky 2007). If genetic bottlenecks are occurring during founding or range expansion, gene flow among colonizing populations will restore diversity and counteract deleterious effects of genetic drift (Dlugosch and Parker 2008; Blackburn et al. 2015). Where gene flow occurs between divergent populations or species, it can result in substantial fitness benefits from the resulting novel combinations of alleles, even in the absence of genetic bottlenecks (Ellstrand and Schierenbeck 2000; Rius and Darling 2014).

There is a general hypothesis that admixture and hybridization may be especially beneficial to colonizers, particularly in the context of introduced and invasive species (Ellstrand and Schierenbeck 2000; Hufbauer 2008, 2017; Verhoeven et al. 2011; Rius and Darling 2014). Among invasive species, multiple introductions of a species into an area are common and often include introductions from different parts of the native range, resulting in admixture (Dlugosch and Parker 2008; Uller and Leimu 2011; Dlugosch et al. 2015a). A recent survey estimated that almost 40% of introduced species studied have been identified as admixed (Barker et al. 2017b). In cases of hybridization between different species, multiple cases of particularly successful invaders of hybrid origin have been identified (Ellstrand and Schierenbeck 2000; Drake 2006; Lavergne and Molofsky 2007; Keller and Taylor 2010). Determining where admixture and hybridization have occurred and their potential contributions to adaptive variation has thus become a key goal of many studies of colonizing species (Rius and Darling 2014).

Population genomics is a powerful tool for identifying and quantifying patterns of gene flow, admixture (gene flow between divergent populations of the same species), and hybridization (gene flow between different species) (Payseur and Rieseberg 2016). As mentioned above, by testing of alternate models of historical demography, gene flow and admixture events can be identified, even where these events were obscure or difficult to pinpoint with previous approaches (Gompert and Buerkle 2013). Moreover, population genomic tools can identify the specific regions of the genome involved in introgression events [e.g., HapMix (Price et al. 2009), RASPberry (Wegmann et al. 2011), popanc (Gompert 2016)]. Determining the sources of individual loci or blocks of loci in genomes is particularly important for identifying alleles that might be under positive selection and the source of adaptations that arise during colonization (Gompert et al. 2016). Assigning ancestry to loci in genomic data is complicated by uncertainty in sequence data, identifying an appropriate model of hybridization and assigning potential source and opportunities for gene flow, and properly accounting for variation in recombination and selection (Gompert et al. 2016; Payseur and Rieseberg 2016). Recent models have worked to better account for these issues in admixed populations (e.g., Gompert and Alex Buerkle 2010; Gompert and Buerkle 2012; Gompert 2016).

There are several genetic mechanisms that can contribute to fitness effects resulting from the mixture of divergent source populations, and population genomic approaches can assist in distinguishing among them. These mechanisms include increases in adaptive variation, rescue of genetic load, overdominance, underdominance, and epistasis (Barker et al. 2017b; Hufbauer 2017). Population genomic

analyses of admixture can indicate that hybrid combinations at particular loci are disfavored (underdominant), favored (overdominant), or interacting among loci (epistatic) and that individual alleles from different sources are favored universally (adaptive, potentially rescuing genetic load) or only under certain environments (locally adaptive). These alternative outcomes can reveal whether admixed/hybrid populations are favorable variants to colonizing populations and whether these mechanisms are likely to be at work in many colonizing species.

For example, a study by Nolte et al. (2009) examined 168 microsatellite loci in a range expanding hybrid species of sculpin (*Cottus perifretum*) and its secondary contact zones with a parental native lineage. Over 20% of loci showed underdominance, which would promote isolation of the lineages, but almost 30% of loci appeared to be under selection for introgression into the colonizer. A small subset of these loci displayed patterns consistent with overdominance or epistasis and patterns different among different contact zones, indicating that some patterns of selection were local. This study nicely demonstrates the ability of genomic analyses to decompose the various potential effects of gene flow on colonization, even in complex demographic scenarios.

5 Selection and Adaptive Evolution

One of the most exciting applications of population genomics is the identification of loci that are likely to be evolving in response to natural selection. Colonizers should often encounter novel environments and associated novel patterns of selection (Waddington 1965). Colonization events create outstanding opportunities to link these differences in the environment to adaptations. Moreover, the success of colonization itself may be enhanced by rapid adaptation to new conditions (Thompson 1998; Sakai et al. 2001). Colonizers are known to experience differences in biotic interactions, climate, availability of resources, and disturbance regimes relative to their source regions, often with opportunities to evolve changes in resource allocation which favor their success (Colautti et al. 2004; Balanyá et al. 2006; Lee and Gelembiuk 2008; Dlugosch et al. 2015b; Koskella 2015). Adaptive evolutionary shifts in response to novel selection regimes may therefore be central to initial establishment and spread after colonization (García-Ramos and Rodríguez 2002; Colautti and Barrett 2013; Colautti and Lau 2015).

Comparisons between colonizing populations and their source populations can allow for powerful inferences of loci under differential selection between ranges. Shifts in allele frequencies between ranges at individual loci, relative to the remainder of the genome, suggest the influence of selection. These approaches will be affected by the completeness with which polymorphisms in the genome have been sampled (i.e., whether loci under selection or linked markers have been captured), the strength of selection and allele frequency divergence at individual loci, and the influence of demographic history, which will affect the likelihood of observing signatures of selection against the genomic background (Oleksyk et al. 2010;

Berg and Coop 2014; Poh et al. 2014; Tiffin and Ross-Ibarra 2014; François et al. 2016; Lowry et al. 2017). Genomic comparisons can focus on identifying loci with evidence of selective sweeps, loci with particularly high divergence between populations associated with divergent adaptations, or loci with allele frequency correlations to particular environmental variables (indicating adaptation to local environments). Each of these three approaches (Fig. 3) can provide insights to

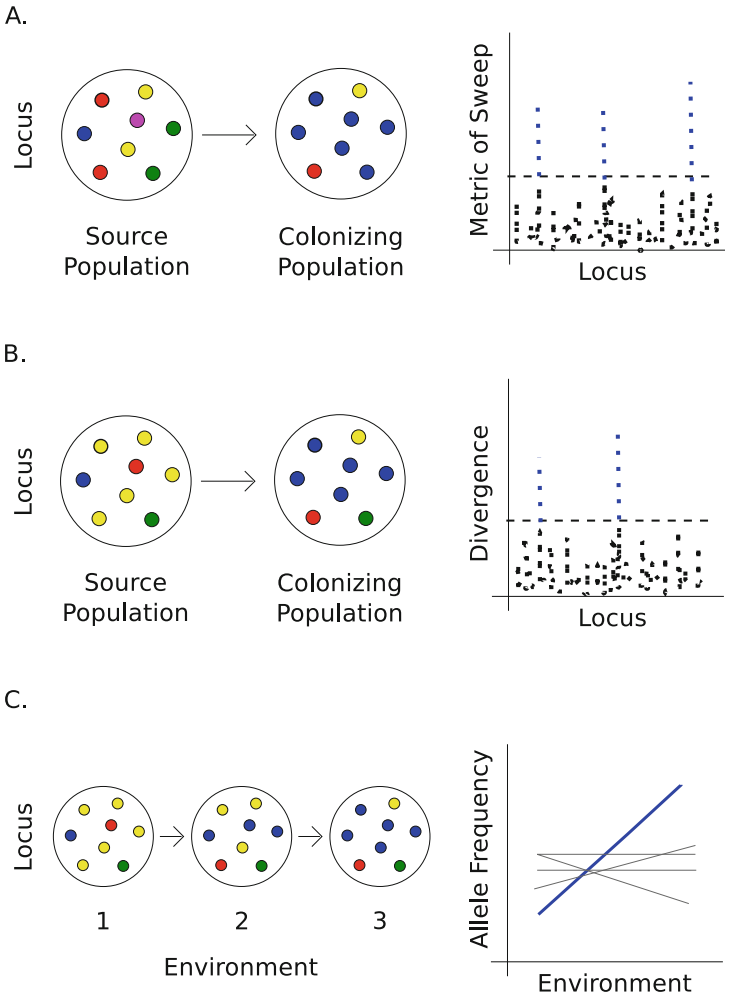

Fig. 3 Population genomic approaches to identifying loci under selection during colonization. The effects of selection on allele frequencies (colored circles) are shown on the left, and their signatures as outliers (blue) relative to patterns among other loci (black) are shown on the right. Approaches include scans for (**a**) selective sweeps, (**b**) allele frequency divergence, and (**c**) associations with environmental variables. Metrics of selective sweeps in (**a**) include losses of diversity, shifts in the allele frequency spectrum, and linkage disequilibrium around a locus

adaptive divergence in different ways and may reveal unique sets of candidate loci involved in different aspects of colonization.

Selective sweeps occur when a rare variant is favored and rises to high frequency in a population, resulting in exceptionally low diversity, high linkage disequilibrium, and shifts in the allele frequency spectrum near the locus under selection (Kaplan et al. 1989; Braverman et al. 1995; Kelly 1997; Fay and Wu 2000; Oleksyk et al. 2010). In a "hard" selective sweep, the favored allele is extremely rare (e.g., occurring as a new mutation or in a single colonist in a founding population) and sweeps to high frequency in a single genetic background. Alleles at nearby variable sites will "hitchhike" along with the sweep, producing a large signal of exceptionally low diversity, long shared haplotypes, high linkage disequilibrium, and a skewed site frequency spectrum in the area of the genome. Hard sweeps are relatively easily detectable by scans for these features, though the strength of the signal will decay with time and recombination events that break up the hitchhiking markers (Thornton and Jensen 2007; Oleksyk et al. 2010; Pritchard and Di Rienzo 2010; Tiffin and Ross-Ibarra 2014). In contrast, selection from standing variation, wherein favored alleles occur in many genetic backgrounds, will produce a "soft" sweep with a much weaker signal in these types of scans for selection (Hermisson and Pennings 2005). Many colonizers might demonstrate particularly clear signatures of selective sweeps, where adaptation has been relatively recent and extant source populations can provide an accurate comparison that is not obscured by large changes in demography or selection since divergence. For example, Li et al. (2017) recently compared nucleotide diversity of two lineages of weedy rice to their crop progenitors and identified several very strong peaks around regions of low relative diversity, indicating areas of the genome that have likely experienced a selective sweep during the evolution of weediness.

Alternatively, adaptation may also be detectable as elevated differentiation between populations at particular loci, rather than as a loss of diversity (particularly for soft sweeps). Any test statistic that quantifies population divergence can be calculated for each locus or region of the genome and used to identify outlier loci putatively under selection, though F_{ST} is the most commonly used metric for these types of analyses (Thornton and Jensen 2007; Porto-Neto et al. 2013; Tiffin and Ross-Ibarra 2014). Neutral expectations and thresholds for identifying outliers must be established by defining expected relationships among diverging populations. Approaches to this include simulating a specific demographic history [e.g., as in FDIST2 (Beaumont and Nichols 1996), Arlequin (Excoffier et al. 2009b)] or assuming divergence from a single common ancestral genepool [such as might be appropriate for direct source-colonist comparisons, e.g., as in BayeScan (Foll and Gaggiotti 2008)]. These methods are sensitive to having identified accurate population structure and are prone to false-positive from bottlenecks and rapid range expansions (Narum and Hess 2011; de Villemereuil et al. 2014; Poh et al. 2014). New methods that infer the history of coancestry between populations show particular promise for addressing some of these issues (Lotterhos and Whitlock 2014).

Finally, candidate loci underlying adaptive variation can also be identified as those that correlate particularly strongly with environmental variables among

colonizing populations, source populations, or both. This can be done using genome-environment association methods, which test for correlations between allele frequencies and ecologically relevant variables, ideally after taking into account population structure (Joost et al. 2007; Coop et al. 2010; Frichot et al. 2013, 2015; Hoban et al. 2016). These methods generally rely on identifying loci with strong linear relationships with environmental variables [e.g., as in BayEnv (Coop et al. 2010) and LFMM (Frichot et al. 2013)], and recent advances have focused on resolving issues of population structure and allele frequency estimation (Günther and Coop 2013; de Villemereuil et al. 2014; Lotterhos and Whitlock 2014; de Villemereuil and Gaggiotti 2015). A very different approach was recently suggested by Fitzpatrick and Keller (2015), who leverage advances in nonlinear community modeling approaches [Generalized Dissimilarity modeling (Ferrier et al. 2007) and Gradient Forests machine learning (Ellis et al. 2012b)] to identify allele-environment relationships rather than species-environment relationships. This latter approach is especially intriguing, because it offers the possibility of identifying nonlinear and threshold adaptations to the environment, which are likely to be common in nature. For example, the authors identified a threshold response of circadian clock alleles to temperature gradients in balsam poplar (*Populus balsamifera*) (Fitzpatrick and Keller 2015).

Several studies have used selection scans to identify candidate loci for adaptation during colonization of a new range, generating insights into the types of genetic variation involved and their potential ecological functions. For example, Puzey and Vallejo-Marín (2014) used whole genome resequencing data to scan for shifts in the site frequency spectrum to detect positive selection in introduced populations of monkeyflower (*Mimulus guttatus*). Regions putatively under selection were associated with flowering time and abiotic and biotic stress tolerance and included regions associated with a chromosomal inversion polymorphism between the native and introduced range. Zenni and Hoban (2015) scanned for loci with high divergence during range expansion of loblolly pine (*Pinus taeda*), using a SNP assay of previously identified polymorphic loci. They identified 25 loci with outlying divergence, most of which were unique to different range expanding populations, and 25% of which were also associated with a climate gradient. As mentioned above regarding genetic drift, White et al. (2013) used a different type of scan for divergent loci in the invasive black vole (*Myodes glareolus*). The authors estimated covariance in allele frequencies (using RADseq) among parallel transects from the core to the edge of the current expansion. This approach will detect loci that have risen to high frequency in all of the edge populations, reflecting parallel adaptation to the same selective environment, and they find support for candidate loci with potential effects on immunological functions and behavior. Their method is particularly notable for taking into account the potential for allele surfing during range expansion, assuming that there is no gene flow connecting edge populations.

Finally, Vandepitte et al. (2014) scanned for high divergence outliers in the introduced range of Pyrenean rocket (*Sisymbrium austriacum* subsp. *chrysanthum*) using SNPs generated through RADseq. Several candidate loci aligned to transcriptome sequences and were associated with flowering phenology. A particularly

powerful aspect of this analysis was the inclusion of historical specimens, an approach available to many studies of recent colonization. Candidate loci in this study showed increasing divergence over time in historical specimens, supporting the hypothesis of response to selection in the colonized range, rather than a founder effect as the source of allele frequency change.

6 Mutation: The Genetic Basis of Adaptive Variation

Whether adaptations often derive from new mutations or standing genetic variation is a major open question in evolutionary biology (Hermisson and Pennings 2005). It seems particularly likely that rapid evolution of colonizing populations will occur most often where there is standing adaptive variation from the source region present in the founders (Barrett and Schluter 2008), though colonizing populations also provide many opportunities for new mutations to arise and become targets of selection (Dlugosch et al. 2015a; Exposito-Alonso et al. 2018). Studies that identify both source location(s) and the alleles that have contributed to adaptation in colonizing populations should shed light on the importance of standing variation vs. new mutations, though identification of standing variation can be obfuscated by limited sampling and accurate determination of homology (Barrett and Schluter 2008).

Population genomic approaches have been the primary avenue for identifying the mutations underlying adaptation in colonizing species (Fig. 4), particularly outside of model systems with well-studied mutation lines. Species that have been involved in colonization events over relatively recent periods of time, such that colonizing populations have not diverged into reproductively isolated species, are in fact especially well-suited to these approaches (Dlugosch et al. 2015a). These species are likely to include adaptive variation within and among extant populations across the range, facilitating genetic mapping, the identification of current targets of selection, and phenotypic observations of genetic variants in native and invaded environments. Ecologically important loci can be identified through a combination of "top-down" (forward genetic) and "bottom-up" ("reverse ecology" and candidate gene) genomic tools (Fig. 4). A top-down genetic approach begins with phenotypic traits that are known to vary between colonizing and source genotypes and are thought to be relevant to adaptation. Association between traits and loci responsible for them can be determined using a genome-wide association study (GWAS; e.g., Hamilton et al. 2015) or quantitative trait locus (QTL; e.g., Linde 2001) mapping population to correlate allelic states with phenotypic differences. Alternatively and complementary, in bottom-up approaches, loci of interest may come from candidate genes showing patterns of divergence or selection in colonizing populations (known as "reverse ecology"; Li et al. 2008) or a priori candidate loci thought to be important from work in model systems (e.g., Krieger and Ross 2002; Nachman et al. 2003; Mueller et al. 2014). Candidate loci may suggest their associated phenotypes based on information from model systems, which can then be tested with GWAS or QTL

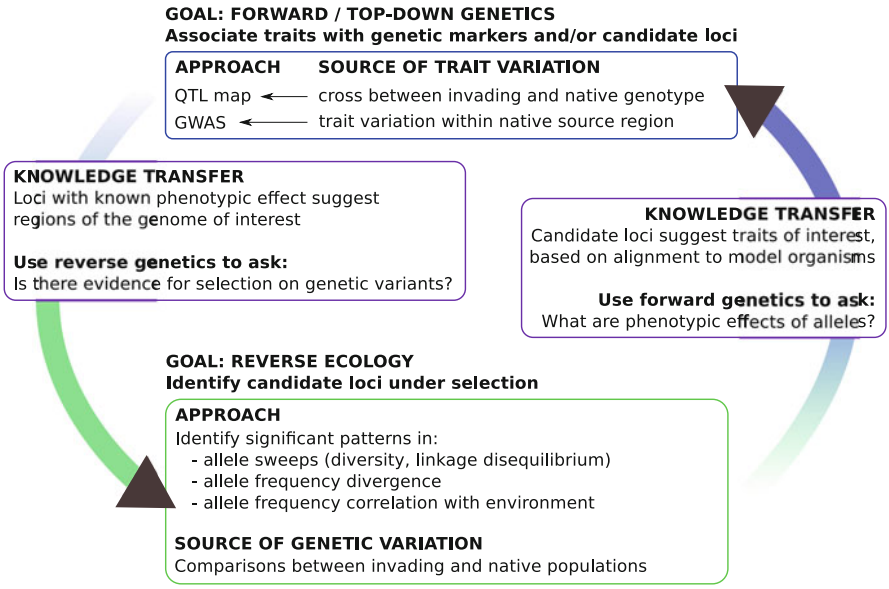

Fig. 4 Approaches to identifying the genetic basis of ecologically important trait variation. Top-down approaches (forward genetics) identify regions of the genome associated with specific traits, while reverse ecology approaches identify regions of the genome that appear to be under selection. The two approaches are complementary and can be used in concert to identify regions with known phenotypic effects that appear to be under selection. Figure modified with permission from Dlugosch et al. (2015a)

approaches, or increasingly through targeted allele substitution experiments (e.g., CRISPR; Hwang et al. 2013).

Once a locus that affects adaptive variation has been identified, the distribution of this variation can be observed across colonization events and its potential for impacts on colonization studied. There are several different categories of genetic variants that may be particularly relevant in the context of colonization:

1. *Large vs. small effect loci*. Quantitative traits can vary as a result of allelic variation at loci with large individual effects and at loci with small individual effects on the phenotype. Mendelian traits vary as a result of alleles at one or a few loci, which therefore have large effects on the phenotype by definition. The differences between large and small effect loci are important to the study of colonization for several reasons. Beneficial large effect mutations should be less likely than small effect mutations to be lost to genetic drift, because they experience stronger selection (Kimura 1985; Orr 1998). Large effect mutations can also facilitate jumps across low fitness valleys in adaptive landscapes, with large resulting effects on ecology (Wright 1932; Whitlock et al. 1995; de Visser and Krug 2014). On the other hand, traits governed by large effect loci will also be the most susceptible to changes of variation due to genetic drift, or increases in variation due to gene flow (Fig. 5; Baker and Stebbins 1965; Dlugosch et al. 2015a).

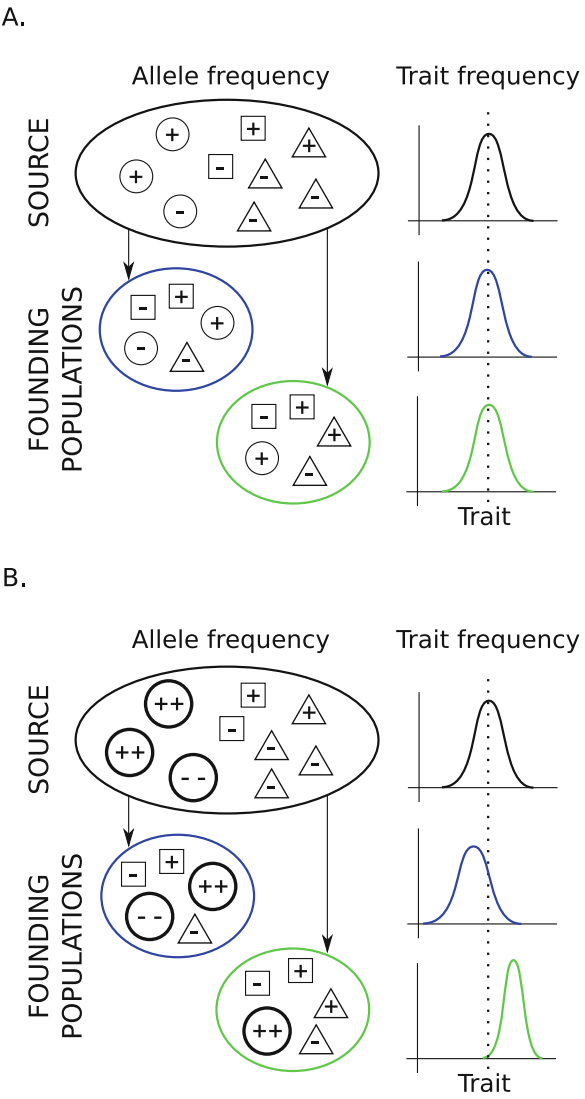

Fig. 5 The genetic architecture of ecologically important traits will shape how population bottlenecks during colonization impacts quantitative trait variation. Allelic variants (acting to increase [+] or decrease [−] the trait value) are shown in proportion to their frequency in a population. (**a**) Traits governed by many loci of small effect are expected to change little in mean or variance. (**b**) Traits that include a locus of large effect may shift in both mean and variance in response to either fixation or frequency shifts at large effect loci. Figure reprinted with permission from Dlugosch et al. (2015a)

Therefore, a large effect allele can amplify the impacts of genetic drift and gene flow on a colonizing phenotype and may therefore have more potential to alter establishment and/or spread (though certainly adaptation can still occur through evolution at loci of small effect; Olson-Manning et al. 2012). Large effect loci appear to be common in natural populations (Louthan and Kay 2011; Olson-Manning et al. 2012), and so they are likely to play a significant role in colonization genetics (Dlugosch et al. 2015a).

2. *Chromosomal inversions*. Chromosomal inversions were some of the first mutations associated with adaptive variation, including in colonizing populations

(Carson 1965; Dobzhansky 1965). These structural rearrangements of chromosomes can place particular loci in close physical proximity and suppress recombination, preserving coadapted gene complexes and/or making a larger effect locus out of multiple loci of smaller effect, which may be particularly advantageous during range expansion (Hoffmann and Rieseberg 2008; Yeaman 2013; Kirkpatrick and Barrett 2015). Both the invasion of Australia by *Drosophila melanogaster* (Hoffmann and Weeks 2007) and the invasion of the Americas by *D. subobscura* (Prevosti et al. 1988; Pascual et al. 2007) show evidence of rapid adaptation in chromosomal inversion frequencies. Three-spined sticklebacks (*Gasterosteus aculeatus*) have an inversion (and several other large effect loci) that contributes to the repeated evolution of freshwater and benthic forms during their post-Pleistocene invasion of freshwater lakes (Jones et al. 2012).

3. *Copy number variants*. Copy number variants (i.e., gene duplications) are now known to be one of the major forms of mutation differentiating closely related species and individuals of the same species (Lynch and Conery 2000; Freeman et al. 2006). Copy number changes occur frequently and appear much more likely than other types of mutations to avoid deleterious effects, because they are copies of existing, functioning genes (Kondrashov 2012; Hirase et al. 2014; Żmieńko et al. 2014). These features suggest that copy number variants could be a major source of both standing variation and new mutations that are beneficial for colonizing species. This area is not well-studied to date, but appropriate genome resequencing data should be increasingly obtainable for non-model organisms (Demuth and Hahn 2009; Tiffin and Ross-Ibarra 2014).
4. *Genome size and transposable element variation*. Genome size has been associated with differences in the rate of DNA replication and in cell size, suggesting the potential for developmental and physiological effects of this kind of genomic variation (Beaulieu et al. 2008, and references therein). Genome variation has been studied most extensively in plants, which span three orders of magnitude in genome size (Tenaillon et al. 2010). In members of this group, genome size has been positively correlated with seed size, minimum generation time, and cell size and negatively correlated with relative growth rate (Grotkopp et al. 2004). Consistent with the idea that colonizers might benefit from fast development times (Baker 1965), plants with small genome sizes are relatively overrepresented among weedy and invasive taxa (Kuester et al. 2014). A recent study of the invasive slender wild oat (*Avena barbata*) found smaller median genome sizes in invading genotypes than in those from the native range (Crosby et al. 2014). Variation in genome size is typically the result of variation in transposable element (TE) content (Tenaillon et al. 2010). TEs have the potential to expand or contract very quickly and therefore may be a key source of variation during colonization (Stapley et al. 2015). In particular, TEs may proliferate in response to conditions often encountered during colonization, including environmental stress and hybridization (Wessler 1996; Kashkush et al. 2003; Grandbastien et al. 2005; Ungerer et al. 2006; Maumus et al. 2009). TEs can have phenotypic effects both through genome size and through insertions into functional regions, and they have been associated with adaptation during range expansions in

Drosophila melanogaster (Aminetzach et al. 2005); invasive knotweed, *Fallopia japonica* (Richards et al. 2012); house sparrow, *Passer domesticus* (Schrey et al. 2012); ants, *Cardiocondyla obscurior* (Schrader et al. 2014); and mosquito, *Toxorhynchites amboinensis* (Zhou et al. 2014). Population genomic information about genome size can be readily obtained using flow cytometry (Doležel et al. 1998), and information about TE variation can be obtained through low coverage genome sequencing (Straub et al. 2012).

5. *Polyploidy*. Colonizing genotypes might benefit from the duplication of entire genomes via polyploidy (Stebbins 1985; te Beest et al. 2012; Welles and Ellstrand 2016). While polyploid formation can involve a large increase in genome size, evidence suggests that polyploids have a positive association with colonization, independent of negative associations between genome size and colonization within ploidy (Pandit et al. 2014). Polyploids may benefit from additional copies of functional regions of the genome as well as novel gene combinations in the case of allopolyploid (hybrid) lineages (te Beest et al. 2012; Pandit et al. 2014). Recent studies have suggested that polyploids may have greater phenotypic flexibility in gene expression in response to environmental differences (Kondrashov 2012; Mattenberger et al. 2017; Mutti et al. 2017), a characteristic which could benefit colonizers experiencing new environments in some cases (Huang and Agrawal 2016; Lohman et al. 2017; Wellband and Heath 2017).

A growing number of studies have identified the genetic basis of phenotypic divergence of colonizing populations (Bock et al. 2015). QTL mapping studies have identified major-effect as well as minor-effect loci influencing the propensity of Johnsongrass (*Sorghum halepense*) to produce asexually via rhizomes (Paterson et al. 1995); flowering time differences among ecotypes of the invasive plant shepherd's purse (*Capsella bursa-pastoris*; Linde 2001); a variety of traits involved in range expansion of the sunflower, *Helianthus annuus texanus* (Whitney et al. 2015); and morphological changes involved in the colonization of freshwater lakes by sticklebacks, *Gasterosteus aculeatus* (Jones et al. 2012). Hamilton et al. (2015) map the genetic basis of fitness using a GWAS approach and compare the phenotypic effects of loci associated with adaptation to environment in the native range of the model *Arabidopsis thaliana* between the native and introduced ranges. Candidate gene approaches have successfully revealed adaptive evolutionary changes in coat color in deer mice (*Peromyscus maniculatus*) colonizing new habitats (Linnen et al. 2013) and social recognition in invasive supercolonies of both the fire ant (*Solenopsis invicta*) and the Argentine ant (*Linepithema humile*; Tsutsui et al. 2000, 2003; Krieger and Ross 2002). In the case of the ants, the loss of variation at these loci during founder events has resulted in decreased conspecific aggression and increased invasiveness, providing some of the most famous examples of the effects of genetic bottlenecks on colonization and invasion.

Many of these studies suggest an important role for standing variation of large effect mutations. As we accumulate studies of the genetic basis of colonizer phenotypes, it may be useful to consider whether certain types of traits are more likely to

have standing variation at large effect loci, influencing the types of traits that might evolve rapidly during colonization. For example, a recent review of QTL studies in plants concluded that large effect loci have been found more often in traits governing biotic interactions than in traits associated with adaptation to abiotic conditions (Louthan and Kay 2011). If true, then this might suggest that there is greater potential for adaptation in some of the very traits hypothesized to be most important to the success of colonizing populations (Keane and Crawley 2002).

7 Conclusions and Future Perspectives

Recent advances in sequencing and population genomics have begun to address many open questions in the biology of colonization and invasion. Reconstructing the introduction history of invasive species using demographic inference models that allow for explicit modeling of genetic bottlenecks, admixture between independent introductions, serial introductions, and unsampled populations has allowed for an increased understanding of patterns of colonization and has provided the basis for making appropriate comparisons to source populations in further evolutionary studies. Recent population genomic studies have begun to tease apart evolution occurring due to founder effects, genetic drift, and gene flow associated with both initial founder events and range expansion. This knowledge is improving our ability to identify loci that are likely targets of positive selection during colonization and to elucidate the genetic basis of adaptive evolution in colonizers.

Determining what loci are under selection gives further traction on several important issues surrounding how evolution might alter the fates of colonizers. Once candidate loci are identified, it is possible to ask whether genetic bottlenecks or other stochastic processes have altered variation available for adaptation, whether adaptive variants have introgressed through admixture or hybridization, and whether loci involved in adaptation are present as standing variation in source populations or whether they might represent critical new mutations. Connecting loci under selection with their phenotypic effects further offers opportunities to understand the type of traits that are shaping colonizer ecology, such as biotic interactions or climatic gradients. Colonizers present both opportunities and challenges in this area, because selection and divergence may be recent and more easily detected and studied, but these species will also be affected by complex nonequilibrium demography, including genetic bottlenecks and allele surfing, that may obscure responses to selection.

The next step forward for population genomics of colonizing and invasive species will be to link shifts in gene frequencies and other population genetic metrics with shifts in evolutionary ecology associated with evolution during expansion (Dlugosch et al. 2015a). Making these strides requires broad collections of genotypes from both the source and colonized ranges, characterization of genomic variation in these genotypes, detailed abiotic and biotic environmental data from habitats across the range, and quantification of phenotype and fitness in these environments. While such combinations of data are not trivial to assemble, the studies highlighted in this

chapter already bring together much of this information. Further, new techniques are becoming accessible for non-model colonizers. For example, ultra-long read single-molecule sequencing is facilitating efficient de novo whole genome assembly, identification of structural variants, and phasing of haplotype information in highly heterozygous outbred genotypes (e.g., Oxford Nanopore; Jain et al. 2018; Michael et al. 2018). Additionally, gene editing technology such as CRISPR/Cas9 (Bortesi and Fischer 2015) combined with forward and reverse approaches to identifying loci of interest (Fig. 4) will advance our ability to connect genotypic variation to its consequences for adaptation in ecologically relevant traits. Indeed, studies that link population genomics with population ecology promise to fundamentally advance our understanding of how ecology might rapidly evolve during the nearly ubiquitous process of colonization in the evolutionary history of species.

Acknowledgments The authors thank B. S. Barker, F. A. Cang, and members of the Dlugosch lab for helpful discussion regarding the information in this chapter. Support was provided by USDA grant #2015-67013-23000 and NSF grant #1550838 to KMD.

References

Allendorf FW, Lundquist LL. Introduction: population biology, evolution, and control of invasive species. Conserv Biol. 2003;17(1):24–30.

Aminetzach YT, Macpherson JM, Petrov DA. Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. Science. 2005;309(5735):764–7.

Avise JC. Molecular markers, natural history and evolution. New York: Springer; 1994.

Baker HG. Self-compatibility and establishment after "long-distance" dispersal. Evolution. 1955;9(3):347–9.

Baker HG. Characteristics and modes of origin of weeds. In: Baker HG, Stebbins GL, editors. The genetics of colonizing species. New York: Academic; 1965.

Baker HG, Stebbins GL, editors. The genetics of colonizing species. New York: Academic; 1965.

Balanyá J, Oller JM, Huey RB, Gilchrist GW, Serra L. Global genetic change tracks global climate warming in *Drosophila subobscura*. Science. 2006;313(5794):1773–5.

Barker BS, Andonian K, Swope SM, Luster DG, Dlugosch KM. Population genomic analyses reveal a history of range expansion and trait evolution across the native and invaded range of yellow starthistle (*Centaurea solstitialis*). Mol Ecol. 2017a;26(4):1131–47.

Barker BS, Cocio JE, Anderson SR, Braasch J, Cang FA, Gillette HD, et al. The prevalence and benefits of admixture during species invasions: a role for epistasis? [Internet]. bioRxiv. 2017b. p. 139709. http://biorxiv.org/content/early/2017/05/18/139709. Accessed 21 May 2017.

Barrett RDH, Schluter D. Adaptation from standing genetic variation. Trends Ecol Evol. 2008;23(1):38–44.

Barrett SCH, Husband BC, Brown AHD, Clegg MT, Kahler AL, Weir BS. The genetics of plant migration and colonization. Sunderland: Sinauer Associates; 1990. p. 254–77.

Barrett SCH, Colautti RI, Dlugosch KM, Rieseberg LH, editors. Invasion genetics: the Baker and Stebbins legacy. Chichester: Wiley; 2017.

Beaulieu JM, Leitch IJ, Patel S, Pendharkar A, Knight CA. Genome size is a strong predictor of cell size and stomatal density in angiosperms. New Phytol. 2008;179(4):975–86.

Beaumont MA, Nichols RA. Evaluating loci for use in the genetic analysis of population structure. Proc Biol Sci. 1996;263(1377):1619–26.

Berg JJ, Coop G. A population genetic signal of polygenic adaptation. PLoS Genet. 2014;10(8): e1004412.
Bertorelle G, Benazzo A, Mona S. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol Ecol. 2010;19(13):2609–25.
Blackburn T, Lockwood JL, Cassey P. The influence of numbers on invasion success. Mol Ecol. 2015;24(9):1942–53.
Bock DG, Caseys C, Cousens RG, Hahn MA, Heredia SM, Hubner S, et al. What we still don't know about invasion genetics. Mol Ecol. 2015;24:2277–97.
Bortesi L, Fischer R. The CRISP R/Cas9 system for plant genome editing and beyond. Biotechnol Adv. 2015;33(1):41–52.
Braverman JM, Hedson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 1995;140(2):183–96.
Briskie JV, Mackintosh M. Hatching failure increases with severity of population bottlenecks in birds. PNAS. 2004;101(2):558–61.
Brumfield RT, Beerli P, Nickerson DA, Edwards SV. The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol. 2003;18(5):249–56.
Carroll SP, Dingle H, Famula TR. Rapid appearance of epistasis during adaptive divergence following colonization. Proc Biol Sci. 2003;270:S80–3.
Carson HL. Chromosomal morphism in geographically widespread species of *Drosophila*. In: Baker HG, Stebbins GL, editors. The genetics of colonizing species. New York: Academic; 1965.
Colautti RI, Barrett SCH. Rapid adaptation to climate facilitates range expansion of an invasive plant. Science. 2013;342(6156):364–6.
Colautti RI, Lau JA. Contemporary evolution during invasion: evidence for differentiation, natural selection, and local adaptation. Mol Ecol. 2015;24(9):1999–2017.
Colautti RI, Ricciardi A, Grigorovich IA, MacIsaac HJ. Is invasion success explained by the enemy release hypothesis? Ecol Lett. 2004;7(8):721–33.
Colautti RI, Maron JL, Barrett SCH. Common garden comparisons of native and introduced plant populations: latitudinal clines can obscure evolutionary inferences. Evol Appl. 2009;2 (2):187–99.
Coop G, Witonsky D, Di Rienzo A, Pritchard JK. Using environmental correlations to identify loci underlying local adaptation. Genetics. 2010;185(4):1411–23.
Cornuet J-M, Santos F, Beaumont MA, Robert CP, Marin J-M, Balding DJ, et al. Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics. 2008;24(23):2713–9.
Cox GW. Alien species and evolution. Washington: Island Press; 2004.
Crawford KM, Whitney KD. Population genetic diversity influences colonization success. Mol Ecol. 2010;19(6):1253–63.
Cristescu ME. Genetic reconstructions of invasion history. Mol Ecol. 2015;24(9):2212–25.
Crosby K, Stokes TO, Latta RG. Evolving California genotypes of Avena barbata are derived from multiple introductions but still maintain substantial population structure. PeerJ. 2014;2:e633.
Csilléry K, Blum MGB, Gaggiotti OE, François O. Approximate Bayesian Computation (ABC) in practice. Trends Ecol Evol. 2010;25(7):410–8.
de Villemereuil P, Gaggiotti OE. A new FST-based method to uncover local adaptation using environmental variables. Methods Ecol Evol. 2015;6(11):1248–58.
de Villemereuil P, Frichot É, Bazin É, François O, Gaggiotti OE. Genome scan methods against more complex models: when and how much should we trust them? Mol Ecol. 2014;23 (8):2006–19.
de Visser JAGM, Krug J. Empirical fitness landscapes and the predictability of evolution. Nat Rev Genet. 2014;15(7):480–90.
Demuth JP, Hahn MW. The life and death of gene families. BioEssays. 2009;31(1):29–39.
Dlugosch KM, Parker IM. Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. Mol Ecol. 2008;17(1):431–49.

Dlugosch KM, Anderson SR, Braasch J, Cang FA, Gillette HD. The devil is in the details: genetic variation in introduced populations and its contributions to invasion. Mol Ecol. 2015a;24:2095–111.
Dlugosch KM, Cang FA, Barker BS, Andonian K, Swope SM, Rieseberg LH. Evolution of invasiveness through increased resource use in a vacant niche. Nat Plants. 2015b;1(6):15066.
Dobzhansky T. "Wild" and "domestic" species of *Drosophila*. In: Baker HG, Stebbins GL, editors. The genetics of colonizing species. New York: Academic; 1965.
Doležel J, Greilhuber J, Lucretti S, Meister A, Lysák MA, Nardi L, et al. Plant genome size estimation by flow cytometry: inter-laboratory comparison. Ann Bot. 1998;82(suppl 1):17–26.
Domingues VS, Poh Y-P, Peterson BK, Pennings PS, Jensen JD, Hoekstra HE. Evidence of adaptation from ancestral variation in young populations of beach mice. Evolution. 2012;66 (10):3209–23.
Drake JM. Heterosis, the catapult effect and establishment success of a colonizing bird. Biol Lett. 2006;2(2):304–7.
Edmonds CA, Lillie AS, Cavalli-Sforza LL. Mutations arising in the wave front of an expanding population. Proc Natl Acad Sci U S A. 2004;101(4):975–9.
Ellis EC, Antill EC, Kreft H. All is not loss: plant biodiversity in the anthropocene. PLoS One. 2012a;7(1):e30535.
Ellis N, Smith SJ, Pitcher CR. Gradient forests: calculating importance gradients on physical predictors. Ecology. 2012b;93(1):156–68.
Ellstrand NC, Schierenbeck KA. Hybridization as a stimulus for the evolution of invasiveness in plants? Proc Natl Acad Sci U S A. 2000;97(13):7043–50.
Elton CS. The ecology of invasions by animals and plants. Chicago: University of Chicago Press; 1958.
Emerson BC, Paradis E, Thébaud C. Revealing the demographic histories of species using DNA sequences. Trends Ecol Evol. 2001;16(12):707–16.
Eriksen RL, Hierro JL, Eren Ö, Andonian K, Török K, Becerra PI, et al. Dispersal pathways and genetic differentiation among worldwide populations of the invasive weed *Centaurea solstitialis* L. (Asteraceae). PLoS One. 2014;9(12):e114786.
Estoup A, Guillemaud T. Reconstructing routes of invasion using genetic data: why, how and so what? Mol Ecol. 2010;19:4113–30.
Excoffier L, Ray N. Surfing during population expansions promotes genetic revolutions and structuration. Trends Ecol Evol. 2008;23(7):347–51.
Excoffier L, Foll M, Petit RJ. Genetic consequences of range expansions. Annu Rev Ecol Evol Syst. 2009a;40(1):481–501.
Excoffier L, Hofer T, Foll M. Detecting loci under selection in a hierarchically structured population. Heredity. 2009b;103(4):285–98.
Exposito-Alonso M, Becker C, Schuenemann VH, Reiter E, Setzer C, Slovak R, Brachi B, Hagmann J, Grimm DG, Chen J, Busch W, Bergelson J, Ness RW, Weigel D. PLoS Genet. 2018;14(2):e100715.
Facon B, Hufbauer RA, Tayeh A, Loiseau A, Lombaert E, Vitalis R, et al. Inbreeding depression is purged in the invasive insect Harmonia axyridis. Curr Biol. 2011;21(5):424–7.
Falush D, van Dorp L, Lawson D. A tutorial on how (not) to over-interpret STRUCTURE/ ADMIXTURE bar plots [Internet]. bioRxiv. 2016. http://biorxiv.org/content/early/2016/07/28/066431.abstract. Accessed 13 May 2017.
Fay JC, Wu C-I. Hitchhiking under positive Darwinian selection. Genetics. 2000;155(3):1405–13.
Ferrier S, Manion G, Elith J, Richardson K. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Divers Distrib. 2007;13(3):252–64.
Fitzpatrick MC, Keller SR. Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. Ecol Lett. 2015;18(1):1–16.

Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics. 2008;180(2):977–93.
François O, Martins H, Caye K, Schoville SD. Controlling false discoveries in genome scans for selection. Mol Ecol. 2016;25(2):454–69.
Frankham R. Resolving the genetic paradox in invasive species. Heredity. 2005;94(4):385.
Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: new insights in genome diversity. Genome Res. 2006;16(8):949–61.
Frichot E, Schoville SD, Bouchard G, François O. Testing for associations between loci and environmental gradients using latent factor mixed models. Mol Biol Evol. 2013;30(7):1687–99.
Frichot E, Schoville SD, de Villemereuil P, Gaggiotti OE, François O. Detecting adaptive evolution based on association with ecological gradients: orientation matters! Heredity. 2015;115 (1):22–8.
García-Ramos G, Rodríguez D. Evolutionary speed of species invasions. Evolution. 2002;56 (4):661–8.
Gerlach JD. How the West was lost: reconstructing the invasion dynamics of yellow starthistle and other plant invaders of western rangelands and natural areas. Proc Calif Exotic Pest Plant Counc Symp. 1997;3:67–72.
Gompert Z. A continuous correlated beta process model for genetic ancestry in admixed populations. PLoS One. 2016;11(3):e0151047.
Gompert Z, Alex Buerkle C. introgress: a software package for mapping components of isolation in hybrids. Mol Ecol Resour. 2010;10(2):378–84.
Gompert Z, Buerkle CA. bgc: software for Bayesian estimation of genomic clines. Mol Ecol Resour. 2012;12(6):1168–76.
Gompert Z, Buerkle CA. Analyses of genetic ancestry enable key insights for molecular ecology. Mol Ecol. 2013;22(21):5278–94.
Gompert Z, Mandeville EG, Buerkle CA. Using genomic data in the analysis of hybrid zones [Internet]. Annu Rev Ecol Evol Syst. 2016. http://annualreviews.org/doi/abs/10.1146/annurev-ecolsys-110316-022652.
Graciá E, Botella F, Anadón JD, Edelaar P, Harris DJ, Giménez A. Surfing in tortoises? Empirical signs of genetic structuring owing to range expansion. Biol Lett. 2013;9(3):20121091.
Gralka M, Stiewe F, Farrell F, Möbius W, Waclaw B, Hallatschek O. Allele surfing promotes microbial adaptation from standing variation. Ecol Lett. 2016;19(8):889–98.
Grandbastien M-A, Audeon C, Bonnivard E, Casacuberta JM, Chalhoub B, Costa A-PP, et al. Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. Cytogenet Genome Res. 2005;110(1–4):229–41.
Gray AJ, Mack RN, Harper JL, Usher MB, Joysey K, Kornberg H. Do invading species have definable genetic characteristics? Philos Trans Biol Sci. 1986;314(1167):655–74.
Grotkopp E, Rejmánek M, Sanderson MJ, Rost TL. Evolution of genome size in pines (*Pinus*) and its life-history correlates: supertree analyses. Evolution. 2004;58(8):1705–29.
Guillemaud T, Beaumont MA, Ciosi M, Cornuet J-M, Estoup A. Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. Heredity. 2010;104(1):88–99.
Günther T, Coop G. Robust identification of local adaptation from allele frequencies. Genetics. 2013;195(1):205–20.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5(10):e1000695.
Hallatschek O, Hersen P, Ramanathan S, Nelson DR. Genetic drift at expanding frontiers promotes gene segregation. Proc Natl Acad Sci U S A. 2007;104(50):19926–30.
Hamilton J, Okada M, Korves T, Schmitt J. The role of climate adaptation in colonization success in Arabidopsis thaliana. Mol Ecol. 2015;24(9):2253–63.
Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 2005;169(4):2335–52.

Hirase S, Ozaki H, Iwasaki W. Parallel selection on gene copy number variations through evolution of three-spined stickleback genomes. BMC Genomics. 2014;15(1):735.
Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, et al. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. Am Nat. 2016;188 (4):379–97.
Hodgins KA, Bock DG, Hahn MA, Heredia SM, Turner KG, Rieseberg LH. Comparative genomics in the Asteraceae reveals little evidence for parallel evolutionary change in invasive taxa. Mol Ecol. 2015;24(9):2226–40.
Hoffmann AA, Rieseberg LH. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? Annu Rev Ecol Evol Syst. 2008;39:21–42.
Hoffmann AA, Weeks AR. Climatic selection on genes and traits after a 100 year-old invasion: a critical look at the temperate-tropical clines in *Drosophila melanogaster* from eastern Australia. Genetica. 2007;129(2):133–47.
Huang Y, Agrawal AF. Experimental evolution of gene expression and plasticity in alternative selective regimes. PLoS Genet. 2016;12(9):e1006336.
Huber CD, Nordborg M, Hermisson J, Hellmann I. Keeping it local: evidence for positive selection in Swedish Arabidopsis thaliana. Mol Biol Evol. 2014;31(11):3026–39.
Hufbauer RA. Biological invasions: paradox lost and paradise gained. Curr Biol. 2008;18(6): R246–7.
Hufbauer RA. Admixture is a driver rather than a passenger in experimental invasions. J Anim Ecol. 2017;86(1):4–6.
Hwang WY, Fu Y, Reyon D, Maeder ML, Kaini P, Sander JD, et al. Heritable and precise zebrafish genome editing using a CRISPR-Cas system. PLoS One. 2013;8(7):e68708.
Jabot F, Faure T, Dumoulin N. EasyABC: performing efficient approximate Bayesian computation sampling schemes using R. Methods Ecol Evol. 2013;4(7):684–7.
Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of the genome with ultra-long reads. Nat Biotechnol. 2018;36:338–45.
Jeschke JM, Strayer DL. Invasion success of vertebrates in Europe and North America. Proc Natl Acad Sci U S A. 2005;102(20):7198–202.
Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484(7392):55–61.
Joost S, Bonin A, Bruford MW, Després L, Conord C, Erhardt G, et al. A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. Mol Ecol. 2007;16(18):3955–69.
Kalinowski ST. The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. Heredity. 2011;106(4):625–32.
Kaplan NL, Hudson R, Lagley C. The "hitchhiking effect" revisited. Genetics. 1989;123 (4):887–99.
Kashkush K, Feldman M, Levy AA. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet. 2003;33(1):102–6.
Keane R, Crawley MJ. Exotic plant invasions and the enemy release hypothesis. Trends Ecol Evol. 2002;17(4):164–70.
Keller SR, Taylor DR. History, chance and adaptation during biological invasion: separating stochastic phenotypic evolution from response to selection. Ecol Lett. 2008;11(8):852–66.
Keller SR, Taylor DR. Genomic admixture increases fitness during a biological invasion. J Evol Biol. 2010;23(8):1720–31.
Kelly JK. A test of neutrality based on interlocus associations. Genetics. 1997;146(3):1197–206.
Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press; 1985. p. 367.
Kirkpatrick M, Barrett B. Chromosome inversions, adaptive cassettes, and the evolution of species' ranges. Mol Ecol. 2015;24:2046–55.

Knowles LL. Statistical phylogeography. Annu Rev Ecol Evol Syst. 2009;40:593–612.
Knowles LL, Maddison WP. Statistical phylogeography. Mol Ecol. 2002;11(12):2623–35.
Kolbe JJ, Glor RE, Rodríguez Schettino L, Lara AC, Larson A, Losos JB. Genetic variation increases during biological invasion by a Cuban lizard. Nature. 2004;431(7005):177–81.
Kondrashov FA. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc Biol Sci. 2012;279(1749):5048–57.
Koskella B. Research highlights for issue 7: the evolution of invasiveness. Evol Appl. 2015;8(7):633–4.
Krieger MJB, Ross KG. Identification of a major gene regulating complex social behavior. Science. 2002;295(5553):328–32.
Kuester A, Conner JK, Culley T, Baucom RS. How weeds emerge: a taxonomic and trait-based examination using United States data. New Phytol. 2014;202(3):1055–68.
Lavergne S, Molofsky J. Increased genetic variation and evolutionary potential drive the success of an invasive grass. Proc Natl Acad Sci U S A. 2007;104(10):3883–8.
Lee CE. Evolutionary genetics of invasive species. Trends Ecol Evol. 2002;17(8):386–91.
Lee CE, Gelembiuk GW. Evolutionary origins of invasive populations. Evol Appl. 2008;1(3):427–48.
Li L-F, Costello JC, Holloway AK, Hahn MW. "Reverse ecology" and the power of population genomics. Evolution. 2008;62(12):2984–94.
Li L-F, Jia Y, Caicedo AL, Olsen KM. Signatures of adaptation in the weedy rice genome. Nat Genet. 2017;49(5):811–4.
Linde M. Flowering ecotypes of *Capsella bursa-pastoris* (L.) Medik. (Brassicaceae) analysed by a cosegregation of phenotypic characters (QTL) and molecular markers. Ann Bot. 2001;87(1):91–9.
Linnen CR, Poh Y-P, Peterson BK, Barrett RDH, Larson JG, Jensen JD, et al. Adaptive evolution of multiple traits through multiple mutations at a single gene. Science. 2013;339(6125):1312–6.
Lohman B, Stutz W, Bolnick D. Gene expression stasis and plasticity following migration into a foreign environment [Internet]. bioRxiv. 2017. p. 121608. http://biorxiv.org/content/early/2017/03/28/121608. Accessed 20 May 2017.
Lopes JS, Balding D, Beaumont MA. PopABC: a program to infer historical demographic parameters. Bioinformatics. 2009;25(20):2747–9.
Lotterhos KE, Whitlock MC. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. Mol Ecol. 2014;23(9):2178–92.
Louthan AM, Kay KM. Comparing the adaptive landscape across trait types: larger QTL effect size in traits under biotic selection. BMC Evol Biol. 2011;11(1):60.
Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, et al. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. Mol Ecol Resour. 2017;17(2):142–52.
Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290(5494):1151–5.
Mattenberger F, Sabater-Muñoz B, Toft C, Fares MA. The phenotypic plasticity of duplicated genes in *Saccharomyces cerevisiae* and the origin of adaptations. G3. 2017;7(1):63–75.
Maumus F, Allen AE, Mhiri C, Hu H, Jabbari K, Vardi A, et al. Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. BMC Genomics. 2009;10:624.
Meudt HM, Clarke AC. Almost forgotten or latest practice? AFLP applications, analyses and advances. Trends Plant Sci. 2007;12(3):106–17.
Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat Commun. 2018;9:541.
Moreau C, Bhérer C, Vézina H, Jomphe M, Labuda D, Excoffier L. Deep human genealogies reveal a selective advantage to be on an expanding wave front. Science. 2011;334(6059):1148–50.
Mueller JC, Edelaar P, Carrete M, Serrano D, Potti J, Blas J, et al. Behaviour-related DRD4 polymorphisms in invasive bird populations. Mol Ecol. 2014;23(11):2876–85.

Mutti JS, Bhullar RK, Gill KS. Evolution of gene expression balance among homeologs of natural polyploids. G3. 2017;7(4):1225–37.
Nachman M, Hoekstra H, D'Agostino S, Kidwell M. The genetic basis of adaptive melanism in pocket mice. Proc Natl Acad Sci U S A. 2003;100(9):5268–73.
Narum SR, Hess JE. Comparison of FST outlier tests for SNP loci under selection. Mol Ecol Resour. 2011;11:184–94.
Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA. Genotyping-by-sequencing in ecological and conservation genomics. Mol Ecol. 2013;22(11):2841–7.
Nei M, Maruyama T, Chakraborty R. The bottleneck effect and genetic variability in populations. Evolution. 1975;29(1):1–10.
Nolte AW, Gompert Z, Buerkle CA. Variable patterns of introgression in two sculpin hybrid zones suggest that genomic isolation differs among populations. Mol Ecol. 2009;18(12):2615–27.
Novembre J. Pritchard, Stephens, and Donnelly on population structure. Genetics. 2016;204 (2):391–3.
Ochocki BM, Miller TEX. Rapid evolution of dispersal ability makes biological invasions faster and more variable. Nat Commun. 2017;8:14315.
Oleksyk TK, Smith MW, O'Brien SJ. Genome-wide scans for footprints of natural selection. Philos Trans Biol Sci. 2010;365(1537):185–205.
Olson-Manning CF, Wagner MR, Mitchell-Olds T. Adaptive evolution: evaluating empirical support for theoretical predictions. Nat Rev Genet. 2012;13(12):867–77.
Orr HA. Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data. Genetics. 1998;149(4):2099–104.
Pandit MK, White SM, Pocock MJO. The contrasting effects of genome size, chromosome number and ploidy level on plant invasiveness: a global analysis. New Phytol. 2014;203(2):697–703.
Pannell JR. Evolution of the mating system in colonizing plants. Mol Ecol. 2015;24(9):2018–37.
Parmesan C, Yohe G. A globally coherent fingerprint of climate change impacts across natural systems. Nature. 2003;421(6918):37–42.
Pascual M, Chapuis MP, Mestres F, Balanyà J, Huey RB, Gilchrist GW, et al. Introduction history of *Drosophila subobscura* in the New World: a microsatellite-based survey using ABC methods. Mol Ecol. 2007;16(15):3069–83.
Paterson AH, Schertz KF, Lin YR, Liu SC, Chang YL. The weediness of wild plants: molecular analysis of genes influencing dispersal and persistence of johnsongrass, *Sorghum halepense* (L.) Pers. Proc Natl Acad Sci U S A. 1995;92(13):6127–31.
Payseur BA, Rieseberg LH. A genomic perspective on hybridization and speciation. Mol Ecol. 2016;25(11):2337–60.
Peischl S, Excoffier L. Expansion load: recessive mutations and the role of standing genetic variation. Mol Ecol. 2015;24(9):2084–94.
Peischl S, Kirkpatrick M, Excoffier L. Expansion load and the evolutionary dynamics of a species range. Am Nat. 2015;185(4):E81–93.
Pierce AA, Zalucki MP, Bangura M, Udawatta M, Kronforst MR, Altizer S, et al. Serial founder effects and genetic differentiation during worldwide range expansion of monarch butterflies [Internet]. Proc Biol Sci. 2014;281(1797). https://doi.org/10.1098/rspb.2014.2230.
Poh Y-P, Domingues VS, Hoekstra HE, Jensen JD. On the prospect of identifying adaptive loci in recently bottlenecked populations. PLoS One. 2014;9(11):e110579.
Porto-Neto LR, Lee SH, Lee HK, Gondro C. Detection of signatures of selection using FST. In: Gondro C, van der Werf J, Hayes B, editors. Genome-wide association studies and genomic prediction, Methods in molecular biology. New York: Humana Press; 2013. p. 423–36.
Prevosti A, Ribo G, Serra L, Aguade M, Balana J, Monclus M, et al. Colonization of America by *Drosophila subobscura*: experiment in natural populations that supports the adaptive role of chromosomal-inversion polymorphism. Proc Natl Acad Sci U S A. 1988;85(15):5597–600.
Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 2009;5(6): e1000519.

Pritchard JK, Di Rienzo A. Adaptation – not by sweeps alone. Nat Rev Genet. 2010;11(10):665–7.
Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59.
Pudlo P, Marin J-M, Estoup A, Cornuet J-M, Gautier M, Robert CP. Reliable ABC model choice via random forests. Bioinformatics. 2016;32(6):859–66.
Puzey J, Vallejo-Marín M. Genomics of invasion: diversity and selection in introduced populations of monkeyflowers (Mimulus guttatus). Mol Ecol. 2014;23(18):4472–85.
Qi X, An H, Ragsdale AP, Hall TE, Gutenkunst RN, Chris Pires J, et al. Genomic inferences of domestication events are corroborated by written records in Brassica rapa [Internet]. Mol Ecol. 2017. https://doi.org/10.1111/mec.14131.
Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 2014;197(2):573–89.
Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A. 2005;102(44):15942–7.
Ramakrishnan AP, Musial T, Cruzan MB. Shifting dispersal modes at an expanding species' range margin. Mol Ecol. 2010;19(6):1134–46.
Richards CL, Schrey AW, Pigliucci M. Invasion of diverse habitats by few Japanese knotweed genotypes is correlated with epigenetic differentiation. Ecol Lett. 2012;15(9):1016–25.
Rius M, Darling JA. How important is intraspecific genetic admixture to the success of colonising populations? Trends Ecol Evol. 2014;29(4):233–42.
Sakai AK, Allendorf FW, Holt JS, Lodge DM, Molofsky J, With KA, et al. The population biology of invasive species. Annu Rev Ecol Syst. 2001;32:305–32.
Schrader L, Kim JW, Ence D, Zimin A, Klein A, Wyschetzki K, et al. Transposable element islands facilitate adaptation to novel environments in an invasive species. Nat Commun. 2014;5:5495.
Schrey AW, Coon CAC, Grispo MT, Awad M, Imboma T, McCoy ED, et al. Epigenetic variation may compensate for decreased genetic variation with introductions: a case study using house sparrows (Passer domesticus) on two continents. Genet Res Int. 2012;2012:979751.
Stapley J, Santure A, Dennis S. Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. Mol Ecol. 2015;24:2241–52.
Stebbins GL. Polyploidy, hybridization, and the invasion of new habitats. Ann Mo Bot Gard. 1985;72(4):824–32.
Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am J Bot. 2012;99(2):349–64.
Szűcs M, Melbourne BA, Tuff T, Hufbauer RA. The roles of demography and genetics in the early stages of colonization. Proc Biol Sci. 2014;281(1792):20141073.
Szűcs M, Melbourne BA, Tuff T, Weiss-Lehman C, Hufbauer RA. Genetic and demographic founder effects have long-term fitness consequences for colonising populations. Ecol Lett. 2017;20(4):436–44.
te Beest M, Le Roux JJ, Richardson DM, Brysting AK, Suda J, Kubesová M, et al. The more the better? The role of polyploidy in facilitating plant invasions. Ann Bot. 2012;109(1):19–45.
Tenaillon MI, Hollister JD, Gaut BS. A triptych of the evolution of plant transposable elements. Trends Plant Sci. 2010;15(8):471–8.
Thompson JN. Rapid evolution as an ecological process. Trends Ecol Evol. 1998;13(8):329–32.
Thornton KR, Jensen JD. Controlling the false-positive rate in multilocus genome scans for selection. Genetics. 2007;175(2):737–50.
Tiffin P, Ross-Ibarra J. Advances and limits of using population genetics to understand local adaptation. Trends Ecol Evol. 2014;29(12):673–80.
Tsutsui ND, Suarez AV, Holway DA, Case TJ. Reduced genetic variation and the success of an invasive species. Proc Natl Acad Sci U S A. 2000;97(11):5948–53.

Tsutsui ND, Suarez AV, Grosberg RK. Genetic diversity, asymmetrical aggression, and recognition in a widespread invasive species. Proc Natl Acad Sci U S A. 2003;100(3):1078–83.
Uller T, Leimu R. Founder events predict changes in genetic diversity during human-mediated range expansions. Glob Chang Biol. 2011;17(11):3478–85.
Ungerer MC, Strakosh SC, Zhen Y. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. Curr Biol. 2006;16(20):R872–3.
Vandepitte K, de Meyer T, Helsen K, van Acker K, Roldán-Ruiz I, Mergeay J, et al. Rapid genetic adaptation precedes the spread of an exotic plant species. Mol Ecol. 2014;23(9):2157–64.
Verhoeven KJF, Macel M, Wolfe LM, Biere A. Population admixture, biological invasions and the balance between local adaptation and inbreeding depression. Proc Biol Sci. 2011;278 (1702):2–8.
Vermeij GJ. Invasion as expectation: a historical fact of life. In: Sax DF, Stachowicz JJ, Gaines SD, editors. Species invasions: insights into ecology, evolution, and biogeography. Sunderland: Sinauer Associates; 2005. p. 315–39.
Waddington CH. Introduction to the symposium. In: Baker HG, Stebbins GL, editors. The genetics of colonizing species. New York: Academic; 1965.
Wagner NK, Ochocki BM, Crawford KM, Compagnoni A, Miller TEX. Genetic mixture of multiple source populations accelerates invasive range expansion. J Anim Ecol. 2017;86 (1):21–34.
Wang J. The computer program structure for assigning individuals to populations: easy to use but easier to misuse [Internet]. Mol Ecol Resour. 2016. https://doi.org/10.1111/1755-0998.12650.
Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinf. 2010;11:116.
Wegmann D, Kessner DE, Veeramah KR, Mathias RA, Nicolae DL, Yanek LR, et al. Recombination rates in admixed individuals identified by ancestry-based inference. Nat Genet. 2011;43 (9):847–53.
Wellband KW, Heath DD. Plasticity in gene transcription explains the differential performance of two invasive fish species [Internet]. Evol Appl. 2017. https://doi.org/10.1111/eva.12463.
Welles SR, Ellstrand NC. Rapid range expansion of a newly formed allopolyploid weed in the genus Salsola. Am J Bot. 2016;103(4):663–7.
Wessler SR. Plant retrotransposons: turned on by stress. Curr Biol. 1996;6(8):959–61.
White TA, Perkins SE, Heckel G, Searle JB. Adaptive evolution during an ongoing range expansion: the invasive bank vole (*Myodes glareolus*) in Ireland. Mol Ecol. 2013;22(11):2971–85.
Whitlock MC, Phillips PC, Moore FB, Tonsor SJ. Multiple fitness peaks and epistasis. Annu Rev Ecol Syst. 1995;26(1):601–29.
Whitney KD, Broman KW, Kane NC, Hovick SM, Randell RA, Rieseberg LH. Quantitative trait locus mapping identifies candidate alleles involved in adaptive introgression and range expansion in a wild sunflower. Mol Ecol. 2015;24(9):2194–211.
Williams JL, Kendall BE, Levine JM. Rapid evolution accelerates plant population spread in fragmented experimental landscapes. Science. 2016;353(6298):482–5.
Wright S. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. Proc Sixth Intl Congr Genetics. 1932;1:356–66.
Yeaman S. Genomic rearrangements and the evolution of clusters of locally adaptive loci. Proc Natl Acad Sci U S A. 2013;110:E1743–51.
Zenni RD, Hoban SM. Loci under selection during multiple range expansions of an invasive plant are mostly population specific, but patterns are associated with climate. Mol Ecol. 2015;24 (13):3360–71.
Zhou X, Rinker DC, Pitts RJ, Rokas A, Zwiebel LJ. Divergent and conserved elements comprise the chemoreceptive repertoire of the nonblood-feeding mosquito *Toxorhynchites amboinensis*. Genome Biol Evol. 2014;6(10):2883–96.
Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M. Copy number polymorphism in plant genomes. Theor Appl Genet. 2014;127(1):1–18.

Population Genomics of Crop Domestication: Current State and Perspectives

Philippe Cubry and Yves Vigouroux

Abstract Genomics has enabled access to unprecedented amounts of genomic and transcriptomic data. Studies of crop domestication have benefited from these datasets for deeper insights into when, where, and how crops were domesticated. Although genomics makes it possible to answer such questions, it also creates new technical and methodological challenges. Such large genomic and transcriptomic datasets provide the opportunity to advance from descriptive to hypothesis testing studies. Several model-based methods are now available to test hypotheses and to trace the history of crops. Studies of gene expression and of ancient DNA are new very active fields which hold great promise. Here, we review some key questions concerning crop domestication and discuss how genomics can help answer these questions and what interesting new approaches could be used in the future. As genomics data continue to become available, domestication studies will advance our knowledge not only of well-known domestication models, such as rice and maize, but also of other currently less widely studied crops. We will then be able to test general hypotheses associated with domestication across species.

Keywords Crop plants · Domestication · Evolution · Genomes · Inference of evolutionary history · Population genomics · Selection

1 Introduction

The transition to agricultural society is a key step in human history and evolution. The study of crop and animal domestication provides valuable information on where and when this transition took place. Studying domestication history also offers the opportunity to tackle adaptation at very short time scales. In the last decade, the production

P. Cubry · Y. Vigouroux (✉)
Institut de Recherche pour le développement, Université de Montpellier, Montpellier, France
e-mail: yves.vigouroux@ird.fr

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_48,

of huge genomic data has enabled a better understanding of the domestication processes that determined where, when, and how crops were domesticated. However, these analyses have not been exempt from controversy. For example, rice (*Oryza sativa*) domestication is considered to be the result of a single event by some (Molina et al. 2011; Huang et al. 2012; Choi and Purugganan 2018) and multiple domestication events by others (Civáň et al. 2015; Wang et al. 2018). Even though genomic data provide a wealth of information, understanding crop evolution requires elaborating and testing models that make it possible to proceed "from storytelling to story testing" (Gerbault et al. 2014). Fortunately, genomic data make such inferences possible. Here, we focus on key issues, methods, and questions concerning crop domestication. Population genomics offers an excellent opportunity to improve our understanding of domestication. The full strength of population genomics methods is still rarely used for the study of crop domestication. We hope this chapter will lead to their wider use.

We first review the particularity of crop genomic sequences and the nature of the variant we are looking at. Most of the inferences made about crop domestication rely on single nucleotide polymorphism (SNP) in the genomes, the quality of which is of paramount importance. How we call SNP might make our analysis useless. Very special care is required at this step, and we believe it to be a step that has been largely neglected up to now. We then review different methods used to infer the evolutionary history of crop domestication, along with their advantages and drawbacks. We believe using model-based hypothesis testing should become a standard approach in the field and, to this end, present the most promising advances in spatial modeling of crop domestication. Domestication is also associated with marked morphological changes. The genetic basis of domestication first benefited from the identification of quantitative trait loci (QTL) and more recently from genome selection scanning. We highlight what we have learned and in which direction the field is moving.

Finally, next-generation sequencing approaches make it possible to obtain ancient DNA gene expression data more easily and to investigate the methylation of DNA and histone. A lot of knowledge about plant domestication can be acquired using these very new techniques. A few recent studies produced very promising results, paving the way for major discoveries in the years to come.

2 Genomic Resources for Crop Domestication Studies

The number of the sequenced genomes available for the study of crop domestication has expanded considerably in the last 10 years. In June 2018, 606 sequence-assembled plant genomes were available in GenBank. However, the genomes are in different states of completeness and different stages of assembly. The difficulty involved in assembling plant genomes is directly linked to their particular complexity: (1) their high genetic diversity and (2) the high proportion of repetitive sequences. Genomic studies of domestication are now mainly associated with resequencing using either whole-genome resequencing (Wang et al. 2018) or partial sequencing of the genome using genotyping by sequencing (Elshire et al. 2011), sequencing of the expressed fraction of the genome using RNAseq (Bellucci et al. 2014; Sarah et al. 2017), or sequence capture (Mariac et al. 2014; da Fonseca et al.

2015). These approaches first require mapping reads to a reference genome and then SNP genotype calling. Depending on the method, the number and quality of SNPs will also vary (Berthouly-Salazar et al. 2016). In this chapter, we focus mainly on whole genome resequencing, but a large number of approaches can easily deal with partial sequencing of the genome.

2.1 Mapping Reads

Different softwares for mapping sequence read to a reference genome exist, for example, BWA (Li and Durbin 2009), SOAP2 (Li et al. 2009), BWA-MEM (Li 2013), Bowtie2 (Langmead and Salzberg 2012), Stampy (Lunter and Goodson 2011), and NGM (Sedlazeck et al. 2013). These mapping programs display different degrees of sensitivity to sequence divergence from the reference genome; for example, NGM and Stampy are better for reads that diverge from the reference than BWA (Lunter and Goodson 2011; Sedlazeck et al. 2013). Consequently for highly divergent species, like maize (*Zea mays*), some authors use both Bowtie and Stampy (Brandenburg et al. 2017). In addition, algorithms that allow high sequence divergence, like Stampy, can be very time-consuming. Choosing good mapping tools is important and depends on the expected diversity among the resequenced individuals. For some applications, using different mapping tools and assessing concordance have been suggested (Kofler et al. 2016).

2.2 SNP Calling: Probabilistic Approach (or Not)?

Partly due to the size of plant genomes, only a handful of plants with high genome coverage are available for genomic studies. For example, roughly twofold genome coverage was used for the study of domestication of Asian rice, *Oryza sativa* (Huang et al. 2012), while one study of maize domestication study used fivefold coverage (Hufford et al. 2012). Low genome coverage has a direct effect on heterozygote genotype calling. With an average depth of 1×, many SNPs in the genome would be covered once or twice, and consequently we would only be able to call one out of the two alleles of a heterozygous genotype. But the other more direct consequence of low genome sequence coverage is dealing with a high percentage of missing data and sequencing errors. To get around the problem, one first has to consider the uncertainty involved in genotype calling and decide which the best strategy is to use. One of the useful strategies for low genome coverage is to circumvent genotype calling and instead use an analysis based on genotype likelihood. During the process of genotype calling, one of the first steps after mapping is to calculate genotype likelihood. Considering a two-allele SNP (which most are), there are ten possible combinations (AA, AC, AG, AT, CC, CT, CG, GG, GG, TT) for a diploid individual. The first step in calling SNP is estimating the likelihood of these different genotypes. This leads to a genotype probability with a simple two-allele marker and a reference nucleotide 0 (could be A, T, C, G) to calculate the genotype

likelihood of genotypes 0/0, 0/1, and 1/1. The genotype calling could be clear-cut with a strong probability of a 0/0 genotype and almost zero for 0/1 and 1/1. But with low genome coverage, probability of two of the three genotypes might be quite high. If one has only one read and it harbors a nucleotide 1, then we do not know if the genotype is 1/1 or 0/1. So either we call the genotype or deal with the probabilistic nature of the genotype. High coverage genome data makes it possible to call SNP with strong clear-cut decision-making rules, but low genome coverage doesn't allow such an unambiguous decision. One strategy has, thus, been to directly consider the uncertainty in the genotype calls by conducting an analysis based on genotype likelihood rather than on the actual genotype. A set of methods has been developed to perform such analyses using ANGSD (Korneliussen et al. 2014; Durvasula et al. 2016). And other analyses can also be conducted using the same type of data. For example, using such an approach, one can assess deviation from the expected neutral diversity distribution using Tajima's D statistics or analysis of population structure with approach using NGSADMIX (Skotte et al. 2013). These strategies take the uncertainty for genotype calling directly into account in the analysis. One of the useful statistics that can be extracted using this type of probabilistic approach is the site frequency spectrum. Such a strategy is particularly useful in the case of low genome coverage.

2.3 *Chloroplast Genome Diversity Studies*

In a resequencing study, it is also possible to study the diversity not only of the nuclear genome but also of the whole chloroplast (Tong et al. 2016). Such a study could throw light on the origin and diffusion of the crop species. Specific pipelines are now available to minimize the noise associated with the identification of the SNP in the chloroplast genome (Scarcelli et al. 2016). This approach should be easy even with less well-known cultivated species (Moreira et al. 2016). Not limiting a study to the nuclear genome is a wise strategy for the study of domestication, as gene flow in pollen and seeds may have differed considerably during the evolutionary history of crops. The whole chloroplast genome analysis helped to disprove local domestication of tree gourd (*Cujete cujete*) in Amazonia (Moreira et al. 2016, 2017a, b). These datasets also prove wild gene flow between Amazonian wild species (*Cujete amazonica*) and tree gourd is used by local population to shape fruit morphology and size (Moreira et al. 2017b).

2.4 *Plant Nuclear Genomes Lead to Assess Diversity in Lowly Repeated Region*

Plant genomes are still not fully annotated or fully assembled. There is a set of known continuous sequences (contigs), assembled into a scaffold and then into a pseudo-chromosome. There is a patchwork of known sequences separated by a large number of unknown sequences. Our assemblies are generally associated with a large number of contigs (Fig. 1) separated by unknown sequences. One of the known complexities contributing to difficulties in assembling plant genomes is the

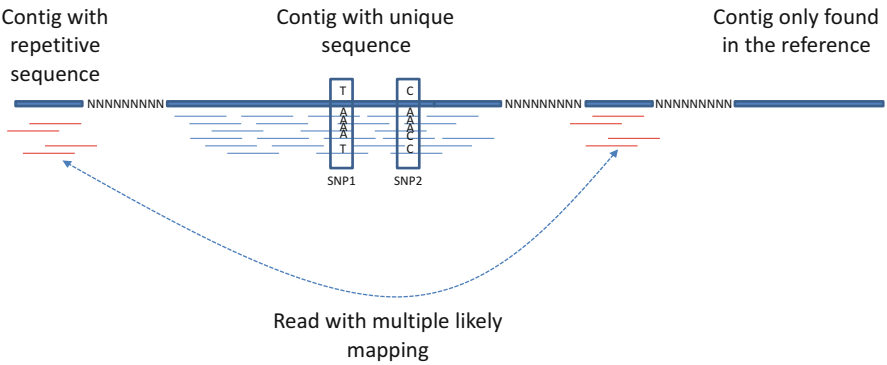

Fig. 1 Mapping and calling variants. The genomes available to date are associated with contigs separated by unknown sequences (NNNN here). Some region/gene might be absent in the plant under study (no mapped reads on some contigs), some reads could be mapped to two locations (in red in the figure), and some reads could map perfectly to only one contig. Only this unique sequence contig is usually available to call the SNP and the genotype. Consequently, our dataset is an island of informed genome sequences with SNPs and regions with no information (no SNP, no data)

occurrence of repeated sequences in the plant genomes. As plant genomes can be very repetitive, a large fraction of short-sequenced reads map to multiple locations. As these reads have multiple possible mapping positions, they are filtered out during SNP calling (Fig. 1). Consequently, we can only clearly identify SNPs in the most unique part of the genome. At present, it is generally still not possible to fully investigate the repetitive parts of the genome. In the future, increasing read length could allow these repeated regions and more structural variation to be revealed and used, but as of now, the most widely used mapping strategy generally focuses on a subset of the genome: lowly repeated regions.

3 Population Genomic Inferences of Evolutionary History of Crops

Identifying centers of the origin of crops has been a long-standing research venue in the study of domestication from the early work of botanist Vavilov et al. (1992). The main questions are: Where did the crop plants originate? Where crop plants were originally domesticated? What are the wild relative's populations closest to the cultivated populations?

3.1 Inferring Changes in Effective Population Size Resulting from Domestication

Genomic SNP data have made it possible to infer changes in effective population size resulting from the domestication events using the pairwise sequentially

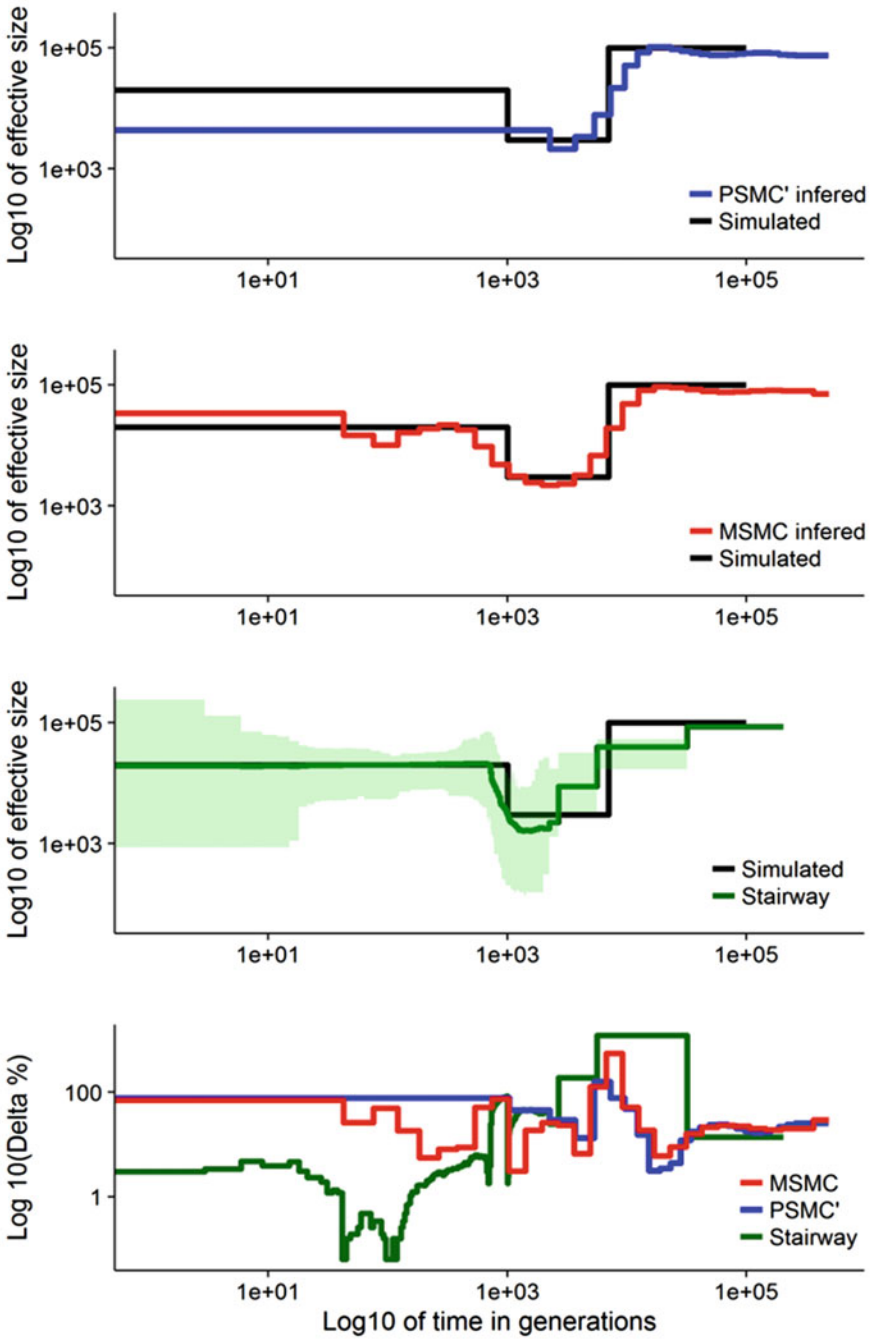

Fig. 2 Comparison of methods for the inference of domestication bottlenecks. We tested several methods for the inference of past demography from genome data. Two methods were based on whole genome data from either 2 (PSMC') or up to 16 (MSMC) haplotypes and use the local density of polymorphisms along the genome to infer coalescence trees, which are then used to estimate the timing and relative importance of changes in the effective population size. The third method (Stairway plot) also relies on coalescence theory to estimate the timing of size changes as well as their related effective size but is based on the site frequency spectrum obtained from

Markovian coalescent model (PSMC) (Li and Durbin 2011) or the multiple sequentially Markovian coalescent (MSMC) (Schiffels and Durbin 2014) algorithm. The underlying model is based on coalescence with recombination. These methods estimate the rate of coalescence, a rate that could be directly translated into effective population size based on coalescent theory including recombination. For PSMC, two-phased chromosomes of a single individual were considered. The fragment of each chromosome in a strictly allogamous species was present in two different individuals in the previous generation and consequently in four different individuals the generation before that and eight different individuals in the generation before that and so on. Thus the two alleles studied in the original individual (rapidly) reveal the history of a whole population. If we consider a coalescence with only two alleles and an effective size N, the probability of a coalescent time t a generation ago is $P(T_2 = t) = 1/2N\ (1–1/2N)^{t-1}$ (Hein et al. 2004). Consequently, the mean time of coalescence is $2N$ (Hein et al. 2004). This method, based on only two chromosomes, has most coalescent time around $2N$ for a population of a fixed size. Hence, inference about the coalescence time will be accurate around this value, but the method will not be very precise for shorter time scales ($<<2N$). Extension to up to eight individuals with MSMC allows more precision at a shorter time scale, but it is still relatively limited. Two approaches have been proposed allowing inference on shorter time scale: the stairway plot (Liu and Fu 2015) or SMC++ (Terhorst et al. 2017). These methods allow a large number of individual to be analyzed. With large numbers of individuals, many coalescence events occur very early and consequently provide a lot of information at this shorter time scale. Consequently, these methods allow more effective inference of effective population size at shorter time scales. The proposed strategy is to combine the short-term (stairway, SMC++) approach and the longer term (PSMC/MSMC) to make inferences concerning population size dynamics (Liu and Fu 2015). Simulation studies modeling a domestication bottleneck illustrate the difference between the methods (Fig. 2). The stairway plot is effective at the short time scale, while PSMC/MSMC methods are effective at longer time

Fig. 2 (continued) genome-wide polymorphisms. To test the accuracy of these methods to study domestication history, we simulated a domestication bottleneck starting 7,000 generations in the past by an instant decrease in the population effective size from an ancestral size of 100,000 to a bottleneck size of 3,000. This effective size was constant for 6,000 generations before there was a significant increase in the population effective size up to 20,000. We simulated 12 independent sequences of 20 Mb, with a mutation rate per base per generation θ=5.2 × 10^{-4} and a recombination rate of 0.8 × θ. We simulated either 2, 16, or 163 haplotypes for analysis using PSMC', MSMC, and Stairway plot, respectively. Each analysis was performed using default parameters. The outcomes of the analysis for the three methods were then plotted on a log scale, with the timing of the events (in generations) on the x-axis and the effective size on the y-axis (PSMC' blue, MSMC red, and Stairway green in the three upper panels) alongside simulated history (black lines). In order to analytically compare the three methods, we then computed an error rate based on the percentage of error in the estimated effective size relative to the simulated one at each point of inference (lower panel). The inference of recent events was more accurate when a large panel of haplotypes was considered (Stairway plot), while accuracy was better with the other methods when looking at more ancient events

scales (Fig. 2). The use of the MSMC approach in maize enabled characterization of the very rapid increase in maize population size after domestication in the last thousand years, from less than 10^5 individuals to roughly 10^{10} individuals (Beissinger et al. 2016). This study revealed a long bottleneck, or at least a long period with a relatively low effective population size, from 1,000 to 10,000 years ago. In African rice, inference based on whole genome data also suggests a very long bottleneck over more than 15,000 years (Meyer et al. 2016; Cubry et al. 2018). This result may seem surprising since the oldest known domestication (wheat, *Triticum* spp.) occurred only around 10,000 to 12,000 years ago. However, recent analyses suggest that the same long decline in effective population size also occurred in the wild population of wild African rice. The long bottleneck observed in wild African rice mirrors the known period of drying of the Sahara (Cubry et al. 2018) and the move of the Poaceae community from the Sahara to the Sahel (Cubry et al. 2018). Consequently, this long period of low effective population size is not specific to the cultivated rice and is not probably directly linked to domestication. A complete domestication of African rice is known to have occurred around 2,800 years before present (Cubry et al. 2018). The conclusion is that drying of the Sahara might have triggered domestication by depletion of wild resources (Cubry et al. 2018). Finally, we have to consider the limits associated with the use of the above methods. One relatively important aspect rarely investigated in these different studies is the impact of the imperfect nature of the genome assemblies and annotations: the number of missing data, phasing error, and low coverage on this analysis. Structure and gene flow will also bias the inferred effective population size, since the hypothesis implied in these methods (PSMC/MSMC/SMC++) is that the population is isolated. So, despite the limits to using this method, recent applications in maize and rice have produced some very interesting inferences as outlined above (Cubry et al. 2018). Used more widely, these approaches could provide valuable insights into crop domestication.

3.2 Origin of Crop Domestication: A Model-Based Approach

The question of the geographical origin of crops is often framed as a dual question: is domestication associated with a single or several origins? Did domestication occur once or several times? While using this way of questioning is interesting, it is also limiting because an increasing number of studies recognize the role wild diversity might have played not only at the beginning of crop domestication but also during diffusion from its original location (Molina et al. 2011).

The study of the origin of domestication in maize (Matsuoka et al. 2002) and wheat (Heun et al. 1997) pinpointed the closest wild populations to the cultivated form. For maize, the region of origin is located in the Balsas basin south of Mexico (Matsuoka et al. 2002). For wheat, South-eastern Turkey was identified as the likely most proximate wild populations (Heun et al. 1997). These studies (Heun et al. 1997; Matsuoka et al. 2002) were mainly based on phylogenetic approaches to assess the

proximity of wild and cultivated samples. The phylogenetic methods assume that gene flow is negligible. Consequently the results from such analyses might sometimes be very difficult to interpret; thus phylogenetic methods are not the most statistically sound approaches. Model-based inferences are certainly one of the best approaches to understand the origin and spread of crop species.

Several model-based inference approaches have been developed in the last decade to study demographic history using either likelihood or pseudo likelihood approaches, e.g., FastSimCoal (Excoffier et al. 2013), $\partial a \partial i$ (Gutenkunst et al. 2009), or more broadly, approximate Bayesian computation [ABC, (Beaumont et al. 2002)]. Inferences based on $\partial a \partial i$ (Gutenkunst et al. 2009) were used to investigate if Asian rice was domesticated once or twice, and the results pointed to a single origin (Molina et al. 2011). The initial study was done with 600 gene fragments (Molina et al. 2011). The result was the subject of wide debate but was finally partly validated with genomic data (Huang et al. 2012). However, the issue is still the subject of debate, and several recent studies suggest up to three independent domestication events (Civáň et al. 2015; Wang et al. 2018). All authors acknowledge the role of wild to cultivated gene flow during rice domestication and diffusion (Choi and Purugganan 2018), which completely reshaped rice diversity. We have to acknowledge that using model-based inference allows a better statistical approach to hypothesis testing (Gerbault et al. 2014) and that such model-based inference will make it possible to (1) better infer crop evolutionary history and (2) assess how much confidence we can have in a given hypothesis. We hope such approaches will be used by the different research groups to allow statistical testing of the different hypotheses around domestication in the future.

Among the different methods available, the ABC approach is particularly suitable for use with more complex models. ABC is based on extensive simulation of a given model and assessment of whether the model output fits the observed data. The fit was originally assessed using the "proximity" of summary statistics, for example, heterozygosity, differentiation, or the number of alleles, but could also be done using the whole site frequency spectrum (SFS). The different steps of the ABC approach (Csilléry et al. 2010) are (1) selecting the parameters of the model from a prior distribution, (2) running the simulation using these parameters, (3) assessing the proximity of the model output (summary statistics or SFS) to the real dataset, and (4) deriving the posterior distribution of the parameters from the improved simulation. In pearl millet (*Pennisetum glaucum*), this approach enabled validation of a model showing a marked increase in population size, gene flow between cultivated and wild species, and an inference about the timing of domestication around 5,000 years ago (Clotault et al. 2012). But the ABC approach can also be used to address more complex models.

3.2.1 Testing Complex Models

One of the interesting aspects of both the FastSimCoal (Excoffier et al. 2013) and the ABC approach is that they both offer the opportunity to test models with gene flow

and the origin from several wild populations. This approach made it possible to prove that during the diffusion of apple (*Malus* spp.) cultivation from Asia to Europe, local European wild species contributed to the European cultivated diversity (Cornille et al. 2012, 2014).

3.2.2 Toward Spatial Models and Inference of Crop Origin

One interesting prospect for this type of approach in crop population genomics concerns the identification of the origin of crops. Indeed, spatial models could be built using geo-referenced genomic data and then used to infer the most likely spatial origin of a given crop. Only a few such studies have been conducted on plants to date (François et al. 2008), but this type of study holds great promise for the future. A set of 76 individuals and a set of 876 nuclear loci were used for the study of the origin of *Arabidopsis* in Europe (François et al. 2008). The spatially explicit diffusion of the crop was simulated using SPLATCHE2 (Ray et al. 2010), which simulates population diffusion. A first simulation of forward-in-time diffusion was run using a stepping stone model (Fig. 3a) in which migration originates from a neighboring population. Each population (deme) is colonized and can then colonize another neighboring population. Parameters that allow migration (or not) can be included in the model as a friction matrix, for example, allowing diffusion in mountain areas (or not) (Fig. 3b). The simulation creates a population migrating from the initial population (Fig. 3c, d) and consequently creating a wave of advancing colonization. Although such studies are very rare in plants, they certainly hold great promise to resolve issues such as single or multiple origins, since both single and multiple origins can be implemented in the model. Such studies will also make it possible to go further by assessing the relative contribution of different origins to the diversity observed today. One recent application of these approaches to the study of the origin of domestication was African rice (Cubry et al. 2018). The model considered an origin of domestication throughout the Sahel, and the ABC spatial approach gave the probable origin of the crop as the inner delta of Niger in what today is Mali (Cubry et al. 2018). Interestingly, the oldest known cultivated archeological remain was found in the same area (Cubry et al. 2018).

Analysis of origin could also be conducted based on the study of successive bottlenecks associated with migration. An original study used a decrease in heterozygosity to trace the roots of human migration (Handley et al. 2007). A recent study used the same idea based on the ψ statistic by measuring the difference in derived allele frequency between two populations (Peter and Slatkin 2013). From one population close to the origin and the second population sampled further away, the statistic increased. Using the same idea as the one used to triangulate the origin of a cellphone call using antennae, it is also possible to infer the origin of diffusion (Peter and Slatkin 2013, 2015). Such a study on plants would also make it possible to check whether or not several origins are possible (Peter and Slatkin 2015). The limits of this method need to be assessed in the context of wild/cultivated gene flow in crops,

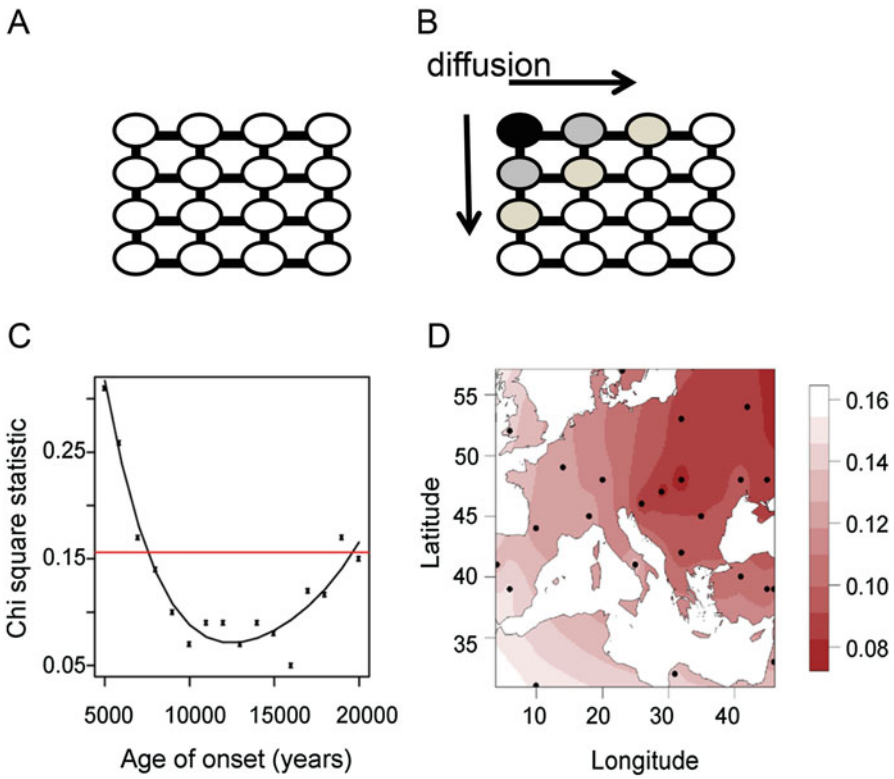

Fig. 3 Simulation of *Arabidopsis thaliana* diffusion. The diffusion of *Arabidopsis thaliana* was simulated using populations or demes that fit the actual map of Europe and are represented by circles in panel (**a**). A map allowing migration to be restricted was built using mountains as the main barrier to diffusion (**b**). Diffusion from a population to the neighboring population and growth of the population were simulated (**c**). The overall simulation created a wave of advancing colonization (**d**). This type of simulation allows inference of the origin of the migration of *Arabidopsis* in Europe and different diffusion routes. Reproduced from François et al. (2008)

but, whatever the results, this type of method will certainly enable better as well as be a complementary inference of the origin of the diffusion of a crop.

4 Selection and Genetic Basis of Domestication

Crop domestication is associated with major phenotypic changes (Doebley et al. 2006). Understanding the genetic basis of these major changes is an important research objective. The first study of the basis of domestication phenotypes was done using controlled crosses and mapping of quantitative trait loci (Doebley et al. 1990; Poncet et al. 2002).

One of the early conclusions of comparison across species is the convergence of the same quantitative trait loci (QTLs) involved in domestication across crop plant species (Paterson et al. 1995). However, when this hypothesis was first proposed, we only knew few of the genes underlying these domestication QTLs. One of the first plant domestication genes cloned following a QTL analysis was a gene that plays a major role in maize aerial ramification, *tb1* (Doebley and Stec 1991; Doebley et al. 1995, 1997; Wang et al. 1999). The insertion of a transposon in the promoter region of its genes changed its regulation, and this allele was selected for reduced plant aerial ramification during maize domestication (Clark et al. 2006; Studer et al. 2011). Interestingly, a recent study showed that the same gene is associated with inflorescence branching in bread wheat (Dixon et al. 2018). In rice, a similar convergence for an erect phenotype is associated with independent loss-of-function mutations of the *PROG1* gene in Asian (Tan et al. 2008) and African (Cubry et al. 2018) rice. Another gene, *PROG7*, was also recently associated with the erect phenotype in African rice (Hu et al. 2018). However, interactions between the two genes (*PROG1* and *PROG7*) are not known and require further investigation. Overall, the original hypothesis of convergent domestication seems to hold true for the identification of the QTLs underlying genes. These results are also true for one of the key traits associated with domestication in cereals, the loss of the shattering or brittle-rachis phenotype. Attached seeds are easily harvested, which was a key step in cereal domestication. Fixation of varieties with almost complete loss of shattering or brittle rachis took several thousand years in rice, wheat, and barley, *Hordeum vulgare* (Purugganan and Fuller 2009). The genetic basis of this phenotype is well known in a large set of cereals (Table 1) and suggests strong convergence of domestication effects. Studies on rice suggest that *SH4* was selected during rice domestication in both Asian and African (Konishi et al. 2006; Wu et al. 2017) rice. The two genes *TfBr1A* and *TfBr1B* found in wheat (Avni et al. 2017) present a loss of function and

Table 1 Discovery of non-shattering genes and allele selected during cereal domestication

Species	Genes	Transcription factor	Type	Pleiotropic effect	
Asian rice	*qSH1*	Yes	Homeobox	Awn length	(Konishi et al. 2006; Magwa et al. 2016; Li et al. 2006)
	SH4	Yes	Trihelix	–	
	qSH3			–	
African rice	*SH3/SH4*	Yes	Trihelix	Grain size	(Win et al. 2017; Wu et al. 2017)
	SH5	Yes	Homeobox	–	Cubry et al. 2018
Wheat	*TtBtr1A-B*	No	–	–	(Avni et al. 2017)
	Q	Yes	AP2	Yield and grain shape	(Xie et al. 2018)
Barley	*Btr1*	No	–	–	(Pourkheirandish et al. 2015)
	Btr2	No	–	–	
Sorghum	*Sh1*	Yes	YABBY	–	(Lin et al. 2012)

are associated with a non-brittle-rachis phenotype. These two genes are homologs of the *Btr1*gene found in barley (Pourkheirandish et al. 2015). In barley, loss of function in one of the two genes *Btr1* and *Btr2* is necessary for the same non-brittle-rachis phenotype (Pourkheirandish et al. 2015). Altogether, the genes identified in the last 10 years point to the occurrence of several convergent selections of the same set of genes in independent domestication events.

Genes selected during and after domestication events share very interesting characteristics. A total of 55 to 63% of the selected genes are transcription factors (Doebley et al. 2006; Meyer and Purugganan 2013). From 30 to 43% of the mutations found in these genes are regulatory changes (Doebley et al. 2006; Meyer and Purugganan 2013). Finally, more than 50% of the mutations found in these genes are loss of function (Doebley et al. 2006; Meyer and Purugganan 2013). One of the hypotheses associated with selection for regulatory change mutations is the reduced pleotropic impact such changes could have on other phenotypic traits (Doebley and Lukens 1998; Lukens and Doebley 2001).

With the new development in genomics, the identification of key genes associated with domestication and after domestication now relies on the detection of selection from genome-wide selection scans. Strong domestication selection led to a signature of diversity for the selected genes. For strongly selected genes, this signature implies a marked decrease in genetic diversity, stronger differentiation from their wild relatives, and stronger linkage disequilibrium around the genes. Some seminal studies began using detection across the genome even before full genomes became available. One such genome-scan study was conducted on maize to identify the signatures of selection during domestication using genic microsatellites (Vigouroux et al. 2002). The study revealed ten microsatellites showing evidence of selection under stringent criteria and provided evidence for selection sweep for MADS box transcriptional regulator gene during maize domestication (Vigouroux et al. 2002). With genomic datasets, the selection tests used now are based on (1) the site frequency spectrum within a given population, like Tajima's D (Tajima 1996) or the composite likelihood test (Nielsen et al. 2005); (2) differentiation which between populations is assessed using the SFS (XP-CLR) or directly by calculating a differentiation index, such as F_{ST} (Chen et al. 2010); and (3) using haplotypes and – more broadly – linkage disequilibrium (Sabeti et al. 2002; Ferrer-Admetlla et al. 2014; Garud et al. 2015). Depending on the intensity of selection and if selection is complete or selection is on standing variation, the magnitude of the selection signal will vary, and each method will have a different power of detection (Vitti et al. 2013). A recent RAID test that combines different statistics seems interesting (Alachiotis and Pavlidis 2018). This composite test enables detection of strong selection, and this powerful method is good for mild bottlenecks (Alachiotis and Pavlidis 2018). The authors also highlight the fact that gene flow seriously challenges the detection of selection (Alachiotis and Pavlidis 2018).

Detection of selection led to a long list of domestication genes, but they do not always pinpoint to a likely selected phenotype. One recent example of the success of genome selection scan approach is the identification of the gene *PROG1* during African rice domestication (Cubry et al. 2018). This gene is associated with an erect

architecture phenotype and showed convergent selection in African and Asian rice domestication (Cubry et al. 2018). In maize, using a genome-scan approach, Hufford et al. (2012) reported that genes involved in a flowering pathway (zagl1 and GRMZM2G448355) and gibberellin pathway (GRMZM2G152354 and GRMZM2G036340) have been under selection before or after domestication. But, not all the genes identified as underlying a QTL linked with domestication were found in a genome-wide selection scan. More than 50% of the genes associated with domestication are loss-of-function alleles (Doebley et al. 2006; Meyer and Purugganan 2013). Mutations leading to a loss-of-function allele are certainly more frequent, and independent mutation could also lead to different nonfunctional alleles. Most of the tests for selection are tailored for identifying strongly selected single new mutation, i.e., hard sweep. If several alleles are present, signatures of selection are soft (Hermisson and Pennings 2017), and detection of selection of hard sweep is impaired. Specific statistics using haplotype homozygosity (Garud et al. 2015) and now machine learning approaches (Schrider and Kern 2016, 2017) are currently being developed for the detection of such soft signatures of selection. Future research should investigate the possible role of soft selection (Hermisson and Pennings 2017) in domestication, which might lead to a better understanding of how functional diversity was shaped during domestication.

Identifying genes under selection is a key to understanding domestication, but key questions about the intensity and the date of the selection are rarely investigated. One seminal paper proposed a method to make such inferences concerning the timing (t) and intensity of selection (s) (Przeworski 2003). The method was used to study the maize *tga1* gene (Wang et al. 2005) and led to a high estimated selection coefficient: $s = 0.035$. The strength of selection across genes associated with domestication across the maize genome was estimated at 0.015 (Hufford et al. 2012). When only genes associated with selection during maize improvement were considered, the average strength of selection was lower: 0.011 (Hufford et al. 2012). New methods have recently been developed to facilitate inferences about the timing of selection using approximate Bayesian computation (Nakagome et al. 2016) or hidden Markovian model (Smith et al. 2018). Beyond the identification of key domestication genes, these methods will advance our understanding of the timing of key adaptations during domestication and will also certainly provide new insights into the process.

5 Ancient DNA and Selection Inference

Ancient DNA (aDNA) has only rarely been investigated in plant domestication studies. In maize, Jaenicke-Després et al. (2003) studied the diversity of three genes, *teosinte branched 1* (*tb1*), *prolamin box binding factor* (*pbf*), *and sugary1* (*su1*), and provided valuable information about when these genes were selected. For *tb1*, the maize allele repressed axillary meristem growth. The *pbf* and *su1* genes were shown to play a role in seed protein storage and starch quality, respectively. A study

of 4,500-year-old archeological remains showed that the alleles of both *tb1* and *pbf* resemble those found in modern maize (Jaenicke-Després et al. 2003), so these alleles were certainly selected early on in the domestication process. For *sugary*1, two mutant alleles only appeared around 1,800 years ago (Jaenicke-Després et al. 2003). The study provided vital information on the timing of the selection process during domestication. Recently, 32 archeological samples made it possible to trace the evolutionary history and selection of maize in the southwest United States (da Fonseca et al. 2015). That study was based on hybridization capture of 348 target genes and helped identify the signatures of selection again notably on *sugary1* during the last 6,000 years. This study gives a glimpse into both the selectively neutral history of maize using archeological remains and the dynamics of selection. Using genomic prediction of polygenic traits, it was even possible to reconstruct the phenotype of a 2,000-year-old corn in the southwestern United States from ancient DNA (Swarts et al. 2017). Such studies will certainly be easier in the future due to improvements in the protocol to study ancient DNA and should be of particular interest for the study of crops growing in dry environments where archeological remains can survive for thousands of years. Results obtained from aDNA studies in the last few years are impressive. Hopefully, the next 10 years will see a wealth of studies using ancient DNA to document domestication.

6 The Cost of Domestication

Fixation of deleterious mutations is a balance between the strength of selection and the strength of the drift parameter $1/2N$ (Ohta 2002). When the absolute value of the selection is stronger than $1/2N$, then selection is a stronger force than drift. Conversely, when the absolute value of the selection coefficient is lower than $1/2N$, then drift becomes a driving force. Knowing that most mutations are deleterious (Eyre-Walker and Keightley 2007), if the effective population size of crop plants is low, the slightly deleterious mutations will not be counter selected. Consequently, they might end up fixed in the population. As cultivated crops have been subject to serious bottlenecks, one would expect several deleterious mutations to be fixed or at least brought to high frequency by the effect of drift. This deleterious mutation load will consequently be shared by all cultivated plants. The term "the cost of domestication" was coined to describe the fixation of such deleterious mutations. There are still only a few studies documenting the cost of domestication associated with plant domestication (Li et al. 2006; Nabholz et al. 2014; Wang et al. 2017; Moyers et al. 2018). This phenomenon is observed in Asian rice, African rice (Lu et al. 2006; Nabholz et al. 2014; Liu et al. 2017), and in maize (Wang et al. 2017). A recent study documents the increased occurrence of deleterious alleles in five domesticated animals and two domesticated plant species (Makino et al. 2018). Such rare deleterious allele variants are also linked to the deregulation of gene expression (Kremling et al. 2018). Altogether, it appears that domestication has its own burdens.

The study of the cost of domestication will be facilitated by the increased availability of genomics data. What also needs to be investigated is, if the existence of domestication costs is confirmed, what are the consequences of gene flow for wild relatives? Does gene flow enable purging of deleterious mutations?

7 Crop Domestication Beyond DNA

The access to inexpensive next-generation sequencing data enables studies of gene expression (RNA sequenced data, mi, and siRNA), methylation of DNA, or of histone linked with plant domestication. Such data will certainly advance our understanding of selection during the domestication of a crop and notably how the consequences of selection of certain key genes led to a cultivated phenotype. For example, we could now ask how the expression of micro and small RNA was reshaped during domestication (Ta et al. 2016). The comparison of African rice inflorescence at different stages in wild *Oryza barthii* and cultivated *O. glaberrima* (Ta et al. 2016) revealed a striking difference in transcript expression during the development of the panicle in the cultivated and the wild relative. A broad category of transRNA was differentially expressed in a way suggesting that a regulator of this transRNA was selected during domestication (Ta et al. 2016). Further study is needed to better understand this pattern, but interesting questions were raised concerning how domestication can reshape the regulation of key organs like the inflorescence in cereals.

Cis-regulation was also the subject of a study associated with crop domestication. A general hypothesis postulates that the *cis*-regulation allele is a preferential target of selection during adaptation. This question was addressed in the context of domestication using an RNAseq approach (Lemmon et al. 2014). This type of experiment requires the creation of an F_1 hybrid between a cultivated and a wild plant to obtain RNAseq data from the two alleles (wild and cultivated) with a similar background. This setup makes it possible to eliminate *trans*-effects on allele expression, because the *trans* regulator allele will be present in the F_1 and consequently will act on both wild and cultivated alleles. In maize, the study by Lemmon et al. (2014) suggests that both *cis*-regulation associated with the domestication allele and *cis*-regulation are more frequent in ear tissue than in the leaf and stem (Lemmon et al. 2014). However, it is also possible to assess relative changes in expression between selected and nonselected genes found to be associated with domestication (Hufford et al. 2012). The variation in gene expression was found to be reduced in maize compared to teosinte (Hufford et al. 2012), but this may reflect the relative loss of *cis*-regulated alleles in cultivated maize, i.e., it could simply be the consequence of the domestication bottleneck. However, one important feature detected was a 7% increase or decrease in expression in cultivated maize compared to teosinte (Hufford et al. 2012). This result could be a sign of selection of alleles with a *cis*-regulation effect during domestication (Hufford et al. 2012), and if this pattern is found across species is now technically possible.

8 Conclusion and Future Outlook

In this chapter, we focused on population genomics of domestication in key cereals like maize and rice, crops for which population genomics studies are the most advanced. However, a large set of species with a reference genome is now available, and with the cost of sequencing going down, genomic datasets could be obtained for non-model species. As more genomics data become available, we will be able to tackle the question in domestication genomics across different crops. Here, we argue that model-based inference should become the standard approach to test hypotheses associated with crop domestication, such as a single versus several origins. With the increased availability of genome sequences, such approaches will be easy to transfer from well-known crops to less widely used models. Such model-based inference will allow us to move from "storytelling to story testing" (Gerbault et al. 2014).

The study of selection associated with domestication needs to take the evolutionary history of crops more into account (Vigouroux et al. 2002; Tenaillon et al. 2004). New methods using machine learning are based on such modeling. Consequently in the near future, detection of selection will certainly also benefit from the development of models to build a neutral expected baseline of diversity to detect outlier loci. Progress in this field will initially be made in human or animal population genomics.

Ancient DNA is still a relatively new field in crop domestication research, and exciting results have been published in the last 5 years (da Fonseca et al. 2015; Swarts et al. 2017). We expect that these approaches will provide extraordinary insights into plant domestication in the years to come.

Finally, with genomics, we now have access to a wealth of data on gene expression, gene regulation, DNA, and histone methylation. These new methods have already added invaluable knowledge to the study of domesticated plants (Kremling et al. 2018), and we are just at the beginning. Altogether, a very exciting era is starting for the study of plant domestication using genomics.

References

Alachiotis N, Pavlidis P. RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. Commun Biol. 2018;1:79. https://doi.org/10.1038/s42003-018-0085-8.

Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, Hale I, Mascher M, Spannagl M, Wiebe K, Jordan KW, Golan G, Deek J, Ben-Zvi B, Ben-Zvi G, Himmelbach A, MacLachlan RP, Sharpe AG, Fritz A, Ben-David R, Budak H, Fahima T, Korol A, Faris JD, Hernandez A, Mikel MA, Levy AA, Steffenson B, Maccaferri M, Tuberosa R, Cattivelli L, Faccioli P, Ceriotti A, Kashkush K, Pourkheirandish M, Komatsuda T, Eilam T, Sela H, Sharon A, Ohad N, Chamovitz DA, Mayer KFX, Stein N, Ronen G, Peleg Z, Pozniak CJ, Akhunov ED, Distelfeld A. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science. 2017;357:93–7. https://doi.org/10.1126/science.aan0032.

Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002;162:2025–35.

Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, Ross-Ibarra J. Recent demography drives changes in linked selection across the maize genome. Nat Plants. 2016;2:16084. https://doi.org/10.1038/nplants.2016.84.

Bellucci E, Bitocchi E, Ferrarini A, Benazzo A, Biagetti E, Klie S, Minio A, Rau D, Rodriguez M, Panziera A, Venturini L, Attene G, Albertini E, Jackson SA, Nanni L, Fernie AR, Nikoloski Z, Bertorelle G, Delledonne M, Papa R. Decreased nucleotide and expression diversity and modified coexpression patterns characterize domestication in the common bean. Plant Cell. 2014;26:1901–12. https://doi.org/10.1105/tpc.114.124040.

Berthouly-Salazar C, Mariac C, Couderc M, Pouzadoux J, Floc'h J-B, Vigouroux Y. Genotyping-by-sequencing SNP identification for crops without a reference genome: using transcriptome based mapping as an alternative strategy. Front Plant Sci. 2016;7:777. https://doi.org/10.3389/fpls.2016.00777.

Brandenburg J-T, Mary-Huard T, Rigaill G, Hearne SJ, Corti H, Joets J, Vitte C, Charcosset A, Nicolas SD, Tenaillon MI. Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. PLoS Genet. 2017;13:e1006666. https://doi.org/10.1371/journal.pgen.1006666.

Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome Res. 2010;20:393–402. https://doi.org/10.1101/gr.100545.109.

Choi JY, Purugganan MD. Multiple origin but single domestication led to Oryza sativa. G3 (Bethesda). 2018;8:797–803. https://doi.org/10.1534/g3.117.300334.

Civáň P, Craig H, Cox CJ, Brown TA. Three geographically separate domestications of Asian rice. Nat Plants. 2015;1:15164. https://doi.org/10.1038/nplants.2015.164.

Clark RM, Wagler TN, Quijada P, Doebley J. A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat Genet. 2006;38:594–7. https://doi.org/10.1038/ng1784.

Clotault J, Thuillet A-C, Buiron M, De Mita S, Couderc M, Haussmann BIG, Mariac C, Vigouroux Y. Evolutionary history of pearl millet (Pennisetum glaucum [L.] R. Br.) and selection on flowering genes since its domestication. Mol Biol Evol. 2012;29:1199–212. https://doi.org/10.1093/molbev/msr287.

Cornille A, Gladieux P, Smulders MJM, Roldán-Ruiz I, Laurens F, Le Cam B, Nersesyan A, Clavel J, Olonova M, Feugey L, Gabrielyan I, Zhang X-G, Tenaillon MI, Giraud T. New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. PLoS Genet. 2012;8:e1002703. https://doi.org/10.1371/journal.pgen.1002703.

Cornille A, Giraud T, Smulders MJM, Roldán-Ruiz I, Gladieux P. The domestication and evolutionary ecology of apples. Trends Genet. 2014;30:57–65. https://doi.org/10.1016/j.tig.2013.10.002.

Csilléry K, Blum MGB, Gaggiotti OE, François O. Approximate Bayesian computation (ABC) in practice. Trends Ecol Evol (Amst). 2010;25:410–8. https://doi.org/10.1016/j.tree.2010.04.001.

Cubry P, Tranchant-Dubreuil C, Thuillet AC, Monat C, Ndjiondjop MN, Labadi K, Cruaud C, Engelen S, Scarcelli N, Rhoné B, Burgarella C, Dupuy C, Larmande P, Wincker P, François O, Sabot F, Vigouroux Y. The rise and fall of African rice cultivation revealed by analysis of 246 new genomes. Curr Biol. 2018;28(14):2274–2282.e6.

da Fonseca RR, Smith BD, Wales N, Cappellini E, Skoglund P, Fumagalli M, Samaniego JA, Carøe C, Ávila-Arcos MC, Hufnagel DE, Korneliussen TS, Vieira FG, Jakobsson M, Arriaza B, Willerslev E, Nielsen R, Hufford MB, Albrechtsen A, Ross-Ibarra J, Gilbert MTP. The origin and evolution of maize in the Southwestern United States. Nat Plants. 2015;1:14003. https://doi.org/10.1038/nplants.2014.3.

Dixon LE, Greenwood JR, Bencivenga S, Zhang P, Cockram J, Mellers G, Ramm K, Cavanagh C, Swain SM, Boden SA. TEOSINTE BRANCHED1 regulates inflorescence architecture and development in bread wheat (Triticum aestivum). Plant Cell. 2018;30:563–81. https://doi.org/10.1105/tpc.17.00961.

Doebley J, Lukens L. Transcriptional regulators and the evolution of plant form. Plant Cell. 1998;10:1075–82.
Doebley J, Stec A. Genetic analysis of the morphological differences between maize and teosinte. Genetics. 1991;129:285–95.
Doebley J, Stec A, Wendel J, Edwards M. Genetic and morphological analysis of a maize-teosinte F2 population: implications for the origin of maize. Proc Natl Acad Sci U S A. 1990;87:9888–92.
Doebley J, Stec A, Gustus C. Teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. Genetics. 1995;141:333–46.
Doebley J, Stec A, Hubbard L. The evolution of apical dominance in maize. Nature. 1997;386:485–8. https://doi.org/10.1038/386485a0.
Doebley JF, Gaut BS, Smith BD. The molecular genetics of crop domestication. Cell. 2006;127:1309–21. https://doi.org/10.1016/j.cell.2006.12.006.
Durvasula A, Hoffman PJ, Kent TV, Liu C, Kono TJY, Morrell PL, Ross-Ibarra J. angsd-wrapper: utilities for analysing next-generation sequencing data. Mol Ecol Resour. 2016;16:1449–54. https://doi.org/10.1111/1755-0998.12578.
Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6:e19379. https://doi.org/10.1371/journal.pone.0019379.
Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS Genet. 2013;9:e1003905. https://doi.org/10.1371/journal.pgen.1003905.
Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet. 2007;8:610–8. https://doi.org/10.1038/nrg2146.
Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol. 2014;31:1275–91. https://doi.org/10.1093/molbev/msu077.
François O, Blum MGB, Jakobsson M, Rosenberg NA. Demographic history of European populations of Arabidopsis thaliana. PLoS Genet. 2008;4:e1000075. https://doi.org/10.1371/journal.pgen.1000075.
Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS Genet. 2015;11:e1005004. https://doi.org/10.1371/journal.pgen.1005004.
Gerbault P, Allaby RG, Boivin N, Rudzinski A, Grimaldi IM, Pires JC, Vigueira CC, Dobney K, Gremillion KJ, Barton L, Arroyo-Kalin M, Purugganan MD, de Casas RR, Bollongino R, Burger J, Fuller DQ, Bradley DG, Balding DJ, Richerson PJ, Gilbert MTP, Larson G, Thomas MG. Storytelling and story testing in domestication. Proc Natl Acad Sci U S A. 2014;111:6159–64. https://doi.org/10.1073/pnas.1400425111.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5:e1000695. https://doi.org/10.1371/journal.pgen.1000695.
Handley LJL, Manica A, Goudet J, Balloux F. Going the distance: human population genetics in a clinal world. Trends Genet. 2007;23:432–9. https://doi.org/10.1016/j.tig.2007.07.002.
Hein J, Schierup M, Wiuf C. Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford: Oxford University Press; 2004.
Hermisson J, Pennings PS. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods Ecol Evol. 2017;8:700–16. https://doi.org/10.1111/2041-210X.12808.
Heun M, Schäfer-Pregl R, Klawan D, Castagna R, Accerbi M, Borghi B, Salamini F. Site of einkorn wheat domestication identified by DNA fingerprinting. Science. 1997;278:1312–4. https://doi.org/10.1126/science.278.5341.1312.
Hu M, Lv S, Wu W, Fu Y, Liu F, Wang B, Li W, Gu P, Cai H, Sun C, Zhu Z. The domestication of plant architecture in African rice. Plant J. 2018;94:661–9. https://doi.org/10.1111/tpj.13887.
Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, Guo Y, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, Lu T,

Feng Q, Qian Q, Li J, Han B. A map of rice genome variation reveals the origin of cultivated rice. Nature. 2012;490:497–501. https://doi.org/10.1038/nature11532.

Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, Lai J, Morrell PL, Shannon LM, Song C, Springer NM, Swanson-Wagner RA, Tiffin P, Wang J, Zhang G, Doebley J, McMullen MD, Ware D, Buckler ES, Yang S, Ross-Ibarra J. Comparative population genomics of maize domestication and improvement. Nat Genet. 2012;44:808–11. https://doi.org/10.1038/ng.2309.

Jaenicke-Després V, Buckler ES, Smith BD, Gilbert MTP, Cooper A, Doebley J, Pääbo S. Early allelic selection in maize as revealed by ancient DNA. Science. 2003;302:1206–8. https://doi.org/10.1126/science.1089056.

Kofler R, Langmüller AM, Nouhaud P, Otte KA, Schlötterer C. Suitability of different mapping algorithms for genome-wide polymorphism scans with Pool-Seq data. G3. 2016;6:3507–15. https://doi.org/10.1534/g3.116.034488.

Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, Yano M. An SNP caused loss of seed shattering during rice domestication. Science. 2006;312:1392–6. https://doi.org/10.1126/science.1126410.

Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014;15:356. https://doi.org/10.1186/s12859-014-0356-4.

Kremling KAG, Chen S-Y, Su M-H, Lepak NK, Romay MC, Swarts KL, Lu F, Lorant A, Bradbury PJ, Buckler ES. Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature. 2018;555:520–3. https://doi.org/10.1038/nature25966.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. https://doi.org/10.1038/nmeth.1923.

Lemmon ZH, Bukowski R, Sun Q, Doebley JF. The role of cis regulatory evolution in maize domestication. PLoS Genet. 2014;10:e1004745. https://doi.org/10.1371/journal.pgen.1004745.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. arXiv:1303.3997 [q-bio].

Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–6. https://doi.org/10.1038/nature10231.

Li C, Zhou A, Sang T. Rice domestication by reducing shattering. Science. 2006;311:1936–9. https://doi.org/10.1126/science.1123604.

Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25:1966–7. https://doi.org/10.1093/bioinformatics/btp336.

Lin Z, Li X, Shannon LM, Yeh C-T, Wang ML, Bai G, Peng Z, Li J, Trick HN, Clemente TE, Doebley J, Schnable PS, Tuinstra MR, Tesso TT, White F, Yu J. Parallel domestication of the Shattering1 genes in cereals. Nat Genet. 2012;44:720–4. https://doi.org/10.1038/ng.2281.

Liu X, Fu Y-X. Exploring population size changes using SNP frequency spectra. Nat Genet. 2015;47:555–9. https://doi.org/10.1038/ng.3254.

Liu Q, Zhou Y, Morrell PL, Gaut BS. Deleterious variants in Asian rice and the potential cost of domestication. Mol Biol Evol. 2017;34:908–24. https://doi.org/10.1093/molbev/msw296.

Lu J, Tang T, Tang H, Huang J, Shi S, Wu C-I. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet. 2006;22:126–31. https://doi.org/10.1016/j.tig.2006.01.004.

Lukens L, Doebley J. Molecular evolution of the teosinte branched gene among maize and related grasses. Mol Biol Evol. 2001;18:627–38. https://doi.org/10.1093/oxfordjournals.molbev.a003843.

Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011;21:936–9. https://doi.org/10.1101/gr.111120.110.

Magwa RA, Zhao H, Yao W, Xie W, Yang L, Xing Y, Bai X. Genomewide association analysis for awn length linked to the seed shattering gene qSH1 in rice. J Genet. 2016;95:639–46.

Makino T, Rubin C-J, Carneiro M, Axelsson E, Andersson L, Webster MT. Elevated proportions of deleterious genetic variation in domestic animals and plants. Genome Biol Evol. 2018;10:276–90. https://doi.org/10.1093/gbe/evy004.

Mariac C, Scarcelli N, Pouzadou J, Barnaud A, Billot C, Faye A, Kougbeadjo A, Maillol V, Martin G, Sabot F, Santoni S, Vigouroux Y, Couvreur TLP. Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. Mol Ecol Resour. 2014;14:1103–13. https://doi.org/10.1111/1755-0998.12258.

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, Doebley J. A single domestication for maize shown by multilocus microsatellite genotyping. Proc Natl Acad Sci U S A. 2002;99:6080–4. https://doi.org/10.1073/pnas.052125199.

Meyer RS, Purugganan MD. Evolution of crop species: genetics of domestication and diversification. Nat Rev Genet. 2013;14:840–52. https://doi.org/10.1038/nrg3605.

Meyer RS, Choi JY, Sanches M, Plessis A, Flowers JM, Amas J, Dorph K, Barretto A, Gross B, Fuller DQ, Bimpong IK, Ndjiondjop M-N, Hazzouri KM, Gregorio GB, Purugganan MD. Domestication history and geographical adaptation inferred from a SNP map of African rice. Nat Genet. 2016;48:1083–8. https://doi.org/10.1038/ng.3633.

Molina J, Sikora M, Garud N, Flowers JM, Rubinstein S, Reynolds A, Huang P, Jackson S, Schaal BA, Bustamante CD, Boyko AR, Purugganan MD. Molecular evidence for a single evolutionary origin of domesticated rice. Proc Natl Acad Sci U S A. 2011;108:8351–6. https://doi.org/10.1073/pnas.1104686108.

Moreira PA, Mariac C, Scarcelli N, Couderc M, Rodrigues DP, Clement CR, Vigouroux Y. Chloroplast sequence of treegourd (Crescentia cujete, Bignoniaceae) to study phylogeography and domestication. Appl Plant Sci. 2016;4:1600048. https://doi.org/10.3732/apps.1600048.

Moreira PA, Aguirre-Dugua X, Mariac C, Zekraoui L, Couderc M, Rodrigues DP, Casas A, Clement CR, Vigouroux Y. Diversity of treegourd (Crescentia cujete) suggests introduction and prehistoric dispersal routes into Amazonia. Front Ecol Evol. 2017a;5:150. https://doi.org/10.3389/fevo.2017.00150.

Moreira PA, Mariac C, Zekraoui L, Couderc M, Rodrigues DP, Clement CR, Vigouroux Y. Human management and hybridization shape treegourd fruits in the Brazilian Amazon Basin. Evol Appl. 2017b;10:577–89. https://doi.org/10.1111/eva.12474.

Moyers BT, Morrell PL, McKay JK. Genetic costs of domestication and improvement. J Hered. 2018;109:103–16. https://doi.org/10.1093/jhered/esx069.

Nabholz B, Sarah G, Sabot F, Ruiz M, Adam H, Nidelet S, Ghesquière A, Santoni S, David J, Glémin S. Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (Oryza glaberrima). Mol Ecol. 2014;23:2210–27. https://doi.org/10.1111/mec.12738.

Nakagome S, Alkorta-Aranburu G, Amato R, Howie B, Peter BM, Hudson RR, Rienzo AD. Estimating the ages of selection signals from different epochs in human history. Mol Biol Evol. 2016;33:657–69. https://doi.org/10.1093/molbev/msv256.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. 2005;15:1566–75. https://doi.org/10.1101/gr.4252305.

Ohta T. Near-neutrality in evolution of genes and gene regulation. Proc Natl Acad Sci U S A. 2002;99:16134–7. https://doi.org/10.1073/pnas.252626899.

Paterson AH, Lin YR, Li Z, Schertz KF, Doebley JF, Pinson SR, Liu SC, Stansel JW, Irvine JE. Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. Science. 1995;269:1714–8. https://doi.org/10.1126/science.269.5231.1714.

Peter BM, Slatkin M. Detecting range expansions from genetic data. Evolution. 2013;67:3274–89. https://doi.org/10.1111/evo.12202.

Peter BM, Slatkin M. The effective founder effect in a spatially expanding population. Evolution. 2015;69:721–34. https://doi.org/10.1111/evo.12609.

Poncet V, Martel E, Allouis S, Devos M, Lamy F, Sarr A, Robert T. Comparative analysis of QTLs affecting domestication traits between two domesticated x wild pearl millet (Pennisetum glaucum L., Poaceae) crosses. Theor Appl Genet. 2002;104:965–75. https://doi.org/10.1007/s00122-002-0889-1.

Pourkheirandish M, Hensel G, Kilian B, Senthil N, Chen G, Sameri M, Azhaguvel P, Sakuma S, Dhanagond S, Sharma R, Mascher M, Himmelbach A, Gottwald S, Nair SK, Tagiri A, Yukuhiro F, Nagamura Y, Kanamori H, Matsumoto T, Willcox G, Middleton CP, Wicker T, Walther A, Waugh R, Fincher GB, Stein N, Kumlehn J, Sato K, Komatsuda T. Evolution of the grain dispersal system in barley. Cell. 2015;162:527–39. https://doi.org/10.1016/j.cell.2015.07.002.

Przeworski M. Estimating the time since the fixation of a beneficial allele. Genetics. 2003;164:1667–76.

Purugganan MD, Fuller DQ. The nature of selection during plant domestication. Nature. 2009;457:843–8. https://doi.org/10.1038/nature07895.

Ray N, Currat M, Foll M, Excoffier L. SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. Bioinformatics. 2010;26:2993–4. https://doi.org/10.1093/bioinformatics/btq579.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419:832–7. https://doi.org/10.1038/nature01140.

Sarah G, Homa F, Pointet S, Contreras S, Sabot F, Nabholz B, Santoni S, Sauné L, Ardisson M, Chantret N, Sauvage C, Tregear J, Jourda C, Pot D, Vigouroux Y, Chair H, Scarcelli N, Billot C, Yahiaoui N, Bacilieri R, Khadari B, Boccara M, Barnaud A, Péros J-P, Labouisse J-P, Pham J-L, David J, Glémin S, Ruiz M. A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. Mol Ecol Resour. 2017;17:565–80. https://doi.org/10.1111/1755-0998.12587.

Scarcelli N, Mariac C, Couvreur TLP, Faye A, Richard D, Sabot F, Berthouly-Salazar C, Vigouroux Y. Intra-individual polymorphism in chloroplasts from NGS data: where does it come from and how to handle it? Mol Ecol Resour. 2016;16:434–45. https://doi.org/10.1111/1755-0998.12462.

Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014;46:919–25. https://doi.org/10.1038/ng.3015.

Schrider DR, Kern AD. S/HIC: robust identification of soft and hard sweeps using machine learning. PLoS Genet. 2016;12:e1005928. https://doi.org/10.1371/journal.pgen.1005928.

Schrider DR, Kern AD. Soft sweeps are the dominant mode of adaptation in the human genome. Mol Biol Evol. 2017;34:1863–77. https://doi.org/10.1093/molbev/msx154.

Sedlazeck FJ, Rescheneder P, von Haeseler A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics. 2013;29:2790–1. https://doi.org/10.1093/bioinformatics/btt468.

Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. Genetics. 2013;195:693–702. https://doi.org/10.1534/genetics.113.154138.

Smith J, Coop G, Stephens M, Novembre J. Estimating time to the common ancestor for a beneficial allele. Mol Biol Evol. 2018;35:1003–17. https://doi.org/10.1093/molbev/msy006.

Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon insertion in the maize domestication gene tb1. Nat Genet. 2011;43:1160–3. https://doi.org/10.1038/ng.942.

Swarts K, Gutaker RM, Benz B, Blake M, Bukowski R, Holland J, Kruse-Peeples M, Lepak N, Prim L, Romay MC, Ross-Ibarra J, Sanchez-Gonzalez JJ, Schmidt C, Schuenemann VJ, Krause J, Matson RG, Weigel D, Buckler ES, Burbano HA. Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. Science. 2017;357:512–5. https://doi.org/10.1126/science.aam9425.

Ta KN, Sabot F, Adam H, Vigouroux Y, De Mita S, Ghesquière A, Do NV, Gantet P, Jouannic S. miR2118-triggered phased siRNAs are differentially expressed during the panicle

development of wild and domesticated African rice species. Rice (N Y). 2016;9:10. https://doi.org/10.1186/s12284-016-0082-9.

Tajima F. The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. Genetics. 1996;143:1457–65.

Tan L, Li X, Liu F, Sun X, Li C, Zhu Z, Fu Y, Cai H, Wang X, Xie D, Sun C. Control of a key transition from prostrate to erect growth in rice domestication. Nat Genet. 2008;40:1360–4. https://doi.org/10.1038/ng.197.

Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS. Selection versus demography: a multilocus investigation of the domestication process in maize. Mol Biol Evol. 2004;21:1214–25. https://doi.org/10.1093/molbev/msh102.

Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. 2017;49:303–9. https://doi.org/10.1038/ng.3748.

Tong W, Kim T-S, Park Y-J. Rice chloroplast genome variation architecture and phylogenetic dissection in diverse Oryza species assessed by whole-genome resequencing. Rice. 2016;9:57. https://doi.org/10.1186/s12284-016-0129-y.

Vavilov NI, Vavylov MI, Vavílov NÍ, Dorofeev VF. Origin and geography of cultivated plants. Cambridge: Cambridge University Press; 1992.

Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y, Doebley J. Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proc Natl Acad Sci U S A. 2002;99:9650–5. https://doi.org/10.1073/pnas.112324299.

Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet. 2013;47:97–120. https://doi.org/10.1146/annurev-genet-111212-133526.

Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, Faller M, Bomblies K, Lukens L, Doebley JF. The origin of the naked grains of maize. Nature. 2005;436:714–9. https://doi.org/10.1038/nature03863.

Wang L, Beissinger TM, Lorant A, Ross-Ibarra C, Ross-Ibarra J, Hufford MB. The interplay of demography and selection during maize domestication and expansion. Genome Biol. 2017;18:215. https://doi.org/10.1186/s13059-017-1346-4.

Wang RL, Stec A, Hey J, Lukens L, Doebley J. The limits of selection during maize domestication. Nature. 1999;398:236–9. https://doi.org/10.1038/18435.

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, Shen F, Cui X, Yu H, Li Z, Chen M, Detras J, Zhou Y, Zhang X, Zhao Y, Kudrna D, Wang C, Li R, Jia B, Lu J, He X, Dong Z, Xu J, Li Y, Wang M, Shi J, Li J, Zhang D, Lee S, Hu W, Poliakov A, Dubchak I, Ulat VJ, Borja FN, Mendoza JR, Ali J, Li J, Gao Q, Niu Y, Yue Z, Naredo MEB, Talag J, Wang X, Li J, Fang X, Yin Y, Glaszmann J-C, Zhang J, Li J, Hamilton RS, Wing RA, Ruan J, Zhang G, Wei C, Alexandrov N, McNally KL, Li Z, Leung H. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018;557:43–9. https://doi.org/10.1038/s41586-018-0063-9.

Win KT, Yamagata Y, Doi K, Uyama K, Nagai Y, Toda Y, Kani T, Ashikari M, Yasui H, Yoshimura A. A single base change explains the independent origin of and selection for the nonshattering gene in African rice domestication. New Phytol. 2017;213:1925–35. https://doi.org/10.1111/nph.14290.

Wu W, Liu X, Wang M, Meyer RS, Luo X, Ndjiondjop M-N, Tan L, Zhang J, Wu J, Cai H, Sun C, Wang X, Wing RA, Zhu Z. A single-nucleotide polymorphism causes smaller grain size and loss of seed shattering during African rice domestication. Nat Plants. 2017;3:17064. https://doi.org/10.1038/nplants.2017.64.

Xie Q, Li N, Yang Y, Lv Y, Yao H, Wei R, Sparkes DL, Ma Z. Pleiotropic effects of the wheat domestication gene Q on yield and grain morphology. Planta. 2018;247:1089–98. https://doi.org/10.1007/s00425-018-2847-4.

Population Genomics of Animal Domestication and Breed Development

Samantha Wilkinson and Pamela Wiener

Abstract Domesticated animals have a rich and complex history, comprising several population-shaping events, which has resulted in an assortment of distinctive phenotypes and highly specialised breeds that meet a variety of human needs. The availability of whole genome sequences and single nucleotide polymorphism (SNP) arrays for the major domestic animal species allows for a thorough interrogation of the genomic landscape of breeds using population genomic approaches. In this chapter, we synthesise insights gained into the processes of domestication and breed development from the patterns of diversity mapped across domestic genomes, with particular focus on cattle (*Bos taurus taurus* and *Bos taurus indicus*), chicken (*Gallus gallus domesticus*), dog (*Canis lupus familiaris*), pig (*Sus scrofa*) and sheep (*Ovis aries*) breeds. First, we evaluate the current state of genome-wide diversity within domestic animals, a topic of importance considering concerns over the continuing erosion of genetic variation within breeds. Second, we review the growing catalogue of selective sweeps found for key phenotypic traits in domestic animals, illustrating that breeds have been intensively selected for a range of breed-defining traits (e.g. coat colour, horn morphology, ear carriage and body size) and production traits (e.g. milk production, muscular conformation, reproduction and meat quality). Finally, we discuss insights into the selection history of domestic animals and the genetic architecture of phenotypic traits and we address the future management of genetic diversity in domestic breeds.

Keywords Coat colour · Dairy breeds · Domestication genetics · Effective population size · Genomic diversity · Meat breeds · Phenotypic traits · Signatures of selection · SNPs

---

The original version of this chapter was revised. In this revised version, Section 4.2, the caption of Figure 7 and some citations to the literature in the reference section have been updated.

S. Wilkinson (✉) · P. Wiener
The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, UK
e-mail: samantha.wilkinson@roslin.ed.ac.uk

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2017_8,

1 Introduction

Domestication and breed development is a complex process that has produced an estimated 7,616 distinctive breeds worldwide, collectively possessing an extraordinary array of phenotypic characteristics (FAO 2015). Domestic animals have long been considered useful models to advance the understanding of the evolutionary processes that accompany population change because they have experienced a large number of population-shaping events over a relatively short evolutionary timescale compared to natural and human populations. Charles Darwin used breeds as examples to support the theory of natural selection in his publication on the domestication in plants and animals, in which he reflected on the "infinite diversity of many varieties of our domesticated productions" and the selective breeding practices of the nineteenth century (Darwin 1868). Breeds have been moulded over time by a combination of human-mediated pressures, including small founder populations, bottlenecks, isolation of reproductive breeding pools, intense artificial selection, cross-breeding and use of favoured sires. These factors can, to varying degrees, influence the evolutionary forces (selection, mutation and random genetic drift) that affect the genetic composition of breeds. Population genetics concepts and approaches, whilst principally designed to characterise the genetic structure of human and natural populations (Hartl and Clark 2007), can be equally applied to domestic breeds. The revolution in genomic technology and increasing availability of thousands (and in some cases, millions) of genome-wide single nucleotide polymorphisms (SNPs) for the majority of domestic animal species provides an unprecedented opportunity to densely probe the variation spread across domestic genomes, thus transporting us into the era of population genomics of domesticated animals.

In this chapter, we review the application of population genomics approaches to domestic animals to advance the knowledge of genetic changes that have accompanied domestication and breed development. In order to introduce the topic, we first provide a brief history of domestication and breed formation to highlight the human societal pressures and evolutionary forces that domestic populations have experienced and subsequently we outline population genomic tools, i.e. genotyping resources and statistical methods, used by animal geneticists. In the main section of the chapter, we describe: (1) the genetic patterns that have arisen from the demographic events experienced by breeds and (2) the genetic variation underlying the selection for desired phenotypes and the genes identified to have a functional influence on the selected phenotypes. Through a series of examples, we aim to convey that population genomics is a highly powerful and informative tool to advance our understanding of the genetics underpinning the complex history of animal domestication and breed development and that new and developing techniques will continue to increase the impact of this research area.

2 Domestic Animals

2.1 A Brief History of Animal Domestication and Breed Development

The initial stages of animal domestication likely involved separating small sets of less aggressive animals from the ancestral populations and gradually subsuming them into the human community, independently for each of the major domesticated species (Clutton-Brock 1999). Subsequently, from the centre points of the original domestication events, animal stock would have accompanied migrating human populations thereby increasing the geographical spread of domestic animals across the world. In new environments, there likely followed a combination of local adaptation and selection for desirable traits, leading to the early emergence of different breeds (FAO 2015). The eighteenth century saw the commencement of a more organised approach to livestock production that led to rapid change in livestock breeds (Hall and Clutton-Brock 1988), pioneered by the likes of Robert Bakewell, one of the first British agriculturists to implement a systematic breeding method for horses (*Equus caballus*), sheep (*Ovis aries*) and cattle (*Bos taurus taurus*). Animals were monitored and carefully selected to breed for improved performance of key production traits from earlier maturation and increased prolificacy to growth rate. Breeds gradually evolved to fill a variety of functional roles from the single-purpose meat-providing pig (*Sus scrofa*) to the multi-purpose meat, dairy and wool sheep and working, hunting and guard dogs (*Canis lupus familiaris*). Furthermore, appearance-related characteristics, such as ear shape, body shape and coat colour, were also subject to selective breeding to create populations exhibiting particular phenotypes. This led to the development of distinctive breeds each possessing a collection of shared phenotypic attributes, as exemplified for pig and cattle breeds in Figs. 1 and 2. From the late eighteenth and early nineteenth century, breed societies and kennel clubs were founded, which can be viewed as the first acknowledgement of the genetic and phenotypic distinctiveness of breeds (Hall and Clutton-Brock 1988). By requiring animals to meet morphological criteria and keeping the herdbooks closed (thereby confining the breeding pool), these organisations were instrumental in maintaining the phenotypic integrity and uniformity within breeds. This complex history of population-shaping events stretching several hundred years led to the development of an extraordinary number of animal breeds, with a current estimated worldwide count of 1,019, 1,514, 694, 543 and 1,155 for cattle, chickens, horses, pigs and sheep, respectively (FAO 2015).

The next major chapter in the history of domestic animals was the introduction of a highly systematic management to achieve greater efficiency and output for meat, dairy and reproduction traits, from circa 1950s onwards. Intensive selection was channelled through the development of sophisticated statistical methods (e.g. estimated breeding values (EBVs)) thereby enabling the identification of animals with superior genetics for desired traits combined with advancements in reproductive technologies, such as artificial insemination. This produced a framework of breeding

Fig. 1 An assemblage of European pig breeds to illustrate the array of phenotypic diversity present in domestic pigs. Clock-wise from *top-left*: Wild boar, Landrace, Large Black, Tamworth, Gloucestershire Old Spots and British Saddleback. Compiled by Ian Hesketh, with photo attributes clockwise from *top-left*: Luc Viatour [GFDL, CC BY-SA 3.0 or CC BY-SA 2.5-2.0-1.0], Zeilog (Own work) [GFDL or CC BY-SA 3.0], Tamsin Slater [CC BY-SA 2.0], Keith Evans [CC BY-SA 2.0], Jon Whitton [CC BY-NC-ND 2.0], Amanda Slater from Coventry, England (Gloucester Old Spot Boar) [CC BY-SA 2.0] and jon smith "una nos lucror" from Stamford, England (saddleback pig) [CC BY-SA 2.0]. All via Wikimedia Commons except for the Large Black by Keith Evans via geograph (www.geograph.org.uk) and the Tamworth by Jon Whitton via Flickr (https://www.flickr.com/photos/jwhittox/2371374460/in/photostream/). License abbreviations correspond as follows: CC BY-SA 2.0, CC BY-SA 2.5-2.0-1.0, CC BY-SA 3.0, CC BY-NC-ND 2.0 and GFDL are https://creativecommons.org/licenses/by-sa/2.0, https://creativecommons.org/licenses/by-sa/2.5-2.0-1.0, https://creativecommons.org/licenses/by-sa/3.0, https://creativecommons.org/licenses/by-nc-nd/2.0/ and www.gnu.org/copyleft/fdl.html

pyramids where the genetics of a few breeding individuals contributed to the gene pools of large commercial populations, particularly for cattle, pigs and chickens (*Gallus gallus domesticus*). Gains were also concentrated on choice high-performing breeds within species, leading to the dominance of certain breeds in the livestock industry, such as Holstein-Friesian cattle for milk production. Although output from the livestock sector grew at an unprecedented rate in the latter half of the twentieth century, the consequence of focusing selection on a small set of breeds was the extinction of breeds considered less productive. There was an estimated loss of 184, 60, 87, 107 and 160 breeds, for cattle, chickens, horses, pigs and sheep, respectively, during the twentieth century (FAO 2015).

Fig. 2 An assemblage of European taurine cattle breeds to illustrate the array of phenotypic diversity present in domestic cattle. Clock-wise from *top-left*: Murray Grey, Aberdeen Angus, Limousin, Jersey, Hereford, Holstein, Charolais and Montbeliarde. Compiled by Ian Hesketh, with photo attributes clockwise from *top-left*: Cgoodwin (Own work) [GFDL or CC BY 3.0], brittgow (CC BY-SA 2.0), Budotradan (Own work) [CC BY-SA 3.0], Storye book (Own work) [CC BY 3.0], User Robert Merkel on en.wikipedia (US Department of Agriculture) [Public domain], Keith Weller/USDA (www.ars.usda.gov: Image Number K5176-3) [Public domain], Robert Scarth [flickr.com ("Taken by me, Robert Scarth")] [CC BY-SA 2.0] and groms78 (photo prise dans une prairie du Haut-Doubs) [GFDL or CC BY-SA 3.0]. All via Wikimedia Commons, except for the Aberdeen Angus photo via Flickr (https://www.flickr.com/photos/brittgow/4782264442). License abbreviations correspond as follows: CC BY-SA 2.0, CC BY-SA 3.0, CC BY 3.0 and GFDL are https://creativecommons.org/licenses/by-sa/2.0, https://creativecommons.org/licenses/by-sa/3.0, https://creativecommons.org/licenses/by/3.0 and www.gnu.org/copyleft/fdl.html

2.2 *Pre-genomic Research on the Genetic Basis of Animal Domestication*

As a result of the fundamental importance of domestic animals for human society, great effort has gone into mapping loci, identifying genes, detecting causative variants and unravelling biological mechanisms associated with production traits, such as growth, milk production and meat quality, and key breed-defining characteristics, such as coat colour (see reviews: Andersson and Georges 2004; Goddard and Hayes 2009; Wiener and Wilkinson 2011). Additionally, domestic animals have long been exploited as model systems to investigate the complexity of genotype–phenotype relationships (Andersson and Georges 2004; Megens and Groenen 2012). This is because the population-shaping events that occurred during the development of domestic breeds happened over a relatively short timescale, and this was followed by a strict maintenance of breed phenotypic uniformity, and in so

doing, the associated genetic changes have been amassed and conserved within contemporary breeds. Prior to the availability of high-density genomic resources, microsatellites were employed in linkage mapping studies of experimental crosses to isolate quantitative trait loci (QTL) (e.g. ear variation in pig breeds, Wei et al. 2007) and diversity studies to characterise the genetic variation amongst breeds (e.g. genetic diversity in chicken breeds, Wilkinson et al. 2012). This foundation of work on the genetic basis of animal domestication and breed development informs the design and interpretation of subsequent population genomic studies.

3 Tools to Characterise Genomic Variation in Domestic Animals

The complexity of domesticated animals, evidenced by the large number of breeds, different production types and diversity of phenotypes, requires a range of population genomic approaches to decipher the genetic basis of animal domestication and breed development. One approach is to conduct large-scale population genomic studies sampling hundreds of commercial and traditional breeds from diverse geographical locations worldwide (e.g. cattle, Bovine HapMap Consortium 2009; sheep, Kijas et al. 2012). Alternatively, the focus can be narrowed to concentrate on the genetics of specific breeds, for example, dominant commercial breeds (e.g. Holsteins, Hayes et al. 2008), commercial lines within a breed (e.g. broiler chicken lines, Stainton et al. 2015), pooled breeds based on a shared phenotype (e.g. ear type in pigs, Wilkinson et al. 2013), production types (e.g. beef versus dairy cattle, Hayes et al. 2009) and breeds from different geographical locations (e.g. African taurine versus European taurine cattle, Orozco-terWengel et al. 2015). Furthermore, the ancestral wild progenitors of some domestic species still exist today in the wild and incorporating ancestral genotypic data into analyses provides a unique opportunity to gain additional insight into the domestication process (Muir et al. 2008).

The two key requirements to conduct population genomic studies on domestic breeds, dense genotypes and statistical methodology, are briefly described below.

3.1 Sequencing and Single Nucleotide Polymorphism Arrays

The technological infrastructure required to obtain high-density genotype data for population genomic studies is well established for domestic animals. High-quality reference genomes have been assembled for most major domestic species (cattle, Bovine Genome Sequencing and Analysis Consortium 2009; chicken, International Chicken Genome Sequencing Consortium 2004; dog, Lindblad-Toh et al. 2005; goat, Bickart et al. 2017; horse, Wade et al. 2009; pig, Archibald et al. 2010; sheep,

Jiang et al. 2014). DNA sequencing of pools of animals led to the discovery of millions of genetic variants in the domestic genomes. For example, roughly 2.8 million SNPs were typed for chickens (International Chicken Polymorphism Map Consortium 2004) and 2.2 million SNPs were detected in cattle (Bovine HapMap Consortium 2009). This has led to the development of custom-made commercial SNP chips containing thousands of genetic variants spread across the genomes for cattle (Matukumalli et al. 2009), chickens (Groenen et al. 2011; Kranis et al. 2013), dogs (Lindblad-Toh et al. 2005), goats (Tosser-Klopp et al. 2014), horses (McCue et al. 2012), pigs (Ramos et al. 2009) and sheep (Kijas et al. 2009). The inexpensive availability of these SNP panels allows for samples to be genotyped quickly, in bulk and at a modest cost. Furthermore, as prices continue to fall, next-generation sequencing is increasingly being used to discover and characterise variants within empirical studies (e.g. cattle breeds, Hayes et al. 2010; pig breeds, Amaral et al. 2011).

3.2 Population Genomic Methods

Levels of genomic diversity in breeds can be characterised by estimating a range of parameters, such as heterozygosity (H) (e.g. chickens, Rubin et al. 2010), nucleotide diversity (π) (e.g. chickens, International Chicken Polymorphism Map Consortium 2004) and Watterson's nucleotide diversity (θ) (e.g. pigs, Amaral et al. 2011). An additional key parameter to consider as a diversity estimate is the effective population size (N_e), which is the number of reproducing individuals in an idealised population. It is an important concept in population genetics because genetic drift, the predicted rate of loss of genetic diversity in a finite population, is directly related to N_e (Kilman et al. 2008). Hill (1975, 1981) also showed that the levels of linkage disequilibrium (LD) between genetic markers are shaped by changes in N_e over time and thus can be used to estimate current and past N_e. LD between closely linked loci reflects the historic population size whilst LD between distant loci is indicative of more recent N_e (Hill 1975, 1981). Extending this concept, Hayes et al. (2003) showed that, assuming a linear increase in population size, LD between loci at a specific recombination distance (c) reflects ancestral N_e $1/2c$ generations in the past. Thus, LD patterns across the genome can be leveraged to chart the demographic history of breeds.

Patterns of localised LD and genetic diversity can also be exploited to identify genomic regions that may have experienced diversifying selection. When a locus is driven towards fixation due to selection, neutral loci in LD with the selected locus will also show this pattern (Maynard Smith and Haigh 1974). This phenomenon is termed as the "hitchhiking effect" and whilst the selected locus is the target, the genomic region as a whole will experience a reduction in genetic diversity due to the hitchhiking effect. This genomic signature will gradually erode as recombination events occur within the region. There are several within- and between-population

selection mapping approaches (i.e. methods to identify genomic regions that show evidence of past or current selection) designed to exploit these genetic patterns, thereby identifying candidate signals associated with selection.

An easily applicable approach is to scan the genomes within breeds for regions that exhibit measureable reductions in genetic diversity (i.e. high levels of homozygosity) by estimating measures such as average observed heterozygosity (H_{obs}) (e.g. dogs, Quilez et al. 2011), nucleotide site diversity (h) (e.g. chickens, Stainton et al. 2017), pooled heterozygosity (H_P) (e.g. chickens, Rubin et al. 2010) and standardised heterozygosity (ZH_P) (e.g. pigs, Rubin et al. 2012). A commonly used between-population approach is to measure genetic differentiation between two or more populations using F_{ST} (Wright 1951) and its derivatives such as d_i (Akey et al. 2010). These methods measure the variation in allele frequencies amongst two or more populations to identify regions of differentiation between populations where SNPs with the highest genetic differentiation are considered candidates of differential selection (Cavalli-Sforza 1966; Lewontin and Krakauer 1973). A regular practice with both F_{ST} and diversity genome scans is to apply a sliding window to smooth out stochastic variation between SNPs that may arise from genetic drift (Weir et al. (2005), who suggested this for F_{ST} single-locus estimates). This is a logical step to adopt in selection mapping analyses as the hitchhiking phenomenon leads to a localised allelic fixation for a set of neighbouring markers and a sliding window helps to distinguish these from stochastic locus-by-locus variation. A large number of studies in dogs and various livestock species have successfully identified isolated signatures of selection by estimating genetic differentiation between breeds (e.g. Flori et al. 2009; Akey et al. 2010; Kijas et al. 2012; Wilkinson et al. 2013; Kemper et al. 2014).

An alternative set of approaches characterises the genetic diversity of extended genomic regions. The "long-range haplotype" (LRH) test (Sabeti et al. 2002) identifies regions of slow decay in homozygosity, which are indicative of alleles that rapidly arose to high frequency due to selection (dragging with it neutral variants via hitchhiking) before recombination could break down the long-range LD. The related "integrated Haplotype Score" (iHS) measures the extent of decay in extended haplotype homozygosity (EHH) around a core haplotype for the derived allele relative to the ancestral allele (Voight et al. 2006). The XP-EHH statistic extends these methods by incorporating a cross-population approach to identify regions of homozygosity in one population relative to polymorphic regions in another population (Sabeti et al. 2007). Examples of the application of these methods in livestock include studies on cattle (e.g. Rothhammer et al. 2013; Kemper et al. 2014), pigs (e.g. Ma et al. 2015; Li et al. 2016), chickens (e.g. Li et al. 2012a; Liu et al. 2016) and sheep (e.g. Lv et al. 2014). There are several additional selection mapping methods also designed to identify signatures across the genome characteristic of the hitchhiking effect: diversity and/or LD patterns (Kelly 1997; Kim and Nielsen 2004; Wiener and Pong-Wong 2011; Jacobs et al. 2016) and extreme allele frequencies or significantly distorted site frequency spectra (Kim and Stephan 2002; Nielsen et al. 2005).

Finally, the genome-wide association approach is also worth mentioning. In a genome-wide association study (GWAS), the genotype (i.e. each SNP) is regressed

onto the phenotype to identify SNPs significantly associated with a particular phenotype, which can include traits important in domestication or breed development (e.g. those related to production or breed characteristics). This technique can be applied within breeds, where quantitative measurements are used as the phenotypes, or across breeds, where breed averages of the measurements are commonly used.

4 Genome-Wide Diversity of Domestic Animals

Knowledge of the demographic trends in animal domestication and breed development suggests that the genomic landscape of domestic breeds has changed markedly over time. Characterisation of genomic patterns such as LD and genome-wide diversity has provided further insights on the historical narrative of domestic animals and on the contemporary genetic diversity of breeds.

4.1 Change in Population Size

The patterns of LD and trajectories of N_e over time largely mirror the changes in population sizes that domestic animals experienced during domestication and breed development. Across a range of domestic species, estimates of N_e for ancestral populations were relatively large prior to domestication and fell slowly after domestication, as subsets of animals were derived from wild progenitor populations during the early stages of domestication. A more rapid decline in N_e followed after breed formation, and this was likely due to the commencement of systematic breeding in the eighteenth century (cattle, Bovine HapMap Consortium 2009; Orozco-terWengel et al. 2015; dogs, Gray et al. 2009; Boyko et al. 2010; Stern et al. 2013; pigs, Uimari and Tapio 2011; Badke et al. 2012; Ai et al. 2013; sheep, Kijas et al. 2012, 2014; horses, McCue et al. 2012).

The intensity of human-mediated pressures has also varied amongst breeds and this is reflected in different estimates of N_e between breeds, as evaluated from differences in LD decay. Disentangling LD patterns amongst taurine cattle breeds reveals that the commercial milking breeds Brown Swiss, Holstein and Jersey have amongst the highest LD at both shorter and longer distances, suggesting small early domesticated populations and subsequent population contraction due to intense selection along with the use of popular sires (Bovine HapMap Consortium 2009). Likewise, the commercial beef taurine breeds, such as Aberdeen Angus, Charolais and Limousin, have high levels of LD at shorter distances (Bovine HapMap Consortium 2009; Hoze et al. 2013). However, the decay in LD with increasing distance is faster for these breeds such that LD levels are on average slightly higher in dairy than beef breeds at intermediate to longer distances (Hoze et al. 2013), suggesting small early domesticated populations but less severe population contraction for the beef breeds. These LD patterns are reflected in the current N_e of these breeds: very

low for the mainstream dairy breeds (<100), whilst slightly higher for the beef breeds (110–174) (Bovine HapMap Consortium 2009). Indicine (*Bos taurus indicus*) breeds, on the other hand, have lower LD levels, especially at shorter distances, suggesting that their ancestral population was far larger than that of taurine cattle (Bovine HapMap Consortium 2009; Pérez O'Brien et al. 2014).

In contrast to cattle, levels of LD between markers at both shorter and longer distances are notably lower in sheep breeds, which is reflected in the relatively high current N_e estimates for sheep breeds (Kijas et al. 2012, 2014). Even commercially important sheep breeds, like the Scottish Blackface and Scottish Texel, have fairly high levels of N_e, estimated at 528 and 305, respectively (Kijas et al. 2012). Kijas et al. (2014) postulated that the relatively higher N_e estimated for sheep breeds is due to a broader sampling of ancestral genetic pools at the early stages of domestication, followed by milder bottleneck events and more extensive gene flow amongst populations during the formation of breeds.

Dog breeds are a compelling example of the pressure that domestication and breed development can have on the genomic landscape. Similar to cattle, dogs possess high levels of LD at short distances (Lindblad-Toh et al. 2005; Boyko et al. 2010; Stern et al. 2013), but as the distance between loci increases, the decay in LD is much slower in dogs compared to other domestic animals (Fig. 3). This persistent and extended level of high LD concurs with the documented history of dogs, with this domestic species experiencing arguably the most intense of population pressures, including extremely small founder populations, an absence of gene flow and breeding of highly related individuals to fix fashionably desired traits. However, the extent of intermediate to long-range LD varies between breeds, as can be seen in Fig. 3, from a very slow decay in LD levels observed in the Boxer to a very sharp decay in the Jack Russell Terrier (Boyko et al. 2010). As with other domestic species, the differences in LD patterns suggest that the magnitude of selection (and with it, bottlenecks) varied between breeds, which may be due in part to human keenness to mould certain breeds.

4.2 Genetic Diversity

The erosion of genetic diversity in domestic breeds caused by long and continued intense selection for genetic improvement is an increasing concern amongst animal geneticists and policy-makers (e.g. DEFRA 2006, 2009; Ajmone-Marsan and GLOBALDIV Consortium 2010; Bruford et al. 2015; FAO 2015). Reductions in estimated N_e over time in domestic animals as well as evidence of diminished genetic diversity are often used to exemplify the consequences of human-mediated selection pressures on population genetic structure (Ajmone-Marsan and GLOBALDIV Consortium 2010; Groeneveld et al. 2010; Bruford et al. 2015).

By exploiting the ancestor–descendant relationship, Muir et al. (2008) investigated the levels of SNP diversity in broilers, layers, non-commercial chicken breeds (i.e. fanciers) and the wild progenitor of the domestic chicken, the Red Jungle Fowl

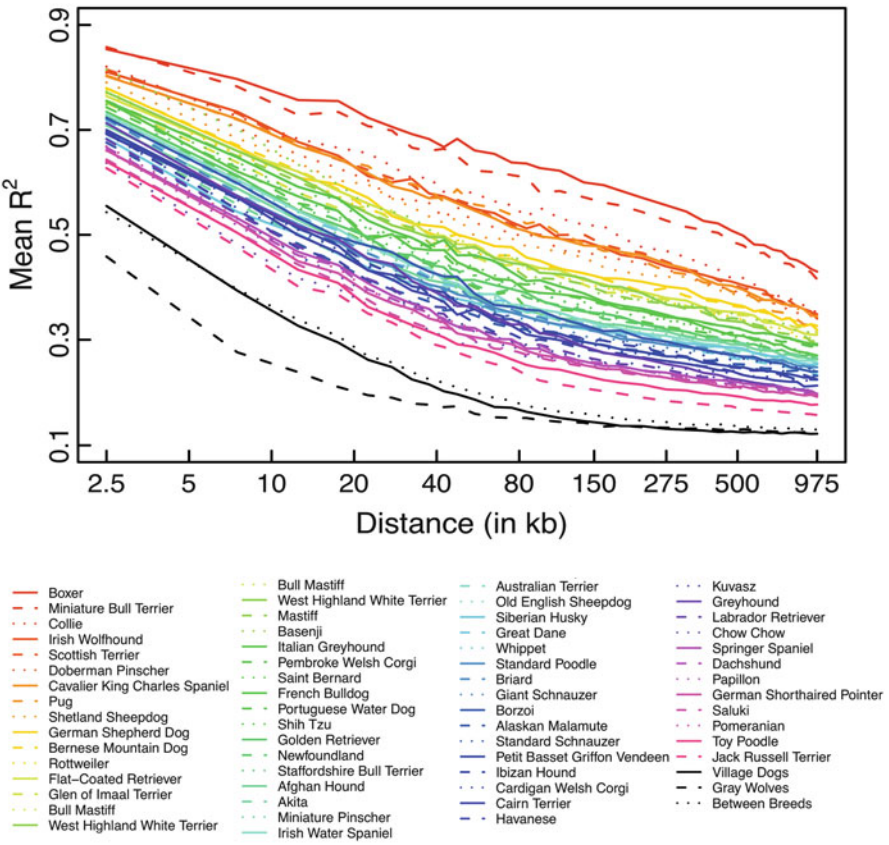

Fig. 3 Patterns of linkage disequilibrium (LD) in dogs shows extensive long-range LD within most breeds but not across breeds. Shown are LD decay curves (mean r^2) within dog breeds, village dogs and gray wolves, and between breeds (calculated from dogs selected from 10 different breeds). This figure is adapted from the article "A simple genetic architecture underlies morphological variation in dogs" by Boyko et al. (2010). PLoS Biology;8:e1000451 (https://doi.org/10.1371/journal.pbio.1000451). The original article is open access distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

(*Gallus gallus*). The study found that an estimated 50% of ancestral alleles were not present in current commercial chicken populations, suggesting a massive loss in ancestral genetic diversity over the course of several bottlenecks. Furthermore, SNP detection through whole genome sequencing found lower levels of heterozygosity in commercial poultry lines (range of 0.74×10^{-3}–3.88×10^{-3}) compared to a pool of Red Jungle Fowl birds (4.07×10^{-3}) (Rubin et al. 2010). Cattle breeds are another example used to demonstrate the effects that intense selection pressures and industrial consolidation may have had on genomic diversity. Despite the global population size of commercial dairy cattle breeds numbering in the millions, the N_e of these

breeds is very low (~150) after a sharp decline from a large ancestral population (N_e of ~90,000) (Bovine HapMap Consortium 2009; Qanbari et al. 2010; OrozcoterWengel et al. 2015). The genomic diversity within dog breeds has also apparently declined, exemplified by the presence of extensive runs of homozygosity (ROHs) extending over hundreds of kilobases and scattered across more than 25% of the genome of many dog breeds (Boyko et al. 2010; Vaysse et al. 2011). Likewise, N_e is extremely low for most dog breeds, where one study examining population size amongst 112 breeds found an extremely narrow range of N_e, with estimates extending from 53 for the Bull Terrier to 230 for the Chihuahua (Dreger et al. 2016).

Loss of genetic diversity may have a negative impact on the biological fitness of individuals and an adverse effect on the long-term viability of populations, as conservation genetic theory stipulates that genetic diversity provides species and populations with adaptive potential (Frankham et al. 2010). Unfavourable health changes that have accompanied selection in some domestic breeds provides evidence that genetic erosion could have a detrimental effect, as discussed for chickens (Dawkins and Layton 2012), cattle (Oltenacu and Broom 2010) and dogs (Bateson and Sargan 2012; Farrell et al. 2015). In commercial dairy cattle, one negative effect that has accompanied genetic gains in milk production is a reduction in fertility occurring in the latter half of the twentieth century (Walsh et al. 2011). Research shows that there was a genetic component to this trend, with a negative genetic correlation between milk production and reproductive performance (Berry et al. 2016). Recognising the problem of reduced fertility for the dairy industry, further study has revealed that genetic variation exists for the trait: although the heritability for fertility is low, at $<10\%$, the coefficient of genetic variation (a unit-free measure of the magnitude of genetic variation) is similar to some production traits (Berry et al. 2016). This suggests that fertility traits can be incorporated into breeding objectives (Wall et al. 2003) and, indeed, improvements in the reproductive performance of dairy cattle have been observed from the early twenty-first century (Pryce et al. 2014).

Despite indications that intense selection pressures and industrial consolidation have negatively impacted genome-wide variation within some breeds, there is also evidence that a healthy reservoir of genetic diversity still exists in many domestic species. First, nucleotide diversity within certain breeds has been found to be reasonably high. Despite the low N_e of cattle breeds, the Bovine HapMap Consortium (2009) found a considerable amount of nucleotide diversity within Aberdeen Angus and Holstein ($\sim 1.4 \times 10^{-3}$), reportedly higher than that found in human populations. Many European pig breeds have been found to have higher or similar levels of genomic diversity to their ancestral progenitor, the European wild boar (Amaral et al. 2011; Bosse et al. 2012). This likely reflects a combination of past events: a reduction in population size of European wild boar and the introgression of Asian pig alleles into the European pig breed gene pool to introduce desired traits. Second, even where genetic homogeneity is observed *within* many breeds, domestic species as a whole (i.e. pooling across breeds) still harbour considerable levels of genetic diversity. Unlike the persistent and extensive LD levels observed at longer distances within dog breeds, across dog breeds (i.e. mean LD decay), there is

a rapid decay in LD with increasing distance (Fig. 3), similar to that found in free-ranging "village dogs" and humans (Lindblad-Toh et al. 2005; Boyko et al. 2010). Additionally, a comprehensive SNP discovery analysis of 24 chicken lines found that whilst a substantial proportion (~23%) of detected SNPs were shared amongst the different chicken types (broiler, layer and inbred lines), a key observation was that a notable percentage of SNPs were unique to each type (9.3%, 20.5% and 0.8%, respectively) (Kranis et al. 2013). A conservation scheme proposed by Muir et al. (2008) for poultry stocks, but equally applicable for other domestic animals, is that lost diversity within breeds can be (partially) recovered by outcrossing to improve within-breed variability.

5 Mapping Genomic Variation Associated with Phenotypic Traits in Domestic Animals

Domestic animals have been selected for a suite of biological characteristics, which can be roughly grouped into two categories: (1) visually striking hallmark physical characteristics that are used to define breeds and (2) production traits that are selected to increase output. Mapping signatures of selection using population genomic approaches has illuminated our understanding of the genetics underpinning phenotypic traits, the strength of selective pressures and the traits preferred during domestication and breed development.

5.1 Breed-Defining Appearance Traits

5.1.1 Coat Colour

Since early domestication, domestic animals have been selected for a striking array of coat colours and patterns (see Figs. 1 and 2; Cieslak et al. 2011). The genetics of coat colour has longed intrigued biologists, including Sewall Wright who addressed pigment production in mammals (Wright 1917). Following detailed genetic dissection of coat colour mutations in mouse (*Mus musculus*) (Jackson 1991), studies have investigated coat colour in domestic animals, looking at modes of inheritance, gene mechanisms, causative mutations, gene interactions and constructing the pathways involved in pigment cell development (Cieslak et al. 2011; Linderholm and Larson 2013).

Population genomics has built on this work, uncovering evidence of diversifying selection at major coat colour genes (e.g. *ASIP*, *EDNRB*, *KIT*, *KITLG*, *MC1R*, *MITF*, *PMEL17* and *TYRP1*; Cieslak et al. 2011) in domestic animals. In cattle, a common coat colour pattern is the pied animal, whereby pigmented spots are displayed on a white background, as seen for black and white pied Holstein and red and white pied Montbeliarde breeds. An F_{ST} contrast between these two pied breeds detected a

signature of selection at the *MC1R* gene (Flori et al. 2009). Hayes et al. (2010) carried out a GWAS of Holstein, with the proportion of black present on the coat as the quantitative phenotype, and detected signals of association at the *KIT* and *MITF* genes, along with a third locus (*PAX5*). For other coat phenotypes, Kemper et al. (2014) found evidence of selection at several coat colour genes in a number of cattle breeds using either within- (iHS) or between-population (F_{ST}) selection mapping methods: *MC1R* (Limousin, Charolais, Aberdeen Angus, Holstein and Murray Grey), *KIT* (Hereford and Holstein), *PMEL* (Charolais, Aberdeen Angus and Murray Grey) and *KITLG* (Hereford). In sheep, a strong F_{ST} signal distinguished white versus non-white breeds at SNPs genotyped within *ASIP* (Li et al. 2013) and in a global F_{ST} scan of a collection of sheep breeds, Kijas et al. (2012) reported signals underlying *KIT* and *MITF*. Likewise, population genomic studies of dog breeds have detected signatures of diversifying selection underlying coat colour genes such as *ASIP*, *MC1R* and *MITF* (Karlsson et al. 2007; Akey et al. 2010; Boyko et al. 2010).

In contrast, identifying signatures of selection at the major coat colour genes *KIT* and *MC1R* has been inconsistent in pig breeds, with some studies reporting sweeps (Amaral et al. 2011; Ma et al. 2015) and others failing to detect signals (Rubin et al. 2012; Wilkinson et al. 2013), although there is evidence that these genes control coat colour in some pig breeds (Andersson and Plastow 2011). Strong signatures of selection have, however, been found at other coat colour genes in pig breeds. Wilkinson et al. (2013) conducted an F_{ST} scan of pig breeds and identified genetic signals of diversifying selection near *EDNRB* in the spotted Gloucestershire Old Spots (Fig. 4) and near *KITLG* in the Berkshire (black with white extremities). It was proposed that the two phenotypes are of an Asian origin, the result of introgression of Asian pigs into the European pig breeding pool from the late eighteenth century. Sequence data for *EDNRB* revealed two non-synonymous changes in the first exon of the endothelin receptor B gene unique to Gloucestershire Old Spots amongst a set of European pig breeds, but shared by some Asian pig breeds. Selection mapping solely in Chinese pig breeds also detected strong signatures of selection at the *EDNRB* locus in white belted pigs (Ai et al. 2013; Wang et al. 2015) and weaker signatures of selection in black pigs with the white extremity phenotype (Lu et al. 2016), supporting the role of this gene in coat colour in Asian pigs.

Beyond colourful coat patterning observed in domestic breeds, the genetic basis of additional coat attributes has been investigated. Cadieu et al. (2009) performed a multi-breed dog GWAS of three furnishing traits: moustache and eyebrows, hair length and hair curl and identified three genomic regions, containing the genes *RSPO2*, *FGF5* and *KRT71*, with these three genes accounting for most (80%) of the variation for these phenotypes.

5.1.2 Body Size

Within domestic animal species, there is an extraordinary range in body size, from the diminutive 70 cm Shetland Ponies to the 2 m tall Percheron horse or the smallest dog breed, the Chihuahua weighing in at ~1.5 kg being towered by the Great Dane,

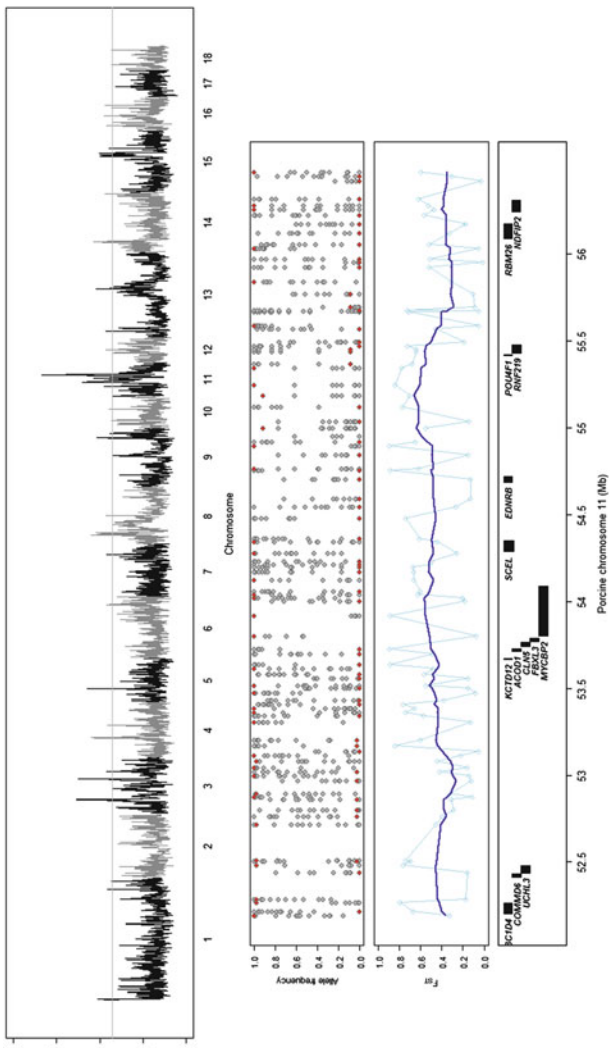

Fig. 4 A genome scan of differential selection in the Gloucestershire Old Spots pig breed revealed a strong selective sweep on porcine chromosome 11, near the coat colour gene *EDNRB*. The *top panel* shows levels of genome-wide genetic differentiation (13-SNP F_{ST} sliding window) between Gloucestershire Old Spots and 12 other European pig breeds, plotted with respect to genomic position, with the horizontal *grey line* representing the 99th percentile. The strongest signal was on porcine chromosome 11 from 52.19 to 56.47 Mb and zooming in on this genomic region: the *second panel* from the top plots the allele frequencies in Gloucestershire Old Spots (*red dots*) and other European pig breeds (*grey dots*), the *third panel* plots 13-SNP F_{ST} sliding window (*dark blue line*) and single SNP (single nucleotide polymorphism) F_{ST} estimates (*light blue line*) and the *bottom panel* shows the genes, including *EDNRB* (54.69–54.72 Mb), present in the selective sweep region

which weighs up to 90 kg. In dogs, a linkage mapping study of Portuguese Water Dogs detected at least six QTLs associated with the genetic architecture of body size, with a QTL of large effect found on chromosome 15 (Chase et al. 2002). A subsequent genome scan revealed an interval of high F_{ST} differentiating small and large dog breeds on chromosome 15 centred around the gene insulin growth factor 1 (*IGF1*) (Sutter et al. 2007), with a 20-SNP haplotype at the gene accounting for 15% of variation in body size. Three multi-breed selective sweep analyses detected genetic variants on several chromosomes associated with body size in dogs (Fig. 5), including the *IGF1* locus (Akey et al. 2010; Boyko et al. 2010; Vaysse et al. 2011), concurring with the earlier linkage mapping study that the genetic basis of skeletal size in dogs comprises several genomic regions. The studies also highlighted several candidate genes from the other chromosomal regions, including *HGMA2*, *SMAD2*, *STC2* and *LCORL*, most of which have been previously associated with body size variation in other mammalian species (e.g. humans, Lango Allen et al. 2010).

The genes associated with body size in dogs have likewise been identified underlying signatures of selection for the same trait in other domestic animals. A multi-breed GWAS of small versus large horse breeds revealed four genomic regions (including the *LCORL*, *HMGA2* and *ZFAT* genes) accounted for 83% of the variation in stature in horses (Makvandi-Nejad et al. 2012). Whilst cattle have not been strongly selected for body size (see Fig. 2), there are differences in stature amongst breeds and a comprehensive review (Randhawa et al. 2014) catalogued 12 strong genetic signals for this trait, with *LCORL*, *PLAG1* and *SMAD2* genes repeatedly identified in multiple cattle breed selection mapping studies (Bovine HapMap Consortium 2009; Flori et al. 2009; Rothhammer et al. 2013; Druet et al. 2013; Kemper et al. 2014; Zhao et al. 2015). Similarly, the genome of commercial pig breeds was scanned using a ZH_P approach and regions of homozygosity were identified that contain *LCORL* and *PLAG1* genes (Rubin et al. 2012).

5.1.3 Facial Profiles

Facial morphology of domestic animals encompasses a number of external features and variation in size, shape and length of different anatomical parts can together result in a montage of diverse facial profiles, as can be seen for pig breeds in Fig. 1. One key breed-defining facial feature is the size and carriage of ears in domestic dogs and pigs. Early linkage mapping studies followed by selective sweep analyses have illuminated the genetic architecture of this trait in pig breeds (for detailed discussion see Case Study 1).

Case Study 1 Unravelling the Genetic Basis of Variation in Ear Morphology in Pig Breeds

One subtle but striking physical characteristic that varies amongst pig breeds is ear morphology. The ancestral state of this phenotypic trait is an upright or

(continued)

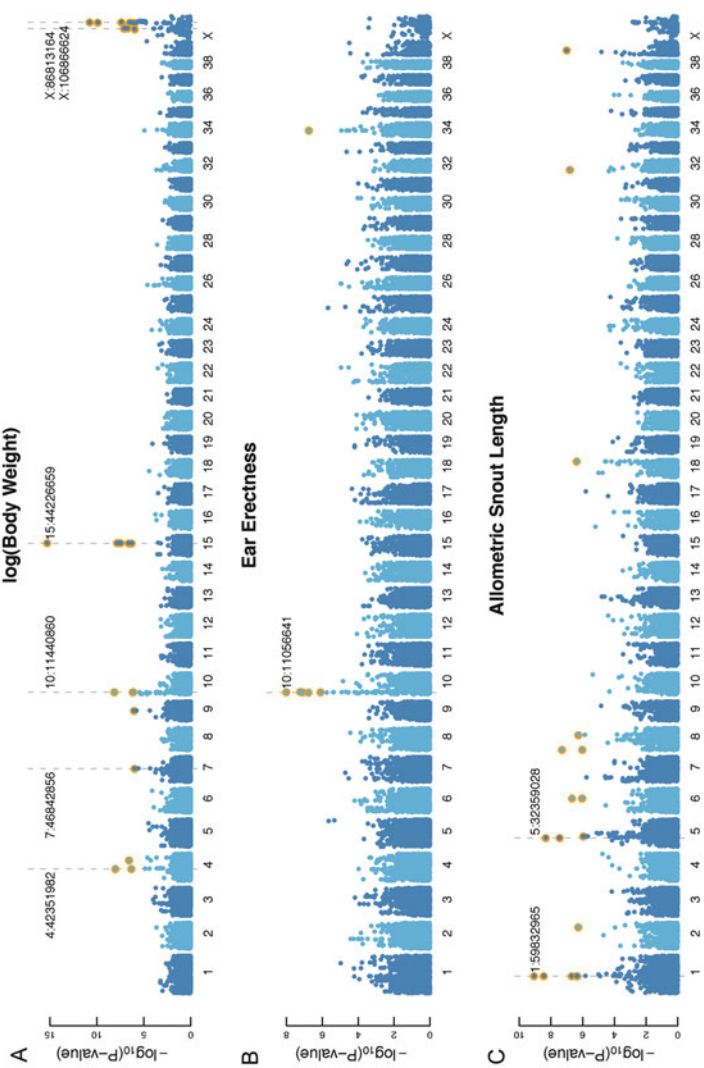

Fig. 5 A genome-wide association study (GWAS) of breed-defining external characteristics of dog breeds revealed a simple genetic architecture for these traits. Manhattan plots are shown for three phenotypic traits traits: log(body weight), ear erectness (floppy versus erect ears) and allometric snout length. Each Manhattan plots displays the $-\log_{10}(P\text{-value})$ of association of each SNP with the phenotype with respect to genomic position on the canine genome. *Orange circles* represent SNPs that passed Bonferroni correction and the *grey dashes* indicate SNPs included in best-fit predictive models. This figure is adapted from the article "A simple genetic architecture underlies morphological variation in dogs" by Boyko et al. (2010). PLoS Biology;8:e1000451 (https://doi.org/10.1371/journal.pbio.1000451). The original article is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

prick ears, as seen in wild boar. In pig breeds, ears vary from an upright stance to a more intermediate phenotype that is partially upright and then to floppy long ears that drape down over the front of the face (see Fig. 1).

Ear phenotypes are a major pig breed-defining characteristics and ascertaining its genetic control will enhance understanding the genes and pathways governing morphological diversity. QTLs associated with ear morphology were first identified in intercrosses between the prick-eared Large White pigs and floppy-eared Meishan pigs (Wei et al. 2007). Amongst several QTLs, those with major effect and significant at the genome-wide level ($P < 0.01$) were detected on chromosomes 5 and 7 of the porcine genome for both ear size and erectness measurements (chromosome 5: at 51 cM and 43 cM and chromosome 7: both at 70 cM) (Wei et al. 2007). A larger study on a cross between intermediate-eared Duroc and floppy-eared Erhualian identified 23 genome-wide significant QTLs, with the largest effect found on chromosome 7 (56–60 cM), explaining more than 40% of the variance in ear weight and area and 15.7% of variance in ear erectness (Ma et al. 2009). The next largest QTL was on chromosome 5 (59–73 cM), explaining more than 12% of the variance in ear weight, area and erectness.

To determine if population genomics could deepen the understanding of the genetic architecture of ear morphology traits, Wilkinson et al. (2013) scanned the porcine genome for regions of genetic differentiation between diverse ear phenotypes in European pig breeds. Pig breeds were grouped into different ear categories and F_{ST} pairwise sliding window mapped three regions showing very strong signatures of selection ($F_{ST} > 0.4$), one on chromosome 5 and two on chromosome 7 (Fig. 6). The region on chromosome 5 extended from 31.74–33.78 Mb and was uncovered in the differentiation between prick- vs. floppy-eared pigs and intermediate- vs. floppy-eared pigs. The first region on chromosome 7 extended from 31.86–34.19 Mb and differentiated only the prick- vs. intermediate-eared pigs. Finally, the second region on chromosome 7 extended from 55.43–58.19 Mb was uncovered in the differentiation between prick- vs. intermediate-eared pigs and intermediate- vs. floppy-eared pigs.

The findings from Wilkinson et al. (2013) are in concordance with those from the earlier linkage mapping studies, that ear morphology in pig breeds is likely controlled by two or three loci of major effect located on chromosomes 5 and 7, with additional QTLs of smaller effect scattered across the remainder of the porcine genome. Furthermore, the signal on porcine chromosome 5 is a strong candidate controlling mammalian ear morphology because its genomic position is orthologous to a region on the canine genome associated with ear morphology in dog breeds (see Fig. 5) (Boyko et al. 2010; Vaysse et al. 2011; Webster et al. 2015).

(continued)

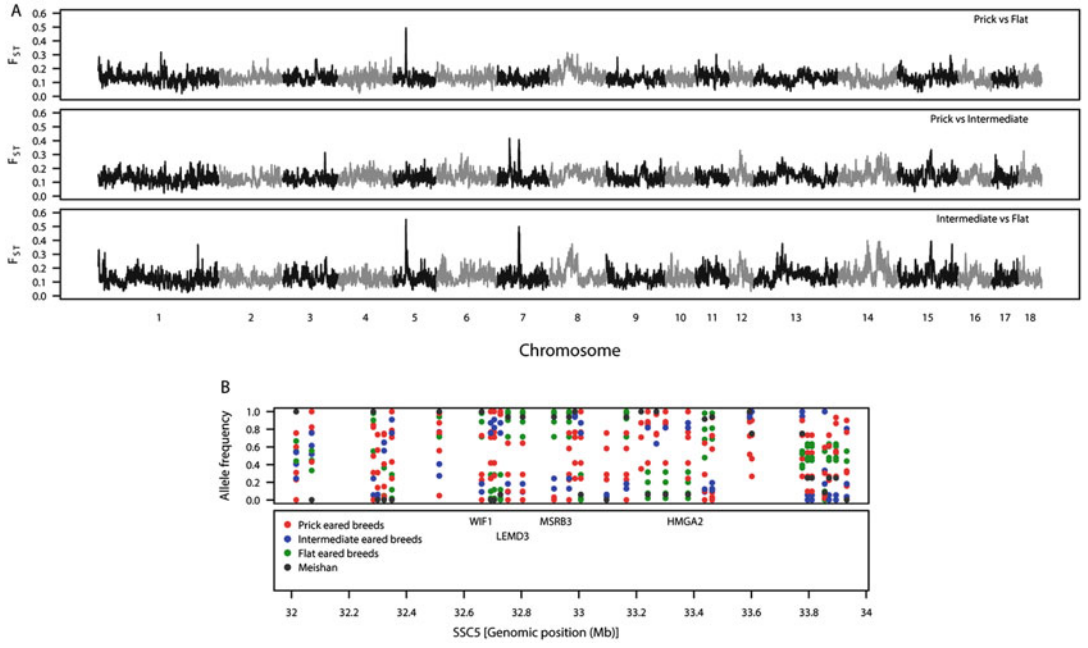

Fig. 6 A scan of the porcine genome for regions of genetic differentiation between ear phenotypes identified three genomic regions associated with ear morphology in European pig breeds. Levels of genome-wide genetic differentiation (13-SNP F_{ST} sliding window) are plotted with respect to genomic position: A: prick-eared contrasted against floppy-eared breeds (*top panel*), prick-eared breeds contrasted against intermediate-eared breeds (*middle panel*) and intermediate-eared breeds against floppy-eared breeds (*bottom panel*). The strongest signal was on porcine chromosome 5 from 31.74 to 33.78 Mb and zooming in on this genomic region in B: variation in the allele frequencies for the prick-, intermediate- and floppy-eared breeds at the candidate region are shown with the positions of candidate genes also highlighted. This figure is reproduced from the article "Signatures of diversifying selection in European pig breeds" by Wilkinson et al. (2013). PLoS Genetics;9(4):e1003453 (https://doi.org/10.1371/journal.pgen.1003453). The original article is open access distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

Both the QTL and selection mapping study in European pig breeds revealed that ear morphology in pigs is of a simple genetic architecture and causative putative genes have been identified. Four genes resided in the region of differentiation on chromosome 5: *WIF1*, *LEMD3*, *MSRB3* and *HGMA2* (Fig. 6) and Wilkinson et al. (2013) proposed the candidate gene *LEMD3* as it was located closest to the peak F_{ST}-window. *HMGA2* was proposed as a contender for pig ear morphology in another study as SNPs in the gene were significantly associated with variation in ear size (Li et al. 2012b). Yet, a GWAS found 35 SNPs on chromosome 5 to be significantly associated with ear size in pig breeds spanning a region of 30.14–40.92 Mb and identified a 450-kilobase interval where genotype frequencies were near fixation, and this covered *WIF1* and *LEMD3* (Zhang et al. 2015). An additional study found three SNPs genotyped in the *MSRB3* gene to be significantly associated with ear size in a pig breed intercross and there was higher mRNA expression of *MSRB3* in pigs with larger ear size (Zhang et al. 2015). The dog GWAS studies suggested *MSRB3* and *HGMA2* as candidate genes due to the proximity of the associated SNPs (Boyko et al. 2010; Vaysse et al. 2011). With little agreement on candidate gene(s), Wilkinson et al. (2013) sequenced the region encompassing the genomic signal on porcine chromosome 5 to determine if variants in the coding sequence could pinpoint a plausible molecular mechanism associated with the diverse ear phenotypes in pigs. However, no non-synonymous variants distinguished prick- and floppy-eared pig breeds, suggesting that regulatory elements may be involved in the genetic architecture underlying ear morphology.

In contrast to chromosome 5, little attention has been turned towards identifying candidate genes and mutations associated with ear morphology on pig chromosome 7. Wilkinson et al. (2013) proposed the gene *ADAMTSL3* as it occurred near the peak F_{ST}-window for the second signal on the chromosome. Ren et al. (2011) mapped a QTL to a 2 cM region on chromosome 7 and a 630-kilobase selective sweep was identified where majority of SNPs were near fixation. Of the nine genes in this region, the gene *PPARD* was considered to be a strong candidate due to its role in lipogenesis, and sequencing led to the discovery of a missense mutation (G32E) in a functionally conserved domain of this gene (Ren et al. 2011). The mutation was significantly associated with variation in ear size and downregulated β-catenin production in the Wnt/β-catenin signalling pathway. However, *PPARD* (36.13–36.21 Mb) fell outside the F_{ST} signals found on chromosome 7 by Wilkinson et al. (2013). Furthermore, it appears that the G32E mutation identified by Ren et al. (2011) is associated with ear morphology in Asian pig breeds but not European pig breeds. Whilst the derived 32E allele was found at high frequency in Asian large floppy-eared breeds, the 32G allele was near fixation for both European and Asian wild boar, Asian small prick-eared breeds and all European pig

(continued)

breeds including the floppy-eared British Saddleback, Gloucestershire Old Spots and Large Black (Ren et al. 2011). Thus, these results suggest that the G32E mutation may have been selected for in Asian floppy-eared pigs but not European floppy-eared pigs. Therefore, *PPARD* cannot be considered as a candidate for the floppy ear phenotype across all pigs. In conclusion, selection mapping has clearly shown that strong selection for certain ear phenotypes has driven associated alleles to fixation, but further sequencing and gene expression analyses are required to establish the biological mechanisms underlying ear morphology in pig breeds.

Phenotypic variation in skull shape, a trait that includes a number of different cranial anatomical features such as skull length, skull width and snout length, has also been widely selected in domestic animals. In dogs, variation in this trait ranges from long-snouted breeds, such as Collies and German Shepherds, to squashed and widened skulls (brachycephaly), such as Bull dog and Boxer, to unique breed-defining features like the "dome head" of the Chihuahua (Schoenenbeck and Ostrander 2013). An initial case-control GWAS design contrasting short- and long-muzzled dog breeds identified multiple significant SNPs on chromosome 1 and observed an appreciable reduction in heterozygosity (as measured using H_P) in brachycephalic breeds centred around the genes *THBS2* and *SMOC2* (Bannasch et al. 2010). More comprehensive studies using linear and geometric measurements took multiple points across the cranium and mapped genetic signals for skull shape to several chromosomes (Fig. 5), including that on chromosome 1 (Boyko et al. 2010; Schoenebeck et al. 2012). Subsequent functional analysis identified a transposable element within *SMOC2* associated with a significant reduction in *SMOC2* gene expression levels in brachycephalic dogs (Marchant et al. 2017). Sequencing of an additional candidate region on chromosome 32 also revealed a missense mutation in the gene *BMP3* that was near fixation in all small brachycephalic dog breeds examined (Schoenebeck et al. 2012).

Whilst there is also considerable phenotypic variation in skull shape amongst the pig breeds, perhaps the most evident difference is between domestic pigs and their ancestor, the wild boar (*S. scrofa*), where domestication has produced a marked reduction in snout and skull length (Fig. 1). By utilising data from the ancestral population, Wilkinson et al. (2013) detected a 2 Mb block of high genetic differentiation on porcine chromosome 2 between European pig breeds and wild boar. This region was orthologous to the one found on canine chromosome 1 associated with brachycephaly in dog breeds, and contained, amongst 17 genes, *THBS2* and *SMOC2*.

A final facial feature to consider, one of both evolutionary and breeding importance, are horns. This trait has a broad spectrum of phenotypes ranging from large elaborate horn structures, particularly for the males, to the absence of horns ("polled") in both sexes. Breeding for the absence of horns in domestic animals was likely because their ancestral function (self-defense and sexual selection) was no

longer a requirement (Zeder 2012) and their presence can cause injury to human handlers.

In sheep, linkage mapping of two-horned morphology identified a single QTL of large effect on chromosome 10 associated with horn length and type in the feral Soay sheep breed (Johnston et al. 2010) and an experimental cross of domestic horned and polled sheep breeds (Montgomery et al. 1996). Following on from this, a GWAS confirmed the association on chromosome 10 and found that a single SNP in the 3′ untranslated region (UTR) of the *RXFP2* gene explained 76% of the genetic variation in horn size in Soay sheep (Johnston et al. 2011). The role of *RXFP2* in controlling two-horned morphology in sheep has subsequently been confirmed in other breeds. Kijas et al. (2012) conducted a selective sweep analysis using F_{ST}, contrasting domestic horned and polled sheep breeds, and identified a genetic signal at the locus *RXFP2* (Fig. 7). Kardos et al. (2015) detected a region of reduced heterozygosity, as measured by ZH_P, surrounding the *RXFP2* locus in the feral Rocky Mountain bighorn sheep (*Ovis canadensis*). Subsequent sequencing of *RXFP2* found a 1.8 kb insertion in the 3′UTR of the gene present in horned sheep, but absent from polled animals (Wiedemar and Drögemüller 2015).

Kijas et al. (2016) posed the question of whether the four-horned phenotype is under the same genetic control as the two-horned phenotype in domestic sheep. A GWAS of two-horned (coded as controls) and four-horned (coded as cases) sheep breeds interestingly found no evidence of association at the *RXFP2* locus for the four-horned phenotype. Instead, a single strong association was identified on chromosome 2, which contained, amongst other candidate genes, a HoxD gene (Kijas et al. 2016). Nonetheless, it appears that both the two-horned and four-horned phenotypes in domestic sheep breeds have a simple genetic architecture.

5.2 Production Traits

5.2.1 Milk Production

Worldwide, milk is one of the most important nutritional daily food sources and this production trait has arguably experienced the strongest recent selective pressure, particularly to meet the demands of an expanding global human population. Intensive breeding for increased milk production commenced in the 1940s and 1950s with the development of artificial insemination techniques and of statistical methodology to estimate breeding values using large pedigrees (examples of dairy cattle breeds are shown in Fig. 2). In less than half a century, the average milk yield from dairy cows almost doubled, with ~56% of this increase attributed to genetics (Van Raden 2004). Genetic studies on milk production traits have identified QTLs (Khatkar et al. 2004) and validated the contribution of several candidate genes (*ABCG2*, *DGAT1*, *GHR*, casein cluster and prolactin genes; Thaller et al. 2003; Cohen-Zinder et al. 2005; Schennink et al. 2007; Banos et al. 2008; Sun et al. 2009).

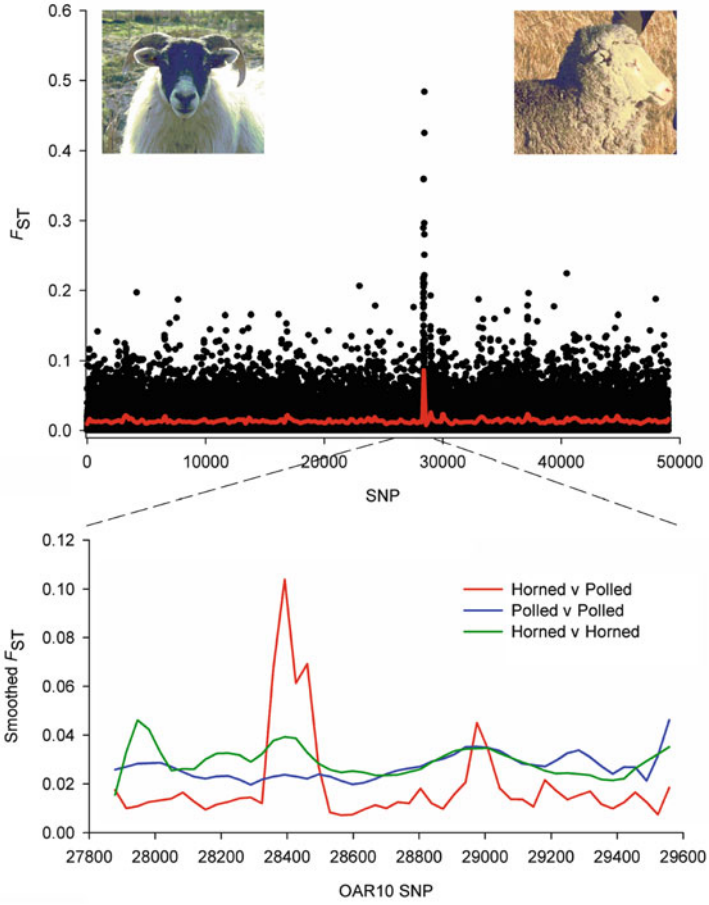

Fig. 7 A genome scan of differential selection between two-horned and polled sheep identifies a strong selective sweep on ovine chromosome 10. The *top panel* shows levels of genome-wide genetic differentiation between two-horned (Dorset Horn and Merino) and polled (Poll Dorset and Poll Merino) sheep breeds, plotted with respect to genomic position: *black dots* represent SNP F_{ST} values and the *red line* corresponds to a smoothed F_{ST}. The strongest signal was on ovine chromosome 10 from 27.87 to 29.47 Mb and zooming in the *bottom panel* shows smoothed F_{ST} values in this genomic region for horned contrasted against polled breeds (*red line*), between a pair of horned breeds (*green line*) and between a pair of polled breeds (*blue line*). This figure is reproduced from the article "A genome wide survey of SNP variation reveals the genetic structure of sheep breeds" by Kijas et al. (2012). PLoS Biology;10(2):e1001258 (https://doi.org/10.1371/journal.pbio.1001258). The original article is open access distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

Holsteins are the globally dominant breed of the dairy industry and initial selection mapping studies focused on this breed to ascertain if breeding for increased milk production had imparted signatures of selection at milk production genes. In this breed, several regions of decay in homozygosity have been found on bovine chromosome 6, including the region covering the *ABCG2* gene (Hayes et al. 2008), and another covering the casein cluster (Qanbari et al. 2010). By calculating allele frequency differences between Holsteins and the Aberdeen Angus beef breed, Hayes et al. (2009) identified strong signals near the *DGAT1* and *GHR* genes. Associations at these genes have also been detected in other dairy breeds; for instance, in a larger study of ten dairy and beef breeds, XP-EHH was calculated between the two cattle types and the genes from the casein cluster as well as *ABCG2*, *DGAT1* and *GHR* were found underneath some of the selective sweeps (Rothhammer et al. 2013). Another multiple breed study (Kemper et al. 2014) found evidence of differential selection at *DGAT1* when Holstein and Jersey dairy breeds were contrasted against beef breeds and within-breed iHS signals were identified at the *GHR* and *ABCG2* loci in the dairy breeds. Similarly, strong associations were found at *DGAT1* and *GHR* in a GWAS investigating specific milk properties (milk, fat and protein yields) in Finnish Ayrshire, Danish Red and Swedish Red dairy cattle breeds, shown in Fig. 8 (Iso-Touru et al. 2016).

Sheep, whilst not prominent in the contemporary dairy industry, were also historically treated as dual purpose and bred for increased milk output. A global F_{ST} scan of worldwide sheep detected a signal of differentiation at a prolactin gene (*PRLR*) (Kijas et al. 2012). Gutiérrez-Gil et al. (2014) adopted a strategy of exploring the levels of genetic differentiation between pairs of related dairy and meat sheep breeds and found some overlap with selection signatures associated with milk production in cattle, including at the locus *ABCG2*, suggesting convergent selection in the two domestic species.

5.2.2 Meat Production

The key trait for improvement in beef cattle is breeding for increased proportion of muscle content relative to overall carcase size. One such phenotype termed “double muscling” is characterised by a substantial increase in muscle content and has been bred across many taurine beef populations. Initial studies revealed the candidate gene and mutations in the coding sequence associated with doubling muscling. Then, selection mapping uncovered evidence of reduced heterozygosity at the candidate locus in certain beef populations (this is discussed in detail in Case Study 2).

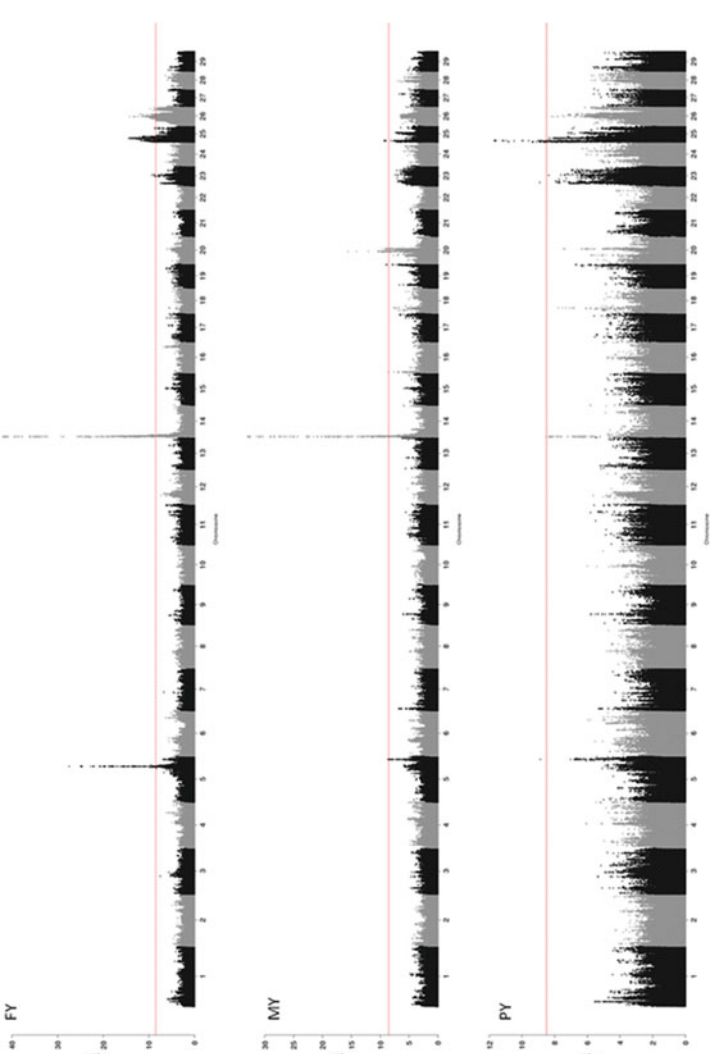

Fig. 8 A GWAS of the milk production traits fat yield (FY), milk yield (MY), protein yield (PY) in the dairy Nordic Red Cattle breed showed strong associations at the regions of the *DGAT1* and *GHR* genes. The $-\log_{10}(P\text{-value})$ of association of each SNP with the phenotype is plotted with respect to genomic position on the bovine genome and the *red line* indicates the genome-wide significance level. The milk genes *DGAT1* (17.95–18.04 Mb) and *GHR* (31.89–32.06 Mb) lie underneath the significantly associated SNPs on chromosome 14 and chromosome 20, respectively. This figure is reproduced from the article "Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants" by Iso-Touru et al. (2016). BMC Genetics;17:55 (https://doi.org/10.1186/1880-6805-31-14).

In addition to overall levels of meat (muscle), meat characteristics are important to the livestock industry because of their effects on eating quality and consumer perception. As a trait, meat quality comprises a collection of characteristics, one of which is fatty acid composition. In pigs, the higher concentrations of saturated and mono-unsaturated fatty acids and lower concentrations of polyunsaturated fatty acids found in the Duroc breed compared to other commercial breeds has been linked with overall improved eating quality (Cameron and Enser 1991; Warriss et al. 1996). A genome scan of differentiation between the Duroc and other European pig breeds mapped a signature of selection to chromosome 14 (Wilkinson et al. 2013). Wilkinson et al. (2013) proposed that this genetic signal was associated with fatty acid composition in the Duroc because it overlapped a previously identified QTL for fatty acid composition in the breed (Uemoto et al. 2012) and contained two genes with known roles in fatty acid synthesis (*ELOVL3* and *SCD*). A GWAS confirmed that this genomic region has a major effect on fatty acid composition in Duroc pigs, associated with a number of fatty acid phenotypes: saturated fatty acid, monounsaturated fatty acid, oleic acid (C18:1) and the ratio of oleic to stearic acid (C18:1/C18:0) (Ros-Freixedes et al. 2016).

Meat characteristics have also been monitored in beef cattle to ensure optimum quality. The amount of intramuscular fat (IMF) that gives a marbled appearance to meat influences its palatability (taste, texture and tenderness). Aberdeen Angus, the British beef breed, is known for its relatively high amount of marbling and there is a major genetic component to the trait (roughly 48% heritability) in Aberdeen Angus, suggesting that IMF can be selected to improve beef palatability (MacNeil et al. 2010). Mapping the allele frequency differences between Aberdeen Angus and Holstein, Hayes et al. (2009) identified a signature of selection at thyroglobulin (*TG*), a gene known to influence IMF content and Rothhammer et al. (2013) using XP-EHH also detected a region of divergent artificial selection near the *TG* gene in Belgian Blue cattle.

Case Study 2 Double Muscling and Signatures of Selection at *GDF-8* in Beef Cattle Breeds

In the early nineteenth century, animals with a heavily sculpted muscular appearance began to emerge in herds of European cattle breeds like Belgian Blue and Piedmontese. This visually striking physical phenotype is commonly known as double muscling (Fig. 9) and these cattle are prolific producers, with large amounts of muscle mass and desirable carcass cuts with a low fat percentage. These characteristics result in an increased final carcass value, with financial implications for beef production. Systematic breeding has made double muscling a widespread phenotype in many contemporary European beef-breed populations.

Double muscling was long known to be heritable in cattle and studies in the 1990s showed that the gene responsible for the phenotype is myostatin

(continued)

Fig. 9 Three beef cattle breeds with the double muscling phenotype. Clockwise from *left*: Belgian Blue, Limousin and Piedmontese. Figure compiled by Ian Hesketh, with photo attributes clockwise from *left*: dieses Foto wurde von mir selbst gemacht [Public domain], Budotradan (Own work) [CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0)] and Vicki Johnson [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], all via Wikimedia Commons

(*MSTN*) or growth and differentiation factor (*GDF-8*) (Grobet et al. 1997; McPherron and Lee 1997). *GDF-8* is a member of the transforming growth factor β superfamily, a group of factors that regulate development and tissue homeostasis, and the myostatin protein acts by repressing muscle cell growth and differentiation. Mutations at the gene inhibit the activity of the protein, leading to an increase in the number of skeletal muscle fibres, thereby producing the double-muscling phenotype in cattle. In Belgian Blue and a number of other breeds, double-muscled animals carry an 11-base pair deletion in the third exon of the gene, which causes a frameshift that leads to a stop codon in the bioactive carboxy-terminal domain resulting in a truncated protein and loss of *GDF-8* function (Kambadur et al. 1997; Grobet et al. 1997). In Piedmontese cattle, double-muscled animals carry a G → A transition in the same exon, which results in an amino acid change from cysteine to tyrosine (Kambadur et al. 1997), thereby altering a disulphide bridge required for proper conformation of the protein (Berry et al. 2002).

Population genomic studies have since been performed to ascertain the impact of intensive selection for double muscling on genetic diversity surrounding the *GDF-8* locus in various beef cattle breeds. Early population genetic studies with markers typed in the region of *GDF-8* showed strong evidence of selection near the gene in several double-muscled cattle breeds (Wiener et al. 2003) and different patterns of diversity observed across breeds suggested that the oldest origin of the 11-bp mutation was in the Belgian Blue breed (Wiener and Gutierrez-Gil 2009). Subsequent genome-wide studies of

(continued)

double-muscled cattle also detected selection-induced patterns of diversity in the region of *GDF-8*. In Belgian Blue, a scan for heterozygosity across the genome uncovered a large selective sweep extending 504 kb containing *GDF-8* where most alleles were at fixation (Fig. 10) (Druet et al. 2013, 2014). By implementing the iHS test to compare the homozygosity at ancestral alleles relative to derived alleles within Piedemontese and Limousin, signatures of selection were identified at *GDF-8* in both breeds (Bovine HapMap Consortium 2009). Other studies have since reaffirmed these findings for Limousin using the iHS (Zhao et al. 2015), EHH (Gurgul et al. 2015) and haplotype homozygosity methods (Kemper et al. 2014).

Heavily muscled phenotypes associated with *GDF-8* are not limited to cattle but are also observed in other mammals, including sheep, horses and dogs. An early study of the muscular Texel sheep breed mapped a QTL associated with the trait to the ovine chromosome that harbours *GDF-8* and

(continued)

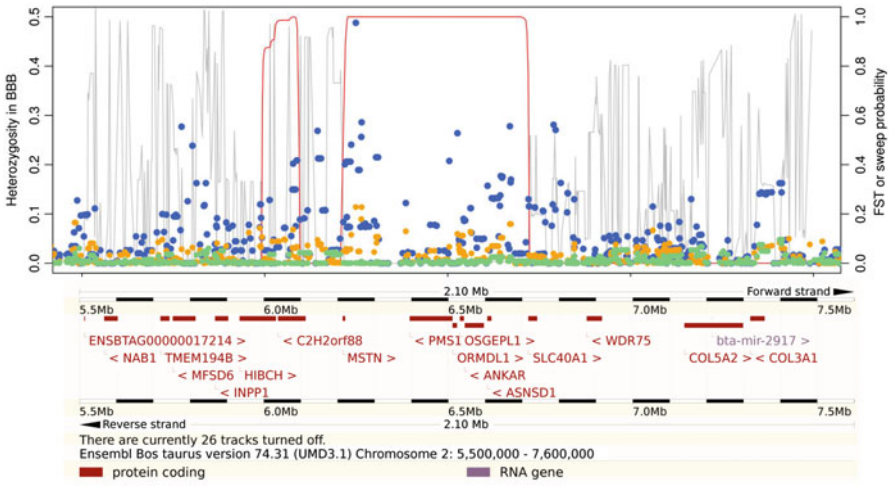

Fig. 10 Multi-method selection mapping in Belgian Blue cattle revealed a strong sweep on bovine chromosome 2 at the *GDF-8* (*MSTN*) gene. The *top panel* shows the sweep probability estimated by Sweepy (*red line*), SNP heterozygosity (*grey line*), genetic differentiation (measured as F_{ST}) between Belgian Blue and the dual purpose "Blancs Bleus Mixtes" (*orange dots*) and dairy Holstein (*blue dots*). The *bottom panel* shows the position of genes, including *GDF-8* (*MSTN*) (6.21–6.22 Mb), taken from Ensembl Bos taurus version 74.1 (UMD3.1). This figure is reproduced from the article "Selection in action: dissecting the molecular underpinnings of the increasing muscle mass of Belgian Blue Cattle" by Druet et al. (2014). BMC Genomics; 15:796 (https://doi.org/10.1186/1471-2164-15-796). The original article is open access distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

went on to show that a point mutation in the 3′UTR of the gene creates a target site for microRNAs involved in expression in the muscle (Clop et al. 2006). Kijas et al. (2012) contrasted three populations of the Texel sheep breed with other global sheep breeds and found the highest F_{ST} values to be associated with SNPs in the ovine genomic region of *GDF-8*. Petersen et al. (2013) applied the F_{ST}-statistic, d_i (Akey et al. 2010), to differentiate the short-distance sprinting muscular Quarter horse against the leaner Paint horse and identified a 5.5 Mb region on equine chromosome 18 encompassing *GDF-8*. Subsequent sequencing of the gene identified a short interspersed element (SINE) insertion in the promoter and an SNP in the first intron, which were significantly associated with muscle fibre composition in the Quarter horse.

These signatures of selection in beef cattle and other domesticated animals illustrate that favoured variants at *GDF-8* have been swept to fixation in certain populations as a result of breeding to increase the proportion of animals with high levels of muscling.

5.2.3 Reproduction

Enhancement of livestock productivity is also addressed through improving the reproductive performance of animals. A feature that distinguishes domestic animals from their wild counterparts is that reproduction is not seasonally regulated in domestic animals as in natural populations. Rubin et al. (2010) conducted a selective sweep analysis of four layer and four broiler lines using ZH_P and near complete homozygosity was observed in a 40 kb segment encompassing the thyroid stimulating hormone receptor (*TSHR*) gene. The thyroid system is known for its role in regulating seasonal reproduction in birds and mammals (Nakao et al. 2008) and sequencing of the gene in domestic chicken lines showed a non-conservative missense mutation in domestic chickens but not in the Red Jungle Fowl. A subsequent study adopted a paleogenomic approach whereby archaeological chicken samples dated from ~280 B.C. to the eighteenth century A.D. were densely genotyped to assess temporal variation at "domestication genes" (Flink et al. 2014). Extensive variation was found at the *TSHR* locus in the ancient samples compared to their modern counterparts, suggesting that selection for year round reproduction of chickens occurred in the last 500 years rather than during early domestication (Flink et al. 2014) (for further discussion on paleogenomics of domestic breeds see Box 1).

Although domestic animals have been bred for improved reproductive performance, as already discussed earlier in this chapter, compromised fertility in commercial dairy cattle breeds is a serious problem for the industry. Genomic regions that have a large effect on reproductive performance in dairy cattle have been mapped (Fortes et al. 2013; Khatkar et al. 2014). A recent GWAS of a fertility index in a group of commercial dairy cattle breeds found the strongest association on chromosome 12 and subsequent sequencing identified a 660 kb deletion on this chromosome in Finnish Ayrshire and Swedish Red dairy cattle breeds (Kadri et al.

2014). The study found that despite the lethality of this deletion, causing embryonic death in homozygotes, it is maintained at surprisingly high frequencies (23% in Swedish and 32% in Finnish Red cattle) because of the positive effect it has on milk production traits. This knowledge has far-reaching practical applications: mating plans that limit the use of carrier sires can be implemented to manage the frequency of such deleterious mutations in breeding populations.

6 Insights from Applying Population Genomics to Domesticated Breeds

The growing number of population genomic studies has contributed significantly to disentangling the genetics underpinning animal domestication and breed development. A cursory examination of the patterns of diversity of whole genomic landscapes suggests that each cattle, chicken, dog, pig and sheep breed has a unique genomic profile, but comparing patterns of genetic variation amongst breeds and species reveals shared attributes. This suggests that whilst different pressures have been imparted on individual species and breeds, they have at times experienced similar selective and demographic forces.

Assembling the signatures of selection identified for a range of different phenotypes in domestic animals has deepened our understanding of the domestication and breed development process, along with providing insights into the genomic architecture of phenotypic traits. There are numerous examples where selective sweeps for phenotypic traits are orthologous across domestic species and different breeds, evidenced by the repeated detection of the same genomic regions and genes for analogous morphological traits (Gutiérrez-Gil et al. 2015). Comparable genetic signals at the dairy-related *ABCG2*, *DGAT1* and *GHR* genes have been found in global populations of the Holstein, Jersey and Guernsey dairy cattle breeds and at *ABCG2* in dairy sheep breeds. Furthermore, there is less evidence of selection at the casein cluster and prolactin genes, suggesting weaker selection pressures acting on the latter group of milk genes in both sheep and cattle breeds. Stature is another good example of consistent genetic signatures across multiple domestic species and breeds. Moreover, the genes identified are also a subset of the 500+ genes implicated in controlling height in humans (Lango Allen et al. 2010). Interestingly, the complex genetic control of human height compared to the relatively simpler genetic architecture of body size in some domestic animals could reflect different selection pressures, such that artificial selection for stature in domestic animals may have been a far stronger force than natural selection acting on human height. A role for convergent evolution is further supported by shared derived phenotypes (e.g. floppy ear morphology) across species. The existence of shared signatures of selection across domestic species and breeds suggests that common breeding goals were, in some cases, independently enforced and selection acted on the same genes.

A narrative of marked differences in the genetic architecture across traits is also emerging from population genomic studies of domestic breeds. Certain traits appear to have a simpler genetic architecture, where the phenotypes are governed by a few loci of large effect. Breed-defining morphological traits, such as coat colour, stature and ear morphology, fall into this category because relatively few but strong signatures of selection have been mapped for these traits. For instance, Makvandi-Nejad et al. (2012) showed that roughly 83% of the genetic variation in body size in horses was captured by four loci and Rimbault et al. (2013) concluded that variants at six genes explained approximately half of the reduction in body size of small dog breeds. Similarly, three very strong genetic signals differentiated flat, intermediate and prick-eared pig breeds suggesting that these few genomic regions play a major role in controlling variation in ear morphology in pigs (Wilkinson et al. 2013). Furthermore, given the simple genetic architecture of these traits, there is a unique opportunity to identify the sequence variants that are likely major determinants of the phenotypes. Many domestic breed studies have taken the next step to sequence selective sweep regions to elucidate the molecular basis of phenotypes, thereby furthering the narrative of the history of breed development and its effect on genomic variation (e.g. sequencing of the *EDNRB* region to isolate the causative variant of coat spotting phenotype in the Gloucestershire Old Spots pig breed, Wilkinson et al. 2013).

In contrast, other traits show more complex genetic architecture, where phenotypes appear to be associated with many loci and the distribution of their effects varies. Unlike the breed-defining morphological traits of domestic breeds, most commercial traits, such as meat, dairy and reproduction traits, appear to have more complex genetic architectures. Signatures of selection have been detected for these traits, like the strong signals found at the milk production genes *ABCG2*, *DGAT1* and *GHR* (and in some studies, the casein genes). However, these genes comprise a small subset of the 344 QTLs/genes linked to milk production, as catalogued by Ogorevc et al. (2009). There are further examples where selection mapping has struggled to uncover signals for quantitative traits, even where a causative variant is known to be a major determinant of a phenotype. Pig breeders have selected for increased muscle mass and a single nucleotide change in the insulin growth factor 2 (*IGF2*) gene has a large effect on muscle growth, with the mutation reported to have swept through the European pig breeds (Van Laere et al. 2003). However, genome scans have failed to pinpoint a signature of selection at *IGF2* in commercial pig breeds selected for increased muscle content (Rubin et al. 2012; Wilkinson et al. 2013). Similarly, pig breeds have been selected for increased reproductive performance (e.g. ovulation rate, number of teats and litter size) yet selection mapping has not uncovered persuasive genetic signals corresponding to established reproductive QTLs and genes (Buske et al. 2006). Comparison of findings from linkage mapping versus signature of selection studies for complex traits was addressed in two cattle studies and both reported a low concordance between the genomic positions of QTLs versus selective sweeps (Wiener et al. 2011; Kemper et al. 2014).

There are several reasons to explain why selection mapping has failed to unravel some genotype–phenotype relationships in domestic animals. First, substantial

sequence complexity may surround a locus that influences a phenotypic trait. The coat colour gene *KIT*, which is a major determinant of several coat types in pigs (wild, belt, patch and dominant white types), incorporates various combinations of duplications and deletions of sequence blocks associated with different coat types (Rubin et al. 2012). Although the haplotype diversity at *KIT* in pigs may result from strong selection, it may be too complex for standard selection mapping methods to uncover. Second, at the core of selection mapping methods is that they search for genomic footprints of reduced heterozygosity or genetic differentiation (Nosil and Buerkle 2010) such that the methods are only likely to detect evidence at strongly divergent genomic regions (see Rajora et al. 2016). For complex traits, selection likely affects hundreds, if not thousands, of loci spread across the genome, causing smaller changes in allele frequencies, each of which has a minor effect on phenotypic variation. Single-locus population genomic approaches are unlikely to detect these widespread signatures of weaker selection. Therefore, multi-locus approaches for deciphering the genetic architecture of polygenic quantitative traits are needed (Rajora et al. 2016). Additionally, the varying success at uncovering loci associated with key production traits versus breed-defining traits could be due to the duration of selection pressures. Evidence suggests that selection for different coat colour phenotypes has occurred since the early stages of domestication (Cieslak et al. 2011) whilst strong selection for meat, dairy and reproduction has only occurred over the last few hundred years.

The intense process of artificial selection for genetic improvement in domestic animals can have far-reaching unfavourable consequences. It is clear that the events experienced by domestic animals, from the small pools of individuals isolated at the early stages of domestication and the intense selection that followed, have diminished N_e considerably in breeds. To some extent, genomic diversity has been eroded with successive generations, which may be related to the negative health problems arising in some domestic breeds. However, where issues are recognised by breeders then steps can be taken to manage the situation. The example of the dairy industry showed that: (1) genetic variation for traits of concern (e.g. fertility) in modern breeds exists and can be harnessed to counter the deleterious effects of intense selection for other traits (e.g. milk production) and (2) the highly structured management system and comprehensive recording in place for commercial breeding allows for monitoring of traits and the flexibility to broaden breeding objectives to manage unfavourable changes. Furthermore, considerable levels of diversity still exist in domestic animal genomes and this genetic resource can be maintained through effective and managed breeding decisions.

7 Future Trends and Perspectives

Over the past decade, the field of population genetics of domestic breeds has been transformed, with breakthroughs in genomics and genotyping technology and development of computational infrastructure. Population-scale genomic datasets have

been produced for many domestic species and breeds, from which variation across genomes has been extensively mined to decipher the genetic basis of domestication and breed development. The coming years will see extensive genotyping and sequencing of additional breeds which will allow further questions on the evolution of phenotypes to be addressed in population genomic studies.

Population genomics has been applied unevenly across domestic breeds worldwide (e.g. Mwai et al. 2015), predominantly focusing on mainstream commercial breeds, unsurprisingly considering their economic importance. Indeed, whilst there have been a number of studies investigating breed diversity worldwide (e.g. sheep, Kijas et al. 2012), genomic and phenotypic resources are still lacking for many non-commercial and traditional breeds, particularly those from the less developed and more inaccessible parts of the globe (Bruford et al. 2015). Population genomic analyses of more indigenous breeds will not only provide a more comprehensive understanding of the global history of animal domestication but also insights into the genetic basis of novel geographic-specific important traits.

Large-scale acquisitioning of non-commercial and traditional breed samples will also supplement current data, increasing the catalogue of genotype–phenotype resources. Such data could be a novel population genomic resource at the disposal of the livestock industry. It is generally recognised that the goals of the livestock industry have moved beyond improvement of milk, meat and reproduction traits (for which monumental strides have already been made over the last several decades) to incorporate additional challenges. Worldwide, the future brings with it the potential for irreversible changes in environmental conditions that may have a detrimental effect on livestock productivity and thus breeds adapted to more extreme climates will provide biological information as well as genetic resources that may benefit other populations. A second area of concern for the livestock industry is the increasing occurrence of infectious diseases and outbreaks in large herds that can have potentially devastating economic consequences. Certain breeds are known to be more resistant to particular diseases and identifying resistance-specific signatures of selection can provide genetic information that could be harnessed to counteract production losses associated with infectious diseases. For instance, N'Dama cattle are considered to be more tolerant to the disease trypanosomiasis, caused by vector-borne protozoan parasites of the genus *Trypanosoma*, compared to other more economically productive African and non-African cattle breeds. A cross-population analysis between N'Dama and less trypanotolerant breeds using a combination of selection mapping approaches identified selective sweeps and genes associated with, amongst other functions, the immune system (Kim et al. 2017). This example nicely illustrates the potential value of more indigenous and less productive breeds as a genomic resource. Thus, a more concerted effort is required to characterise the genome-wide diversity of traditional breeds and beneficial phenotypes from developing countries, through multi-national collaborations, to build long-term resources for the livestock industry.

The experimental designs of population genomic studies of domestic breeds are also progressing beyond population-scale datasets of one or a set of breeds to incorporate *a priori* links between genotype data and particular phenotypes. For

example, Moradi et al. (2012) genotyped thin and fat tailed sheep breeds with the aim of solely identifying selective sweeps (assessed using F_{ST}) associated with one trait, fat deposition. A design such as this in population genomic selection mapping studies, where breeds that exhibit extremes of a phenotype are sampled, thereby linking genotypes and phenotypes, allows the detection of trait-specific genetic signals that may not be otherwise uncovered in general large-scale population genomic breed studies.

Another emerging area of population genomics that is proving powerful in deciphering the animal domestication process is paleogenomics, the use of ancient DNA to draw inferences about genetic processes (MacHugh et al. 2017). As suggested by the name, paleogenomic analyses involve sequencing (or genotyping) of mitochondrial DNA, candidate genes or whole genomes for ancient wild and domesticated samples. These sequences can then be compared to modern sequences or, where samples are available, analysed at multiple time-points in the past. Ancient DNA has already made significant contributions to the understanding of the evolution of phenotypes and domestication, particularly for the horse but also for other livestock and dogs (see Box 1 detailing a few studies), and it is likely that this approach will be increasingly important in future studies on domesticated species.

Box 1 Paleogenomics of Domestic Breeds

Paleogenomics involves comparing the genomic profiles of ancient and contemporary samples to characterise genetic changes that have occurred along an evolutionary timescale. This allows the various stages of the domestication process to be explored for the occurrence of population-shaping events (admixture, introgression, selection, origins and migration), making paleogenomics a highly powerful approach that can provide a unique insight into animal domestication.

Paleogenomics is a growing section within the field of population genomics of domestic breeds (see Sect. 7) and it has already started to illuminate the history of domestication, from the demographic trends of early domestic populations to the evolution of phenotypes. Phylogeographical analyses of ancient DNA have produced genetic networks revealing the geographic locales of domestication centres (e.g. cattle and pigs), routes of migration (e.g. pigs and chickens) and levels of admixture between wild and domestic ancestors (e.g. cattle, pigs and horses) (described in detail by MacHugh et al. 2017).

Specific genes associated with phenotypic traits considered of importance in animal domestication have been studied through the examination of allele frequency patterns over time. In archaeologically preserved chicken specimens dating from ~280 B.C. to the eighteenth century A.D., extensive genetic variation was found at two genes reputedly selected for during chicken domestication, *TSHR*, associated with reproduction, and *BCDO2*, associated

(continued)

Box 1 (continued)

with yellow skin colour (Flink et al. 2014) (see Sect. 5.2.3). These genes were previously hypothesised to be selected for during the early stages of chicken domestication because there is evidence of strong selective sweeps at *TSHR* and *BCDO2* in contemporary chickens, whilst the derived alleles are absent in the Red Junglefowl (Rubin et al. 2010). However, as genetic diversity is evident at these genes in chickens dated up to the sixteenth to eighteenth century, it suggests that the selection at these loci occurred fairly recently (Flink et al. 2014).

Coat colour is another phenotypic trait that has been the focus of a number of paleogenomic domestic breed studies, as it is considered a key desired trait since the early stages of domestication (Cieslak et al. 2011). In horse samples dated from the Late Pleistocene to the Iron Age, allelic variation at eight coat colour genes suggests an increase in coat colours early in domestication (Ludwig et al. 2009). A subsequent study found that the allelic frequency distribution changed in coat colour genes over time, showing an increase in spotted coats later in domestication and a return to predominance of horses with solid coats in medieval times (Wutke et al. 2016).

Beyond candidate genes, whole-genome sequencing of ancient and modern samples to carry out genome-wide selection scans has provided insight into additional historic selective pressures and levels of introgression. Park et al. (2015) examined modern cattle against the ancestral British wild auroch genome, revealing evidence for ancient introgression of local aurochs into domesticated cattle of the British Isles. Librado et al. (2017) compared the genomes of ancient horses sampled from burial sites across the Central Asian steppes with their modern counterparts, identifying an extinct lineage that contributed to current populations and found fewer deleterious mutations in pre-domestication horses than modern animals.

8 Conclusions

The combined application of genome-wide SNP markers and population genetics methodology has proved to be a powerful and easily applicable approach to advance our understanding of the history of domesticated animals. Population genomics has revealed the impact that demographic and selective forces have had on variation across the genomes of domestic breeds. In addition, this area of research has deepened our understanding of the genetic architecture of phenotypic traits, the strength of selective pressures and the suite of characteristics desired in domestic breeds. It is also important to remember that population genomics is one step on the road of unravelling the genetic basis of phenotypic traits. After uncovering a genomic region showing evidence of selection, gene(s) and causative variants that play a functional role in determining the phenotype should be identified, followed by

functional studies to determine the specific effect(s) of variants. With increasingly inexpensive and efficient genotyping and whole-genome sequencing options, combined with the application to more breeds and phenotypes, population genomics will continue to be a powerful tool to contribute to the understanding of the history of domesticated animals and the genetic dissection of important traits.

References

Ai H, Huang L, Ren J. Genetic diversity, linkage disequilibrium and selection signatures in Chinese and western pigs revealed by genome-wide SNP markers. PLoS One. 2013;8(2):e56001. https://doi.org/10.1371/journal.pone.0056001.

Ajmone-Marsan P, GLOBALDIV Consortium. A global view of livestock biodiversity and conservation–GLOBALDIV. Anim Genet. 2010;41(Suppl 1):1–5. https://doi.org/10.1111/j.1365-2052.2010.02036.x.

Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, et al. Tracking footprints of artificial selection in the dog genome. Proc Natl Acad Sci U S A. 2010;107(3):1160–5. https://doi.org/10.1073/pnas.0909918107.

Amaral AJ, Ferretti L, Megens HJ, Crooijmans RP, Nie H, Ramos-Onsins SE, et al. Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. PLoS One. 2011;6:e14782. https://doi.org/10.1371/journal.pone.0014782.

Andersson L, Georges M. Domestic-animal genomics: deciphering the genetics of complex traits. Nat Rev Genet. 2004;5(3):202–12. https://doi.org/10.1038/nrg1294.

Andersson L, Plastow G. Molecular genetics of coat colour variation. In: Rothschild MF, Ruvinsky A, editors. The genetics of the pig. Oxon, UK: CAB International; 2011. p. 38–50.

Archibald AL, Bolund L, Churcher C, Fredholm M, Groenen MAM, Harlizius B, et al. Pig genome sequence – analysis and publication strategy. BMC Genomics. 2010;11:438. https://doi.org/10.1186/1471-2164-11-438.

Badke YM, Bates RO, Ernst CW, Schwab C, Steibel JP. Estimation of linkage disequilibrium in four US pig breeds. BMC Genomics. 2012;13:24. https://doi.org/10.1186/1471-2164-13-24.

Bannasch D, Young A, Myers J, Truvé K, Dickinson P, Gregg J, et al. Localization of canine brachycephaly using an across breed mapping approach. PLoS One. 2010;5(3):e9632. https://doi.org/10.1371/journal.pone.0009632.

Banos G, Woolliams JA, Woodward BW, Forbes AB, Coffey MP. Impact of single nucleotide polymorphisms in leptin, leptin receptor, growth hormone receptor, and diacylglycerol acyltransferase (DGAT1) gene loci on milk production, feed, and body energy traits of UK dairy cows. J Dairy Sci. 2008;91:3190–200. https://doi.org/10.3168/jds.2007-0930.

Bateson P, Sargan DR. Analysis of the canine genome and canine health: a commentary. Vet J. 2012;194(3):265–9. https://doi.org/10.1016/j.tvjl.2012.09.001.

Berry C, Thomas M, Langley B, Sharma M, Kambadur R. Single cysteine to tyrosine transition inactivates the growth inhibitory function of piedmontese myostatin. Am J Physiol Cell Physiol. 2002;283(1):C135–41.

Berry DP, Friggens NC, Lucy M, Roche JR. Milk production and fertility in cattle. Annu Rev Anim Biosci. 2016;4:269–90. https://doi.org/10.1146/annurev-animal-021815-111406.

Bickart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Cjam S, et al. Single-molecular sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. Nat Genet. 2017;49:643–50. https://doi.org/10.1038/ng.3802.

Bosse M, Megens HJ, Madsen O, Paudel Y, Frantz LA, Schook LB, et al. Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. PLoS Genet. 2012;8(11):e1003100. https://doi.org/10.1371/journal.pgen.1003100.

Bovine Genome Sequencing and Analysis Consortium. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science. 2009;324:522–8. https://doi.org/10.1126/science.1169588.

Bovine HapMap Consortium. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science. 2009;324:528–32. https://doi.org/10.1126/science.1167936.

Boyko A, Quignon P, Li L, Schoenenbeck J, Degenhardt J, Lohmueller K, et al. A simple genetic architecture underlies morphological variation in dogs. PLoS Biol. 2010;8:e1000451. https://doi.org/10.1371/journal.pbio.1000451.

Bruford MW, Ginja C, Hoffmann I, Joost S, Orozco-terWengel P, Alberto FJ, et al. Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025. Front Genet. 2015;6:314. https://doi.org/10.3389/fgene.2015.00314.

Buske B, Sternstein I, Brockmann G. QTL and candidate genes for fecundity in sows. Anim Reprod Sci. 2006;95:167–83. https://doi.org/10.1016/j.anireprosci.2005.12.015.

Cadieu E, Neff MW, Quignon P, Walsh K, Chase K, Parker HG, et al. Coat variation in the domestic dog is governed by variants in three genes. Science. 2009;326:150–3. https://doi.org/10.1126/science.1177808.

Cameron ND, Enser MB. Fatty acid composition of lipid in Longissimus dorsi muscle of Duroc and British Landrace pigs and its relationship with eating quality. Meat Sci. 1991;29:295–307. https://doi.org/10.1016/0309-1740(91)90009-F.

Cavalli-Sforza LL. Population structure and human evolution. Proc R Soc Lond Biol Sci. 1966;164:362–79.

Chase K, Carrier DR, Adler FR, Jarvik T, Ostrander EA, Lorentzen TD, Lark KG. Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. Proc Natl Acad Sci U S A. 2002;99:9930–5. https://doi.org/10.1073/pnas.152333099.

Cieslak M, Reissmann M, Hofreiter M, Ludwig A. Colours of domestication. Biol Rev Camb Philos Soc. 2011;86:885–99. https://doi.org/10.1111/j.1469-185X.2011.00177.x.

Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, Bibe B, et al. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat Genet. 2006;38(7):813–8. https://doi.org/10.1038/ng1810.

Clutton-Brock J. A natural history of domesticated mammals. Cambridge: Cambridge University Press; 1999.

Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Evertsvan der Wind A, Lee JH, et al. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Res. 2005;15(7):936–44. https://doi.org/10.1101/gr.3806705.

Darwin C. The variation of animals and plants under domestication. London: John Murray; 1868.

Dawkins MS, Layton R. Breeding for better welfare: genetic goals for broiler chickens and their parents. Anim Welf. 2012;21:147–55. https://doi.org/10.7120/09627286.21.2.147.

DEFRA. UK national action plan on farm animal genetic resources. London: DEFRA; 2006.

DEFRA. Review of molecular characterisation studies relating to UK farm animal genetic resources. London: DEFRA; 2009.

Dreger DL, Rimbault M, Davis BW, Bhatnagar A, Parker HG, Ostrander EA. Whole-genome sequence, SNP chips and pedigree structure: building demographic profiles in domestic dog breeds to optimize genetic-trait mapping. Dis Model Mech. 2016;9(12):1445–60. https://doi.org/10.1242/dmm.027037.

Druet T, Pérez-Pardal L, Charlier C, Gautier M. Identification of large selective sweeps associated with major genes in cattle. Anim Genet. 2013;44(6):758–62. https://doi.org/10.1111/age.12073.

Druet T, Ahariz N, Cambisano N, Tamma N, Michaux C, Coppieters W, et al. Selection in action: dissecting the molecular underpinnings of the increasing muscle mass of Belgian Blue Cattle. BMC Genomics. 2014;15:796. https://doi.org/10.1186/1471-2164-15-796.

FAO. In: Scherf BD, Rome PD, editors. The second report on the state of the world's animal genetic resources for food and agriculture. Rome: FAO Commission on Genetic Resources for Food and Agriculture Assessments; 2015.

Farrell LL, Schoenebeck JJ, Wiener P, Clements DN, Summers KM. The challenges of pedigree dog health: approaches to combating inherited disease. Canine Genet Epidemiol. 2015;2:3. https://doi.org/10.1186/s40575-015-0014-9.

Flink GL, Allen R, Barnett R, Malmstrom H, Peters J, Eriksson J, et al. Establishing the validity of domestication genes using DNA from ancient chickens. Proc Natl Acad Sci U S A. 2014;111 (17):6184–9. https://doi.org/10.1073/pnas.1308939110.

Flori L, Fritz S, Jaffrezic F, Boussaha M, Gut I, Heath S, et al. The genome response to artificial selection: a case study in dairy cattle. PLoS One. 2009;4(8):e6595. https://doi.org/10.1371/journal.pone.0006595.

Fortes MR, Deatley KL, Lehnert SA, Burns BM, Reverter A, Hawken RJ, et al. Genomic regions associated with fertility traits in male and female cattle: advances from microsatellites to high-density chips and beyond. Anim Reprod Sci. 2013;141(1-2):1–19. https://doi.org/10.1016/j.anireprosci.2013.07.002.

Frankham R, Ballou JD, Briscoe DA. Introduction to conservation genetics. 2nd ed. Cambridge, UK: Cambridge University Press; 2010.

Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009;10:381–91. https://doi.org/10.1038/nrg2575.

Gray MM, Granka JM, Bustamante CD, Sutter NB, Boyko AR, Zhu L, et al. Linkage disequilibrium and demographic history of wild and domestic canids. Genetics. 2009;181(4):1493–505. https://doi.org/10.1534/genetics.108.098830.

Grobet L, Poncelet D, Royo LJ, Brouwers B, Pirottin D, Michaux C, et al. Molecular definition of an allelic series of mutations disrupting the myostatin function and causing double-muscling in cattle. Mamm Genome. 1997;9(3):210–3.

Groenen MAM, Megens H-J, Zare Y, Warren WC, Hillier LW, et al. The development and characterization of a 60 K SNP chip for chicken. BMC Genomics. 2011;12(1):274. https://doi.org/10.1186/1471-2164-12-274.

Groeneveld LF, Lenstra JA, Eding H, Toro MA, Scherf B, Pilling D, et al. Genetic diversity in farm animals –a review. Anim Genet. 2010;41(suppl. 1):6–31. https://doi.org/10.1111/j.1365-2052.2010.02038.x.

Gurgul A, Pawlina K, Frys-Żurek M, Bugno-Poniewierska M. Identification of differential selection traces in two Polish cattle breeds. Anim Sci J. 2015;86(1):17–24. https://doi.org/10.1111/asj.12242.

Gutiérrez-Gil B, Arranz JJ, Pong-Wong R, García-Gámez E, Kijas J, Wiener P. Application of selection mapping to identify genomic regions associated with dairy production in sheep. PLoS One. 2014;9(5):e94623. https://doi.org/10.1371/journal.pone.0094623.

Gutiérrez-Gil B, Arranz JJ, Wiener P. An interpretive review of selective sweep studies in Bos taurus cattle populations: identification of unique and shared selection signals across breeds. Front Genet. 2015;13(6):167. https://doi.org/10.3389/fgene.2015.00167.

Hall S, Clutton-Brock J. Two hundred years of British farm livestock. London: British Museum (Natural History); 1988.

Hartl D, Clark A. Principles of population genetics. 4th ed. Sunderland, MA: Sinauer Associates; 2007.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 2003;13(4):635–43. https://doi.org/10.1101/gr.387103.

Hayes BJ, Lien S, Nilsen H, Olsen HG, Berg P, Maceachern S, et al. The origin of selection signatures on bovine chromosome 6. Anim Genet. 2008;39:105–11. https://doi.org/10.1111/j.1365-2052.2007.01683.x.

Hayes BJ, Chamberlain AJ, Maceachern S, Savin K, McPartlan H, MacLeod I, et al. A genome map of divergent artificial selection between Bos taurus dairy cattle and Bos taurus beef cattle. Anim Genet. 2009;40(2):176–84. https://doi.org/10.1111/j.1365-2052.2008.01815.x.

Hayes BJ, Pryce J, Chamberlain AJ, Bowman PJ, Goddard ME. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein

cattle as contrasting model traits. PLoS Genet. 2010;6(9):e1001139. https://doi.org/10.1371/journal.pgen.1001139.
Hill WG. Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor Popu Biol. 1975;8(2):117–26. https://doi.org/10.1016/0040-5809(75)90028-3.
Hill WG. Estimation of effective population-size from data on linkage disequilibrium. Genet Res. 1981;38:209–16. https://doi.org/10.1017/S0016672300020553.
Hoze C, Fouilloux MN, Venot E, Guillaume F, Dassonneville R, Fritz S, et al. High-density marker imputation accuracy in sixteen French cattle breeds. Genet Sel Evol. 2013;45:33. https://doi.org/10.1186/1297-9686-45-33.
International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004;432(7018):695–716. https://doi.org/10.1038/nature03154.
International Chicken Polymorphism Map Consortium. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature. 2004;432(7018):717–22. https://doi.org/10.1038/nature03156.
Iso-Touru T, Sahana G, Guldbrandtsen B, Lund MS, Vilkki J. Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. BMC Genet. 2016;17:55. https://doi.org/10.1186/s12863-016-0363-8.
Jackson I. Mouse coat colour mutations: a molecular genetic resource which spans the centuries. Bioessays. 1991;13(9):439–6. https://doi.org/10.1002/bies.950130903.
Jacobs GS, Sluckin TJ, Kivisild T. Refining the use of linkage disequilibrium as a robust signature of selective sweeps. Genetics. 2016;203(4):1807–25. https://doi.org/10.1534/genetics.115.185900.
Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, et al. The sheep genome illuminates biology of the rumen and lipid metabolism. Science. 2014;344(6188):1168–73. https://doi.org/10.1126/science.1252806.
Johnston SE, Beraldi D, McRae AF, Pemberton JM, Slate J. Horn type and horn length genes map to the same chromosomal region in Soay sheep. Heredity. 2010;104(2):196–205. https://doi.org/10.1038/hdy.2009.109.
Johnston SE, McEwan JC, Pickering NK, Kijas JW, Beraldi D, Pilkington JG, et al. Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. Mol Ecol. 2011;20(12):2555–66. https://doi.org/10.1111/j.1365-294X.2011.05076.x.
Kadri NK, Sahana G, Charlier C, Iso-Touru T, Guldbrandtsen B, Karim L, et al. A 660-kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. PLoS Genet. 2014;10(1):e1004049. https://doi.org/10.1371/journal.pgen.1004049.
Kambadur R, Sharma M, Smith TP, Bass JJ. Mutations in myostatin (GDF8) in double-muscled Belgian Blue and Piedmontese cattle. Genome Res. 1997;7(9):910–6.
Kardos M, Luikart G, Bunch R, Dewey S, Edwards W, McWilliam S, et al. Whole-genome resequencing uncovers molecular signatures of natural and sexual selection in wild bighorn sheep. Mol Ecol. 2015;24(22):5616–32. https://doi.org/10.1111/mec.13415.
Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NHC, Zody MC, Anderson N, et al. Efficient mapping of mendelian traits in dogs through genome-wide association. Nat Genet. 2007;39(11):1321–8. https://doi.org/10.1038/ng.2007.10.
Kelly JK. A test of neutrality based on interlocus associations. Genetics. 1997;146(3):1197–206.
Kemper KE, Saxton SJ, Bolormaa S, Hayes BJ, Goddard ME. Selection for complex traits leaves little or no classic signatures of selection. BMC Genomics. 2014;15:246. https://doi.org/10.1186/1471-2164-15-246.
Khatkar MS, Thomson PC, Tammen I, Raadsma HW. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genet Sel Evol. 2004;6:163–90. https://doi.org/10.1051/gse:2003057.

Khatkar MS, Randhawa IAS, Raadsma HW. Meta-assembly of genomic regions and variants associated with female reproductive efficiency in cattle. Livest Sci. 2014;166:144–57. https://doi.org/10.1016/j.livsci.2014.05.015.

Kijas JW, Townley D, Dalrymple BP, Heaton MP, Maddox JF, McGrath A, et al. A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS One. 2009;4(3):e4668. https://doi.org/10.1371/journal.pone.0004668.

Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, et al. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS Biol. 2012;10(2):e1001258. https://doi.org/10.1371/journal.pbio.1001258.

Kijas JW, Porto-Neto L, Dominik S, Reverter A, Bunch R, McCulloch R, et al. Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. Anim Genet. 2014;45(5):754–7. https://doi.org/10.1111/age.12197.

Kijas JW, Hadfield T, Naval Sanchez M, Cockett N. Genome-wide association reveals the locus responsible for four-horned ruminant. Anim Genet. 2016;47(2):258–62. https://doi.org/10.1111/age.12409.

Kilman R, Sheehy B, Schultz J. Genetic drift and effective population size. Nat Ed. 2008;1(3):3.

Kim Y, Nielsen R. Linkage disequilibrium as a signature of selective sweeps. Genetics. 2004;167(3):1513–24. https://doi.org/10.1534/genetics.103.025387.

Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. 2002;160(2):765–77.

Kim S-J, Ka S, Ha J-W, Kim J, Yoo D, Kim K, et al. Cattle genome-wide analysis reveals genetic signatures in trypanotolerant N'Dama. BMC Genomics. 2017;18:371. https://doi.org/10.1186/s12864-017-3742-2.

Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, Talbot R, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 2013;14:59. https://doi.org/10.1186/1471-2164-14-59.

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467(7317):832–8. https://doi.org/10.1038/nature09410.

Lewontin RC, Krakauer J. Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. Genetics. 1973;74:175–95.

Li DF, Liu WB, Liu JF, Yi GQ, Lian L, Qu LJ, et al. Whole-genome scan for signatures of recent selection reveals loci associated with important traits in White Leghorn chickens. Poult Sci. 2012a;91(8):1804–12. https://doi.org/10.3382/ps.2012-02275.

Li P, Xiao S, Wei N, Zhang Z, Huang GR, et al. Fine mapping of a QTL for ear size on porcine chromosome 5 and identification of high mobility group AT-hook 2 (*HMGA2*) as a positional candidate gene. Genet Sel Evol. 2012b;44:6. https://doi.org/10.1186/1297-9686-44-6.

Li MH, Tiirikka T, Kantanen J. A genome-wide scan study identifies a single nucleotide substitution in ASIP associated with white versus non-white coat-colour variation in sheep (Ovis aries). Heredity. 2013;112(2):122–31. https://doi.org/10.1038/hdy.2013.83.

Li Z, Chen J, Wang Z, Pan Y, Wang Q, Xu N, Wang Z. Detection of selection signatures of population-specific genomic regions selected during domestication process in Jinhua pigs. Anim Genet. 2016;47(6):672–81. https://doi.org/10.1111/age.12475.

Librado P, Gamba C, Gaunitz C, Der Sarkissian C, Pruvost M, Albrechtsen A, et al. Ancient genomic changes associated with domestication of the horse. 2017;356(6336):442–5. https://doi.org/10.1126/science.aam5298.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005;438(8):803–19. https://doi.org/10.1038/nature04338.

Linderholm A, Larson G. The role of humans in facilitating and sustaining coat colour variation in domestic animals. Semin Cell Dev Biol. 2013;24(6-7):587–93. https://doi.org/10.1016/j.semcdb.2013.03.015.

Liu Z, Sun C, Qu L, Wang K, Yang N. Genome-wide detection of selective signatures in chicken through high density SNPs. PLoS One. 2016;11(11):e0166146. https://doi.org/10.1371/journal.pone.0166146.

Lu MD, Han XM, Ma YF, Irwin DM, Gao Y, Deng J-K, et al. Genetic variations associated with six-white-point coat pigmentation in Diannan small-ear pigs. Sci Rep. 2016;6:27534. https://doi.org/10.1038/srep27534.

Ludwig A, Pruvost M, Reissmann M, Benecke N, Brockmann GA, Castaños P, et al. Coat color variation at the beginning of horse domestication. Science. 2009;324:484–6. https://doi.org/10.1126/science.1172750.

Lv FH, Agha S, Kantanen J, Colli L, Stucki S, Kijas JW, et al. Adaptations to climate-mediated selective pressures in sheep. Mol Biol Evol. 2014;31(12):3324–43. https://doi.org/10.1093/molbev/msu264.

Ma J, Qi W, Ren D, Duan Y, Qiao R, Guo Y, et al. A genome scan for quantitative trait loci affecting three ear traits in a White Duroc x Chinese Erhualian resource population. Anim Genet. 2009;40(4):463–7. https://doi.org/10.1111/j.1365-2052.2009.01867.x.

Ma Y, Wei J, Zhang Q, Chen L, Wang J, Liu J, Ding X. A genome scan for selection signatures in pigs. PLoS One. 2015;10(3):e0116850. https://doi.org/10.1371/journal.pone.0116850.

MacHugh DE, Larson G, Orlando L. Taming the past: ancient DNA and the study of animal domestication. Annu Rev Anim Biosci. 2017;5:329–51. https://doi.org/10.1146/annurev-animal-022516-022747.

MacNeil MD, Nkrumah JD, Woodward BW, Northcutt SL. Genetic evaluation of Angus cattle for carcass marbling using ultrasound and genomic indicators. J Anim Sci. 2010;88(2):5170–522. https://doi.org/10.2527/jas.2009-2022.

Makvandi-Nejad S, Hoffman GE, Allen JJ, Chu E, Gu E, Chandler AM, et al. Four loci explain 83% of size variation in the horse. PLoS One. 2012;7(7):e39929. https://doi.org/10.1371/journal.pone.0039929.

Marchant TW, Johnson EJ, McTeir L, Johnson CI, Gow A, Liuti T, et al. Canine bracycephaly is associated with a retrotransposon-mediated missplicing of SMOC2. Curr Biol. 2017;27(11):1573–1584.e6. https://doi.org/10.1016/j.cub.2017.04.057.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, et al. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One. 2009;4(4):e5350. https://doi.org/10.1371/journal.pone.0005350.

Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974;23:23–35. https://doi.org/10.1017/S0016672300014634.

McCue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E, Binns MM, et al. A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. PLoS Genet. 2012;8:e1002451. https://doi.org/10.1371/journal.pgen.1002451.

McPherron AC, Lee SJ. Double muscling in cattle due to mutations in the myostatin gene. Proc Natl Acad Sci U S A. 1997;94(23):12457–61.

Megens H-J, Groenen MAM. Domesticated species form a treasure-trove for molecular characterization of Mendelian traits by exploiting the specific genetic structure of these species in across-breed genome wide association studies. Heredity. 2012;109(1):1–3. https://doi.org/10.1038/hdy.2011.128.

Montgomery GW, Henry HM, Dodds KG, Beattie AE, Wuliji T, Crawford AM. Mapping the horns (Ho) locus in sheep: a further locus controlling horn development in domestic animals. J Hered. 1996;87(5):358–63.

Moradi MH, Nejati-Javaremi A, Moradi-Shahrbabak M, Dodds KG, McEwan JC. Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. BMC Genet. 2012;13:10. https://doi.org/10.1186/1471-2156-13-10.

Muir WM, Wong GK-S, Zhang Y, Wang J, Groenen MAM, Crooijmans RPMA, et al. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of

rare alleles in commercial breeds. Proc Natl Acad Sci U S A. 2008;105(45):17312–7. https://doi.org/10.1073/pnas.0806569105.

Mwai O, Hanotte O, Kwon Y-J, Cho S. African indigenous cattle: unique genetic resources in a rapidly changing world. Asian-Australas J Anim Sci. 2015;28(7):911–21.

Nakao N, Ono H, Yoshimura T. Thyroid hormones and seasonal reproductive neuroendocrine interactions. Reproduction. 2008;136(1):1–8. https://doi.org/10.1530/REP-08-0041.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. 2005;15(11):1566–75. https://doi.org/10.1101/gr.4252305.

Nosil P, Buerkle A. Population genomics. Nat Ed. 2010;1:8.

Ogorevc J, Kunej T, Razpet A, Dovc P. Database of cattle candidate genes and genetic markers for milk production and mastitis. Anim Genet. 2009;40(6):832–51. https://doi.org/10.1111/j.1365-2052.2009.01921.x.

Oltenacu PA, Broom DM. The impact of genetic selection for increased milk yield on the welfare of dairy cows. Anim Welf. 2010;19(S):39–49.

Orozco-terWengel P, Barbato M, Nicolazzi E, Biscarini F, Milanesi M, Davies W, et al. Revisiting demographic processes in cattle with genome-wide population genetic analysis. Front Genet. 2015;6:191. https://doi.org/10.3389/fgene.2015.00191.

Park SD, Magee DA, McGettigan PA, Teasdale MD, Edwards CJ, Lohan AJ, et al. Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. Genome Biol. 2015;16:234. https://doi.org/10.1186/s13059-015-0790-2.

Pérez O'Brien AM, Mészáros G, Utsunomiya YT, Sonstegard TS, Garcia JF, Van Tassel CP, et al. Linkage disequilibrium levels in Bos indicus and Bos taurus cattle using medium and high density SNP chip data and different minor allele frequency distributions. Livest Sci. 2014;166:121–32. https://doi.org/10.1016/j.livsci.2014.05.007.

Petersen JL, Mickelson JR, Rendahl AK, Valberg SJ, Andersson LS, Axelsson J, Bailey E, et al. Genome-wide analysis reveals selection for important traits in domestic horse breeds. PLoS Genet. 2013;9(1):e1003211. https://doi.org/10.1371/journal.pgen.1003211.

Pryce JE, Woolaston R, Berry DP, Wall E, Winters M, Butler R, Shaffer M. World trends in dairy cow fertility. In: Proceedings of 10th world congress of genetics applied to livestock production. 2014. https://asas.org/docs/default-source/wcgalp-proceedings-oral/154_paper_10356_manuscript_1630_0.pdf?sfvrsn=2. Accessed 1 Dec 2016

Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H. A genome-wide scan for signatures of recent selection in Holstein cattle. Anim Genet. 2010;41(4):377–89. https://doi.org/10.1111/j.1365-2052.2009.02016.x.

Quilez J, Short AD, Martinez V, Kennedy LJ, Ollier W, Sanchez A, et al. A selective sweep of >8 Mb on chromosome 26 in the Boxer genome. BMC Genomics. 2011;12:339. https://doi.org/10.1186/1471-2164-12-339.

Rajora OP, Eckert AJ, Zinck JWR. Single-locus versus multilocus patterns of local adaptation to climate in eastern white pine (*Pinus strobus*, Pinaceae). PLoS One. 2016;11(7):e0158691. https://doi.org/10.1371/journal.pone.0158691.

Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, et al. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS One. 2009;4(8):e6524. https://doi.org/10.1371/journal.pone.0006524.

Randhawa IAS, Khatkar MS, Thomson PC, Raadsma HW. Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. BMC Genet. 2014;15:34. https://doi.org/10.1186/1471-2156-15-34.

Ren J, Duan Y, Qiao R, Yao F, Zhang Z, Yang B, et al. A missense mutation in PPARD causes a major QTL effect on ear size in pigs. PLoS Genet. 2011;7(5):e1002043. https://doi.org/10.1371/journal.pgen.1002043.

Rimbault M, Beale HC, Schoenebeck JJ, Hoopes BC, Allen JJ, Kilroy-Glynn P, et al. Derived variants at six genes explain nearly half of size reduction in dog breeds. Genome Res. 2013;23 (12):1985–95. https://doi.org/10.1101/gr.157339.113.

Ros-Freixedes R, Gol S, Pena RN, Tor M, Ibáñez-Escriche N, Dekkers JCM, Estany J. Genome-wide association study singles out SCD and LEPR as the two main loci influencing intramuscular fat content and fatty acid composition in Duroc pigs. PLoS One. 2016;11(3):e0152496. https://doi.org/10.1371/journal.pone.0152496.

Rothhammer S, Seichter D, Forster M, Medugorac I. A genome-wide scan for signatures of differential artificial selection in ten cattle breeds. BMC Genomics. 2013;14(1):908. https://doi.org/10.1186/1471-2164-14-908.

Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 2010;464 (7288):587–91. https://doi.org/10.1038/nature08832.

Rubin C-J, Megens H-J, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. Proc Natl Acad Sci U S A. 2012;109 (48):19529–36. https://doi.org/10.1073/pnas.1217149109.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419 (6909):832–7. https://doi.org/10.1038/nature01140.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449(7164):913–8. https://doi.org/10.1038/nature06250.

Schennink A, Stoop WM, Visker MHPW, Heck JML, Bovenhuis H, Van der Poel JJ, et al. DGAT1 underlies large genetic variation in milk-fat composition of dairy cows. Anim Genet. 2007;38 (5):467–73. https://doi.org/10.1111/j.1365-2052.2007.01635.x.

Schoenenbeck JJ, Ostrander EA. The genetics of skull shape variation. Genetics. 2013;193(2):317–25. https://doi.org/10.1534/genetics.112.145284.

Schoenebeck JJ, Hutchinson SA, Byers A, Beale HC, Carrington B, Faden DL, et al. Variation of BMP3 contributes to dog breed skull diversity. PLoS Genet. 2012;8(8):e1002849. https://doi.org/10.1371/journal.pgen.1002849.

Stainton JJ, Charlesworth B, Haley CS, Kranis A, Watson K, Wiener P. Detecting signatures of selection in nine distinct lines of broiler chickens. Anim Genet. 2015;46(1):37–49. https://doi.org/10.1111/age.12252.

Stainton JJ, Charlesworth B, Haley CS, Kranis A, Watson K, Wiener P. Use of high-density SNP data to identify patterns of diversity and signatures of selection in broiler chickens. J Anim Breed Genet. 2017;134(2):87–97. https://doi.org/10.1111/jbg.12228.

Stern JA, White SN, Meurs KM. Extent of linkage disequilibrium in large-breed dogs: chromosomal and breed variation. Mamm Genome. 2013;24(9-10):409–15. https://doi.org/10.1007/s00335-013-9474-y.

Sun D, Jia J, Ma Y, Zhang Y, Wang Y, Yu Y, Zhang Y. Effects of *DGAT1* and *GHR* on milk yield and milk composition in the Chinese dairy population. Anim Genet. 2009;40(6):997–1000. https://doi.org/10.1111/j.1365-2052.2009.01945.x.

Sutter NB, Bustamante CD, Chase K, Gray MM, Zhao K, Zhu L, et al. A single IGF1 allele is a major determinant of small size in dogs. Science. 2007;316(5829):112–5. https://doi.org/10.1126/science.1137045.

Thaller G, Kramer W, Winter A, Kaupe B, Erhardt G, Fries R. Effects of DGAT1 variants on milk production traits in German cattle breeds. J Anim Sci. 2003;81(8):1911–8.

Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, et al. Design and characterization of a 52K SNP chip for goats. PLoS One. 2014;9(1):e86227. https://doi.org/10.1371/journal.pone.0086227.

Uemoto Y, Soma Y, Sato S, Ishida M, Shibata T, Kadowaki H, et al. Genome-wide mapping for fatty acid composition and melting point of fat in a purebred Duroc pig population. Anim Genet. 2012;43(1):27–34. https://doi.org/10.1111/j.1365-2052.2011.02218.x.

Uimari P, Tapio M. Extent of linkage disequilibrium and effective population size in Finnish Landrace and Finnish Yorkshire pig breeds. J Anim Sci. 2011;89(3):609–14. https://doi.org/10.2527/jas.2010-3249.

Van Laere A-S, Nguyen M, Braunschweig M, Nezer C, Collette C, Moreau L, et al. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature. 2003;425(6960):832–6. https://doi.org/10.1038/nature02064.

Van Raden PM. Invited review: selection on net merit to improve lifetime profit. J Dairy Sci. 2004;87(10):3125–31. https://doi.org/10.3168/jds.S0022-0302(04)73447-5.

Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, et al. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLoS Genet. 2011;7(10):e1002316. https://doi.org/10.1371/journal.pgen.1002316.

Voight BF, Kudaravalli S, Wen XQ, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4(3):446–58. https://doi.org/10.1371/journal.pbio.0040072.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, et al. Genome sequence, comparative analysis and population genetics of the domestic horse (Equus caballus). Science. 2009;326(5924):865–7. https://doi.org/10.1126/science.1178158.

Wall E, Brotherstone S, Woolliams JA, Banos G, Coffey MP. Genetic evaluation of fertility using direct and correlated traits. J Dairy Sci. 2003;86(12):4093–102. https://doi.org/10.3168/jds.S0022-0302(03)74023-5.

Walsh SW, Williams EJ, Evans ACO. A review of the causes of poor fertility in high milk producing dairy cows. Anim Reprod Sci. 2011;123(3-4):127–38. https://doi.org/10.1016/j.anireprosci.2010.12.001.

Wang C, Wang H, Zhang Y, Tang Z, Li K, Liu B. Genome-wide analysis reveals artificial selection on coat colour and reproductive traits in Chinese domestic pigs. Mol Ecol Resour. 2015;15:414–24. https://doi.org/10.1111/1755-0998.12311.

Warriss PD, Kestin SC, Brown SN, Nute GR. The quality of pork from traditional pig breeds. Meat Focus Int. 1996;5:179–82.

Webster MT, Kamgari N, Perloski M, Hoeppner MP, Axelsson E, Hedhammar Å, et al. Linked genetic variants on chromosome 10 control ear morphology and body mass among dog breeds. BMC Genomics. 2015;16:474. https://doi.org/10.1186/s12864-015-1702-2.

Wei WH, de Koning DJ, Penman JC, Finlayson HA, Archibald AL, Haley CS. QTL modulating ear size and erectness in pigs. Anim Genet. 2007;38(3):222–6. https://doi.org/10.1111/j.1365-2052.2007.01591.x.

Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG. Measures of human population structure show heterogeneity among genomic regions. Genome Res. 2005;15(11):1468–76. https://doi.org/10.1101/gr.4398405.

Wiedemar N, Drögemüller C. A 1.8-kb insertion in the 3′-UTR of RXFP2 is associated with polledness in sheep. Anim Genet 2015;46(4):457–61. https://doi.org/10.1111/age.12309.

Wiener P, Gutierrez-Gil B. Assessment of selection mapping near the myostatin gene (GDF-8) in cattle. Anim Genet. 2009;40(5):598–608. https://doi.org/10.1111/j.1365-2052.2009.01886.x.

Wiener P, Pong-Wong R. A regression-based approach to selection mapping. J Hered. 2011;102(3):294–305. https://doi.org/10.1093/jhered/esr014.

Wiener P, Wilkinson S. Deciphering the genetic basis of animal domestication. Proc Biol Sci. 2011;278(1722):3161–70. https://doi.org/10.1098/rspb.2011.1376.

Wiener P, Burton D, Ajmone-Marsan P, Dunner S, Mommens G, Nijman IJ, et al. Signatures of selection? Patterns of microsatellite diversity on a chromosome containing a selected locus. Heredity. 2003;90(5):350–8. https://doi.org/10.1038/sj.hdy.6800257.

Wiener P, Edriss MA, Williams JL, Waddington D, Law A, Woolliams J, Gutierrez-Gil B. Information content in genome-wide scans: concordance between patterns of genetic differentiation and linkage mapping associations. BMC Genomics. 2011;12:65. https://doi.org/10.1186/1471-2164-12-65.

Wilkinson S, Wiener P, Teverson D, Haley CS, Hocking PM. Characterization of the genetic diversity, structure and admixture of British chicken breeds. Anim Genet. 2012;43(5):552–63. https://doi.org/10.1111/j.1365-2052.2011.02296.x.

Wilkinson S, Lu ZH, Megens H-J, Archibald AL, Haley C, Jackson IJ, et al. Signatures of diversifying selection in European pig breeds. PLoS Genet. 2013;9(4):e1003453. https://doi.org/10.1371/journal.pgen.1003453.

Wright S. Color inheritance in mammals: cattle. J Hered. 1917;8:521–7.

Wright S. The genetical structure of populations. Ann Eugen. 1951;15:323–54.

Wutke S, Benecke N, Sandoval-Castellanos E, Döhle HJ, Friederich S, Gonzalez J, et al. Spotted phenotypes in horses lost attractiveness in the Middle Ages. Sci Rep. 2016;6:38548. https://doi.org/10.1038/srep38548.

Zeder MA. Pathways to animal domestication. In: Gepts P, Famula TR, Bettinger RL, Brush SP, Damania AB, McGuire PE, Qualset CO, editors. Biodiversity in agriculture: domestication, evolution and sustainability. Cambridge: Cambridge University Press; 2012. p. 227–59.

Zhang Y, Liang J, Zhang L, Wang L, Liu X, Yen H, et al. Porcine *methionine sulfoxide reductase B3:* molecular cloning, tissue-specific expression profiles, and polymorphisms associated with ear size in *Sus scrofa*. J Anim Sci Biotech. 2015;6:60. https://doi.org/10.1186/s40104-015-0060-x.

Zhao F, McParland S, Kearney F, Du L, Berry DP. Detection of selection signatures in dairy and beef cattle using high-density genomic information. Genet Sel Evol. 2015;47:49. https://doi.org/10.1186/s12711-015-0127-3.

Population Genomics of Domestication and Breed Development in Canines in the Context of Cognitive, Social, Behavioral, and Disease Traits

Kristopher J. L. Irizarry and Elton J. R. Vasconcelos

Abstract Dogs are unique because they are known to be the first species domesticated by humans, have the greatest morphological variation among terrestrial mammals, and exhibit unique bonds with humans. Yet, until very recently, the history of domestication and the associated consequences of this artificial selection have been a matter of speculation. Domesticated dogs are the ideal organism to study population genomics of domestication and the impact artificial selection has had on cognitive, social, behavioral, and disease traits. Because dogs have been associated with humans for tens of thousands of years, they are uniquely suited to investigate the genetic basis of selection for dietary adaptation during the agricultural revolution. Through a variety of large-scale genomics approaches, the history and consequences of dog domestication are no longer a matter of speculation. This chapter delves into the ancient origins of human-canine interactions and follows the domestication of wolves into dogs with a particular focus on (a) the selection of phenotypes underlying the strong bond between humans and their companion dogs, (b) the morphological variation underlying dog breeds, and (c) the genetic basis of canine diseases. The historical picture that is beginning to emerge provides a genomics framework for understanding why and how the dog became "our best friend."

Keywords Behavioral phenotypes · Canine · Canine genetic diseases · Cognitive phenotypes · Disease phenotypes · Dog breeding · Dog breeds · Domestication · Genomics · Human-animal bond

K. J. L. Irizarry (✉) · E. J. R. Vasconcelos
The Applied Genomics Center, College of Veterinary Medicine, Western University of Health Sciences, Pomona, CA, USA
e-mail: kirizarry@westernu.edu

Om P. Rajora (ed.), *Population Genomics: Concepts, Approaches and Applications*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_43,

1 Introduction

Dogs (*Canis lupus familiaris*) are fondly referred to as "our best friends," and among all organisms on this planet are the species most closely associated with humans. They live in our houses, sleep in our beds, ride in our cars, and even cuddle with us on the couch, while we read and relax. Our shared history extends back tens of thousands of years. This is by far, the longest ever genetic experiment and continues today, with designer dogs selected from crosses of established breeds to produce new and unique combinations of traits. For much of our shared coexistence, the actual impact of domestication and artificial selection has been a matter of speculation.

A wide-held belief is that domestication simply caused dogs to lose their fear of humans. However, the footprints of selection (meaning the specific versions of particular genes selected for in the dog genome during artificial selection) can be detected using population genomics methods. Furthermore, the role these genes play in physiology and biochemistry can be determined using information from prior studies in humans, mice, and other "model" organisms. For example, if one were to identify a gene in wolves (*Canis lupus*) associated with a particular trait and observe that this gene has many different variants in the wolf population, but only a tiny fraction of variants within domesticated dogs, this might provide support for the hypothesis that selection for a particular variant of this gene occurred during artificial selection.

Questions such as "*How are dogs different from wolves?*" and "*What regions of the dog genome encode the traits humans selected for when dogs were domesticated?*" are within the realm of scientific investigation. Population genomics methods provide strategies for decoding the phenotypic consequences of patterns of genetic variation in specific populations. By comparing patterns of genetic variation between two populations (such as wolves and domestic dogs or between Chihuahuas and Great Danes) specific phenotypic differences between the populations can be associated with precise regions of the genome. Such approaches, coupled with comparative genomics and bioinformatics methods, enable us to uncover the particular genes selected during artificial selection and identify the traits these genes encode in the dog genome. Together, this information provides answers to questions about domestication and breed formation.

This chapter presents an informative review detailing genomics aspects of canine domestication and breed formation, with a particular emphasis on cognitive, social, behavioral, and disease traits. The goal of this chapter is to provide a framework for understanding how population genetics and genomics methods have been used to decipher the domestication history and resulting phenotypes that are observed in dogs today. The chapter opens with a brief review of some archeological samples of ancient canids and the results of their genetic analysis. Dogs are the oldest domesticated species and therefore have the longest shared history with humans among all life on the planet. Subsequently, the chapter explores the cognitive and behavioral changes that dogs underwent during their domestication and discusses

a number of studies that have identified the genes underlying these augmented phenotypes that connect them to humans as both companion and working animals.

Next, dog domestication from wolves is presented in the context of sequencing datasets and polymorphic marker analyses. These studies helped elucidate the early history of the dog domestication process. Dogs are known for exhibiting a tremendous amount of phenotypic variation within the species. The chapter then explores the morphological variation and methods employed to deduce the genetic mechanisms underlying this morphological variation. Finally, the chapter delves into clinically relevant phenotypes between specific dog breeds and genes, mutations, and genomic regions underlying these breed-associated diseases. Ultimately, this chapter presents the culmination of our current genetic understanding of canid domestication and provides numerous examples of the specific phenotypes underlying the transformation of ancestral wolves into the dogs we live with today.

2 Time and Place of Dog Domestication

Population genomics methods have offered an unprecedented opportunity to unravel the mysteries underlying dog domestication. These powerful and data-dense genetic approaches have refined our understanding of how dogs transformed from wolves into the hundreds of breeds that exist today. Moreover, through these studies the genomic basis underlying morphological variation between dog breeds is emerging. Through a combination of genetic association studies, whole genome sequencing, and gene expression studies, the veil covering our evolutionary history with dogs has finally been lifted, and the initial discoveries consist of many surprises that, when viewed in the context of "our best friend," make a lot of sense.

2.1 Archeological Evidence

One of the most fundamental and frequently contemplated questions relating humans and dogs is "*When were dogs initially domesticated*?" This is a particularly important question that lies at the heart of the human-animal bond. Ovodov et al. (2011) describe the discovery of 33,000 years old incipient dog remains within the Altai Mountains of Siberia including a complete skull and mandible that were excavated from the site in 1975 (Fig. 1). Evidence of human occupation within the vicinity of the skull and mandible date back approximately 50,000–100,000 years ago and correspond to hunter gatherers that remained in a single location for multiple months at a time (Ovodov et al. 2011).

Fig. 1 33,000-year-old dog skull and mandible represent early stage of canine domestication. (**a**) Aerial view, (**b**) profile, (**c**) palate, (**d**) left mandible, (**e**) left lower tooth row (scale on ruler in cm). Subtriangular hole in the skull is the place of initial sampling for carbon-14 dating in 2007. Originally published in Ovodov et al. (2011)

2.2 Genetic Analysis of Archeological Samples

In 2015 Lee et al. reported the sequencing and phylogenetic analysis of a particular mitochondrial (mt) genomic region, a polymorphic portion of the canine mitochondrial genome that exhibits a 10 bp repeat region that varies by both number of copies and sequence variation between individuals. The data was derived from a 360,000- to 400,000-year-old *Canis* cf. *variabilis* mandible (Fig. 2) obtained from a region in Siberia (Fig. 3) from which multiple ancient and contemporary canid samples have been identified. The study yielded mtDNA region sequence data for all samples investigated leading to the discovery of nine haplotypes. Phylogenetic analysis of the data indicated that the *Canis* cf. *variabilis* sample clustered with other wolf samples from Asia and Russia.

Of particular interest in the Lee et al. (2015) study was the analysis of the haplotypes across ancient wolf samples and contemporary dog breeds. The results indicated that haplotypes obtained from 8,750-year-old samples (site 1 on the map) and 28,000-years-old samples (site 2 on the map) are indistinguishable from

Fig. 2 A 360,000- to 400,000-year-old *Canis* cf. *variabilis* mandible obtained from Siberia. Originally published in Lee et al. (2015)

haplotypes observed in geographically diverse dog breeds that exist today (Fig. 3). A surprising result was that the haplotype observed in the 47,000-year-old canid sample was quite distinct from other wolf haplotypes but differed by only a few mutations from haplotypes observed in the present-day dogs. Taken together, these results provide support for the idea that genetic contributions of ancient Siberian wolves, including possibly *Canis* cf. *variabilis*, may have contributed to the genetic structure of the domestic dog gene pool.

Interestingly, dog domestication appears to have occurred in multiple locations at different times. For example, Thalmann et al. (2013) sequenced the complete mitochondrial genomes of 18 prehistoric canids and compared the results to modern dogs and wolves using maximum likelihood, coalescence, and Bayesian approaches to ascertain phylogenetic relationships. Their findings suggest that contemporary dogs derive their mitochondrial genomes from European canids (Thalmann et al. 2013).

A 2015 study reported by Shannon et al. employed a 185,805-marker genotyping array to investigate the population structure of 4,676 purebred dogs (representing over 160 breeds) and 549 free-ranging village dogs representing 38 countries. The results identified certain geographical subsets of village dogs that appear to be derived almost exclusively from European origins, while village dogs from countries such as Vietnam, India, and Egypt have trace amounts of European admixture, supporting an origin of domestication within Central Asia instead (Shannon et al. 2015).

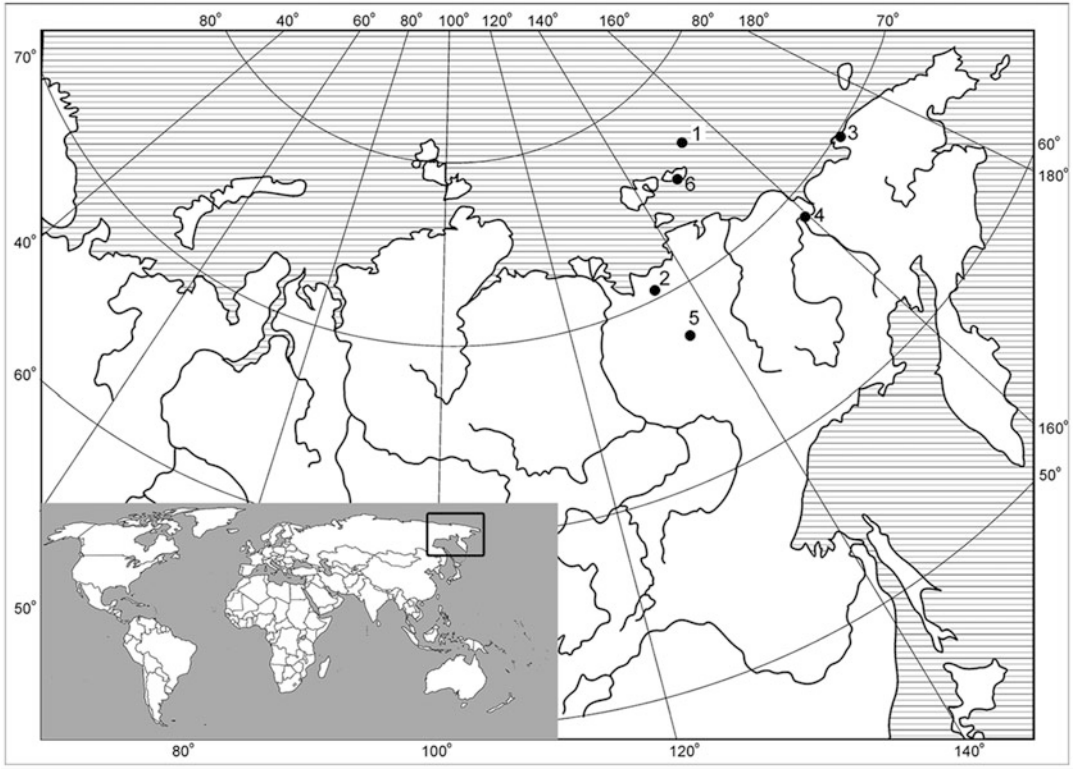

Fig. 3 Region of Siberian Arctic where numerous ancient canid samples have been identified. A number of canid samples have been obtained from six specific sites within this region including: 8,750-year-old *Canis* sp. (site 1 – Zhokhov Island, New Siberian Islands), 28,000-year-old *Canis lupus* (site 2 – Yana RHS, Lower Kolyma River), 1,750-year-old *Canis* sp. (site 3 – Aachim, East Siberian Sea Coast), 47,000-year-old *Canis lupus* (site 4 – DuvanyYar, Lower Kolyma River), 360,000- to 400,000-year-old *Canis* cf. *variabilis* (site 5 – Ulakhan-Suller, Adycha River), and contemporary *Canis lupus* (site 6 – New Siberian Islands). Originally published in Lee et al. (2015)

3 Domestication of Dogs from Wolves

The phenotypic variation among domestic dogs is a consequence of the artificial selection imposed during their domestication and subsequent morphological phenotypic variation that occurred during stratification into different breeds. As of 2018, the American Kennel Club (AKC) recognizes close to 200 distinct dog breeds with additional breeds added each year (http://www.akc.org/). In comparison, the United Kennel Club (UKC) recognizes more than 300 different breeds (https://www.ukcdogs.com) and adds new breeds to the list over time. Similarly, the largest kennel club in the world, Fédération Cynologique Internationale (FCI), currently recognizes close to 350 unique dog breeds (http://www.fci.be). Interestingly, there are dog breeds that are not formally recognized by a breed club. Recently, designer dogs, which are crosses between dogs of different breeds, have gained in popularity. These breeds, lineages, and designer dogs represent pools of dogs that share subsets of genetic variation and together represent one of the most phenotypically diverse species on the planet.

3.1 Early Dog History and Models of Dog Domestication

A recent study of early dog history attempted to characterize the ancestral relationships between dogs and wolves (Lindblad-Toh et al. 2005). This approach used deep genome sequencing of (a) three gray wolves (each centered on a geographical region presumed to correspond to a geographical dog domestication site), (b) two basal dog lineages (the Dingo and Basenji), and (c) the golden jackal (Freedman et al. 2014). In this study the investigators also had access to the Boxer genome because the initial dog genome sequenced published in 2005 was obtained from a female Boxer (Lindblad-Toh et al. 2005). The distribution of samples is illustrated in Fig. 4.

Fig. 4 Geographical distribution of canid samples in genome sequencing study of early dog history. Originally published in Freedman et al. (2014)

Among the data generated across the six canid genomes were 11.2 billion sequencing reads producing over 10.2 million single nucleotide polymorphisms. From this data, the authors inferred effective population sizes based on genome-wide heterozygosity within each genome using a pairwise sequential Markovian coalescent method. By considering an average mutation rate of 1×10^{-8} per generation, the investigators suggest that the dog population underwent a 16-fold reduction over the past 50,000 years. To determine admixture among the genomes, the authors used the nonparametric "ABBA-BABA" test for gene flow between divergent populations. The results of the study were used to construct three models of dog domestication (Fig. 5) that each includes estimates of population divergence and post-divergent gene flow between sample populations.

The three models (Fig. 5) differ in how the ancestral population of wolves ultimately gave rise to different wolf populations and dogs such as the Boxer, the Basenji, and the Dingo. Figure 5a illustrates the model most consistent with the

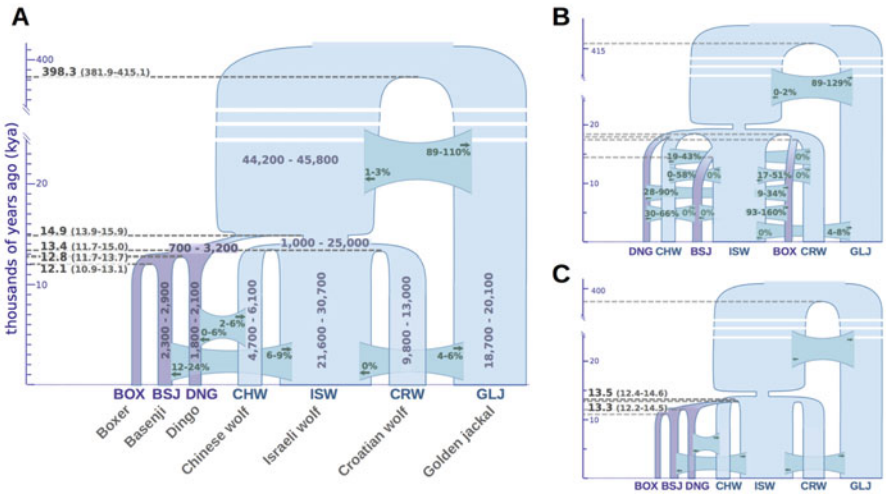

Fig. 5 Models of canine domestication derived from genome sequencing study of early dog history. Divergence times, effective population sizes (N_e), and post-divergence gene flow inferred by *G-PhoCS* in joint analysis of the Boxer reference genome and the sequenced genomes of two basal dog breeds, three wolves, and a golden jackal. The width of each population branch is proportional to inferred population size, and stated ranges of parameter estimates indicate 95% Bayesian credible intervals. Horizontal gray dashed lines indicate timing of lineage divergences, with associated means in bold, and 95% credible intervals in parentheses. Migration bands are shown in green with associated values indicating estimates of total migration rates, which are equal to the probability that a lineage will migrate through the band during the time period when the two populations co-occur. Panels show parameter estimates for (**a**) the population tree best supported by genome-wide sequence divergence, (**b**) a regional domestication model, and (**c**) a single wolf lineage origin model in which dogs diverged most recently from the Israeli wolf lineage (similar star-like divergences are found assuming alternative choices for the single wolf ancestor). Estimated divergence times and effective population sizes are calibrated assuming an average mutation rate of 1×10^{-8} substitutions per generation and an average generation time of 3 years. Originally published in Freedman et al. (2014)

genome-wide sequencing divergence. In this model, the Boxer, Basenji, and Dingo exhibit a lineage from the ancestral wolf population that is distinct from the Chinese wolf, the Israeli wolf, and the Croatian wolf. Figure 5b shows a model in which the Dingo splits off from the Chinese wolf lineage, the Basenji splits from the Israeli wolf lineage, and the Boxer splits from the Croatian wolf lineage. Finally, Fig. 5c represents a model in which all three dogs (Basenji, Dingo, and Boxer) split from the Chinese wolf rather than the ancestral wolf population (as shown in Fig. 5a).

Shearin and Ostrander (2010) provide a quantifiable measure of similarity between dogs and wolves stating that domestic dog differs from its closest ancestor, the gray wolf, by just 0.04% in nuclear protein-coding DNA sequence. In other words, dogs share 99.06% of their protein coding genome with wolves.

4 Evolution and Selection of Cognitive and Behavioral Traits During Canine Domestication

A long-standing question many have asked relating to human domestication of dogs is "*During artificial selection was there any selection for cognitive, behavioral, or communication phenotypes that may have contributed to a strong interspecific bond between humans and their companion dogs*?" The strength of the human-animal bond is so strong that dogs are fondly referred to as humans' "best friends." Subsequently, it seems plausible that artificial selection during domestication may have contributed to divergent phenotypes from wolves that underlie the social interactions between dogs and humans.

4.1 Gene Expression Differences in Brains of Dogs and Wolves

Saetre et al. (2004) investigated the mRNA expression levels of 7,762 genes in dogs, wolves, and coyotes (*Canis latrans*) in three regions of the brain: the hypothalamus, the amygdala, and the frontal lobe. Interestingly, the RNA was obtained from postmortem brains and hybridized to human microarrays. Cross-species microarray hybridization is inherently challenging, and the extent of sequence divergence between the species (human, dog, wolf, and coyote) contributes to interspecific variation in hybridization efficiency. Nonetheless, the investigators chose to focus on a set of genes that exhibited brain region specificity for one of the three brain regions compared to the other two. Specifically, the selected inclusion criteria required that at least a twofold difference in expression was necessary to consider a gene as brain region specific (Saetre et al. 2004).

In the first set of the gene expression experiments, 156 genes were identified as having region-specific expression in all three species. In a second set of experiments, 114 genes exhibiting expression differences between species within each brain region were identified. Next average interspecies expression differences were determined for all 114 genes. These findings led to the observation that in the amygdala and frontal lobe, average differences in expression were close to 30% and similar across all three species. However, the average expression difference in the hypothalamus was around 20% with a difference between coyotes and wolves of merely 13% (background "noise" was 9%). When wolves and dogs were compared for hypothalamus gene expression, there was an average difference of 24%, and the difference between dogs and coyotes was 22%. Gene Ontology (GO) analysis was performed for the 25 genes that exhibited GO annotation. The results indicated that 25 genes shared annotation of overrepresented GO terms (and only 2 were expected by random chance alone). The enriched terms included neurogenesis, cell-cell signaling, and neurotransmission. Among the genes exhibiting such annotation, many were downregulated in the hypothalamus of dogs. Two of these genes are the neuropeptides *NPY* and *CALCB* implicated in energy regulation, feeding behavior, and the hypothalamic pituitary adrenal (HPA)-associated neuroendocrine stress response. Perhaps domestication of dogs occurred, in part, through genetic variation that modulates gene expression levels in particular regions of the brain underlying stress response phenotypes (Saetre et al. 2004).

4.2 *Population Differentiation Between Native Dogs and Wolves*

A similar study by Li et al. (2013) employed a pairwise population differentiation between Chinese native dogs and gray wolves. Chinese native dogs are dogs that live as human commensals and were included in the study to capture the genetic structure of dogs prior to the recent stratification associated with breed creation. Furthermore, the authors chose to compare genome-wide divergence between the Chinese native dogs and German Shepherds obtained from Germany. A total of 48,455 SNPs were selected after filtering, and the average distance between the SNPs across the genome was 23 kb. A final set of 1,878 SNPs were identified, corresponding to the top 5% of the distribution, with $F_{ST} > 0.05$ and mean $F_{ST} = 0.63$ between Chinese native dogs and wolves. These SNPs can be considered to be under strong selection (Li et al. 2013).

Gene Ontology biological process enrichment analysis revealed that 347 genes were associated with behavior, and of those, 224 were associated with locomotor behavior. The analysis of SNPs exhibiting highest F_{ST} values between Chinese native dogs and German Shepherds lacked the extent of exacerbated brain expression that was observed among the genes identified between the Chinese native dogs and the wolves. The authors make the case that human artificial selection during the

primary splitting of dogs from wolves was associated with rapid brain evolution. Furthermore, they connect the emergence of dog-specific behaviors during domestication with altered gene expression changes in their brains (Li et al. 2013).

4.3 Whole Genome Sequence Differences Between Dogs and Wolves

Li et al. (2014) compared the published resequenced genomes of three wolves and ten dogs (five ancient dogs, five contemporary dogs) to an additional three wolves and three Chinese native dog genomes that the group sequenced to identify regions of the genome exhibiting the most dramatic differences between dogs and wolves. A common hypothesis associated with dog domestication is that human artificial selection resulted in altered stress response phenotypes, which facilitated dogs and humans living in closer proximity than wolves and humans. Li et al. argue that if stress-response phenotype was "selected" during domestication, one would expect to see evidence of fixed alleles within genes mediating the phenotype to remain fixed today (Li et al. 2014).

Surprisingly, fixed SNPs within the genes *GRIK3*, *GABRA5*, *GRIK2*, *BCL2*, and *MECP2* were identified in the analysis, and GO enrichment identified the following biological processes as the most significantly enriched: adenylate-cyclase-inhibiting G-protein-coupled receptor activity and glutamate receptor signaling pathway. Glutamate is the brain's main excitatory neurotransmitter and regulates behaviors, emotions, cognitive abilities, as well as learning and memory. The gene expression analysis of the GRIK2 gene indicated that it exhibited greater levels of expression in dog prefrontal cortex compared to wolf prefrontal cortex ($p = 0.0006$) (Li et al. 2014).

Although not statistically significant, *BCL2* and *GABRA5* also exhibit changes that distinguish the dog from the wolf. A weighted gene co-expression network analysis revealed that *GRIK2*, *GRIK3*, *GABRA5*, and *MECCP2* exhibit co-expression patterns that place them all in the same coregulatory network. The authors make the case that, during the early stages of domestication, wolves with better learning and memory phenotypes would "come close to human settlements more frequently, acquire greater food resources, and thus had greater opportunities to survive (with little disadvantage). These individuals would exhibit nonaggressive response because they would understand that the presence of humans was harmless, and thus would have a weakened fear response." The authors propose that instead of reduced fear response, domestication of dogs occurred via selection for excitatory synaptic plasticity, which would alter dog behavior and cognition to the point where dogs could learn the meaning of human gestures and respond more favorably to human commands (Li et al. 2014). The idea that artificial selection during domestication altered the canine brain to enhance dog memory is an exciting and potentially transformative event in human-animal history.

5 Genetic Effects of Dog Domestication

Domestication events can create bottlenecks and consequently reduce genetic diversity, reduce effective population size, and increase inbreeding. Understanding the relationship between the genomic signals observed in the data and the evolutionary mechanisms that contributed to those signals is critical if one hopes to understand how the domestication and selective breeding history of contemporary dog breeds exploited the morphological plasticity encoded in the ancestral canine genome. Boyko et al. (2010) suggest long runs of homozygosity (ROHs) are the result of inbreeding associated with recent selection events, such as breed formation. In contrast, the authors attribute haplotype diversity and linkage disequilibrium (LD) occurring across genomic scales less than a megabase as indicative of more ancestral population processes (Boyko et al. 2010).

The first question addressed by Boyko et al. (2010) was to investigate genomic signatures of canine demographic history by analyzing (1) the pairwise SNP LD, (2) the haplotype diversity across 15-SNP windows, and (3) the extent of ROHs greater than a megabase. They discovered that although the LD extends over 1 Mb within every breed assessed across the entire population of dogs, it decays very quickly. This observation implies that identity-by-descent (IBD) segments are shared across numerous breeds and are quite small. The ROHs observed were longer and occurred more frequently in breed dogs than wolves or the village dogs. Individuals from almost all breeds exhibited between 10 and 50 ROHs greater than 10 Mb. The exception was the Jack Russell Terriers, which showed fewer ROHs and higher levels of genetic diversity than the other breeds (Boyko et al. 2010).

Autozygosity, which occurs when both chromosomes in a diploid organism are derived from the same ancestor, indicates that inbreeding has occurred. Current models suggest "inbreeding depression" is an increase in autozygosity coupled with an increased risk in homozygosity at rare, partially recessive, deleterious mutations. To investigate the impact of autozygosity, it is important to accurately identify real autozygous ROHs from the larger set of often non-autozygous ROHs in a sample (Boyko et al. 2010). Non-autozygous ROHs, stretches of homozygous SNPs that are actually heterozygous at unmeasured variants, are less likely to contain rare, partially recessive, deleterious mutations in homozygous form. Subsequently, an important criterion for defining ROHs – rather than SNP-by-SNP homozygosity – is to assess autozygosity. It is important to identify ROHs that are not autozygous and are identical-by-state (IBS) from ROHs that are autozygous and are identical-by-descent (IBD) (Howrigan et al. 2011).

According to Boyko et al. (2010), autozygosity was detected at high levels in all breeds with Jack Russell Terriers having the lowest average autozygosity (7.5%) and Boxers having the highest (51%). Interestingly, only a few breeds contained genomic regions that were autozygous in all breed members genotyped at the megabase scale. The exception was Basenjis, which showed evidence of high haplotype diversity coupled with high autozygosity. Together these two conditions are suggestive of a recent genetic bottleneck following breed formation that caused

greater levels of inbreeding than would otherwise be expected in the population. According to the breed history, Basenjis in the United States were derived from a relatively small founder population. Linkage disequilibrium (LD), associated with regions of chromosomes encoding shared alleles from common ancestors along a chromosome, is known to extend greater genomic distances within breeds than it does among breeds or within wolves. The analysis performed by the authors indicates that the between-breed LD is much greater than wolf LD which provides support for a bottleneck in dogs during domestication (Boyko et al. 2010).

These results support the idea that dramatic genomic selection occurred within the dog genome on multiple time scales. One time scale, for example, corresponds to an ancient domestication selection process when dogs were selected for affiliation with humans. Afterwards, a more recent breed-radiation selection process occurred where closed breeding pools were created to transform the ancestral genetic variation into breed-specific pockets of genetic and morphological phenotypic uniformity.

6 How Did Domestication-Modulated Oxytocin Mediated Phenotypes

6.1 Oxytocin-Mediated Social Phenotypes in Dogs

The neuropeptide hormone, oxytocin, has a well-established role underlying social bonding in mammals where, through evolution, it has mediated hierarchical social relationships as well as organization of social interactions. In humans, oxytocin coordinates parental responses after physical contact with offspring, interactions between sexual partners, interactions with friends, and empathetic interactions with strangers (Feldman 2017).

Romero et al. (2014) described a prosocial role for oxytocin in dogs. They suggested that oxytocin facilitates prosocial interactions among dogs and humans. Furthermore, they make the point that evolutionary selection pressure may have contributed to the maintenance of neurological mechanisms associated with social bonding due to the adaptive value of long-lasting social relations (Romero et al. 2014).

6.2 Genetic Variation in Dog Oxytocin Receptor

The role of oxytocin signaling in the human-animal bond suggests that it is possible that domesticated dogs were artificially selected for more affiliative relationships with humans through allelic variation within genes mediating oxytocin signaling. And indeed there is evidence for considerable genetic variation within the oxytocin receptor in canids, as well among different dog breeds (Kis et al. 2014; Bence et al. 2017).

Kis et al. (2014) investigated three polymorphisms within the receptor located within either the 5′ UTR or the 3′ UTR of the gene. They genotyped 103 Border Collies (46 males, 57 females), consisting of two subpopulations (59 from Hungary and 44 from Belgium). Additionally, they genotyped a single population of 104 German Shepherd dogs (58 male, 46 female) and assessed behavioral phenotypes across five specific tests: (1) greeting the dog, (2) separation from owner, (3) problem-solving, (4) threatening stranger, and (5) owner hiding from dog. The study results demonstrated evidence of an association between the G-allele of -212AG polymorphism and the behavioral phenotype of decreased owner proximity seeking in both breeds. Additionally, the authors report an association of the rs8679684 polymorphism with friendliness; however, the breeds exhibited divergent associations with the *A*-allele in German Shepherds exhibiting higher friendliness phenotype scores, while in Border Collies, the *A*-allele was linked to decreased friendliness (Kis et al. 2014). Note that the -212AG polymorphism was subsequently renamed to the -213 AG polymorphism as the genomic coordinates for the canine oxytocin receptor were refined.

Bence et al. (2017) characterized nine oxytocin receptor polymorphisms in four different canid species. Their study included three novel oxytocin receptor polymorphisms identified through direct sequencing of the gene and regulatory regions in two Eurasian gray wolves, four North American timber wolves, three Beagles, three Border Collies, three German Shepherds, three Golden Retrievers, and three Siberian Huskies. This sequencing led to the identification of -74C/G, 18575C/T, and a microsatellite marker occurring between positions 18772–18792. They also included the three polymorphisms reported in 2014 by Kis et al., -213A/G, 19208A/G (previously called -212A/G and 19131A/G, respectively), and rs8679684. Additional three polymorphisms were identified in public database searches (Bence et al. 2017). Allele frequencies were assessed in 689 purebred dogs (70 Beagles, 144 Border Collies, 128 German Shepherds, 43 Golden Retrievers, 22 Groenendaels, 32 Hungarian Vizslas, 49 Labrador Retrievers, 40 Malinois dogs, 138 Siberian Huskies, and 23 Tervurens) as well as 42 wolves (34 Eurasian gray, 6 North American timber, 2 Alaskan), 6 golden jackals, 8 Dingos, and 45 Asian street dogs.

The results revealed that only the -213A/G G-allele, -94C/T C-allele, -74C/G C-allele, -50C/G C-allele, rs22927829 T-allele, rs8679684 T-allele, and 19208A/G G-allele were detected in all four species. Interestingly, -213A/G A -allele, -50C/G G-allele, and 19208A/G A-allele are only found in wolf and dog, with the wolf having a higher allele frequency than the dog in each case. The rs22927829 A-allele was only detected in dog and Dingo, while the rs8679684 A-allele was found only in dogs. Across the dog breeds and wolf, Bence et al. (2017) reported that only two of the polymorphisms exhibited evidence of both alleles in Border Collie, Golden Retriever, Labrador Retriever, Hungarian Vizsla, Beagle, Tervuren, Groenendael, Malinois, German Shepherd, Husky, and wolf (-94T/C, -74C/G). The -213A/G polymorphism, for which the G-allele was implicated in owner proximity seeking (Kis et al. 2014), lacked evidence of the G-allele in Tervuren and Groenendael breeds. These results underscore the notion that phenotypic variation in social behavior may exist across dog breeds (Bence et al. 2017).

6.3 Visual Communication and Oxytocin

Nagasawa et al. (2015) investigated the physiological consequence of gazing behavior between dog and owner. The rationale for this study was based on the idea that human-like modes of communicating, such as mutual gaze, may have been selected in dogs during domestication by humans. The authors refer to maternal oxytocin levels rising in human mothers when mother-infant gazing occurs. They designed experiments to test the hypothesis that an oxytocin positive feedback loop may be induced by gaze between dogs and their human owners. From the results of their experiments, the authors suggest the existence of a self-perpetuating positive feedback loop mediated by oxytocin in the human-dog bond. The authors characterize the human-dog bond as being similar to the maternal-infant bond because both bonds are associated with oxytocin positive feedback loops across the bond members. Nagasawa et al. (2015) extrapolate from their results and suggest that gazing behavior between dog and owner over thousands of years of domestication and cohabitation conferred social rewarding effects to both humans and dogs. They further point out that this oxytocin release, in both the dog and the human, would result in a deepening of the mutual relationship and further promote interspecies bonding (Nagasawa et al. 2015). They also examined whether an oxytocin loop may have been acquired during dog domestication or whether it is shared among canids that did not undergo domestication by employing hand-raised wolves in their research (Nagasawa et al. 2015). The wolves did not exhibit long periods of gazing at humans. The authors interpret this finding to mean that wolves do not engage in mutual gaze as a means of social communication and interaction with humans. Furthermore, the authors point out that in wolves, eye contact is considered a threat among conspecifics and wolves generally avoid eye contact with humans (Nagasawa et al. 2015).

6.4 Interbreed Differences in Oxytocin-Mediated Phenotypes

Dog breeds differ in social behavior in response to oxytocin. Kovacs et al. (2016) demonstrated the existence of interbreed differences in social behavior associated with intranasal oxytocin in two dog breeds (Siberian Husky and Border Collie) representing distinct genetic lineages. Kovacs et al. genotyped the dogs (18 Siberian Huskies and 16 Border Collies) on the -213A/G oxytocin receptor polymorphism and identified an association between the dog's genotype and social behavior (Kovacs et al. 2016).

The path to canine domestication resulted in selection for traits contributing to enhanced social bonding with humans and increased perception of human communication and nonverbal gestures. The acquisition of these traits allowed dogs to inhabit a unique social niche among humans. Persson et al. (2016) employed a high-density SNP chip and identified *SEZ6L* as a gene exhibiting an association with

variation in social traits. *SEZ6L* has been implicated in autism, a phenotype in which social interaction and communication deficits occur. Other genes located in proximity to the identified haplotype block in the study include *ARVCF*, which has been linked to schizophrenia, and *TXNRD2* and *COMT*, two genes that play roles in schizophrenia and social disorders (Persson et al. 2016).

7 Regions of the Dog Genome Exhibiting Evidence of Positive Selection During Domestication

7.1 Identification of Positively Selected Genes in Dogs Compared to Wolves

Wang et al. (2013) performed a genomic analysis to identify genes that exhibit evidence of positive selection. They highlight the point that artificial selection acting on dogs occurred in two phases. The first phase was defined by the domestication of dogs from wild canids. These descendants of wolves shared living environments with humans and subsequently shared human dietary resources. The second phase was much more recent, occurring over the last few hundred years when morphological variation was created leading to the diverse array of breeds and the physical phenotypes that define them. Wang et al. (2013) suggest that genes selected during the first phase should be shared among all dogs today and designed the experimental approach in this context. Specifically the authors looked for regions of the genome that contain relatively low levels of diversity between dogs and high levels of diversity between wolves and dogs. Regions of the dog genome that contained low levels of diversity in wolves were excluded from the analysis to prevent the identification of genomic regions exhibiting low diversity in dogs that were inherited directly from wolves without selection during domestication (Wang et al. 2013).

Among a set of 17,661 orthologous gene pairs between dogs and humans, 1,708 and 233 genes exhibited evidence of positive selection for humans and dogs, respectively. Gene Ontology enrichment analysis identified terms such as "regulation of digestion," "negative regulation of intestinal phytosterol absorption," "regulation of lipid transport," "axon," "neuron projection," "cell projection," "gamete generation," "sexual reproduction," and "reproductive process in a multicellular organism." These terms are particularly interesting because they reflect three major themes of evolutionary selection during the initial phase of dog domestication: (1) digestion, (2) reproduction, and (3) neurological process (Wang et al. 2013).

Strikingly, among these three functional categories, the authors identified orthologous genes between dogs and humans that show evidence of positive selection in both species. Those genes include *ABCG5*, *ABCG8*, *PLA2G10*, and *PRSS1* associated with nutrition. The genes *GRM8* and *SLC6A4* were identified within the neurological process group. Among the genes implicated in reproduction, *BFAR*, *BRE*, *ITGB1*, *MET, STK17B*, and *ZMYM2* were identified as being positively

selected in both dogs and humans and are involved in cancer, apoptosis, and cell cycle as genes within the reproductive category. Of particular interest are the neurological genes, *GRIM8* and *SLC6A4*, which correspond to glutamate receptor metabotropic 8 and the serotonin transporter, respectively. These genes modulate phenotypes associated with autism and personality traits in humans (Wang et al. 2013).

7.2 *Selection for Enhanced Starch Digestion in Domestic Dogs*

A major event in the domestication of dogs was the selection for a starch diet. Axelsson et al. (2013) used approximately four million SNPs to identify multiple genes associated with starch digestion and fat metabolism that exhibit evidence of selection in dogs (Axelsson et al. 2013). Specifically, the authors identified ten genes implicated in digestion and fat metabolism that were associated with specific mutations found in dogs. These results provide genetic evidence that domesticated dogs adapted to survive on starch-rich diets compared to the carnivorous diets of their wolf ancestors.

In a follow-up study, Arendt et al. (2014) found that high amylase activity in dogs was correlated with pancreatic amylase (*AMY2B*) copy numbers in the genome. The authors characterized the distribution of *AMY2B* copy numbers across 20 breeds and showed that considerable heterogeneity in *AMY2B* copy number exists across dog breeds, ranging from 6 to 14 copies per genome. Dogs living with humans that were exposed to agricultural advances during the prehistoric rise of agriculture benefitted from these dietary resources. Arendt et al. (2016) determined that adaptation to starch diets did not occur early in dog domestication but rather occurred in subpopulations that were exposed to starch-rich diets. Their results show high levels of *AMY2B* copy numbers in most domesticated dogs but relatively few in dogs originating from the Arctic. This is consistent with the historical geographic spread of agriculture (Arendt et al. 2016).

Reiter et al. (2016) demonstrated that positive selection continued to act on dogs that were exposed to starch-rich diets well after dog domestication had occurred. The authors analyzed the relationship between starch-rich diets and dog breeds to gain a better understanding of the relationship the dietary starch played in *AMY2B* copy numbers. Their results demonstrate that dogs exposed to dietary starch exhibit higher allele frequencies of diploid *AMY2B* repeats. This relationship can be seen within specific dog breeds, such as the Shar Pei and Pekingese (exposed to high-starch diets) compared to the Siberian Husky and Alaskan Malamute (exposed to low-starch diets) as illustrated in Fig. 6 (Reiter et al. 2016).

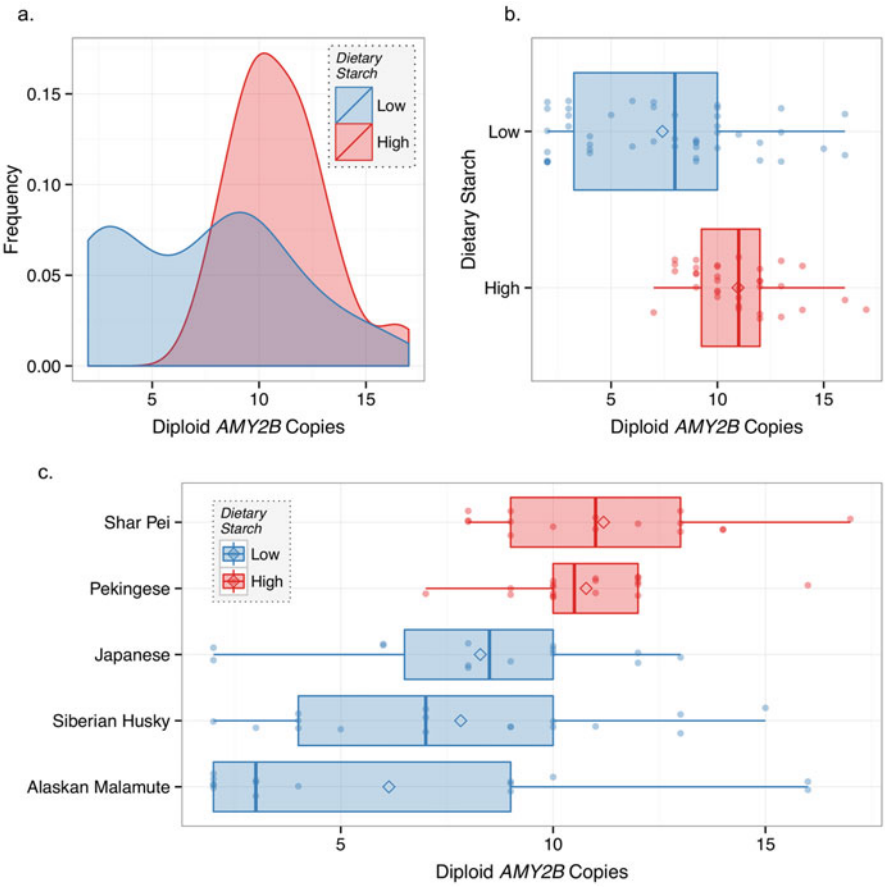

Fig. 6 Diet and AMY2B copy number variation. (**a**) Density plot of ddPCR diploid AMY2B copy number for dogs that traditionally consumed high-starch diets and low-starch diets. Density reflects frequency with which a given diploid copy number appears in each population. (**b**) Tukey boxplot of diploid AMY2B copy number for dogs that traditionally consumed high-starch diets and low-starch diets. (**c**) Tukey boxplot of diploid AMY2B copy number for specific dog breeds that traditionally consumed high-starch diets and low-starch diets. Originally published in Reiter et al. (2016)

7.3 Functional Polymorphisms Exhibiting Fixed Alternative Alleles in Dogs and Wolves

Another study investigating genomic regions targeted by selection during dog domestication was described by Cagan and Blass (2016). Their approach leveraged searching a comprehensive canine polymorphism database to identify polymorphic markers that are highly differentiated between wolves and dogs. Their approach led to the identification of 11 genes for which functional variants are fixed for alternative alleles in dogs and wolves. A pathway analysis of the genomic regions

containing the polymorphic markers with $F_{ST} > 0.75$ identified "adrenaline and noradrenaline biosynthesis pathway," "axon guidance mediated by netrin," "dopamine receptor-mediated pathway," "nicotine pharmacodynamics pathway," "alpha adrenergic pathway," and "gonadotropin-releasing hormone receptor pathway." The authors point out that each pathway was represented by multiple genes. Furthermore, computational analysis suggested that within each of the pathways, there are genes with putatively functional variants (Cagan and Blass 2016).

The authors state that domestication of dogs likely selected for reduced fight or fight responses, which are, in part, mediated by pathways such as "adrenaline and noradrenaline biosynthesis pathway" (nine genes with potentially functional variants), "dopamine receptor-mediated signaling pathway" (eight potential functional variant genes), "alpha adrenergic receptor signaling pathway" (five potential functional variant genes). The identification of neuro-related pathways further lends support to the idea that behavioral phenotypes were selected during the initial phase of dog domestication when wolves and dogs first began diverging (Cagan and Blass 2016).

8 Genetic Structure of Dog Breeds

After dog domestication, the next most frequently pondered questions about dogs are: "*How were the different breeds created*?" and "*What components of the genome are responsible for the morphological phenotypes that define these breeds?*" Answers to these questions lie at the heart of many population genetics/genomics studies carried out on dogs.

8.1 Dog Genome Sequence and Genetic Diversity

Studies revealing the sequence of the dog genome and canine genetic variation have provided considerable information about the population structure of purebred dogs and the relationship between different breeds. The dog genome sequence, derived from a female Boxer, was published in 2005 (Lindblad-Toh et al. 2005). The Boxer was selected due to the decreased heterozygosity within the breed and an expected easier genome assembly process than would be expected for a dog with much greater heterozygosity (Lindblad-Toh et al. 2005).

The genome was sequenced with the whole genome shotgun approach resulting in over 31 million sequence reads corresponding to 7.5× coverage of the ~2.4 billion base pair genome. The assembly was anchored to dog chromosomes with data derived from previously constructed cytogenetic and radiation hybrid maps. The resulting genome sequence enabled the identification of an initial set of 19,300 protein-coding genes. An analysis of 13,816 1:1:1 orthologs between human, mouse, and dog provided lineage-specific data on synonymous (K_S) and

non-synonymous (K_A) changes. This allowed the investigators to calculate the K_A/K_S ratio, which provides a measure of the strength of selection acting on protein coding genes. As part of their analysis, the authors determined the median K_A/K_S ratios and discovered that the ratio differed substantially across each of the lineages. Their results placed the K_A/K_S ratio within the dog lineage in between the mouse and human lineages. The authors relate this finding to the population genetics theory that associates strength of purifying selection with increased effective population size. Their results are consistent with this theory as smaller mammals (such as mouse) tend to have larger effective population sizes (Lindblad-Toh et al. 2005).

To better understand canine genetic diversity, three distinct SNP datasets were analyzed. Lindblad-Toh et al. (2005) identified a total of 770,000 SNPs within the Boxer genome. The authors also compared a previously assembled 1.5× coverage draft sequence of the poodle genome (Kirkness et al. 2003) to their sequence of the Boxer. The comparison identified 1,460,000 SNPs between the two dog breeds (Kirkness et al. 2003). Additionally, Lindblad-Toh et al. (2005) generated shotgun sequencing data from 9 diverse dog breeds, 4 gray wolves, and 1 coyote using 22,000 sequencing reads from each that resulted in a set of 440,000 SNPs. A 1,283 subset of these SNPs were validated by resequencing which indicated a true positive rate of 96% (Lindblad-Toh et al. 2005).

8.2 *Single Nucleotide Polymorphisms in the Dog Genome and Inference of Bottleneck Events*

A comprehensive SNP map was constructed from the above three SNP datasets resulting in a final SNP map of more than 2.5 million SNPs. On average, any two dogs will have a single nucleotide polymorphism within approximately every thousand base pairs between members of different breeds, while members of the same breed will have a SNP within 1,500 bp of their genomes. According to their analysis, the gray wolf (1/580 bp) and the coyote (1/420 bp) exhibit greater genetic variation than the Boxer. Within the Boxer assembly itself, a SNP occurs within roughly every 3,000 bp. Based on their identification and analysis of SNPs, the authors conclude that a set of 10,000 SNPs is sufficient for genetic association studies in dogs (Lindblad-Toh et al. 2005).

As part of their analysis, Lindblad-Toh et al. (2005) modeled the population history of the domestic dog. Specifically, they built a mathematical model in which a dog population experienced both an ancient and a recent bottleneck. The results of their coalescent method fit well with their genetic data when they set the ancient bottleneck to 9,000 generations ago (27,000 years ago), with a population size of 13,000 and an inbreeding coefficient of $F = 0.12$, and to the more recent breed-creation bottleneck 30–90 generations ago (90–270 years ago). The authors also used the modeling approach to generate estimates of breed-specific bottlenecks that were consistent with known histories of the breeds. The breed that exhibited the

poorest fit to the two bottleneck model was the Akita which was created in Japan about 450 generations ago and then underwent a subsequent bottleneck in the 1940s when it was introduced into the United States (Lindblad-Toh et al. 2005).

8.3 Number of Dog Breeds

Worldwide, there are over 400 recognized dog breeds. The American Kennel Club recognizes 192 dog breeds as of 2018, and each year one million dogs are registered by the AKC with over half of all annual AKC dog registrations corresponding to just 10 breeds.

8.4 Microsatellite Analysis of the Genetic Structure of 85 Dog Breeds

An initial analysis of 85 dog breeds (genotyping 5 unrelated dogs from each breed) was conducted using 96 microsatellite markers that spanned the canine genome with an average density of approximately 30 Mb (Parker and Ostrander 2005). The results indicated that a purebred dog could be assigned to its breed of origin 99% of the time. During the analysis, it was discovered that almost 40% of all genetic variation occurring in dogs is detectable when comparing dogs across breeds, for example, when comparing a Great Dane to a Chihuahua versus comparing one Chihuahua to another Chihuahua. This is considerably greater than what has been observed in humans, where just 5–10% of all human genetic variation occurs between populations and races. The genotyping data was used to cluster the 85 breeds based on genetic similarity. Although most breeds mapped cleanly to a single cluster, some breeds such as Australian Shepherd, Bichon Frise, Flat-Coated Retriever, Great Dane, Lhasa Apso, and Pug mapped to more than one cluster (Fig. 7) (Parker and Ostrander 2005).

8.5 Genetic Diversity Differences in Dog Breeds

Quignon et al. (2007) assessed the extent of genetic diversity inherent in Bernese Mountain Dogs (BMD), Flat-Coated Retrievers (FCR), Golden Retrievers (GR), and Rottweilers (ROT) sampled in equal proportions from the United States and Europe. The goal of the study was to better understand how genetic variation within dogs of the same breed varies by geographic location. Genetic studies in dogs can be confounded by population stratification resulting in false-positive associations when population substructure exists within a breed. This can be particularly problematic when studies are designed assuming that all dogs within a breed share the same level

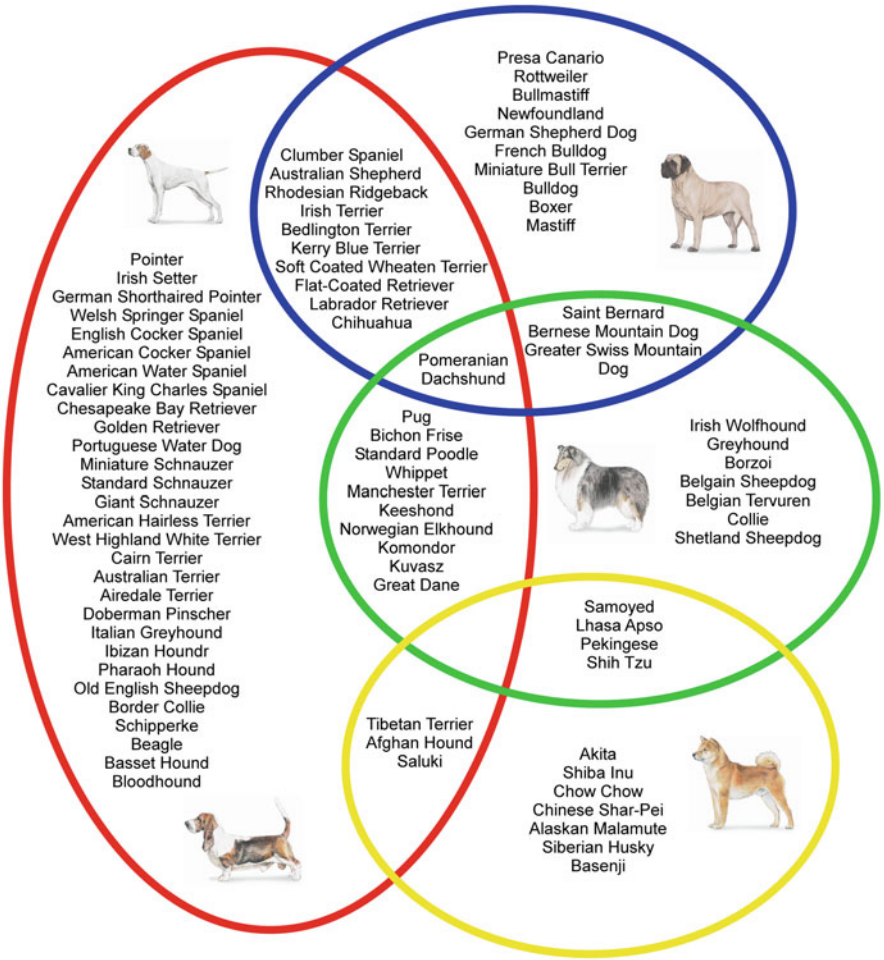

Fig. 7 The population structure of 85 dog breeds. The dataset includes five unrelated dogs from each of the 85 breeds that have been genotyped using 96 (CA)n repeat-based microsatellites that spanned the dog genome at an average density of 30 Mb. Clusters were obtained using the computer program Structure, which implements a Bayesian model-based clustering algorithm that attempts to identify genetically distinct subpopulations based on patterns of allele frequencies. Four distinct clusters described by Parker et al. are depicted as colored circles: cluster one is yellow, cluster two is blue, cluster three is green, and cluster four is red. Breeds associated with each cluster are listed within the appropriate circle, and examples of breeds are shown in the pictures. Some breeds show similarity to more than one cluster and are listed in the overlapping space. Originally published in Parker and Ostrander (2005)

and type of genetic variation. A set of 722 SNPs from four loci on chromosome 1 was genotyped in 120 dogs (Quignon et al. 2007). The investigators determined that the GR exhibited the greatest number of polymorphic SNPs (66.6%), while the fewest polymorphic SNPs were detected in the BMD. The FCR had 57.7% polymorphic SNPs, and the ROT had 54.4% polymorphic SNPs (Quignon et al. 2007).

The finding that dog breeds are not homogenous populations underscores the importance of population substructure when considering case-control genetic association studies. The authors state that variation in allele frequencies can arise through a population's genetic history, ancestral geographical distributions, mating practices, and both reproductive expansions and bottlenecks. Moreover, Quignon et al. indicate that besides population stratification arising from variation in geographical origin, artificial selection during breeding for phenotypic traits such as coat color, herding, hunting, olfactory capabilities, memory, and cognitive ability can also result in undetected population structure when those breeds are used in genetic studies (Quignon et al. 2007). This study highlights the fact that although members of a dog breed may share similar physical traits, each dog is genetically a unique individual.

8.6 Genome-Wide Genetic Structure and Evolution of Dogs Versus Wolves

Vonholdt et al. (2010) carried out a genome-wide analysis of 48,000 SNPs in 912 dogs (representing 85 breeds) and 225 gray wolves (across 11 globally distributed populations). The goal of the study was to gain a better understanding of the evolutionary and geographical history that gave rise to the dramatic diversification of phenotypes observed in dogs today. The authors used Bayesian clustering methods to identify any dog breeds that may have evidence of admixture with wolves. A relatively small set of breeds, considered ancient dog breeds, were identified and include breeds such as Afghan Hound, Akita, Alaskan Malamute, Basenji, Chinese Shar Pei, Chow Chow, Dingo, and Siberian Husky to name a few. Based on historical information, these ancient dog breeds have origins dating back more than 500 years ago (Vonholdt et al. 2010).

To determine the main contribution of genetic diversity in domestic dogs, Vonholdt et al. (2010) considered whether a single wolf population clustered with dogs in neighbor-joining trees by taking into account allele sharing of individual SNPs, 5-SNP haplotypes, and longer multi-SNP haplotypes for individuals and breed groupings. Their results indicated that only for individual SNPs and 5-SNP haplotypes Middle and Near Eastern gray wolves clustered with dogs. Moreover, in this analysis all other wolves clustered together as a single genetic entity separate from dogs. Then they tested whether haplotypes sharing of modern and ancient dog breeds could be associated with any distinct wolf populations. For this analysis, North American wolves were used as a negative control based on existing models of dog domestication excluding North America as the center of dog domestication (Vonholdt et al. 2010).

The results demonstrated that the extent of shared haplotypes between dogs and North American wolves was lower than sharing between dogs and Old-World wolves. More importantly, they discovered that for 5-SNP haplotypes, sharing was greater between Middle Eastern wolves and modern dog breeds than between other

populations of wolves. For longer multi-SNP haplotypes, the authors report that most breeds exhibit the greatest haplotype sharing with Middle Eastern wolves, including geographically diverse breeds such as Basenji, Bassett Hound, Borzoi, and Chihuahua. In separate analysis, the Akita, Chinese Shar Pei, Chow Chow, and Dingo shared most strongly with Chinese wolves. Finally, the authors note that for the 5-SNP and longer multi-SNP haplotype analyses, the Basenji shared the most haplotypes with Middle Eastern gray wolves than any other domestic dogs. It is worthwhile to point out that Basenjis are a dog breed having a Middle Eastern origin. The authors interpret this result as indicating that the Basenji had a large effective population size early in domestication or, alternatively, they have been recently backcrossed with wolves. Taken together, the authors conclude that the Middle East is the main source of genetic diversity in dogs with possible minor contributions derived from Europe and Asia (Vonholdt et al. 2010).

Vaysse et al. (2011) described a comprehensive high-density genotyping analysis of genomic regions exhibiting evidence of selection in 509 dogs across 46 diverse breeds and 15 wolves using 170,000 evenly spaced SNPs. Evolutionary relationships between the sampled subjects were assessed by building a neighbor-joining tree from the genetic distances in the comprehensive genotyped dataset (Fig. 8). Visualizing this tree led to the following conclusions: (1) dogs from the same breed clustered together as is expected from closed gene pool breeding groups, (2) relatively no structure is present within the breeds which is consistent with modern dog breeds arising from a common set of ancestors rather quickly, and (3) the internal branches for Boxer and wolf are longer than those for other breeds which make sense because SNP discovery occurred using genomic sequence data from the Boxer genome and the longer wolf branches likely represent greater evolutionary distance compared to the other dog breeds (Vaysse et al. 2011).

9 Genomic Basis for Morphological Variation Between Dog Breeds

Although dog domestication began at least 15,000 years ago, it wasn't until the Victorian era, roughly 200 years ago, that artificial selection for breed standards in dogs first began. The phenotypes observed in the breeds of today represent extremes of morphological variation (Fig. 9) (Shearin and Ostrander 2010).

Phenotypic variation across breeds is the consequence of a variety of physical traits associated with numerous anatomical regions. Variation in skeletal morphology is associated with differences in body size, leg size, and skull shape between breeds. Tremendous variation in hair phenotypes gives rise to differences in coat texture, length, and color within different breeds (Fig. 9).

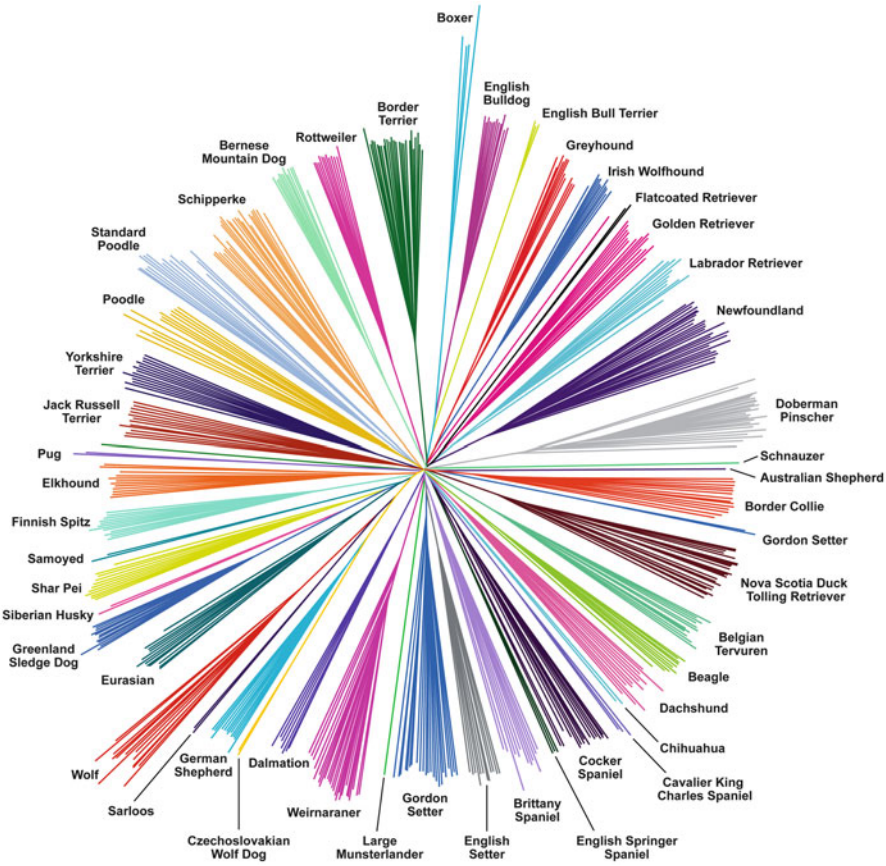

Fig. 8 Neighbor-joining tree constructed from raw genetic distances representing relationships between samples. 170,000 SNPs were genotyped in 46 diverse dog breeds plus wolves using the Canine HD array. The Boxer branches are longer, which likely represent the influence of ascertainment bias, as the SNPs were discovered from sequence alignments involving the Boxer reference sequence. Originally published in Vaysse et al. (2011)

9.1 Head Phenotype

Brachycephaly is a phenotype resulting in a dramatic decrease in muzzle length accompanied by decreased length of the related bones (Fig. 10). Additionally, brachycephalic dog breeds, such as the Boxer, Bulldog, French Bulldog, and Pekingese, have slightly shortened and widened skulls.

An "across-breed" study was designed to investigate the genetic basis of the brachycephalic phenotype. This genome-wide association study design required control breeds lacking the brachycephalic phenotype and included dolichocephalic (long muzzle) and mesaticephalic (intermediate muzzle length) breeds. The dolichocephalic and mesaticephalic breeds included Akitas, Belgian Tervurens, Black

Fig. 9 Morphological variation in the dog. Dog breeds display extremes of morphological variation including body size and proportion, head size and shape, coat texture, color, and patterning. Clockwise from the left: the Bloodhound, the Chinese crested, the Dandie Dinmont Terrier, the Scottish Deerhound, the long-haired Chihuahua, and the French Bulldog (Original Image: Mary Bloom, American Kennel Club). Originally published in Shearin and Ostrander (2010)

Russian Terriers, Bloodhounds, Dalmatians, German Shepherds, and Great Danes. Bannasch et al. (2010) identified the location of the dog genomic region responsible for the brachycephalic phenotype using an across-breed genome-wide association approach. Using the Affymetrix Version 2 Custom Canine SNP arrays to generate genotype calls, the authors successfully identified a brachycephalic head locus that mapped to a region of chromosome 1 between 59.5 and 59.8 Mb (Bannasch et al. 2010). To more clearly delineate the region of association, the investigators used 88 affected dogs and 185 unaffected dogs to genotype a set of 49 SNPs overlapping the most significantly associated region of the originally identified interval. The results of this genotyping revealed a smaller 31 kb genomic interval that overlapped a homozygous haplotype encoding a single gene, THBS2 within brachycephalic breeds (Bannasch et al. 2010).

Schoenebeck et al. (2012) searched for additional genes modulating the multigenic phenotype and cranioskeletal features differentiating dolichocephalic skulls from brachycephalic skulls. In order to more completely characterize the anatomical and geometric differences associated with phenotypic variation in canine skull shape, the authors digitally captured 51 stereotyped anatomical landmarks from 533 skulls obtained from museums representing 120 breeds and 4 gray

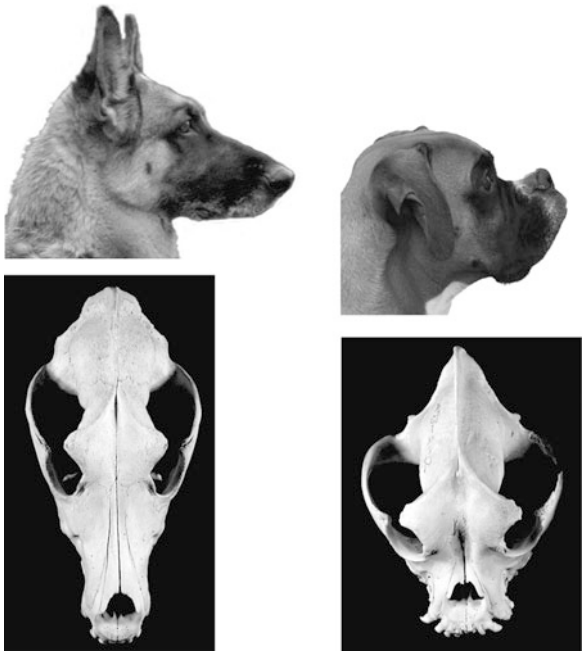

Fig. 10 Brachycephaly in dogs. Comparison of photographs (Photos Mary Bloom, courtesy of AKC) and skulls from a German Shepherd dog with a wild-type skull shape (non-brachycephalic) and a brachycephalic Boxer. Originally published in Bannasch et al. (2010)

wolf subspecies. The variance captured in Principal Component 1 (PC1) (59% variation) corresponds to anatomical differences in rostrum length and angle, palate and zygomatic arch width, and depth of neurocranium, which comprise the cranioskeletal features giving rise to either a brachycephalic or dolichocephalic skull phenotype (Fig. 11).

Schoenebeck et al. (2012) used one set of breed samples for phenotypic measurements (the museum specimens) and another set of breed samples (the DNA samples) for genotyping because purebred dogs conform to a specific morphological standard that is shared among members of the breed. Morphological phenotypes, such as skull shape, are uniformly constrained by the breed. Strong genotype associations were found with PC1 (i.e., variations in skull morphology differentiating brachycephalic skull phenotype from dolichocephalic skull phenotype) associated with polymorphic markers located at specific locations within domestic dog, *Canis familiaris*, chromosomes (denoted CFA): CFA5.36476657, CFA24.26359293, CFA30.35656568, and CFA32.8384767. Some additional markers were weakly implicated on CFA9, CFA13, and CFA30 and another one on CFAX (Schoenebeck et al. 2012).

Schoenebeck et al. (2012) reasoned that skull shape variation is a consequence of artificial selection, and therefore they expected the major loci to exhibit reduced observed heterozygosity (H_o) and elevated genetic differentiation (F_{ST}), both of which are strong indicators of selective sweeps. The CFA32 quantitative trait locus (QTL) was selected as a major focus because it was in the top 2 most associated non-allometric loci that showed strong evidence of selection. The shared

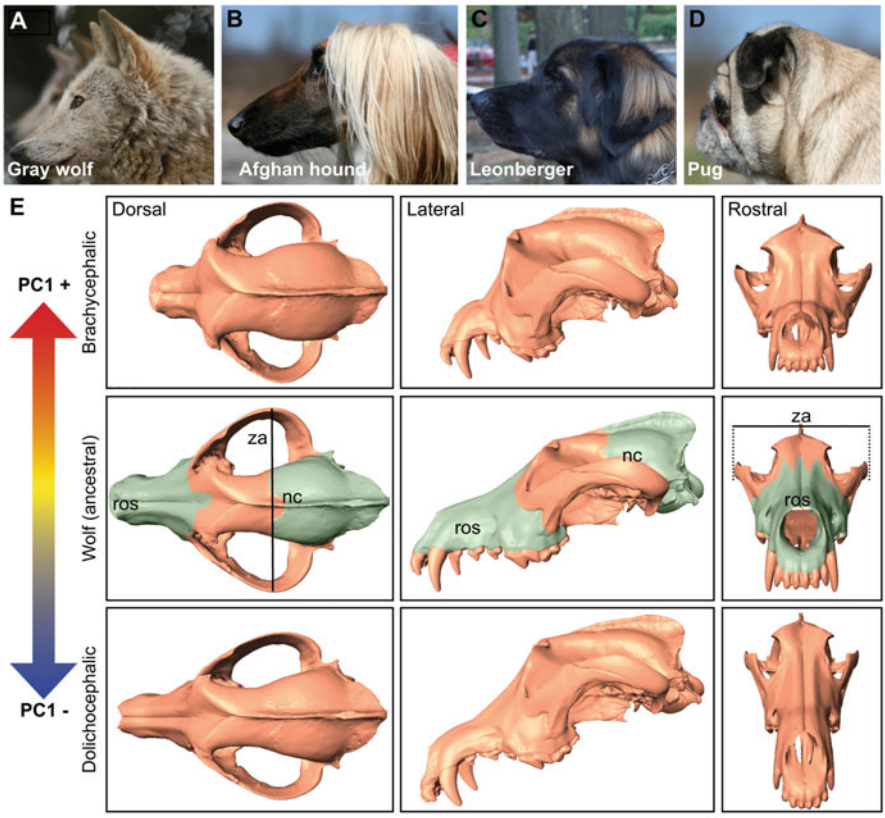

Fig. 11 Quantitative and qualitative assessments of PC1 on canine cranioskeletal shape. (**a**) Gray wolf (mesocephalic, ancestor to dogs) (**b**) Afghan hound (dolichocephalic), (**c**) Leonberger (mesocephalic), (**d**) Pug (brachycephalic). (**e**) Surface scans of a gray wolf skull illustrate morphological changes associated with PC1. Columns (left to right) are dorsal, lateral, and rostral views. Top row: a gray wolf skull morphed by positive PC1. Middle row: a gray wolf skull (no morphing). Bottom row: a gray wolf skull morphed by negative PC1. Pseudocoloring of the gray wolf skull indicates rostrum (ros) and neurocranium (nc). Line indicates width of the zygomatic arches (za). Originally published in Schoenebeck et al. (2012)

haplotypes for CFA32 QTL among six of the seven most brachycephalic breeds (Boston Terrier, Bulldog, Brussels Griffon, French Bulldog, Pekingese, and Pug) defined a 190 kb genomic region in between 8.15 and 8.34 Mb, within which two genes (*PRKG2* and *BMP3*) were located (Schoenebeck et al. 2012).

In order to ascertain genotype-phenotype association within this interval, Schoenebeck et al. (2012) performed whole genome sequence survey from 11 dogs of diverse skull phenotype (including the brachycephalic breeds of Bulldog and Pekingese breeds). The authors identified the SNP at position 8,196,098 that causes a missense mutation in *BMP3* in which a phenylalanine is changed into a leucine (F452L mutation). The substitution of leucine in place of phenylalanine was

predicted to be disruptive to the *BMP3* functional structure. Upon comprehensive genotyping of 842 dogs across 113 breeds, it was found that the F452L mutation is almost always fixed in brachycephalic breeds. This suggests that the missense polymorphism may be the underlying cause for the brachycephalic phenotype (Schoenebeck et al. 2012).

9.2 Genomic Basis of Breed-Associated Morphological Traits

A fundamental question in canine genomics is "*What genomic mechanism enabled selective breeding to produce the tremendous diversity of morphological phenotypes observed in present day dog breeds*?" Boyko et al. (2010) addressed this question by genome-wide scans of SNP variation and genome-wide association mapping of morphological traits using 60,968 SNP genotypes of 915 dogs covering 80 domestic dog breeds coupled with 83 wild canids and 10 outbred African shelter dogs. The genotype map was combined with external measurements using breed standards, museum specimens, and individual dogs to identify genomic regions associated with breed-specific phenotypic variation among 57 morphological traits. One of the purposes of the study was to assess whether most breed-associated phenotypic variation is the consequence of large-effect QTLs or whether most of the observed phenotypic differences arise via the action of many QTLs of relatively weak effects. The answer to this question will provide a better understanding of how domestication and artificial selection have impacted the dog's genome (Boyko et al. 2010).

Boyko et al. (2010) performed a genome-wide scan to detect signatures of recent selection and allele sharing between dog breeds. Because the data supports the idea that relatively little sharing of IBD segments occurs among individuals from different breeds, it is reasonable to expect that when coincident sharing occurs between breeds with a similar phenotype, the shared genomic segments are likely encoding the genetic variation for that trait. The top 11 most extreme F_{ST} regions of the dog genome contained SNPs with $F_{ST} \geq 0.57$ and having a minor allele frequency (MAF) ≥ 0.15 (Boyko et al. 2010). Among the 11 regions detected with high F_{ST}, 6 are tightly linked to genetic variation known to affect canine morphological phenotypes. For example, the 167 bp insertion in *RSPO2* was associated with the fur growth and texture phenotype; the IGF1 haplotype was associated with small body size; an inserted retrogene (*Fgf4*) was associated with short limb length; and three genes modulating coat color phenotypes in dogs were also associated with the identified intervals *ASIP*, *MC1R*, and *MITF*. Additional regions with high F_{ST} were identified: CFA10.11465975 (associated with body weight) and CFA1.97045173 (associated with muzzle length) (Boyko et al. 2010).

Boyko et al. (2010) performed the genome-wide association scans by measuring 55 morphological parameters in order to identify genotype-phenotype associations, especially morphological traits that vary between dog breeds. Additionally, the authors included genomic regions contributing to variation in body size (variation is greatest across dog breeds than any other terrestrial species) as well as ear

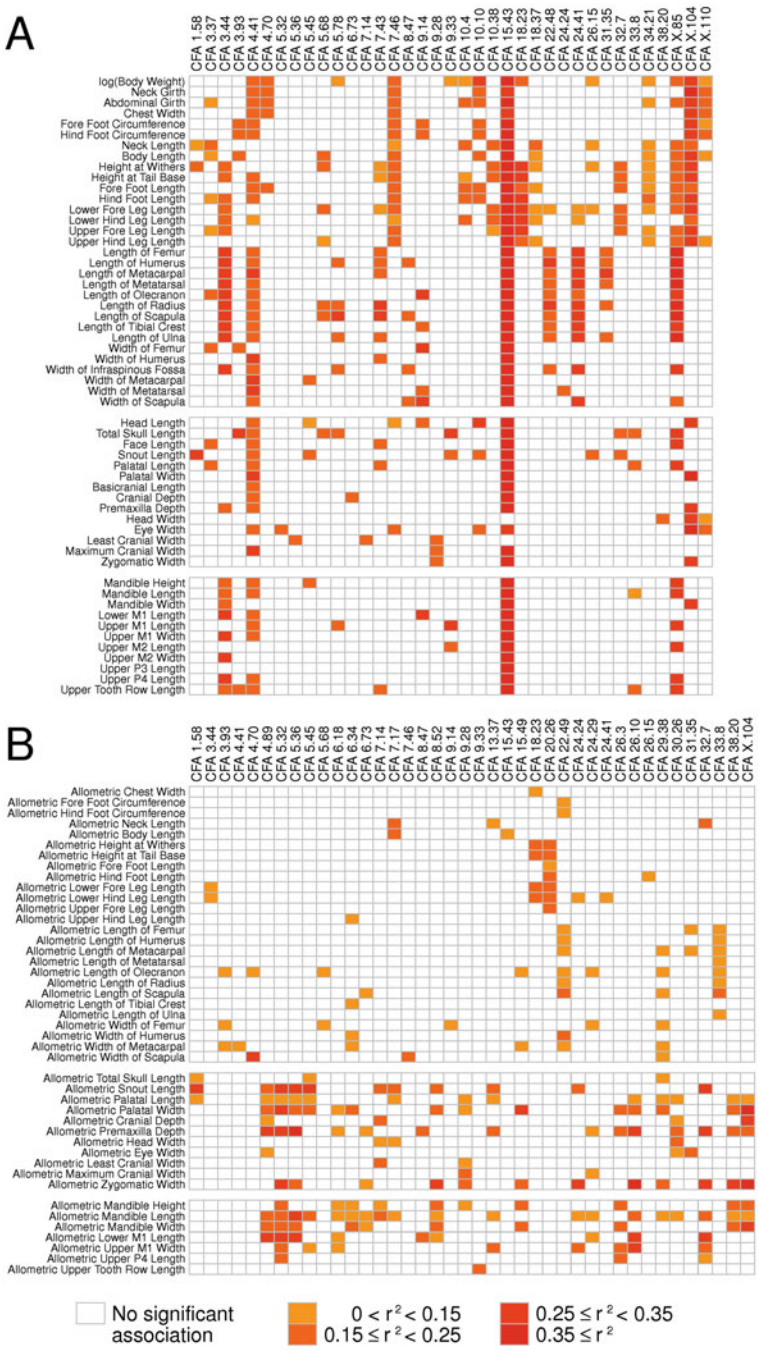

Fig. 12 Summary of associations across genomic regions for multiple traits. Each row corresponds to a trait [either (**a**) absolute or (**b**) proportional], and each column corresponds to a genomic region that has been found associated with at least one trait. The shading of each rectangle shows the R^2

floppiness. The genomic scan for body size [where body size = log (body weight)] resulted in the identification of multiple significant associations. The six strongest signals occurred at CFA15.44226659, CFAX.106866624, CFA10.11440860, CFAX.86813164, CFA4.42351982, and CFA7.46842856. Interestingly, the first four signals identified in the body size variation scan correspond to some of the highest F_{ST} values identified in the genome, along with CFA4 which has an $F_{ST} = 0.46$, consistent with diversifying selection among breeds for body size. Interestingly, in all six regions, wolves are not highly polymorphic (MAF < 0.1), and except for the CFA10 signal, the derived allele is at highest frequency in small breeds (Boyko et al. 2010).

Another trait that exhibits considerable variation across breeds is ear type. All adult wild canids have erect ears, yet dog breeds are fixed for a variety of ear positions including floppy ears. This juvenile type trait is retained by adults of certain breeds in a variety of domesticated mammals, such as dogs, cattle, goats, and rabbits. SNPs associated with breeds fixed for erect or floppy ears were identified and shown to be associated with a single interval on CFA10 that may underlie the ear position phenotype (Boyko et al. 2010). A third trait of interest in the Boyko et al. (2010) study was muzzle length, which varies tremendously across dog breeds. Similar to floppy ears, short snout is another paedomorphic trait. The strongest association signals were CFA1.59832965 and CF5.32359028, having F_{ST} values of 0.55 and 0.42, respectively. These polymorphisms are only found in brachycephalic breeds at high allele frequency (Boyko et al. 2010).

Boyko et al. (2010) constructed a multi-SNP predictive model for each trait. For the models of body weight, ear type, as well as most of the measured traits, the majority of the breed-associated variance was observed in fewer than four loci (Fig. 12). Correlated traits, such as femur length and humerus length, exhibited similar SNP associations. For the set of 55 measured traits, the average proportion of variance explained by the top 1, 2, and 3 SNP models was $R^2 = 0.52$, 0.63, and 0.67, respectively. The authors made the case that, after controlling for body size, mean proportion of variance explained by the models was still considerable, with $R^2 = 0.21$, 0.32, and 0.4, respectively. It is worth mentioning that the most significant genomic regions were similar even using naïve association scans that did not control for population structure. In terms of breed mapping, relatively little population structure was shared among the breeds. Subsequently, whatever portion of the population structure, which might have been shared among the breeds, was small enough to avoid biasing the association inferences (Boyko et al. 2010).

Boyko et al. (2010) state, for the majority of traits investigated, that a few QTLs of large effect determined the phenotype differences between breeds. These QTLs mapped to specific locations on *Canis familiaris* chromosomes (CFA). As an

Fig. 12 (continued) statistic of the single marker model for the trait for all significant associations ($p < 5.0e{-}5$ for absolute external traits, $p < 1.0e{-}4$ for skeletal and proportional traits after correcting for population structure). When multiple SNPs in the region are significant, the largest value of the R^2 statistics is reported. Originally published in Boyko et al. (2010)

example, they site the proportional height at withers for which they identified a large-effect QTL on chromosome CFA18, where they had previously determined a *fgf4* retrogene that confers the phenotype associated with the chondrodysplasia disproportional dwarfism in Basset Hounds, Corgis, and Dachshunds. Similarly, skull shapes were largely determined by genomic regions on CFA1, CFA5, CFA26, and CFA32, along with CFAX.105274087–106866624 region (also associated with body size). Most of these regions were also associated with dental phenotypes along with a strong association on CAF16. It seems that the relationship between phenotypes and associated genomic intervals in domestic dog breeds can be best described as a set of related phenotypes under the direct control of a few genomic regions (Boyko et al. 2010).

10 Genes, Mutations, and Genomic Regions Contributing to Clinically Relevant Phenotypes (Disease Conditions) in Dog Breeds

In the past few years, new advances have been achieved using genome-wide association studies (GWAS) and high throughput sequencing to unveil novel mutations in dog populations associated with clinically relevant phenotypes. These phenotypes span numerous organs, cell types, and body systems. Some interesting examples across a variety of body systems and dog breeds are described below.

10.1 Cardiovascular

Cardiovascular disease affects different dog breeds including the Newfoundland, Whippet, and Doberman Pinscher. Mitral valve degeneration is the most prevalent type of heart disease in dogs and is acquired during aging as degenerative lesions accumulate on the mitral valve. Over time, these lesions result in abnormal valve morphology and function. In severe cases, the mitral valve may prolapse and cause undesirable phenotypes, such as mitral regurgitation and left-sided congestive heart failure.

Stern et al. (2015) used the 170,000 canine high-density (HD) genotyping SNP chip and identified a region in the vicinity of position 57,770,326 on canine chromosome 15, which is near the interval of 58,506,916 and 60,140,841 that was also associated with mitral valve disease compared to normal dogs lacking evidence of mitral valve disease. Within this region is follistatin-related protein 5 precursor as well as some other genes including neuropeptide Y receptors. A region on chromosome 2 also exhibited partial evidence of association peaking at 37,628,875 which is in proximity to rho GTPase-activating protein 26. In the discussion, the authors implicate follistatin-related protein 5 with another gene (*WFIKKN2*) that is involved

in metalloproteinase inhibition activity. Because metalloproteinase activity has been considered a part of the mitral valve disease pathophysiological mechanisms, these two genes represent viable candidates for the undesirable clinical trait of mitral valve disease (Stern et al. 2015).

The Doberman Pinscher is one of the most commonly reported canine breeds with familial dilated cardiomyopathy, which has been linked to congestive heart failure and sudden cardiac death. Meurs et al. (2012) performed a GWAS using a commercial "Canine Genome Array" containing 49,663 SNP markers and identified a locus on CFA14 (Meurs et al. 2012). Fine-mapping of additional SNPs localized a potential haplotype at 23,774,190–23,781,919 region from the same chromosome. DNA sequencing identified a 16 bp deletion in the 5′ donor splice site of intron 10 from the gene encoding the mitochondrial pyruvate dehydrogenase kinase 4 (*PDK4*) in affected dogs. The authors next demonstrated that *PDK4* transcripts derived from the homozygous deletion genotype exhibit decreased expression of exons 10 and 11. This study tested 232 animals, with 66 affected and 66 unaffected Doberman Pinschers, plus 100 healthy dogs from 11 other breeds. The target mutation was identified in 54 out of 66 affected dogs (82%, with 45 heterozygotes and 9 homozygotes) and 26 out of 66 of unaffected dogs (39%, with 18 heterozygotes and 8 homozygotes). Some of the 100 unaffected dogs, representing 11 other breeds, appeared to show the mutated allele as well. Electron microscopy of myocardium from affected dogs demonstrated several mitochondrial disorganization features, suggesting a dysfunction of *PDK4* enzyme due to the mutation (Meurs et al. 2012). The fact that the presence of an associated allele may not always correlate with the associated phenotype underscores the complexity of genetics.

The Irish Wolfhound is another breed that is predisposed to cardiac disease, specifically dilated cardiomyopathy, with up to 20% of dogs in the breed exhibiting the undesirable clinical phenotype. Philipp et al. (2012) performed a genome-wide association study using 190 Irish Wolfhounds. Dilated cardiomyopathy phenotypes were diagnosed with echocardiographic exams. Control dogs were at least 7 years old with no signs of the dilated phenotype. The authors identified six loci corresponding to CFA1 at 123,630,555; CFA10 at 24,159,608 (*ARHGAP8* gene); CFA15 at 61,260,406 (*FSTL5* gene); CFA17 at 58,604,566; CFA21 at 40,670,543 (*PDE3B* gene); and CFA37 at position 31,801,266. The authors report that their associated regions overlapped with genes known to cause dilated cardiomyopathy in humans (Philipp et al. 2012). The human form of dilated cardiomyopathy is a cause for heart transplants, and in the absence of transplantation, chronic heart failure can occur. About half of human cases are inherited, and more than 60 genes have been implicated in the pathology (Toro et al. 2016).

In another example of cardiovascular phenotypes in dogs, Stern et al. (2014) used a pedigree analysis of 45 Newfoundlands, of which 9 exhibited a subvalvular aortic stenosis (SAS) phenotype. Twelve additional dogs in the pedigree displayed systolic heart murmur phenotypes along with either evidence of aortic insufficiency or a subvalvular ridge or both. When dogs with the aortic insufficiency and/or subvalvular ridge phenotypes were bred to normal dogs, offspring displayed undesirable

cardiac phenotypes. A genome-wide association study followed by genomic sequencing identified a mutation in the exonic region of the phosphatidylinositol-binding clathrin assembly protein gene (*PICALM*). Interestingly, *PICALM* is involved in morphogenesis of the heart. Stern et al. (2014) report that the phenotype is likely caused by a 3 bp exonic insertion in the *PICALM* (599K_600LinsL mutation) that was detected and associated with the development of SAS in that breed. Immunohistochemistry validated the presence of *PICALM* protein in the canine myocardium and area of the subvalvular ridge. Overall, 96.1% of the SAS-affected Newfoundland dogs displayed the codon insertion mutation (34.6% homozygous and 61.5% heterozygous), while only 26% of non-affected ones possessed the mutation (4.3% homozygous and 21.7% heterozygous). The authors state that none of 180 control dogs of 30 different breeds possessed the mutation in any form (Stern et al. 2014).

Following the report by Stern et al. in 2014, Drogemuller et al. (2015) provided evidence suggesting that the mutation reported by Stern et al. may not in fact be the causative allele associated with subvalvular aortic stenosis in Newfoundlands. Among the evidence presented, Drogemuller et al. (2015) question the experimental design that was used, pointing out that (a) the number of cases and controls used in the association study would not provide the expected power needed to identify a locus associated with a nondominant mode of inheritance (Drogemuller et al. 2015). Furthermore, Drogemuller et al. (2015) report a replication of portions of the original study and fail to reproduce the findings reported by Stern et al. (2014).

10.2 Endocrinology

An endocrine phenotype of clinical interest is obesity. Obesity and greater food motivation were found as a genetic predisposed disorder in Labrador retrievers (Raffan et al. 2016). The associated gene is pro-opiomelanocortin (*POMC*) that encodes a pro-protein which is cleaved into several bioactive peptides, including b-MSH (melanocyte-stimulating hormone) and b-endorphin. The associated genotype is a 14 bp deletion responsible for a frameshift after the glutamate at the position 188 (p.E188fs). It is predicted to disrupt the coding sequence of *POMC* and cause loss of production of b-MSH and b-endorphin which results in increased body weight with a mean effect size of 1.90 kg per deletion allele. Therefore, it indicates a dominant dosage effect trait. Adiposity and food motivation were polymorphism associated phenotypes in both Labrador Retrievers and the closely related Flat-Coat Retrievers (FCRs). The mutation is significantly more common in Labradors selected to become assistance dog breeding stock (allele frequency = 0.45) than those selected to be companions (allele frequency = 0.12) (Raffan et al. 2016). In humans, *POMC* mutations that produced aberrant forms of b-MSH reveal that this is an important hormone in controlling appetite and obesity development (Challis et al. 2002; Lee et al. 2006). Mice selectively lacking b-endorphin are hyperphagic and obese (Appleyard et al. 2003). Taken together, these findings suggest that the

loss of both neuropeptides in dogs carrying *POMC* p.E188fs could contribute to the observed obese phenotype.

10.3 Ophthalmology

Progressive retinal atrophy (PRA) is a group of inherited eye diseases characterized by retinal degeneration that culminates to blindness in dogs and is often described as the equivalent of retinitis pigmentosa (RP) in humans. It is noteworthy that PRA in dogs has been reported in over 100 dog breeds. Three studies, two in Golden Retrievers and one in Shelties, have uncovered three PRA-related genes. The first study in Golden Retrievers leveraged a genome-wide association study design to ultimately identify a frameshift mutation within the canine solute carrier gene *SLC4A3*. The undesirable allele was present in 56% of PRA dogs and exhibited recessive inheritance with 100% penetrance (Downs et al. 2011).

The second study in Golden Retrievers (GRs), used GWAS in 10 PRA cases and 16 controls, identifying an association of a 737 kb chromosome 8 (CFA8) locus containing six genes with a clinical ocular phenotype. Two of the genes (*TTC8* and *SPATA7*) have already been described as RP-associated in humans. *TTC8* encodes a protein that is a part of the BBSome complex which is responsible for ciliary membrane biogenesis. Affected dogs showed a single nucleotide deletion in *TTC8* exon 8. The frameshift mutation is predicted to cause a premature stop codon. In the investigated cohort, this genotype (*TTC8* c.669delA) is recessive, segregating correctly in 75.9% of the tested cases (22/29), whereas none of the PRA controls are homozygous for the mutation, only 3.5% carry the PRA-associated allele, and 96.5% are homozygous wild type (Downs et al. 2014).

Identifying genes associated with PRA provides a mechanism for developing breeding programs that can eventually remove these undesirable alleles from affected breeds. The pathophysiology and clinical progression of PRA have been well characterized within the Swedish Vallhund dogs by Cooper et al. (2014). A third study reported by Wiik et al. (2015) identified the *CNGA1* gene on CFA13 as a novel PRA-related locus using a genome-wide association approach with 15 Shetland Sheepdog (Sheltie) cases and 14 controls. *CNGA1* is also known to be involved in human RP. This gene encodes a protein involved in phototransduction, by forming cGMP-gated cation channel in the plasma membrane that allows depolarization of rod photoreceptors. Sequencing of this gene in affected Shelties identified a 4 bp deletion in exon 9 (c.1752_1755delAACT). Similar to the *TTC8* mutation in Golden Retrievers, *CGNA1* also alters the translation frame and generates a truncated protein caused by premature termination codon (Wiik et al. 2015).

Besides PRA, other ocular phenotypes affect dogs, such as glaucoma. Two metalloprotease genes *ADAMTS10* and *ADAMTS17* are implicated in primary open angle glaucoma (POAG) in dogs: the former in Beagle (Kuchtey et al. 2013) and Norwegian Elkhound breeds (Ahonen et al. 2014) and the latter in Basset Hound and Basset Fauve de Bretagne breeds (Oliver et al. 2015). Regarding the latter study, 226 Basset Hounds and 27 Basset Fauve de Bretagne dogs were provided an

ophthalmic examination and diagnosed for POAG. The affected Basset Hounds displayed homozygosity for a 19 bp deletion in *ADAMTS17* exon 2 that leads to a frameshift predicted to form a truncated protein. Fifty clinically unaffected Basset Hounds were genotyped for this mutation as either heterozygous or homozygous for the wild-type allele. The affected Basset Fauve de Bretagne dogs contained a nonsynonymous substitution in *ADAMTS17* exon 11 causing a glycine to serine amino acid exchange (G519S) in the disintegrin-like domain that might be related to protein dysfunction. Unaffected Basset Fauve de Bretagne dogs were either heterozygous for the mutation (5/24) or homozygous for the wild-type allele (19/24). Therefore, evidence suggests that both independent POAG-associated mutations are recessive in the two different breeds examined (Oliver et al. 2015).

10.4 Craniofacial

Wolf et al. (2015) described a mutation on the dog's chromosome 27, encoding a frameshift mutation within the *ADAMTS20* metallopeptidase gene (c.1360_1361delAA or p.Lys453Ilefs*3), that leads to a cleft lip with or without cleft palate (CL/P) phenotype in the Nova Scotia Duck Tolling Retriever (NSDTR). This undesirable phenotype exhibits a recessive mode of inheritance (Wolf et al. 2015). CL/P is the most commonly occurring craniofacial congenital disorder. Interestingly, the same study that found *ADAMTS20* as the CL/P-target gene in NSDTR dogs has also reported a suggestive association of the same gene to CL/P human cases in a family-based association analysis (DFAM) using a Guatemalan cohort composed of 25 CL/P phenotypes, 420 unaffected relatives, and 392 controls. In dogs, the mutation alters the reading frame and generates a premature stop codon within the metalloprotease domain of *ADAMTS20* protein. In humans it seems to be associated with the SNP rs10785430 within *ADAMTS20*, but further studies are required to assure whether it alters the protein function.

10.5 Dermatology

Canine atopic dermatitis (CAD) is a chronic inflammatory skin disease triggered by environmental allergens that react with epithelial and immune cells. GWAS and fine-mapping analyses revealed a 9-SNP-containing haplotype overlapping *PKP2* gene that predisposes German Shepherd dogs to CAD. PKP2 encodes plakophilin-2 protein, which is involved in the synthesis of desmosomes, a cell adhesion structure (Tengvall et al. 2016). The haplotype spans ~280 kb on chromosome 27 (CFA27) which encompasses a rare ~48 kb locus shared only with other high-risk CAD breeds. Transient transfections followed by luciferase reporter assays indicated that seven out of the nine CAD-associated SNPs within that haplotype appeared to have enhancer activity with allelic differences in either epithelial or immune cells. These cells include Madin-Darby canine epithelial cell line from Cocker Spaniel (MDCK),

human keratinocyte cell line (HaCaT), human T cell line (Jurkat), and human erythromyeloblastoid leukemia cell line (K562). A top SNP (CFA27:19,086,778) displayed high activity in keratinocytes with 11-fold induction of luciferase transcription by the risk allele (T/T) versus 8-fold by the control allele (C/C) ($p = 0.003$). It also mapped close (~3 kb) to an ENCODE skin-specific enhancer region. Those experiments suggest that GSDs' predisposition to CAD is associated with multiple variants combined in a risk haplotype that may contribute to an altered expression of the PKP2 gene in immune and epithelial cells (Tengvall et al. 2016).

10.6 Pigmentation

A recessive genotype, within the solute carrier family 45, member 2 gene (*SLC45A2*), is responsible for albinism in dogs. The *SLC45A2* protein is found in melanocytes, and, although its exact function is still being studied, it is likely to be involved in melanin synthesis. A large deletion (g.27,141_31,223del) in *SLC45A2* was associated with oculocutaneous albinism (OCA) in Doberman Pinschers (Fig. 13) that were homozygous for that mutation, whereas the albino Lhasa Apso showed homozygosity for a nonsynonymous substitution in the seventh exon of *SLC45A2* (c.1478G > A) that resulted in a switch from glycine to aspartate (p.G493D) (Wijesena and Schmutz 2015). This same study revealed that an albino Pekingese, two albino Pomeranians, and one albino mixed breed dog that was small and long-haired were also homozygous for the 493D allele. Colored offspring from those small long-haired albinos were heterozygous for this allele, clearly indicating that it is a recessive genetic trait. Structural bioinformatics investigation has predicted that the 11th transmembrane domain (where the 493rd amino acid is located) from the *SLC45A2* (p.G493D) protein has an altered structure, which might be deleterious for the proper protein function and, consequently, leads to the albino phenotype due to the lack of melanin production. However, an albino Pug was genotyped as homozygous for the 493G allele, indicating that although 493D allele is related to albinism in some small, long-haired dog breeds, it does not explain all albinism in dogs (Wijesena and Schmutz 2015).

10.7 Musculoskeletal

Mosher et al. (2007) identified the myostatin gene as the cause of increased muscle mass in Whippets. Interestingly, Whippets, like Greyhounds, are bred for racing. The Whippet is a small dog breed weighing approximately 9 kg. Within the population of race-bred Whippets, a "Bully Whippet" phenotype emerged in which heavily muscled Whippets were produced by breeders (Fig. 14). Although owners report that the Bully Whippets are healthy with some incidents of muscle cramping, they are never the less euthanized as they do not conform to the breed standard. The authors report that a total of 22 Whippets were sequenced across the

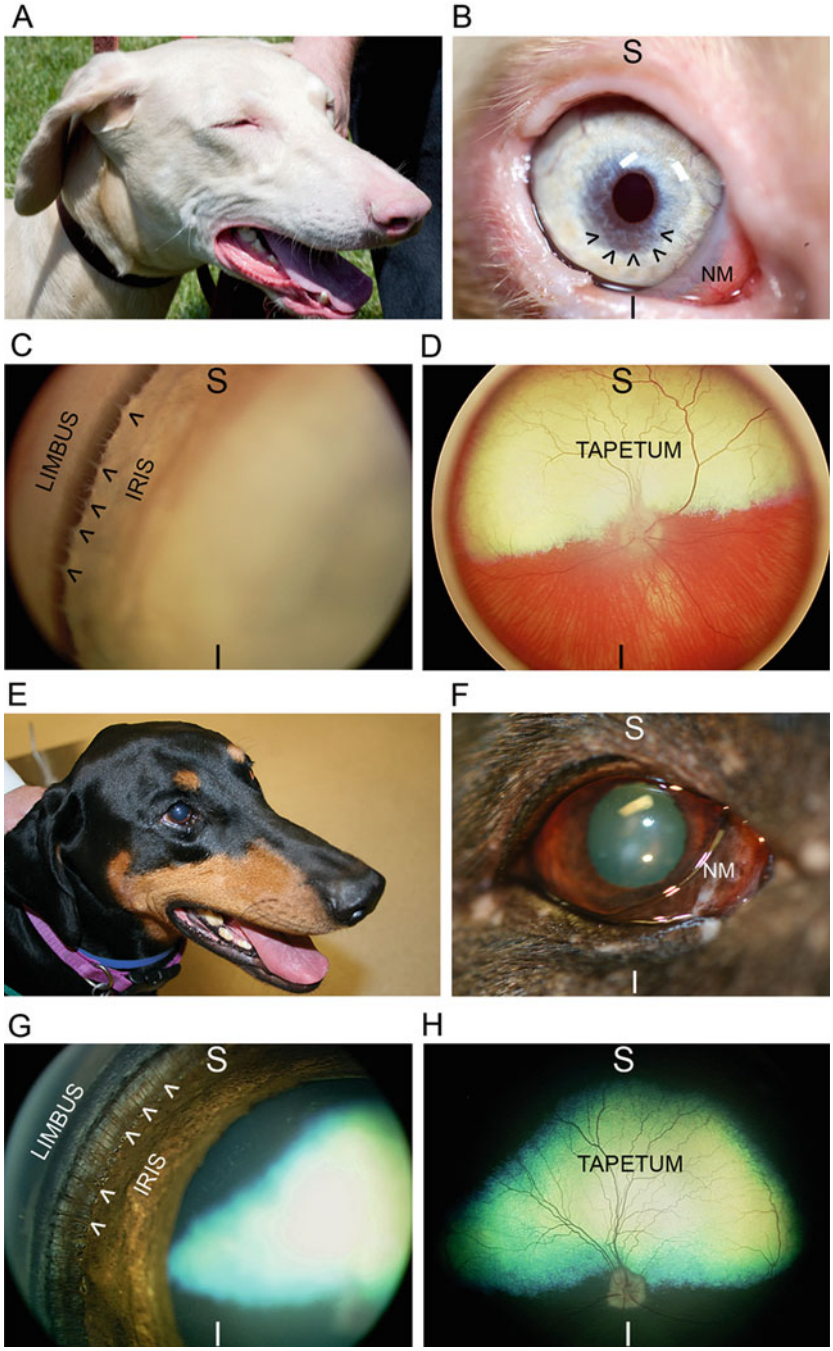

Fig. 13 Ocular phenotype of white Doberman Pinschers. Images taken from white Doberman Pinschers (top row) and black standard-color Doberman Pinscher (bottom row). An image of white Doberman Pinscher head (**a**) demonstrates lightly pigmented nose, lips, and eyelid margins

three exons and most of the introns in the myostatin gene. Among those sequenced, all four with the Bully Whippet phenotype were homozygous for a 2-bp deletion within the third exon that removes nucleotides 939 and 940 resulting in a premature stop codon. Of the five dogs that sired or whelped a Bully Whippet, all were heterozygous for the 2-bp deletion mutation. None of the remaining 13 Whippets, which all lacked the bully phenotype and had no familial history of the phenotype, carried the 2-bp deletion mutation (Mosher et al. 2007).

The authors determined that the bully phenotype displayed a simple autosomal mode of inheritance. Furthermore Mosher et al. (2007) provided statistical support for the idea that heterozygous Whippets contain, on average, 17% more mass per centimeter of height compared to homozygous wild-type Whippets (p-value = 0.00017). When the authors analyzed the genotypes of 85 racing dogs, for which racing results were available, an association between the mutation and racing performance was detected. Specifically, among dogs that were heterozygous for the mutation ($N = 12$), 66% were classified as top racers, while less than 17% of wild-type dogs received the same top ranking ($n = 72$). The Bully Whippets are too heavily muscled to perform well in races, while the heterozygotes exhibit ideal racing performance associated with lean muscle. The authors ultimately sequenced 15 different breeds and determined the haplotypes spanning the myostatin gene (Mosher et al. 2007).

10.8 Neoplasia

Cancers are genetically inherited diseases that occur in multiple species including dogs and humans. Identifying tumorigenesis-associated mutations is of great importance in veterinary medicine; dog's neoplasias are also valuable spontaneous

Fig. 13 (continued) compared with the same darkly pigmented structures in SDP (**e**). A closeup image of WDP eye (**b**) shows nonpigmented leading edge of the nictitating membrane (NM), tan-colored iris base transitioning to blue at pupillary margin, and oval-shaped dyscoric pupil aperture. The black arrowheads (in **b**) demarcate a region of significant iridal stromal thinning that was noted on examination to transilluminate (not shown in image) with retroillumination by light reflected from the tapetum lucidum. SDP eye (**f**) shows darkly pigmented margin of the nictitating membrane (NM) and brown iris with a round pupil aperture. WDP gonioscopy image (**c**), which allows visualization of structures lying within the iridocorneal angle (in images **c** and **g**, this region lies between the words "LIMBUS" and "IRIS"), shows that fibers of the pectinate ligament (demarcated by black arrowheads) are of a similar tan color to the iris base, whereas fibers of the pectinate ligament (demarcated by white arrowheads) are dark brown in SDP (**g**). WDP fundus image (**d**) shows yellow-colored tapetum lucidum (labeled "TAPETUM") and significant hypopigmentation of the retinal pigment epithelium and choroid allowing visualization of the choroidal vasculature. SDP fundus image (**h**) shows green-colored tapetum lucidum (labeled "TAPETUM") and heavy pigmentation of the non-tapetal fundus. For orientation purposes, images taken at higher magnification (**b**–**d** and **f**–**h**) have the superior (S) and inferior (I) globe positions labeled. Originally published in Winkler et al. (2014)

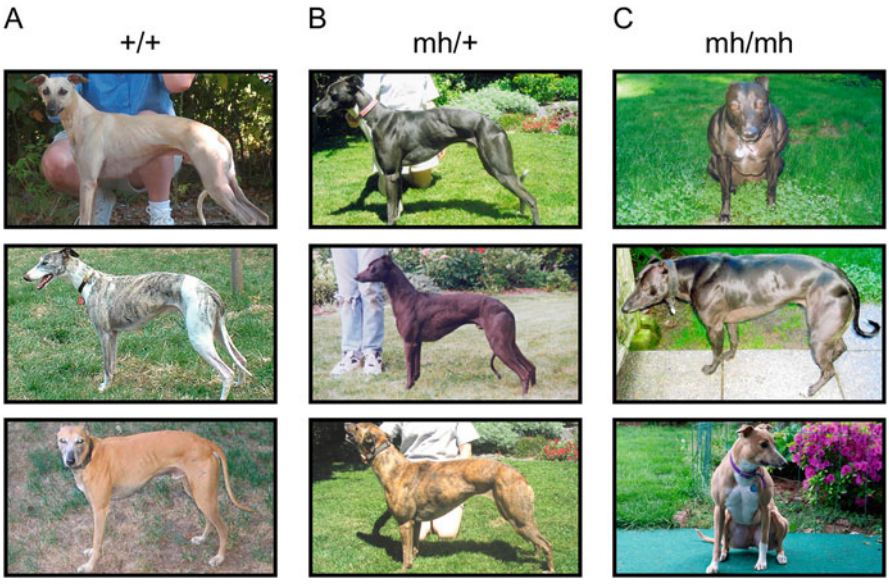

Fig. 14 Whippets with each of the three potential myostatin genotypes. (**a**) Dogs have two copies of the wild-type allele (+/+). (**b**) Dogs are heterozygous with one wild-type allele and one mutant cys → stop allele (mh/+). (**c**) Dogs are homozygous for the mutant allele with two copies of the cys → stop mutation (mh/mh). All photos represent unique individuals except for the top and middle panels in the right-hand column. Originally published in Mosher et al. (2007)

models for better understanding human cancer. The same GWAS approach can also be applied in cancer. For instance, a GWAS containing 39 dog glioma cases and 141 controls from 25 dog breeds identified a significant locus on chromosome 26 (CFA26) (Truvé et al. 2016). Resequencing of a 3.4 Mb target region was performed, revealing 56 SNPs that best fit the association pattern between the resequenced cases and controls. Three candidate genes were highly associated with glioma susceptibility: a calcium-/calmodulin-dependent protein kinase 2 (*CAMKK2*), a P2X ligand-gated ion channel 7 (*P2RX7*), and an mRNA translation reinitiation factor (*DENR*) that influences the migration of cerebral cortical neurons in mice (Haas et al. 2016).

Similarly, an investigation into canine mast cell tumors (CMCT) made use of GWAS in Golden Retrievers from two continents [127 from the United States (70 cases and 57 controls) and 146 from Europe (71 cases and 75 controls)], identifying different regions in the genome associated with risk of CMCT in the two populations (Arendt et al. 2015). Sequencing of GWAS-rescued regions and subsequent fine-mapping identified a *GNAI2* SNP associated with development of CMCT. The *GNAI2* gene encodes an alpha subunit of guanine nucleotide-binding proteins (G proteins) that are transducers in various transmembrane signaling systems and play a role in cell division. The identified SNP introduces an alternative splice form that gives rise to a truncated protein. In addition, CMCT-associated

haplotypes harboring the hyaluronidase genes *HYAL4, SPAM1*, and *HYALP1* on CFA14 and *HYAL1, HYAL2*, and *HYAL3* on CFA20 were identified as separate risk factors in US GRs and European GRs. This suggests that turnover of hyaluronic acid is important for the development of CMCT (Arendt et al. 2015).

It appears that tumorigenesis and cancer associated phenotypes may arise through a variety of mechanisms within the dog. Borge et al. (2015) assessed copy number variations using microarrays to assess genotypes within 117 canine mammary tumors obtained from 69 dogs. The authors point out that cancer cell genomes differ from the host genome through single nucleotide polymorphisms, gain/loss of large chromosomal regions via duplication/deletion of large genomic segments, and expanded/contracted copy numbers of certain loci. Borge et al. (2015) employed the Illumina 170 K canine HD array. Their analysis identified a number of genes with known cancer associations in humans that were frequently amplified or deleted in canine mammary tumors. Some of the genes frequently amplified in the tumors included *BCL6*, *FGFR2*, *MITF*, *MYC*, and *NPM1*, while genes exhibiting deletion loss within canine mammary tumors included *PTEN*, *BMPR1A*, *KDM5C*, *KDM6A*, and *PRF1* (Borge et al. 2015).

Squamous cell carcinoma of the digit (SCCD) in Standard Poodle (STPO) is a locally aggressive cancer that affects only dark coat color individuals. GWAS in 31 STPO SCCD cases and 34 unrelated black STPO controls detected a SNP peak on canine chromosome 15 (Karyadi et al. 2013). Fine-mapping pinpointed a region on the KIT Ligand (*KITLG*) locus. *KITLG* is a pleiotropic factor that acts in the development of both germ and neural cells as well as in hematopoiesis, which is involved in cell migration. Interestingly, the polymorphism within this locus implicated in modulating risk for squamous cell carcinoma appears to be a copy number variant within the transcriptional control region of the KIT locus that is predicted to contain regulatory enhancer elements (Karyadi et al. 2013).

Other mechanisms underlying susceptibility to cancer have been identified. Ferraresso et al. (2014) conducted an in-depth analysis of canine diffuse large B-cell lymphoma (DLBCL) and identified the downregulation of tissue factor pathway inhibitor 2 (*TFPI-2*) as a hallmark of lymph nodes associated with DLBCL. Moreover, the authors demonstrated that hypermethylation of the *TFPI-2* promoter, which increased as a function of age, correlated with decreased expression levels of the gene and demonstrated the age-dependent epigenetic alterations associated with canine DLBCL (Ferraresso et al. 2014).

Melin et al. (2016) performed a GWAS and identified three regions within the canine genome associated with mammary tumors in English Springer Spaniels. The study design consisted of 332 individuals, corresponding to 188 cases and 144 controls. The most significant genomic region was located on chromosome 11 and exhibited a complex architecture of numerous haplotypes spanning the centrosomal cell cycle regulator CDK5 regulatory subunit-associated protein 2 (*CDK5RAP2*). The genomic region spanned 700 kb and was refined to a smaller region of 446 kb. Within this region numerous SNPs, some of which are non-

synonymous and may alter protein function, were identified. Melin et al. (2016) assessed the relationship between the observed haplotypes using a phylogenetic tree approach and then calculated the frequency of cases and controls among the different haplotype groups. The cases within haplotype group 1 exhibited a lower frequency than in haplotype groups 2 and 3. The authors report that within this region of the genome, there are numerous noncoding RNAs such as miRNAs and snoRNAs, potentially implicating RNA-mediated interactions as contributing to mammary tumor susceptibility within this breed (Melin et al. 2016).

10.9 Many Clinically Relevant Traits in German Shepherd

Interestingly, the amount of genetic information about individual dog breeds is continuing to grow rapidly. The German Shepherd dog has been the focus of numerous genetic studies, and the results have opened the door to identification of genetic markers implicated in a significant number of phenotypes, many of which are associated with clinically relevant traits, such as atopic dermatitis and degenerative myelopathy (Table 1). Such knowledge provides opportunities for employing genotyping technology in the artificial selection of next-generation German Shepherds.

11 Conclusions and Future Perspectives

The tremendous wealth of dog genetics and genome information elucidated over the last couple of decades has dramatically altered our understanding of how the dog was domesticated and how artificial selection shaped it into the companion we live with today. There is no doubt that the 30,000 years of selective breeding have given rise to the dogs of today through the selection for specific traits that contribute to the dog's social fitness within human environments. Unfortunately, that same selection has contributed to undesirable clinical phenotypes in dogs as well. The tools of genomics have opened up the possibilities of inferring evolutionary history of dogs as well as the resulting impacts on the genome. Through the lens of genetics, we are able to discern exactly what biochemical molecules were altered in specific breeds during the domestication process. Furthermore, this window into the genome has allowed us to carefully begin to dissect the molecular events contributing to specific morphological phenotypes within particular breeds as well as the undesirable phenotypes associated with disease. These results, taken together, provide clear evidence that selection occurs in the presence of selective pressure and that artificial selection in dogs is an ongoing process. It is interesting to contemplate how dogs will continue to evolve in the future. No doubt it will be at the hands of humans; however, the tools available for aiding the artificial selection process are exponentially more powerful than they were during the original domestication and breed radiation

Table 1 Representative genetic markers associated with German Shepherd dog phenotypes

Phenotype	Gene	Genomic location	Polymorphism	Mode of inheritance	Reference
Activity-impulsivity	DRD4	Chromosome 18	Repeat polymorphism in exon 3	Complex trait (genes and environment)	Hejjas et al. (2007)
Activity-impulsivity	TH	Chromosome 18	36 bp short allele versus 72 bp long allele	Complex trait (genes and environment)	Kubinyi et al. (2012)
Brown liver spots	TYRP1	Chromosome 11	Non-synonymous SNP	Autosomal recessive	Monteagudo and Tejedor (2015)
Canine anal furunculosis	ADAMTS16 and CTNND2	Chromosome 34 (both genes)	Multiple non-synonymous SNPs	Polygenic	Massey et al. (2014)
Canine atopic dermatitis (CAD)	PKP2	Chromosome 27	Multiple upstream regulatory SNPs	Unknown	Tengvall et al. (2016)
Canine atopic dermatitis (CAD)	PKP2	Chromosome 27	~209 kb region with 2 haplotype blocks	Unknown	Tengvall et al. (2013)
Canine degenerative myelopathy (CDM)	SOD1	Chromosome 31	Alters amino acid encoded at position 118	Autosomal recessive	Holder et al. (2016)
Canine inflammatory disease	NOD2	Chromosome 2	Four non-synonymous SNPs in exon 3	Autosomal overdominant	Kathrani et al. (2014)
Chronic superficial keratitis (CSK)	DLA-DRB1	Chromosome 12	Upstream regulatory region SNP	Autosomal recessive	Barrientos et al. (2013)
Hemophilia A	Factor VIII	Chromosome X	Non-synonymous SNP in exon 11	X-linked	Christopherson et al. (2014)
Hereditary multifocal renal cystadenocarcinoma and nodular dermatofibrosis (RCND)	BHD	Chromosome 5	Exon 7 change highly conserved amino acid	Autosomal dominant	Lingaas et al. (2003)

(continued)

Table 1 (continued)

Phenotype	Gene	Genomic location	Polymorphism	Mode of inheritance	Reference
Ichthyosis	ASPRV1	Chromosome 10	Non-synonymous SNP	De novo SNP (not inherited)	Bauer et al. (2017)
Inflammatory bowel disease	TLR4 and TLR5	Chromosome 11 (TLR4) and chromosome 38 (TLR5)	2 SNPs in TLR4 and 1 SNP in TLR5	TLR4 SNP recessive, TLR5 additive	Kathrani et al. (2010)
Olfactory ability	OR10H1-like and OR2K2-like	Chromosome 20 and chromosome 11	SNPs in exonic regions	Polygenic	Yang et al. (2016)
Pancreatic acinar atrophy	DLA-88	Chromosome 12	2 amino acid changes in protein	Autosomal recessive previously, now considered more complex	Tsai et al. (2013)
Pituitary dwarfism	LHX3	Chromosome 9	Deletion of a single 7 bp repeat in intron 5	Autosomal recessive	Voorbij et al. (2011)
White spotting	KIT	Chromosome 13	1 bp insertion in second exon	Autosomal dominant	Wong et al. (2013)

events. Prior to the advent of genomics and genome-wide association studies, artificial selection relied on phenotyping specific animals and breeding them for purposely bred traits. However, the combination of genotyping technology with genomic markers associated with phenotypes of interest allows genetically informed breeding plans to be developed to simultaneously maximize the phenotypes of interest while minimizing the time to achieve the desired artificial selection.

Many breed fanciers are actively working with kennel clubs and geneticists to try to breed out specific undesirable clinical phenotypes like cancer from their lines. This is a challenging process, and consequences of such approaches may result in unintended losses of heterozygosity and alleles within the breed. However, these consequences must be weighed against the backdrop of health for each breed. As the number of genetic markers implicated in dog traits continues to grow, the opportunities for breeding dogs with unique combinations of phenotypes will also increase. Novel breeds may emerge, that have a significantly reduced incidence of undesirable clinical phenotypes. Additionally, it is equally likely that designer dog breeds may be produced that possess unique combinations of morphological phenotypes that previously never co-occurred within the same breed. Combinatorial possibilities are quite literally endless.

Recently, designer dogs (hybrids of two different breeds) have come into fashion. Some dog fanciers view these emerging breeds as a destruction of the underlying breeds. However, others view these dogs as valuable companions and worthwhile pets. One example of such a designer dog is the Labradoodle, a dog produced by a cross of the Labrador Retriever with the Poodle. Considering the combinatorial explosion of pairs that can be crossed from 300 or 400 distinct dog breeds, there are between 44,850 and 79,800 distinct 2-breed designer dog breeds that can be produced from these 300 or 400 breeds, respectively. Furthermore, combining four different breeds to produce a hybrid dog results in more than one billion distinct four-breed combinations.

The demonstrated plasticity of the dog genome represents a powerful mechanism for creating and selecting phenotypes. It is likely that within another 1,000 years, dogs will be selected for combinations of phenotypes and traits that were once thought impossible. It will be truly exciting and breathtaking to witness the evolutionary journey humans will take with dogs.

Although to date dogs appear to have gone through two distinct selection processes, (1) an initial domestication followed by (2) an expansion of breeds more recently, beginning right now, dogs are entering the third selection process, one that will be carried out with the full scientific capability of the human species and where dogs end up will be anyone's guess.

The discoveries made in dog population genomics have been achieved using technology, such as genome sequencing, genotyping arrays, and gene expression arrays. This technology was developed in the past few decades. However, new genomics technology such as RNA sequencing, which provides advantages over microarray-based expression studies, will further open the window to understand complex patterns of gene expression associated with dog domestication, health, and disease. Additionally, the emerging tools associated with epigenetics will

Fig. 15 The human-animal bond was formed through the domestication of wolves into the companion animals we call dogs. Today, millions of dogs are members of human families. The strength of the human-animal bond is frequently represented in media, art, songs, movies, novels, paintings, sculptures, and family photos (as shown here)

undoubtedly provide a greater understanding of how phenotypic variation in dogs can arise through epigenetic regulation of genes. This information will elucidate the underlying mechanisms contributing to gene silencing and clarify why individuals with the same genotypes may exhibit strikingly different phenotypes.

In conclusion, the journey from speculation to knowledge has been very exciting. Moreover, although we have learned some new and important things about dogs, we still have much more to learn. Because dogs are considered to be the first species domesticated by humans, they are the ideal organism to study population genomics and unravel the mysteries underlying domestication and the impact artificial selection has had on anatomical, cognitive, dietary, social, behavioral and disease traits. Through thousands of years living among humans, dogs and humans have shared an extremely strong social bond (Fig. 15). The behavioral and cognitive basis for this bond is beginning to emerge from numerous studies aimed at deciphering the footprints of selection in the dog genome. This is a very exciting time for genomics and for dogs. As we gain a more detailed understanding of our interspecific relationship that evolved over the millennia, we will undoubtedly gain a scientific appreciation for what our hearts already know, and what we already know is that dogs are our best friends.

Acknowledgments Dr. Irizarry acknowledges the role his father and mother had in inspiring him to write this chapter by introducing Dr. Irizarry to the human-animal bond through special relationships with family dogs. Furthermore, Dr. Irizarry wishes to acknowledge the many conversations he had with his parents about cognition, dogs, domestication, and what makes dogs "our best friend."

Those experiences and conversations ultimately paved the path to this chapter. The authors thank Dr. Om Rajora for reading the manuscript and making suggestions during the editing process. Dr. Rajora is acknowledged for simplifying, helping improve text, and favorably upgrading canine knowledge in our chapter. The authors wish to thank Chris Vander Veen for taking time to help prepare documents, without which, this chapter could not be published. The authors are most thankful for the invention of computers, the Internet, word processing programs, and internet-accessible storage platforms which have greatly enhanced collaborative production and sharing of this chapter. The authors fully recognize the hard work of all the people who contributed to the IT environment at Western University of Health Sciences and on Earth. Their commitment to technology ultimately allowed this chapter to be created and reviewed across multiple countries. Without their hard work, this chapter would never see the light of day. The authors are also most appreciative of the hundreds of veterinarians, doctors, scientists, and authors that have published the papers and figures that contributed to the content of this chapter. We truly thank, respect, and value each one of you. Finally, the authors would like to acknowledge the wonderful dogs each of us has known that have inspired and motivated us to learn more about their origins, their health, and their cognition.

References

Ahonen SJ, Kaukonen M, Nussdorfer FD, Harman CD, Komaromy AM, Lohi H. A novel missense mutation in ADAMTS10 in Norwegian Elkhound primary glaucoma. PLoS One. 2014;9:e111941.

Appleyard SM, Hayward M, Young JI, Butler AA, Cone RD, Rubinstein M, Low MJ. A role for the endogenous opioid beta-endorphin in energy homeostasis. Endocrinology. 2003;144:1753–60.

Arendt M, Fall T, Lindblad-Toh K, Axelsson E. Amylase activity is associated with AMY2B copy numbers in dog: implications for dog domestication, diet and diabetes. Anim Genet. 2014;45:716–22.

Arendt ML, Melin M, Tonomura N, Koltookian M, Courtay-Cahen C, Flindall N, Bass J, Boerkamp K, Megquir K, Youell L, Murphy S, McCarthy C, London C, Rutteman GR, Starkey M, Lindblad-Toh K. Genome-wide association study of golden retrievers identifies germ-line risk factors predisposing to mast cell tumours. PLoS Genet. 2015;11:e1005647.

Arendt M, Cairns KM, Ballard JW, Savolainen P, Axelsson E. Diet adaptation in dog reflects spread of prehistoric agriculture. Heredity (Edinb). 2016;117:301–6.

Axelsson E, Ratnakumar A, Arendt ML, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar A, Lindblad-Toh K. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 2013;495:360–4.

Bannasch D, Young A, Myers J, Truvé K, Dickinson P, Gregg J, Davis R, Bongcam-Rudloff E, Webster MT, Lindblad-Toh K, Pedersen N. Localization of canine brachycephaly using an across breed mapping approach. PLoS One. 2010;5:e9632.

Barrientos LS, Zapata G, Crespi JA, Posik DM, Díaz S, It V, Peral-García P, Giovambattista G. A study of the association between chronic superficial keratitis and polymorphisms in the upstream regulatory regions of DLA-DRB1, DLA-DQB1 and DLA-DQA1. Vet Immunol Immunopathol. 2013;156:205–10.

Bauer A, Waluk DP, Galichet A, Timm K, Jagannathan V, Sayar BS, Wiener DJ, Dietschi E, Müller EJ, Roosje P, Welle MM, Leeb T. A de novo variant in the ASPRV1 gene in a dog with ichthyosis. PLoS Genet. 2017;13:e1006651.

Bence M, Marx P, Szantai E, Kubinyi E, Ronai Z, Banlaki Z. Lessons from the canine Oxtr gene: populations, variants and functional aspects. Genes Brain Behav. 2017;16:427–38.

Borge KS, Nord S, Van Loo P, Lingjærde OC, Gunnes G, Alnæs GI, Solvang HK, Lüders T, Kristensen VN, Børresen-Dale AL, Lingaas F. Canine mammary tumours are affected by

frequent copy number aberrations, including amplification of MYC and loss of PTEN. PLoS One. 2015;10:e0126371.

Boyko AR, et al. A simple genetic architecture underlies morphological variation in dogs. PLoS Biol. 2010;8:e1000451.

Cagan A, Blass T. Identification of genomic variants putatively targeted by selection during dog domestication. BMC Evol Biol. 2016;16:10.

Challis BG, Pritchard LE, Creemers JW, Delplanque J, Keogh JM, Luan J, Wareham NJ, Yeo GS, Bhattacharyya S, Froguel P, White A, Farooqi IS, O'Rahilly S. A missense mutation disrupting a dibasic prohormone processing site in pro-opiomelanocortin (POMC) increases susceptibility to early-onset obesity through a novel molecular mechanism. Hum Mol Genet. 2002;11:1997–2004.

Christopherson PW, Bacek LM, King KB, Boudreaux MK. Two novel missense mutations associated with hemophilia A in a family of Boxers, and a German Shepherd dog. Vet Clin Pathol. 2014;43:312–6.

Cooper AE, Ahonen S, Rowlan JS, Duncan A, Seppälä EH, Vanhapelto P, Lohi H, Komáromy AM. A novel form of progressive retinal atrophy in Swedish vallhund dogs. PLoS One. 2014;9:e106610.

Downs LM, Wallin-Håkansson B, Boursnell M, Marklund S, Hedhammar Å, Truvé K, Hübinette L, Lindblad-Toh K, Bergström T, Mellersh CS. A frameshift mutation in golden retriever dogs with progressive retinal atrophy endorses SLC4A3 as a candidate gene for human retinal degenerations. PLoS One. 2011;6:e21452.

Downs LM, Hitti R, Pregnolato S, Mellersh CS. Genetic screening for PRA-associated mutations in multiple dog breeds shows that PRA is heterogeneous within and between breeds. Vet Ophthalmol. 2014;17:126–30.

Drogemuller M, Jagannathan V, Dolf G, Butenhoff K, Kottmann-Berger S, Wess G, Leeb T. A single codon insertion in the PICALM gene is not associated with subvalvular aortic stenosis in Newfoundland dogs. Hum Genet. 2015;134:127–9.

Feldman R. The neurobiology of human attachments. Trends Cogn Sci. 2017;21:80–99.

Ferraresso S, et al. Epigenetic silencing of TFPI-2 in canine diffuse large B-cell lymphoma. PLoS One. 2014;9:e92707.

Freedman AH, et al. Genome sequencing highlights the dynamic early history of dogs. PLoS Genet. 2014;10:e1004016.

Haas MA, Ngo L, Li SS, Schleich S, Qu Z, Vanyai HK, Cullen HD, Cardona-Alberich A, Gladwyn-Ng IE, Pagnamenta AT, Taylor JC, Stewart H, Kini U, Duncan KE, Teleman AA, Keays DA, Heng JI. De novo mutations in DENR disrupt neuronal development and link congenital neurological disorders to faulty mRNA translation re-initiation. Cell Rep. 2016;15:2251–65.

Hejjas K, Vas J, Topal J, Szantai E, Ronai Z, Szekely A, Kubinyi E, Horvath Z, Sasvari-Szekely M, Miklosi A. Association of polymorphisms in the dopamine D4 receptor gene and the activity-impulsivity endophenotype in dogs. Anim Genet. 2007;38:629–33.

Holder AL, Price JA, Adams JP, Volk HA, Catchpole B. A retrospective study of the prevalence of the canine degenerative myelopathy associated superoxide dismutase 1 mutation (SOD1:c.118G > A) in a referral population of German Shepherd dogs from the UK. Canine Genet Epidemiol. 2016;1:10.

Howrigan DP, Simonson MA, Keller MC. Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. BMC Genomics. 2011;12:460.

Karyadi DM, Karlins E, Decker B, vonHoldt BM, Carpintero-Ramirez G, Parker HG, Wayne RK, Ostrander EA. A copy number variant at the KITLG locus likely confers risk for canine squamous cell carcinoma of the digit. PLoS Genet. 2013;9:e1003409.

Kathrani A, House A, Catchpole B, Murphy A, German A, Werling D, Allenspach K. Polymorphisms in the TLR4 and TLR5 gene are significantly associated with inflammatory bowel disease in German Shepherd dogs. PLoS One. 2010;5:e15740.

Kathrani A, Lee H, White C, Catchpole B, Murphy A, German A, Werling D, Allenspach K. Association between nucleotide oligomerisation domain two (Nod2) gene polymorphisms and canine inflammatory bowel disease. Vet Immunol Immunopathol. 2014;161:32–41.

Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, Rusch DB, Delcher AL, Pop M, Wang W, Fraser CM, Venter JC. The dog genome: survey sequencing and comparative analysis. Science. 2003;301:1898–903.

Kis A, Bence M, Lakatos G, Pergel E, Turcsán B, Pluijmakers J, Vas J, Elek Z, Brúder I, Földi L, Sasvári-Székely M, Miklósi A, Rónai Z, Kubinyi E. Oxytocin receptor gene polymorphisms are associated with human directed social behavior in dogs (Canis familiaris). PLoS One. 2014;9:e83993.

Kovacs K, Kis A, Pogany A, Koller D, Topal J. Differential effects of oxytocin on social sensitivity in two distinct breeds of dogs (Canis familiaris). Psychoneuroendocrinology. 2016;74:212–20.

Kubinyi E, Vas J, Hejjas K, Ronai Z, Brúder I, Turcsán B, Sasvari-Szekely M, Miklósi A. Polymorphism in the tyrosine hydroxylase (TH) gene is associated with activity-impulsivity in German Shepherd dogs. PLoS One. 2012;7:e30271.

Kuchtey J, Kunkel J, Esson D, Sapienza JS, Ward DA, Plummer CE, Gelatt KN, Kuchtey RW. Screening ADAMTS10 in dog populations supports Gly661Arg as the glaucoma-causing variant in beagles. Invest Ophthalmol Vis Sci. 2013;54:1881–6.

Lee YS, Challis BG, Thompson DA, Yeo GS, Keogh JM, Madonna ME, Wraight V, Sims M, Vatin V, Meyre D, Shield J, Burren C, Ibrahim Z, Cheetham T, Swift P, Blackwood A, Hung CC, Wareham NJ, Froguel P, Millhauser GL, O'Rahilly S, Farooqi IS. A POMC variant implicates beta-melanocyte-stimulating hormone in the control of human energy balance. Cell Metab. 2006;3:135–40.

Lee EJ, Merriwether DA, Kasparov AK, Nikolskiy PA, Sotnikova MV, Pavlova EY, Pitulko VV. Ancient DNA analysis of the oldest canid species from the Siberian Arctic and genetic contribution to the domestic dog. PLoS One. 2015;10:e0125759.

Li Y, Vonholdt BM, Reynolds A, Boyko AR, Wayne RK, Wu DD, Zhang YP. Artificial selection on brain-expressed genes during the domestication of dog. Mol Biol Evol. 2013;30:1867–76.

Li Y, Wang GD, Wang MS, Irwin DM, Wu DD, Zhang YP. Domestication of the dog from the wolf was promoted by enhanced excitatory synaptic plasticity: a hypothesis. Genome Biol Evol. 2014;6:3115–21.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ 3rd, Zody MC, Mauceli E, Xie X, Breen M, Wayne RK, Ostrander EA, Ponting CP, Galibert F, Smith DR, PJ DJ, Kirkness E, Alvarez P, Biagi T, Brockman W, Butler J, Chin CW, Cook A, Cuff J, Daly MJ, DeCaprio D, Gnerre S, Grabherr M, Kellis M, Kleber M, Bardeleben C, Goodstadt L, Heger A, Hitte C, Kim L, Koepfli KP, Parker HG, Pollinger JP, Searle SM, Sutter NB, Thomas R, Webber C, Baldwin J, Abebe A, Abouelleil A, Aftuck L, Ait-Zahra M, Aldredge T, Allen N, An P, Anderson S, Antoine C, Arachchi H, Aslam A, Ayotte L, Bachantsang P, Barry A, Bayul T, Benamara M, Berlin A, Bessette D, Blitshteyn B, Bloom T, Blye J, Boguslavskiy L, Bonnet C, Boukhgalter B, Brown A, Cahill P, Calixte N, Camarata J, Cheshatsang Y, Chu J, Citroen M, Collymore A, Cooke P, Dawoe T, Daza R, Decktor K, DeGray S, Dhargay N, Dooley K, Dooley K, Dorje P, Dorjee K, Dorris L, Duffey N, Dupes A, Egbiremolen O, Elong R, Falk J, Farina A, Faro S, Ferguson D, Ferreira P, Fisher S, FitzGerald M, Foley K, Foley C, Franke A, Friedrich D, Gage D, Garber M, Gearin G, Giannoukos G, Goode T, Goyette A, Graham J, Grandbois E, Gyaltsen K, Hafez N, Hagopian D, Hagos B, Hall J, Healy C, Hegarty R, Honan T, Horn A, Houde N, Hughes L, Hunnicutt L, Husby M, Jester B, Jones C, Kamat A, Kanga B, Kells C, Khazanovich D, Kieu AC, Kisner P, Kumar M, Lance K, Landers T, Lara M, Lee W, Leger JP, Lennon N, Leuper L, LeVine S, Liu J, Liu X, Lokyitsang Y, Lokyitsang T, Lui A, Macdonald J, Major J, Marabella R, Maru K, Matthews C, McDonough S, Mehta T, Meldrim J, Melnikov A, Meneus L, Mihalev A, Mihova T, Miller K, Mittelman R, Mlenga V, Mulrain L, Munson G, Navidi A, Naylor J, Nguyen T, Nguyen N, Nguyen C, Nguyen T, Nicol R, Norbu N, Norbu C, Novod N, Nyima T, Olandt P, O'Neill B, O'Neill K, Osman S, Oyono L, Patti C, Perrin D,

Phunkhang P, Pierre F, Priest M, Rachupka A, Raghuraman S, Rameau R, Ray V, Raymond C, Rege F, Rise C, Rogers J, Rogov P, Sahalie J, Settipalli S, Sharpe T, Shea T, Sheehan M, Sherpa N, Shi J, Shih D, Sloan J, Smith C, Sparrow T, Stalker J, Stange-Thomann N, Stavropoulos S, Stone C, Stone S, Sykes S, Tchuinga P, Tenzing P, Tesfaye S, Thoulutsang D, Thoulutsang Y, Topham K, Topping I, Tsamla T, Vassiliev H, Venkataraman V, Vo A, Wangchuk T, Wangdi T, Weiand M, Wilkinson J, Wilson A, Yadav S, Yang S, Yang X, Young G, Yu Q, Zainoun J, Zembek L, Zimmer A, Lander ES. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005;438:803–19.

Lingaas F, Comstock KE, Kirkness EF, Sørensen A, Aarskaug T, Hitte C, Nickerson ML, Moe L, Schmidt LS, Thomas R, Breen M, Galibert F, Zbar B, Ostrander EA. A mutation in the canine BHD gene is associated with hereditary multifocal renal cystadenocarcinoma and nodular dermatofibrosis in the German Shepherd dog. Hum Mol Genet. 2003;12:3043–53.

Massey J, Short AD, Catchpole B, House A, Day MJ, Lohi H, Ollier WE, Kennedy LJ. Genetics of canine anal furunculosis in the German Shepherd dog. Immunogenetics. 2014;66:311–24.

Melin M, Rivera P, Arendt M, Elvers I, Murén E, Gustafson U, Starkey M, Borge KS, Lingaas F, Häggström J, Saellström S, Rönnberg H, Lindblad-Toh K. Genome-wide analysis identifies germ-line risk factors associated with canine mammary tumours. PLoS Genet. 2016;12:e1006029.

Meurs KM, Lahmers S, Keene BW, White SN, Oyama MA, Mauceli E, Lindblad-Toh K. A splice site mutation in a gene encoding for PDK4, a mitochondrial protein, is associated with the development of dilated cardiomyopathy in the Doberman pinscher. Hum Genet. 2012;131:1319–25.

Monteagudo LV, Tejedor MT. The b(c) allele of TYRP1 is causative for the recessive brown (liver) colour in German Shepherd dogs. Anim Genet. 2015;46:588–9.

Mosher DS, Quignon P, Bustamante CD, Sutter NB, Mellersh CS, Parker HG, Ostrander EA. A mutation in the myostatin gene increases muscle mass and enhances racing performance in heterozygote dogs. PLoS Genet. 2007;3:e79.

Nagasawa M, Mitsui S, En S, Ohtani N, Ohta M, Sakuma Y, Onaka T, Mogi K, Kikusui T. Social evolution. Oxytocin-gaze positive loop and the coevolution of human-dog bonds. Science. 2015;348:333–6.

Oliver JA, Forman OP, Pettitt L, Mellersh CS. Two independent mutations in ADAMTS17 are associated with primary open angle glaucoma in the Basset Hound and Basset Fauve de Bretagne breeds of dog. PLoS One. 2015;10:e0140436.

Ovodov ND, Crockford SJ, Kuzmin YV, Higham TF, Hodgins GW, van der Plicht J. A 33,000-year-old incipient dog from the Altai Mountains of Siberia: evidence of the earliest domestication disrupted by the last glacial maximum. PLoS One. 2011;6:e22821.

Parker HG, Ostrander EA. Canine genomics and genetics: running with the pack. PLoS Genet. 2005;1:e58.

Persson ME, Wright D, Roth LS, Batakis P, Jensen P. Genomic regions associated with interspecies communication in dogs contain genes related to human social disorders. Sci Rep. 2016;6:33439.

Philipp U, Vollmar A, Haggstrom J, Thomas A, Distl O. Multiple Loci are associated with dilated cardiomyopathy in Irish wolfhounds. PLoS One. 2012;7:e36691.

Quignon P, Herbin L, Cadieu E, Kirkness EF, Hédan B, Mosher DS, Galibert F, André C, Ostrander EA, Hitte C. Canine population structure: assessment and impact of intra-breed stratification on SNP-based association studies. PLoS One. 2007;2:e1324.

Raffan E, Dennis RJ, O'Donovan CJ, Becker JM, Scott RA, Smith SP, Withers DJ, Wood CJ, Conci E, Clements DN, Summers KM, German AJ, Mellersh CS, Arendt ML, Iyemere VP, Withers E, Söder J, Wernersson S, Andersson G, Lindblad-Toh K, Yeo GS, O'Rahilly S. A deletion in the canine POMC gene is associated with weight and appetite in obesity-prone labrador retriever dogs. Cell Metab. 2016;23:893–900.

Reiter T, Jagoda E, Capellini TD. Dietary variation and evolution of gene copy number among dog breeds. PLoS One. 2016;11:e0148899.

Romero T, Nagasawa M, Mogi K, Hasegawa T, Kikusui T. Oxytocin promotes social bonding in dogs. Proc Natl Acad Sci U S A. 2014;111:9085–90.

Saetre P, Lindberg J, Leonard JA, Olsson K, Pettersson U, Ellegren H, Bergström TF, Vilà C, Jazin E. From wild wolf to domestic dog: gene expression changes in the brain. Brain Res Mol Brain Res. 2004;126:198–206.

Schoenebeck JJ, Hutchinson SA, Byers A, Beale HC, Carrington B, Faden DL, Rimbault M, Decker B, Kidd JM, Sood R, Boyko AR, Fondon JW 3rd, Wayne RK, Bustamante CD, Ciruna B, Ostrander EA. Variation of BMP3 contributes to dog breed skull diversity. PLoS Genet. 2012;8:e1002849.

Shannon LM, Boyko RH, Castelhano M, Corey E, Hayward JJ, McLean C, White ME, Abi Said M, Anita BA, Bondjengo NI, Calero J, Galov A, Hedimbi M, Imam B, Khalap R, Lally D, Masta A, Oliveira KC, Pérez L, Randall J, Tam NM, Trujillo-Cornejo FJ, Valeriano C, Sutter NB, Todhunter RJ, Bustamante CD, Boyko AR. Genetic structure in village dogs reveals a Central Asian domestication origin. Proc Natl Acad Sci U S A. 2015;112:13639–44.

Shearin AL, Ostrander EA. Canine morphology: hunting for genes and tracking mutations. PLoS Biol. 2010;8:e1000310.

Stern JA, White SN, Lehmkuhl LB, Reina-Doreste Y, Ferguson JL, Nascone-Yoder NM, Meurs KM. A single codon insertion in PICALM is associated with development of familial subvalvular aortic stenosis in Newfoundland dogs. Hum Genet. 2014;133:1139–48.

Stern JA, Hsue W, Song KH, Ontiveros ES, Luis Fuentes V, Stepien RL. Severity of mitral valve degeneration is associated with chromosome 15 loci in whippet dogs. PLoS One. 2015;10:e0141234.

Tengvall K, Kierczak M, Bergvall K, Olsson M, Frankowiack M, Farias FH, Pielberg G, Carlborg Ö, Leeb T, Andersson G, Hammarström L, Hedhammar Å, Lindblad-Toh K. Genome-wide analysis in German Shepherd dogs reveals association of a locus on CFA 27 with atopic dermatitis. PLoS Genet. 2013;9:e1003475.

Tengvall K, Kozyrev S, Kierczak M, Bergvall K, Farias FH, Ardesjö-Lundgren B, Olsson M, Murén E, Hagman R, Leeb T, Pielberg G, Hedhammar Å, Andersson G, Lindblad-Toh K. Multiple regulatory variants located in cell type-specific enhancers within the PKP2 locus form major risk and protective haplotypes for canine atopic dermatitis in German Shepherd dogs. BMC Genet. 2016;17:97.

Thalmann O, Shapiro B, Cui P, Schuenemann VJ, Sawyer SK, Greenfield DL, Germonpré MB, Sablin MV, López-Giráldez F, Domingo-Roura X, Napierala H, Uerpmann HP, Loponte DM, Acosta AA, Giemsch L, Schmitz RW, Worthington B, Buikstra JE, Druzhkova A, Graphodatsky AS, Ovodov ND, Wahlberg N, Freedman AH, Schweizer RM, Koepfli KP, Leonard JA, Meyer M, Krause J, Pääbo S, Green RE, Wayne RK. Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. Science. 2013;342:871–4.

Toro R, Pérez-Serra A, Campuzano O, Moncayo-Arlandi J, Allegue C, Iglesias A, Mangas A, Brugada R. Familial dilated cardiomyopathy caused by a novel frameshift in the BAG3 gene. PLoS One. 2016;11:e0158730.

Truvé K, Dickinson P, Xiong A, York D, Jayashankar K, Pielberg G, Koltookian M, Murén E, Fuxelius HH, Weishaupt H, Swartling FJ, Andersson G, Hedhammar Å, Bongcam-Rudloff E, Forsberg-Nilsson K, Bannasch D, Lindblad-Toh K. Utilizing the dog genome in the search for novel candidate genes involved in glioma development-genome wide association mapping followed by targeted massive parallel sequencing identifies a strongly associated locus. PLoS Genet. 2016;12:e1006000.

Tsai KL, Starr-Moss AN, Venkataraman GM, Robinson C, Kennedy LJ, Steiner JM, Clark LA. Alleles of the major histocompatibility complex play a role in the pathogenesis of pancreatic acinar atrophy in dogs. Immunogenetics. 2013;65:501–9.

Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, Fall T, Seppälä EH, Hansen MS, Lawley CT, Karlsson EK, Consortium LUPA, Bannasch D, Vilà C, Lohi H, Galibert F, Fredholm M, Häggström J, Hedhammar A, André C, Lindblad-Toh K,

Hitte C, Webster MT. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLoS Genet. 2011;7:e1002316.

Vonholdt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, Quignon P, Degenhardt JD, Boyko AR, Earl DA, Auton A, Reynolds A, Bryc K, Brisbin A, Knowles JC, Mosher DS, Spady TC, Elkahloun A, Geffen E, Pilot M, Jedrzejewski W, Greco C, Randi E, Bannasch D, Wilton A, Shearman J, Musiani M, Cargill M, Jones PG, Qian Z, Huang W, Ding ZL, Zhang YP, Bustamante CD, Ostrander EA, Novembre J, Wayne RK. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature. 2010;464:898–902.

Voorbij AM, van Steenbeek FG, Vos-Loohuis M, Martens EE, Hanson-Nilsson JM, van Oost BA, Kooistra HS, Leegwater PA. A contracted DNA repeat in LHX3 intron 5 is associated with aberrant splicing and pituitary dwarfism in German Shepherd dogs. PLoS One. 2011;6:e27940.

Wang GD, Zhai W, Yang HC, Fan RX, Cao X, Zhong L, Wang L, Liu F, Wu H, Cheng LG, Poyarkov AD, Poyarkov NA Jr, Tang SS, Zhao WM, Gao Y, Lv XM, Irwin DM, Savolainen P, Wu CI, Zhang YP. The genomics of selection in dogs and the parallel evolution between dogs and humans. Nat Commun. 2013;4:1860.

Wiik AC, Ropstad EO, Ekesten B, Karlstam L, Wade CM, Lingaas F. Progressive retinal atrophy in Shetland sheepdog is associated with a mutation in the CNGA1 gene. Anim Genet. 2015;46:515–21.

Wijesena HR, Schmutz SM. A missense mutation in SLC45A2 is associated with albinism in several small long haired dog breeds. J Hered. 2015;106:285–8.

Winkler PA, Gornik KR, Ramsey DT, Dubielzig RR, Venta PJ, Petersen-Jones SM, Bartoe JT. A partial gene deletion of SLC45A2 causes oculocutaneous albinism in Doberman pinscher dogs. PLoS One. 2014;9:e92127.

Wolf ZT, Brand HA, Shaffer JR, Leslie EJ, Arzi B, Willet CE, Cox TC, McHenry T, Narayan N, Feingold E, Wang X, Sliskovic S, Karmi N, Safra N, Sanchez C, Deleyiannis FW, Murray JC, Wade CM, Marazita ML, Bannasch DL. Genome-wide association studies in dogs and humans identify ADAMTS20 as a risk variant for cleft lip and palate. PLoS Genet. 2015;11:e1005059.

Wong AK, Ruhe AL, Robertson KR, Loew ER, Williams DC, Neff MW. A de novo mutation in KIT causes white spotting in a subpopulation of German Shepherd dogs. Anim Genet. 2013;44:305–10.

Yang M, Geng GJ, Zhang W, Cui L, Zhang HX, Zheng JL. SNP genotypes of olfactory receptor genes associated with olfactory ability in German Shepherd dogs. Anim Genet. 2016;47:240–4.

Index

Om P. Rajora (ed.), *Population Genomics*, Population Genomics,
https://doi.org/10.1007/978-3-030-04589-0