Population Genomics

Martin F. Polz
Om P. Rajora *Editors*

Population Genomics: Microorganisms

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.

Martin F. Polz • Om P. Rajora
Editors

Population Genomics: Microorganisms

Springer

Editors
Martin F. Polz
Department of Civil and Environmental Engineering
Massachusetts Institute of Technology
Cambridge, MA, USA

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 978-3-030-04755-9 ISBN 978-3-030-04756-6 (eBook)
https://doi.org/10.1007/978-3-030-04756-6

Library of Congress Control Number: 2018965734

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

Dedicated to my mentors from whom I have learnt, and the student and postdoctoral colleagues from whom I continue to learn.

Martin F. Polz

Respectfully dedicated to my educators and mentors.

Om P. Rajora

Preface

Genomics has revolutionized many fields of biology. For microbes, in particular, it has revealed the enormous scope of diversity coexisting in most environments. Not surprisingly, efforts in microbial genomics have, to a large extent, been directed towards understanding the phylogenetic and functional diversity encompassed by microbes. Although much of microbial diversity remains to be uncovered, there is also a more recent focus on analysis of closely related genomes. This effort was, at least initially, driven by the need to better understand the evolution and epidemiology of pathogenic viruses and bacteria. The continuous decline in sequencing cost has, however, enabled a broader focus on nonhuman pathogens, and environmental and industrial microbes to better understand how microevolutionary processes create variation within populations. Hence, the field of microbial population genomics has come of age, and we believe, it is time for a book that summarizes current developments and future perspectives in this novel but important field.

Population genomics of microorganisms is most commonly understood to encompass the analysis of entire genomes of intraspecific and interspecific closely related individuals using phylogenetic and population genetics concepts and tools. Population genomics, therefore, deals broadly with the analysis of evolutionary forces that both create and remove variation among members of populations, and, perhaps most importantly, lead to adaptation to environmental niches or hosts. Simply put, population genomics is population genetics empowered by genomics. This definition can, however, vary somewhat according to the organisms studied so that many authors within this book provide their own, more nuanced definitions of microbial population genomics. Moreover, availability of data varies greatly for different types of organisms. Not surprisingly, the genomic analysis of human viral and bacterial pathogens is most advanced and although other fields are catching up, for many types of organisms, population genomics of microorganisms represents a nascent field emerging from comparative genomics of closely related organisms. The availability of large collections of closely related strains is, however, bound to rapidly increase in the next few years since standard genetic characterization of

isolates is increasingly done by whole genome rather than single-marker gene sequencing.

In this book, we have tried to cover all major groups of microorganisms for which at least some population genomics studies have been undertaken. The chapters, thus, span the whole spectrum of diversity encompassed by microbes, including bacteria, archaea, fungi, and viruses. And for pathogens, there is further subdivision according to the hosts infected. The result is an in-depth analysis of microbial population genomics that allows comparison among fields. Our hope is that this structure will enable the reader to find commonalities and differences among organisms, and that such comparison will outline a roadmap for new investigators in the field of microorganism population genomics. Because crosstalk between fields is always mediated by common methods, we have included a chapter that explicitly deals with computational tools for microorganism population genomics. However, many of the individual chapters cover additional methods, often developed for specific purposes but often more broadly relevant. Finally, because many microbes remain hard to culture and are only accessible by metagenomics, the book contains a chapter that deals explicitly with the opportunities and challenges in applying population genomics to uncultured organisms.

Talking about population genomics implies that we know how to define and delineate populations. In many cases, we have good intuition of what a population might be, such as in the analysis of highly clonal pathogens or sexually isolated eukaryotes. How to demarcate population boundaries is, however, often not easy. In particular, for bacteria and archaea, as well as for some viruses, the potential for horizontal gene transfer and the vast coexisting genetic diversity exemplify this difficulty. In fact, the term population is often used loosely in microbiology, describing from cells in a culture tube to diverse microbes coexisting in environmental samples. Several of the chapters, therefore, explicitly tackle the issue of how to define populations and how populations split into distinct units in the process of speciation. Based on the sophistication of the analysis, we predict that the next few years will see tremendous advances in theory about how to define microbial populations.

It is an exciting time for a book on microbial population genomics as the field takes shape and is expanding into new areas of research. We thank all the distinguished authors who have taken the time to contribute to this effort and we hope that all have been rewarded by the timeliness and quality of this book.

Cambridge, MA, USA — Martin F. Polz
Fredericton, NB, Canada — Om P. Rajora

Contents

Contributors

Philip Arevalo Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

David A. Baltrus School of Plant Sciences, University of Arizona, Tucson, AZ, USA

School of Animal and Comparative Biomedical Sciences, University of Arizona, Tucson, AZ, USA

Bridget Marie Barker The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA

Harald Brüssow Division of Animal and Human Health Engineering, Laboratory of Gene Technology, University of Leuven, Leuven, Belgium

Guillaume Charron Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Marcus de Melo Teixeira The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA

Vincent J. Denef Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA

Kanny Diallo Department of Zoology, University of Oxford, Oxford, UK

Centre pour les Vaccins en Développement, Bamako, Mali

Xavier Didelot Department of Infectious Disease Epidemiology, Imperial College London, London, UK

Éléonore Durand Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Chris Eberlein Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Fernando García-Arenal Centro de Biotecnología y Genómica de Plantas UPM-INIA, Universidad Politécnica de Madrid, Madrid, Spain

E.T.S. Ingeniería Agronómica, Alimentaria y de Biosistemas, Madrid, Spain

Fernando González-Candelas Joint Research Unit "Infection and Public Health" FISABIO-Universitat de València, Institute for Integrative Systems Biology, I2SysBio (CSIC-UV), Valencia, Spain

CIBER in Epidemiology and Public Health, Madrid, Spain

Mathieu Hénault Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Daniel Raymond Kollath The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA

David J. Krause Laboratory of Genetics, Genome Center of Wisconsin, Wisconsin Energy Institute, J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, WI, USA

Christian R. Landry Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Martin C. J. Maiden Department of Zoology, University of Oxford, Oxford, UK

Hélène Martin Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Lou Nielly-Thibault Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systémes, PROTEO, Université Laval, Quebec City, QC, Canada

Israel Pagán Centro de Biotecnología y Genómica de Plantas UPM-INIA, Universidad Politécnica de Madrid, Madrid, Spain

E.T.S. Ingeniería Agronómica, Alimentaria y de Biosistemas, Madrid, Spain

Juan Ángel Patiño-Galindo Joint Research Unit "Infection and Public Health" FISABIO-Universitat de València, Institute for Integrative Systems Biology, I2SysBio (CSIC-UV), Valencia, Spain

CIBER in Epidemiology and Public Health, Madrid, Spain

Martin F. Polz Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

B. Jesse Shapiro Department of Biological Sciences, University of Montreal, Montreal, QC, Canada

Carlos Valiente-Mullor Joint Research Unit "Infection and Public Health" FISABIO-Universitat de València, Institute for Integrative Systems Biology, I2SysBio (CSIC-UV), Valencia, Spain

CIBER in Epidemiology and Public Health, Madrid, Spain

David VanInsberghe Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Rachel J. Whitaker Department of Microbiology, Carl R. Woese Institute for Genomic Biology, University of Illinois Urbana-Champaign, Champaign, IL, USA

Part I
Concepts and Approaches

Computational Methods in Microbial Population Genomics

Xavier Didelot

Abstract Whole genome sequencing is frequently applied to hundreds of samples within a single microbial population study. The resulting datasets are large and need to be analysed using computationally efficient methods, the development of which is an active research field. Here we review the current state of the art in terms of computation methods used in microbial population genomics. This includes software for assembly and alignment of core genomic regions, which is usually a pre-requirement for analysing the ancestry of the genomes, via phylogenetic on non-phylogenetic methods. We also review additional techniques aimed at combining genomic data with temporal, geographical or other types of metadata, as well as pan-genome methods of analysis that go beyond the core genome.

Keywords Alignment • Assembly • Computation methods • Microbial population genomics • Pan-genome analysis • Phylodynamics • Phylogenetics • Phylogeography • Recombination

1 Introduction

With the advent of new genome sequencing technologies, the cost and time required to sequence whole microbial genomes have decreased to such a point that research studies are now able to include hundreds or even thousands of newly sequenced genomes. Analysis of such large datasets requires the use of specific computational methods, which are reviewed in this chapter, but are still the subject of active development. Section 2 describes how to prepare genomic data for

X. Didelot (✉)
Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, London W2 1PG, UK
e-mail: x.didelot@imperial.ac.uk

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2017_3,

analysis, including identification of core and accessory genomic regions, assembly and alignment. Section 3 summarises methods for analysing the ancestry of the genomes, which can broadly be divided into phylogenetic and non-phylogenetic approaches. Section 4 describes how temporal information about the sampling dates of the genomes can be combined with the genomic data to paint a more complete picture of the evolutionary process. Section 5 covers the use of the geographic locations from which the genomes originate to describe the geographic structure of the population. Section 6 describes how other types of metadata can be integrated into a microbial population genomics study to investigate the distribution and evolution of various properties of interest. Finally, Sect. 7 explains how analysis of the pan-genome can be carried out.

2 Preparing Genomic Data for Analysis

2.1 Core and Accessory Genome

When comparing genomic data from members of a microbial population, it is useful to identify the genomic regions that are present in all the genomes, and which collectively are called the core genome. The remaining regions, which are found in some but not all the genomes, are collectively called the accessory genome, while the sum of core and accessory genome is often called the pan-genome. Analysis of microbial population genomic data typically requires an alignment of the core genomic regions, and this section describes how to prepare such an alignment. The separation of core and accessory genome regions is especially relevant for bacterial population genomics, because bacterial genomes within a population often exhibit significant variation in genomic content, whereas this is not usually the case for viral populations. In bacterial genomics, analysis of non-core regions can be important too, and this subject is treated in Sect. 7.

2.2 Reference-Based Assembly

Sequencing data from current sequencing instruments (reviewed in Loman and Pallen 2015; Goodwin et al. 2016) comes in the form of a large number of reads of length 100–250 bp which are highly redundant, so that each individual genomic position is covered by several reads. The average number of reads covering genomic positions is called the coverage depth and is a good indication of how accurate the final genome sequence will be, for example depth of 40× and above. Assembly is the process whereby reads are put together to reconstruct the genome sequence. There are broadly two forms of assembly: reference-based assembly and de novo assembly, each with their specific strengths and weaknesses.

Reference-based assembly requires that a whole genome from the population (or at least species) under study has been previously sequenced, which is called the reference genome. Each read is then aligned against the reference genome, and popular algorithms to perform this include BWA (Li and Durbin 2009), SMALT (http://www.sanger.ac.uk/science/tools/smalt-0), Stampy (Lunter and Goodson 2011) and Bowtie (Langmead et al. 2009). The next step is called variant calling, which is often done using SamTools (Li et al. 2009), FreeBayes (Garrison and Marth 2012) and/or GATK (McKenna et al. 2010). For each position along the reference genome, the alignment of reads at that position is considered. If there are enough aligned reads and they are in good agreement, the corresponding nucleotide of the target genome is called, otherwise it is left undetermined. The latter happens mostly for regions that are present in the reference genome but not in the target genome and for regions that are repetitive, or if the sequencing quality is low. Reference-based assembly has the advantage that each assembled genome is aligned against the same reference and hence all aligned against each other and directly comparable. Drawbacks include the need for a pre-sequenced reference genome, and the fact that only regions found in the reference genome can be assembled, which is sufficient to study the core genome but not the accessory genome.

2.3 De Novo Assembly

The alternative to reference-based assembly is to assemble each genome de novo, that is by directly comparing and aligning the reads against each other. Popular softwares for de novo assembly include Velvet (Zerbino and Birney 2008), SPAdes (Bankevich et al. 2012), IDBA (Peng et al. 2012) and A5 (Tritt et al. 2012). The output is typically a set of long genomic regions called contigs, which occur along the genome in an undetermined order. De novo assembled genomes need to be aligned against each other before they can be compared. A first approach is to perform a multiple alignment of the whole genome which accounts for possible genomic rearrangements, but even the best software using this strategy such as progressiveMauve (Darling et al. 2010) or MUGSY (Angiuoli and Salzberg 2010) cannot deal with more than ~50 genomes. An alternative is to align each de novo assembled genome against a single reference, for example using MUMmer (Kurtz et al. 2004), but this shares the disadvantages of reference-based assembly described above. A third approach to using de novo assembled data is to search for previously defined genes throughout the contigs using for example BLAST (Altschul et al. 1997), as implemented for example in the BIGSdb platform (Jolley and Maiden 2010). Finally, instead of using predefined genes it is possible to annotate each de novo assembled genome separately, using for example RAST (Overbeek et al. 2014), Prokka (Seemann 2014) or Prodigal (Hyatt et al. 2010) and to search for orthologs between the genomes using a pipeline involving BLAST to compare the genes versus each other, for example OrthoMCL (Li et al. 2003),

LS-BSR (Sahl et al. 2014) or Roary (Page et al. 2015). Once ortholog genes have been found in all genomes, they can be aligned separately using for example Muscle (Edgar 2004).

Reference-based and de novo assemblies are complementary approaches which are often used side by side to compare results in ambiguous regions and exploit the strengths of both strategies, especially the reconstruction of core-genome alignments that are directly analysable in reference-based assembly and the reconstruction of non-core regions in de novo assembly. After applying either or both approaches, an alignment of the genomes is created which contains all core regions (or core genes if a de novo gene-based approach was used). Such an alignment is required as input for the analytical methods described in the next sections.

2.4 Simulation

Analysis of simulated microbial genomic data in parallel with real genomic data can often be useful. This allows for example to test the fit of an evolutionary model to the data, to build empirical distributions of expected quantities, or to estimate evolutionary parameters informally by progressive tuning of the simulation parameters until it resembles the real data. A more formal use of simulated datasets is to use Approximate Bayesian Computation techniques, also known as likelihood-free methods since they do not require to calculate the probability of the data given evolutionary parameters, but instead rely on simulation and comparison of the simulated and real data on a set of summary statistics (Marin et al. 2012). Simulation is also useful on its own (i.e. without combination with real data), to test the accuracy of analytical methods on datasets for which the correct answer is known.

The most popular and powerful approaches to simulate microbial genomic data are based on the coalescent model (Kingman 1982) under which it is possible to simulate the evolutionary history of a sample of genomes without simulating the evolution of the whole population. One of the first methods to be implemented based on this principle was Hudson's ms (Hudson 2002), and it remains popular to date due to the wide range of scenarios that can be simulated using this software. Extensions have also been released, for example msHOT (Hellenthal and Stephens 2007) which allows for the presence of mutational hotspots. Another popular software is fastsimcoal (Excoffier and Foll 2011), which uses an efficient approximation to simulate crossover recombination, allowing the simulation of longer genomes from sexual populations. Clonal organisms such as bacteria undergo a recombination process akin to gene conversion rather than crossover, which can be simulated for example in ms but for which separate algorithms have been specifically implemented. SimMLST (Didelot et al. 2009b) was aimed at simulating multi-locus sequence typing data, where sequence is available for only a handful of short (~400 bp) gene fragments (Maiden et al. 1998). It has recently been superseded by SimBac (Brown et al. 2016) which is 100 times faster and therefore much better suited to simulate whole genome sequence data.

3 Description of Microbial Population Ancestry

3.1 Phylogenetics Ignoring Recombination

The most frequently used method to represent patterns of relationships between a set of microbial genomes is to draw a phylogenetic tree (Fig. 1). Closely related genomes should have fewer differences between them and be more closely clustered together on the tree compared to more distantly related genomes. A tree should always be read along the axis from root to leaves, bearing in mind that the other axis is arbitrary so that two genomes can be next to each other and yet be separated by a long branch (Baum et al. 2005). Phylogenetic methods are computational techniques that use as input an alignment of genomic data like the ones described in the previous section, and produce in output a phylogenetic tree. Most phylogenetic methods assume that no recombination happened, which is appropriate for example to analyse data from bacterial pathogens do not recombine much, e.g. *Mycobacterium tuberculosis* (Comas et al. 2013), or data in which recombinant regions have been previously detected and filtered out (cf next section).

The simplest phylogenetic methods rely on first building a distance matrix between all pairs of genomes, for example UPGMA, Neighbor-Joining (Saitou and Nei 1987) and BIONJ (Gascuel 1997). These methods are not very popular because they do not exploit the full data but only the distance matrix. They are however very quick and therefore still frequently used to provide a starting point for other methods. Parsimony methods are based on the whole genomic data and attempt to reconstruct the tree that minimises the number of substitutions on branches to produce the data (Fitch 1971). Parsimony methods are not currently

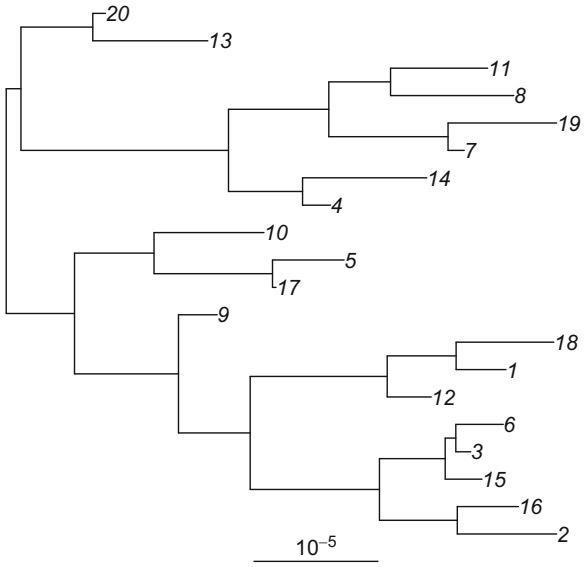

Fig. 1 Example of a phylogenetic tree. The x-axis represents evolutionary distance whereas the y-axis is arbitrary. It is important to 'read' the tree along the correct axis, for example genomes 4 and 10 appear next to other but are not especially closely related

frequently used to analyse microbial genomic data. Maximum likelihood techniques are based on an explicit probabilistic model of how substitutions accumulate on a tree, such that it is possible to define the likelihood, that is the probability of the genetic data given a tree. An efficient algorithm for computing the likelihood is the so-called pruning algorithm (Felsenstein 1981), which leaves the problem of exploring the space of all possible trees to find the one that maximises the likelihood. Powerful algorithms to do so have been developed that are implemented for example in the popular software phyml (Guindon et al. 2010), RAxML (Stamatakis 2006) and FastTree (Price et al. 2009, 2010). For any dataset with more than a handful of genomes, the number of possible trees is too large to allow a complete exploration of all trees, so that the analysis relies on heuristics which are not guaranteed to always return the best tree, but should still return one of the most likely trees.

Bayesian phylogenetic methods are based on an explicit evolutionary model like maximum likelihood but with two important differences. Firstly, the suitability of a tree is not measured in terms of the likelihood but of the posterior probability, which is the product of the likelihood and a prior probability. This term represents how appropriate a tree is deemed to be, only based on a tree model without reference to the genomic data. A prior tree model needs therefore to be specified, for example using the coalescent model (Kingman 1982), and this prior model can include parameters and be used to explore various evolutionary scenarios. Secondly, instead of finding a single maximising tree, the Bayesian approach returns a sample of trees that may have generated the data, also known as a posterior sample of trees. Comparisons between these trees can be performed to assess the statistical confidence in the phylogenetic reconstruction. In non-Bayesian phylogenetic techniques uncertainty measurement is usually achieved approximately and expensively using bootstrapping (Felsenstein 1985), but Bayesian phylogenetics provides a more natural way to do this. Popular software packages to perform Bayesian phylogenetics include MrBayes (Ronquist et al. 2012), RevBayes (Höhna et al. 2016), BEAST (Drummond and Rambaut 2007) and BEAST2 (Bouckaert et al. 2014).

3.2 Phylogenetics Accounting for Recombination

The phylogenetic techniques described in the previous section all assume that no recombination affected the data, so that a single tree applies for all sites. However, many microbes experience significant rates of recombination as they evolve. When this is the case, applying a method that assumes no recombination can lead to incorrect phylogenetic reconstructions (Schierup and Hein 2000; Hedge and Wilson 2014). A first sign of the effect of recombination can be obtained by estimating separate trees for different parts of the genome (for example, a tree for each gene). If recombination had not occurred, we would expect all such trees to look very similar, up to the randomness of the mutation process affecting each gene.

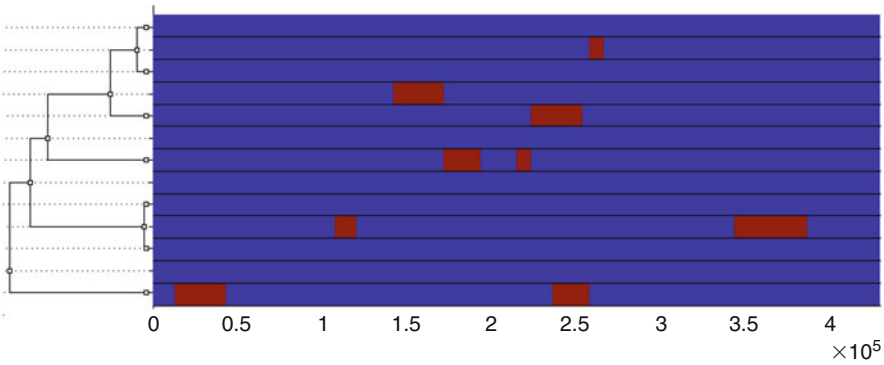

Fig. 2 Example of a phylogenetic tree with recombination events shown as a matrix on the right. To each terminal and internal branch of the tree corresponds a row of the matrix, with positions along the genome alignment shown on the x-axis of the matrix. For any given branch, unrecombined regions are shown in blue and recombined regions are shown in red

In some microbes, recombination happens exclusively as a gene conversion process, that is with a strong asymmetry between the two parents involved in recombination: the recipient cell contributes the vast majority of the resulting genome whereas the donor cell only contributes a short fragment. This is true of all bacterial species for example, irrespective of whether recombination was caused by conjugation, transduction or transformation (Didelot et al. 2010). In this case, recombination can be integrated into the phylogenetic tree reconstruction process by identifying the recombined fragments that happened on every branches of the phylogeny called clonal genealogy (Fig. 2). This clonal genealogy represents the ancestry process obtained by following the line of descent of the recipient at each recombination event, that is the line followed by the majority of the genetic material. A first software following this principle was ClonalFrame (Didelot and Falush 2007), which was originally designed for multilocus sequence typing data (Maiden et al. 1998) but can also work with limited (up to ~100) number of whole genomes, as was demonstrated for example by applications to *Escherichia coli* (Didelot et al. 2012b) and *Chlamydia trachomatis* (Joseph et al. 2011, 2012). For larger whole genome datasets, a newer version has been released which uses maximum likelihood optimisation techniques and is called ClonalFrameML (Didelot and Wilson 2015). ClonalFrameML has been applied for example to large genomic datasets of *Neisseria gonorrhoeae* (De Silva et al. 2016) and *Escherichia coli* (Ingle et al. 2016). A similar tool is Gubbins (Croucher et al. 2015) which operates through an iterative process of building a phylogenetic tree using standard recombination-unaware techniques, finding recombinant regions that do not fit the tree and repeating. Examples of application of Gubbins have been published on *Streptococcus pneumoniae* (Croucher et al. 2011) and *Chlamydia trachomatis* (Harris et al. 2012).

Instead or in addition to this gene conversion process, some microbes undergo recombination akin to crossing-over in higher organisms, that is where both parents

contribute large amounts of DNA. In this case, it is not possible to identify a recipient and donor for recombination events, and therefore there is not a defined clonal genealogy as above that can be targeted for phylogenetic reconstruction. This situation arises for many viruses, for example HIV. A solution is then to try and identify the breakpoints along the alignment where significant recombination events have occurred, and to reconstruct a separate phylogeny for each genomic region between two consecutive breakpoints. Computational software exploiting this idea include TOPALi (Milne et al. 2004, 2009), stepBrothers (Bloomquist et al. 2009), GARD (Pond et al. 2006) and RDP4 (Martin et al. 2015). A special recombination scenario occurs in the evolution of the influenza virus. The genome is made of eight segments, and recombination proceeds by replacement of whole segments, also known as reassortment. Techniques have therefore been developed to exploit this specific process, for example the GiRaF software (Nagarajan and Kingsford 2011) which reconstructs trees for each segment separately and considers the reassortment events that would be needed to reconcile them.

3.3 Non-phylogenetic Ancestry

A phylogeny is not always the best way to represent the ancestry of a sample of individuals. This is especially true for microbes that recombine a lot, as for example *Helicobacter pylori* in which 40% of genes can be affected by recombination within 3 years of within-host evolution (Kennemann et al. 2011). An alternative is to consider that there is a number (K) of underlying populations, with each individual either belonging to a population, or being a genetic mixture of the different populations (Fig. 3). One of the first algorithms to be based on this principle was STRUCTURE (Pritchard et al. 2000) and the linkage option (Falush et al. 2003) within it (as opposed to the non-admixture and admixture options) is especially useful to analyse sequence data since it models the correlation in the ancestry of sites near each other along the genome. For example, two sites next to each other have a high probability of having the same ancestry, since otherwise the boundary of a recombination event would have had to occur exactly between these two sites. The computational cost of running STRUCTURE does not scale well with the length and number of sequence being analysed though, and it is challenging to determine the number (K) of ancestral populations that should be considered in the model. Consequently, its current use in microbiology is limited to very specific situations, for example to quantify the admixture between the two bacterial species *Campylobacter jejuni* and *E. coli* (Sheppard et al. 2013a). Other softwares based on a similar population admixture principle include ADMIXTURE (Alexander et al. 2009) and BAPS (Tang et al. 2009) which is popular to determine population clusters amongst bacterial genomes, for example *Streptococcus pneumoniae* (Chewapreecha et al. 2014). Another non-phylogenetic approach is BratNextGen (Marttinen et al. 2012) which does not cluster individuals into populations as the previously mentioned software, but instead identifies the genomic fragments that

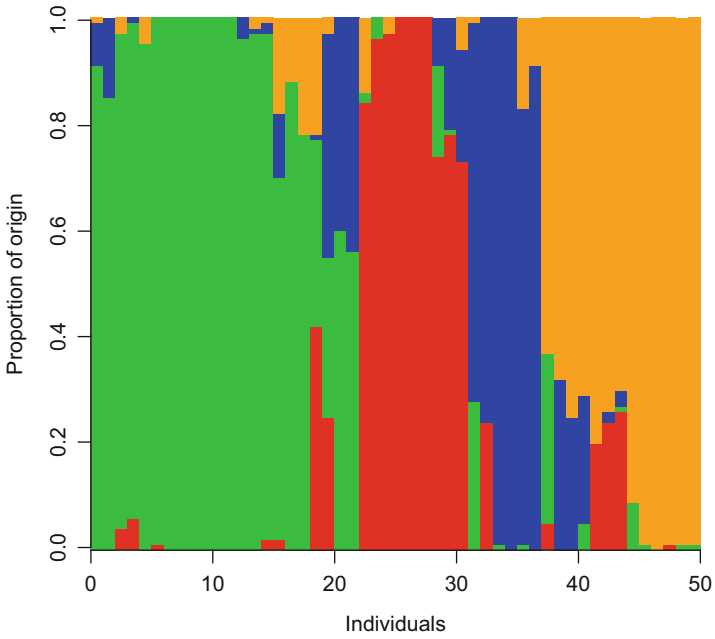

Fig. 3 Example of a barplot representation of population structure. The analysis includes 50 individuals shown on the x-axis and four populations have been detected, each of which corresponds to a colour (red, blue, green, orange). For each individual, the proportion of genomic material originating from each of the four populations is illustrated on the y-axis. The ordering of the individuals on the x-axis is arbitrary and often chosen to group together the individuals with similar profiles

are likely to have come from sources external to the population under consideration. BratNextGen is therefore usually applied to genomes from a single bacterial lineage to identify recombination events coming from other lineages, for example *Streptococcus pneumoniae* PMEN1 (Marttinen et al. 2012) or *Staphylococcus aureus* ST239 (Castillo-Ramírez et al. 2012).

FineStructure (Lawson et al. 2012) is another non-phylogenetic method to reconstruct the population structure. The algorithm proceeds in two steps. First each genome is considered in turn and reconstructed as a mosaic of all other genomes using a copying model (Li and Stephens 2003): each site is copied from one of the genome and copying occurs in blocks so that two neighbouring sites are likely to come from the same genome. The number of blocks copied by each genome from each other genome is then counted and summarised in a so-called co-ancestry matrix. A clustering method is then used to group together the individuals with similar co-ancestry rows into populations. Thus FineStructure reveals both the population of origin of each individual, and the fragments that have been imported from elsewhere, making it comparable to the previously mentioned linkage model of STRUCTURE (Falush et al. 2003). The computational cost of

running FineStructure is however much lower than that of running STRUCTURE, so that very large datasets can be analysed in a manner of hours. The problem of estimating the number of ancestral populations (K) is also resolved by the two-step approach. FineStructure was originally designed for human genetics, but has also proven useful in bacterial genomics, having been applied for example to *Helicobacter pylori* (Yahara et al. 2013), *Vibrio parahaemolyticus* (Cui et al. 2015) and *Myxococcus xanthus* (Wielgoss et al. 2016). An extension called orderedPainting has been developed specifically for detecting recombination hotspots in bacterial genomes (Yahara et al. 2014).

4 Integrating Temporal Data

4.1 Temporal Data in Microbial Genomics

The dates on which the microbes have been isolated are usually known, and it can often be interesting to integrate this information into the microbial genomic analysis. A first approach for doing so, which can be used in both phylogenetic and non-phylogenetic frameworks, is to simply annotate the reconstructed population ancestry with the dates, to see if some lineages or populations seem to have emerged more recently than others (for example, see Haase et al. 2014, Fig. 2b, d). In a phylogenetic framework, however, there is a more powerful approach available which is to try and reconstruct a timed tree (Fig. 4). In a timed tree, branch lengths are measured in a time unit (for example, days or years) rather than a genetic unit (for example, number of substitutions per site). Each tip represents a microbial genome and is aligned with its known date of isolate. Each internal node represents the most

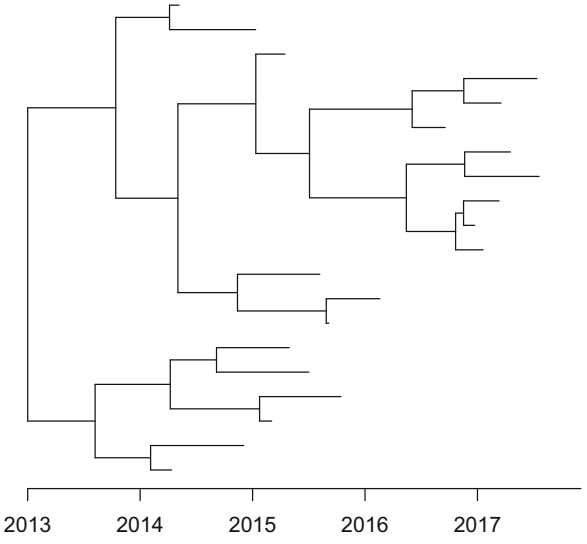

Fig. 4 Example of a timed tree. The interpretation is the same as for a standard phylogenetic tree, except that the time scale (x-axis) is measured in years rather than genetic distance. Each genome is aligned on the x-axis with its known date of isolation. Each internal node of the tree is aligned on the x-axis with the inferred date of existence of the last common ancestor of the genomes underneath

recent common ancestor between the set of genomes descended from the node, and is aligned with the date when it occurred, which is unknown but estimated by the phylogenetic procedure. In particular, the root of the tree represents the most recent common ancestor (MRCA) of the whole set of microbial genomes and it is aligned with the time to the most recent common ancestor (TMRCA) of the whole set. A timed tree therefore allows more natural interpretations to be drawn, especially when the research questions of interest are of an epidemiological or ecological nature, since the dating of all branches is included in the tree. Correctly reconstructing such a timed tree from a set of microbial genomes and associated isolation dates is therefore an important methodological concern.

4.2 Molecular Clock and Building a Timed Tree

Building a timed tree requires an estimate of the molecular clock rate, that is the rate at which substitutions are accumulated over time on genomes and measured for example in units of substitutions per year per site. Let us assume that there is such a rate and that it is relatively constant over the evolutionary history considered. This assumption is called the strict molecular clock assumption. Sometimes this rate has been estimated by previous studies and can be used directly to build the timed tree. For example, in a study of *Clostridium difficile*, genomes sampled longitudinally from the same hosts were compared to estimate the evolutionary clock rate, which was then used to produce timed trees (Didelot et al. 2012a). Otherwise, when the clock rate is unknown, it needs to be estimated from the data at hand. A simple approach for doing so is called root-to-tip method, where a non-timed phylogenetic tree is estimated, and a scatter plot is formed with a dot for each genome, the x-axis corresponding to the known isolation dates and the y-axis to the length of the path from root to the genome in the phylogeny (Fig. 5). If the strict clock assumption holds approximately, and that the range of sample dates is large enough relative to the age of the root, then a linear correlation should be found in this scatter plot. The slope of this linear regression is an estimate of the molecular clock rate, while the value on the x-axis at which the linear regression crosses the x-axis is an estimate of the age of the root of the phylogeny. This method was for example used in *Streptococcus pneumoniae* and showed much better results when based on a phylogeny that had been corrected for recombination compared to one that had not (Croucher et al. 2011). This root-to-tip method is useful to establish whether the temporal signal in the data is strong enough to consider applying the methods described below for reconstructing a timed tree. An implementation of the root-to-tip technique is provided by the software TempEst (Rambaut et al. 2016).

The most popular method to reconstruct a timed tree is that implemented in the softwares BEAST (Drummond et al. 2012) and BEAST2 (Bouckaert et al. 2014), relying on Bayesian statistics to jointly estimate the molecular clock, the timed tree and uncertainties around them. Reconstructing timed trees using BEAST has been especially popular for analysing viral genetic data, for example in influenza (Smith et al. 2009), HIV (Worobey et al. 2008) and Ebola (Gire et al. 2014), but more

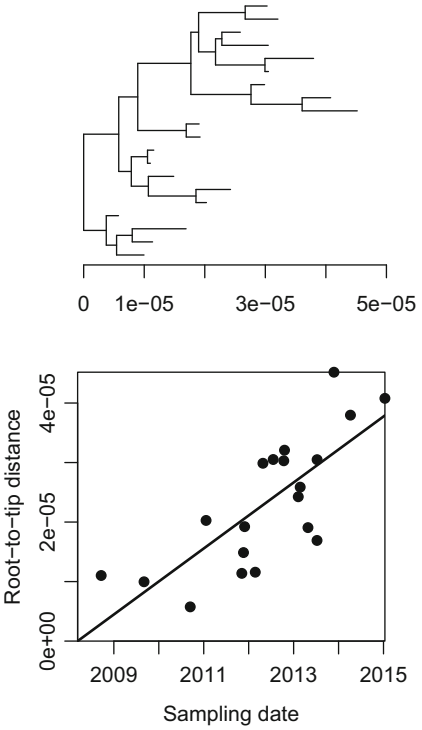

Fig. 5 Example of application of the root-to-tip method. The top panel shows the phylogenetic tree reconstructed for the genomes of interest. On the bottom panel, there is a dot for each of these genomes, with the x-axis representing the known date of isolation of the genome and the y-axis representing the length of the path from root to tip in the phylogenetic tree. A linear regression can then be attempted on the scatter plot, which if statistically well supported can be used to estimate both the molecular clock rate (slope of the regression) and the time of the most recent common ancestor for the whole set of genomes (intersect of the linear regression with the x-axis, here 2008)

recently has also gained in popularity for bacterial genomics (Biek et al. 2015), for example in the study of *Yersinia pestis* (Cui et al. 2013), *Shigella sonnei* (Holt et al. 2012) and *Escherichia coli* (Stoesser et al. 2016). BEAST also implements options to use instead of the strict molecular clock described so far, a relaxed molecular clock where the rate of evolution is allowed to vary to some extent between the different branches of the tree (Drummond et al. 2006; Drummond and Suchard 2010). An alternative to BEAST is LSD (To et al. 2016) which is faster and able to deal with larger datasets as was demonstrated for example recently in an analysis of thousands of simulated HIV genomes (Ratmann et al. 2017).

4.3 Phylodynamics

Past changes in population size affect what a timed genealogy is likely to look like (Griffiths and Tavare 1994). For example, if the population size has been increasing significantly, it will result in longer terminal branches and shorter internal branches compared to a tree under a constant or declining population size. It is also possible to turn this stochastic relationship around, meaning that a reconstructed timed phylogeny is informative about past population size dynamics. Phylodynamics is

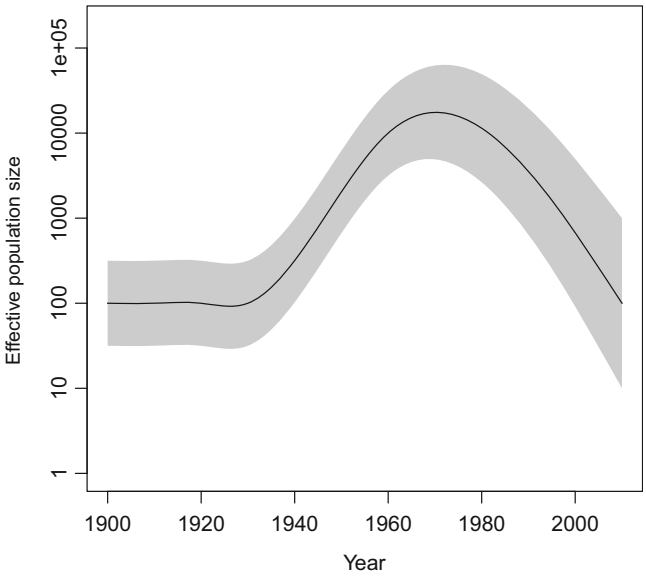

Fig. 6 Example of a skyline plot. The black line indicates the estimated population size over time, with the grey shading representing the 95% credibility interval. Here we see that the population size was stable from 1900 until 1940, increased significantly up until 1965 after which it started to decline back to its original level

the branch of phylogenetics that exploits this property. Following their implementation into the BEAST framework, starting with the Bayesian skyline plot (Drummond et al. 2005), these techniques have become increasingly popular to analyse microbial genomic data. The typical result is a plot with time on the x-axis and the effective population size on the y-axis (often measured on a log scale), with a line indicating the mean estimated population size variations and shading representing the 95% credibility interval over time (Fig. 6). Phylodynamics is very popular to investigate viral population size dynamics, for example in an analysis of rabies in North American raccoons where the skyline plot is in good agreement with epidemiological information about the spread of the disease (Biek et al. 2007). It is also sometimes used in bacterial genomics, for example in a study of the emergence of *Staphylococcus aureus* ST225 in Germany and the Czech Republic (Nübel et al. 2010).

5 Integrating Spatial Data

5.1 Using a Descriptive Approach

When the spatial origin of the genomes is known and varied, using this information in the context of a microbial population genomic analysis can help to reveal the

geographical structuring of the population and the potential occurrence of migrations between locations. The spatial data used for such a phylogeographic analysis can occur at any scale, including between patches of land separated by only a few centimetres (Wielgoss et al. 2016), between different body parts within a single host (Didelot et al. 2016), between different regions of a single country or between different countries throughout the world (Croucher and Didelot 2015).

The simplest approach to investigate the geographical pattern of the origins of the microbial genomes is to plot the geographical data side-by-side with the results of the analysis of population ancestry. If a non-phylogenetic, clustering method was used for the analysis of population ancestry, then the distributions of geographical origins can be compared between inferred clusters. If a phylogenetic method was used, the leaves of the tree can be annotated according to spatial origin, for example by using a different colour for each location. For example, these two types of annotations (non-phylogenetic and phylogenetic) were both used in a genomic analysis of *Streptococcus suis* (Weinert et al. 2015) in Figs. 1c and 5, respectively. This purely descriptive approach can already reveal interesting features and, in the phylogenetic context, the extent to which genomes from each location form clusters in the tree is noteworthy. Such clustering is indicative of the strength of the geographical structure and exceptions where a genome falls into the "wrong" cluster can represent recent migrations, as was shown for example in a global genomic analysis in *Staphylococcus aureus* ST239 (Harris et al. 2010). Likewise, when the aim is to investigate the source of an isolate, simply looking at the origins of its nearest relatives can be highly suggestive, as was used for example to uncover the South-East Asian origin of the 2010 Haiti cholera outbreak (Chin et al. 2011). The Microreact web interface (Argimón et al. 2016) provides a user-friendly way of studying side-by-side the origin of isolates on a map and their genomic relationships, including the ability to interactively explore subsets of isolates defined by geographical or genomic criteria.

5.2 Using an Inferential Approach

A natural next step beyond annotating the leaves of a tree with spatial sources is to try to annotate the internal nodes or branches (Fig. 7). However, doing this requires an algorithm to infer the ancestral locations since this is only known about the leaves. The most widespread approach for doing so is to consider the location as a discrete trait that evolves along the branches of the tree, with mutations of the discrete trait corresponding to migrations from one location to another. Migrations occur according to an unknown matrix of rates from any location to any other, which may be constrained to reduce the number of parameters to estimate, for example by considering that migration from location A to location B happen at the same rate as from location B to location A, so that the migration rate matrix becomes symmetric. Joint inference of the migration matrix and ancestral locations can be performed under such a model using ancestral state reconstruction

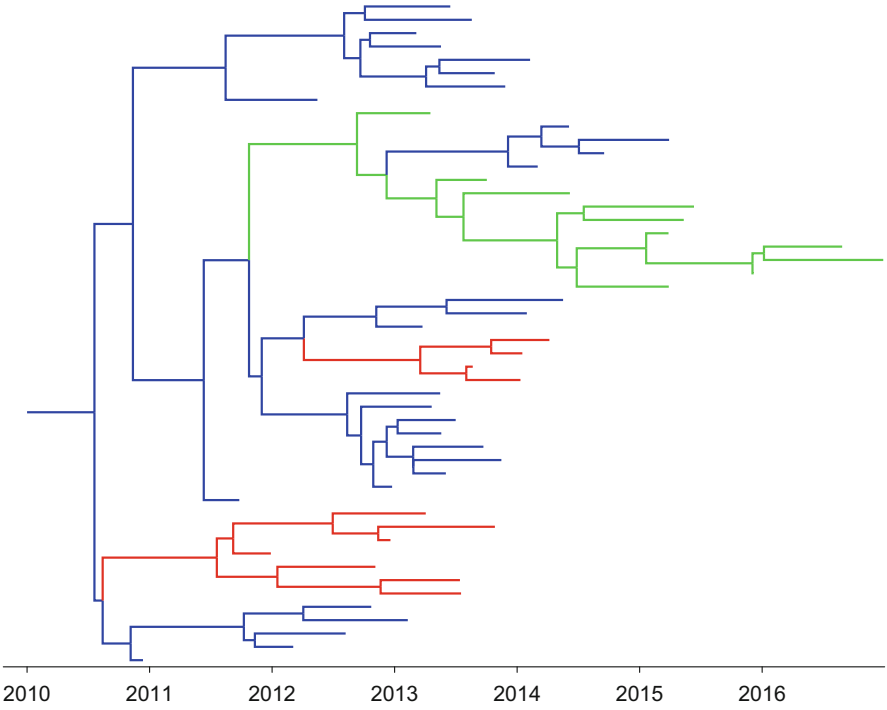

Fig. 7 Example of tree coloured by geographical location. A colour is assigned to each of the locations (here for example three countries are shown in red, green and blue). The location of origin of each genome is known and shown by colouring the corresponding terminal branch with the appropriate colour. The location of ancestors is not known but can be inferred using algorithms as described in the main text and this can then be shown by colouring internal branches of the tree accordingly

techniques (Joy et al. 2016). For example, the ancestral location of *Shigella sonnei* lineages was reconstructed (Holt et al. 2012) by maximum likelihood estimation using the ace command from the R package ape (Paradis et al. 2004). Once the ancestral locations have been reconstructed, the full history of past migrations is revealed since changes in location along a branch or from one branch to its descendent branch can be interpreted as a migration from one location to another. When combined with temporal information (see previous section), this approach can reveal the spatio-temporal spread of a microbe, for example the global spread of the current pandemic of cholera in three waves that all originated from South-East Asia (Mutreja et al. 2011; Didelot et al. 2015).

Phylogeographic analysis can also be performed within the BEAST and BEAST2 frameworks (Drummond et al. 2012; Bouckaert et al. 2014), either using the discrete trait modelling approach described above (Lemey et al. 2009) or a continuous space version (Lemey et al. 2010). The latter has the advantage to analyse the ancestral locations at the same time as the phylogenetic space is being

explored, so that phylogenetic uncertainty is accounted for in the phylogeographic analysis. This technique was originally applied to Avian Influenza A H5N1 (Lemey et al. 2009) and rabies (Lemey et al. 2010) and has since become very popular mostly for viral phylogeography studies (Bloomquist et al. 2010) but also to investigate bacterial phylogeography such as *Mycobacterium tuberculosis* (Comas et al. 2013) and *Clostridium difficile* (He et al. 2013). Powerful interactive visualisation techniques have also been developed to explore the ancestral reconstructions output by these analytical methods (Bielejec et al. 2011, 2016). Within BEAST2 (Bouckaert et al. 2014) a separate algorithm called BASTA has recently been developed which is based on an approximation of the structured coalescent and can lead to more accurate ancestral reconstructions, especially when sampling is highly biased between locations (De Maio et al. 2015).

6 Integrating Other Types of Data

6.1 Application of Ancestral State Reconstruction

Non-genomic metadata can be integrated with phylogeny to provide insight into the evolutionary history of populations. When performing a microbial population genomics study, there are often additional non-genomic metadata that it can be interesting to integrate into the analysis to investigate their relationship with the evolutionary history of the population. The last two sections described specifically the case of temporal and spatial data, and this section discusses the use of other types of data. Depending on the system under study, this metadata may include virulence measurements, antimicrobial resistance profiles, host species of origin, tissue of origin, conditions of isolation, results of in vitro experiments, etc.

Many of the methods described in the previous section for the analysis of spatial data can be applied to other types of metadata, because they are based on models of discrete or continuous trait evolution that are not specific to phylogeographic analysis. The evolutionary history of traits of interest can thus be revealed in the form of changes in the metadata value along branches of the tree (when working in a phylogenetic framework), or significant differences between populations (when working in a clustering framework). For example, maximum likelihood estimation of ancestral state, as implemented for instance in the ace command of the R package ape (Paradis et al. 2004), was used to reconstruct the evolutionary history of pathogenicity in *Clostridium difficile* (Dingle et al. 2014). Likewise, even though the discrete trait analysis methodology implemented in BEAST (Lemey et al. 2009) was originally developed with phylogeography in mind, it has since been applied to other non-spatial traits. In bacterial population genomics, examples include studies of host species in *Campylobacter jejuni* (Dearlove et al. 2015) and host sexual orientation in *Neisseria gonorrhoeae* (Grad et al. 2014). In viral population genomics, examples include studies of host species in rabies (Faria et al. 2013) and antigenic diversity in influenza (Zinder et al. 2013).

6.2 Uncovering Populations and Associations

Beyond the reconstruction of ancestral trait evolution, metadata can also be used to define units within the population, with the rationale that metadata is relatively uniform within units and different between units. If the traits are markers of the ecological environment in which the isolates were sampled, the units thus defined may represent ecologically adapted lineages, otherwise known as ecotypes (Cohan and Perry 2007). For example, AdaptML (Hunt et al. 2008) uses information about the ecological metadata to define clusters on a phylogeny each of which corresponds to a limited number of habitats. AdaptML was originally applied to *Vibrionaeceae* (Hunt et al. 2008), and has more recently been used to investigate ecological adaptation in *Staphylococcus aureus* (Shepheard et al. 2013) and *Escherichia coli* (McNally et al. 2016). More generally, it is possible to consider the evolution of a probabilistic distribution on a phenotype of interest, rather than the phenotype itself, which is especially useful when the phenotype is not perfectly inherited and depends on non-genetic factors (Visscher et al. 2008). Under such a model, changes in the phenotype itself may happen just by chance but changes in the phenotype distribution represent important evolutionary events, which can be used to define units in the population. This approach was recently implemented in TreeBreaker (Ansari and Didelot 2016).

Integrating non-genomic metadata with genomic data is also needed to perform Genome-Wide Association Studies (GWAS) where the aim is to determine the genetic causative basis for a phenotype of interest. GWAS has a long history of being used in human population genetics but has also more recently become of interest to microbial geneticists (Read and Massey 2014). One of the earliest examples of successful microbial GWAS uncovered the role played by a vitamin B5 biosynthesis island in adaptation of *Campylobacter jejuni* to infecting either cattle or chicken (Sheppard et al. 2013b). A key challenge is to account for the population structure, which if not correctly done can lead to spurious results (Balding 2006). Because microbial populations are typically much more structured than the human population, methods popular for human GWAS such as PLINK (Purcell et al. 2007) may not be directly applicable to perform microbial GWAS. New methods are starting to emerge specifically designed for microbial GWAS including bugwas (Earle et al. 2016), SEER (Lees et al. 2016) and treeWAS (Collins and Didelot 2017). These methods are specifically tailored to account in the association analysis for the otherwise potentially confounding properties of clonality, population structure and recombination that occur in microbial populations.

7 Pan-Genome Analysis

7.1 Description of Genomic Content Variations

All computational methods described so far have been focusing on the analysis of the core genome, that is the set of genomic regions that is shared amongst all the genomes under study. However, analysing non-core regions can be important too, especially in bacterial genomics where even closely related isolates can differ in their genomic content, with gain and loss of genes being an important evolutionary force. The first step to analyse non-core regions is to reconstruct the pan-genome, that is the set of regions shared by subsets of the genomes. As such, the size of the pan genome is dependent on the set of genomes being used, and the boundaries of the population from which these genomes are drawn. Following de novo assembly of all the genomes, the pan-genome can be reconstructed either using genes as a unit of content, for example using the Roary pipeline (Page et al. 2015), or purely based on genomic sequences, for example using progressiveMauve (Darling et al. 2010), see Sect. 2 for more details on these different approaches. The gene-based approach can more easily deal with datasets of various diversity levels, whereas the complexity of the gene-free approach increases when the genomes are not closely related, since it becomes more difficult to accurately align the genomes. On the other hand, a gene-free approach has the advantage to inform also about non-coding regions such as promoters, and to exploit genome synteny to reconstruct homologous relationships, which is particularly useful to deal with shorter or highly variable loci.

One of the first pan-genome analysis was conducted in *Streptococcus agalactiae* (Tettelin et al. 2005) and this introduced many of the concepts that were used in subsequent studies (Medini et al. 2005). In particular, it is useful to plot accumulation curves of how many genes are found in the core and pan-genome as more and more genomes are being considered. The accumulation curve for the core genome decreases since genes found in previous genomes but not in a new genome are being removed from the core, whereas the accumulation curve for the pan genome increases as new genes not found in previous genomes are discovered in the newly added genomes (Fig. 8). The core genome curve always decreases to a plateau, which represents the set of genes that are vital for the survival of the bacteria. The pan genome curve on the other hand can take two forms, depending on whether a plateau will eventually be reached or not as new genomes are being considered, leading to two types of pan genomes: closed if all genes would eventually be characterised when enough genomes have been considered, or open if new genes will always be discovered no matter how many genomes have already been considered. Comparative analysis in nine bacterial species found five of them had open pan-genomes and four had closed pan-genome (Tettelin et al. 2008). Pan-genome analysis using accumulation curves has since become a popular approach (Vernikos et al. 2015) and has also been applied at higher levels of diversity, up to the whole bacterial kingdom (Lapierre and Gogarten 2009). A

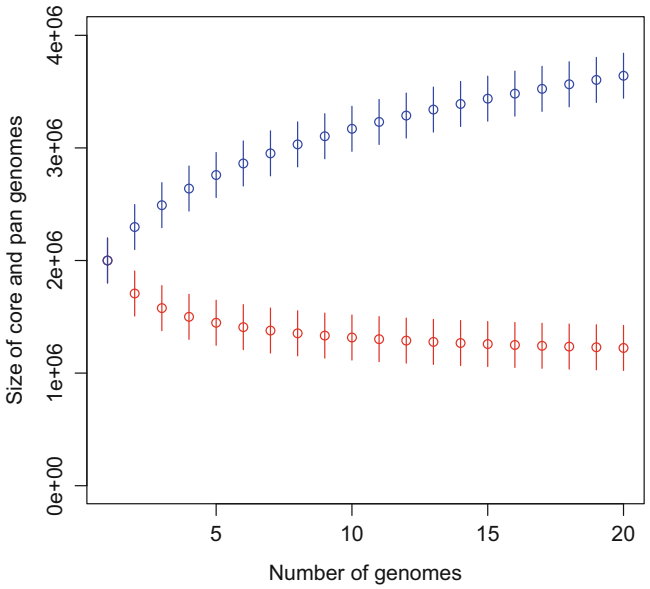

Fig. 8 Example of accumulation curves for the core genome (red) and the pan-genome (blue). The x-axis shows the number of genomes being considered, and the y-axis shows the length of the genomic regions found in all (core, red) or at least one (pan, blue) of the genomes

related measure is the genome fluidity index, which is equal to the average pairwise gene content differences between strains (Kislyuk et al. 2011). For example, a fluidity value of 0.2 indicates that on average any two strains within the population under study share 80% of their genes.

7.2 Inference of Gains and Losses of Genomic Regions

If genes (or genomic regions) were gained and lost in a clock-like manner in the same way as mutations arise along the genome, a UPGMA dendrogram based on the presence/absence of genes should look similar to a phylogenetic tree reconstructed from nucleotide polymorphism data, and these two trees are sometimes plotted next to each other to test this hypothesis (Didelot et al. 2012b; Chaudhari et al. 2016). Similarly, pairwise distances between genomes can be compared when measured in terms of gene content similarity versus homology of the core genome (Wielgoss et al. 2016). To visualise how the presence of specific genes relates to the ancestral relationships between genomes, it is common to plot a matrix of gene presence/absence side-by-side with a phylogenetic tree (Croucher et al. 2014). Beyond simple graphical representation, the presence/absence data can be interpreted in terms of gain/loss of genes using the same discrete trait analysis

methods described in Sect. 5.2. This approach was used for example to investigate gene content evolution in *Escherichia coli* (Touchon et al. 2009), and the gain and loss of antibiotic resistance genes and virulence determinants in *Staphylococcus aureus* (Ward et al. 2014). Since genes and genetic regions are often gained and lost in groups, for example through the integration of phage in the genome, the gain of a plasmid or integrative conjugational elements, it can be important to relax the assumption of a fixed rate for the gain and loss events along the branches of the phylogeny, and significant temporal and lineage-specific variations in these rates have been demonstrated in *Francisella turalensis*, *Streptococcus pyogenes* and *Escherichia coli* (Didelot et al. 2009a). The genetic elements known to be generally gained and lost in one unit can also be treated as the unit of gain and loss along the branches when performing the discrete trait analysis. This approach was used for example to reconstruct the gain and loss events of bacteriophages, plasmids and integrative conjugational elements in *Salmonella enterica* serovar Agona (Zhou et al. 2013).

References

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64. https://doi.org/10.1101/gr.094052.109.

Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.

Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics. 2010;27:334–42.

Ansari MA, Didelot X. Bayesian inference of the evolution of a phenotype distribution on a phylogenetic tree. Genetics. 2016;204:89–98. https://doi.org/10.1101/040980.

Argimón S, Abudahab K, Goater RJE, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genomics. 2016;2:e000093. https://doi.org/10.1099/mgen.0.000093.

Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006;7:781–91. https://doi.org/10.1038/nrg1916.

Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77. https://doi.org/10.1089/cmb.2012.0021.

Baum DA, Smith SD, Donovan SSS. The tree-thinking challenge. Science. 2005;310:979–80. https://doi.org/10.1126/science.1117727.

Biek R, Henderson JC, Waller LA, et al. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. Proc Natl Acad Sci U S A. 2007;104:7993–8. https://doi.org/10.1073/pnas.0700741104.

Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. Measurably evolving pathogens in the genomic era. Trends Ecol Evol. 2015;30:306–13. https://doi.org/10.1016/j.tree.2015.03.009.

Bielejec F, Rambaut A, Suchard MA, Lemey P. SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. Bioinformatics. 2011;27:2910–2. https://doi.org/10.1093/bioinformatics/btr481.

Bielejec F, Baele G, Vrancken B, et al. SpreaD3: interactive visualization of spatiotemporal history and trait evolutionary processes. Mol Biol Evol. 2016;33:2167–9. https://doi.org/10.1093/molbev/msw082.

Bloomquist EWEEW, Dorman KSKSK, Suchard MA. StepBrothers: inferring partially shared ancestries among recombinant viral sequences. Biostatistics. 2009;10:106–20. https://doi.org/10.1093/biostatistics/kxn019.

Bloomquist EW, Lemey P, Suchard MA. Three roads diverged? Routes to phylogeographic inference. Trends Ecol Evol. 2010;25:626–32. https://doi.org/10.1016/j.tree.2010.08.010.

Bouckaert R, Heled J, Kühnert D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014;10:e1003537. https://doi.org/10.1371/journal.pcbi.1003537.

Brown T, Didelot X, Wilson DJ, De Maio N. SimBac: simulation of whole bacterial genomes with homologous recombination. Microb Genomics. 2016;2. https://doi.org/10.1099/mgen.0.000044.

Castillo-Ramírez S, Corander J, Marttinen P, et al. Phylogeographic variation in recombination rates within a global clone of methicillin-resistant Staphylococcus aureus. Genome Biol. 2012;13:R126. https://doi.org/10.1186/gb-2012-13-12-r126.

Chaudhari NM, Gupta VK, Dutta C. BPGA- an ultra-fast pan-genome analysis pipeline. Sci Rep. 2016;6:24373. https://doi.org/10.1038/srep24373.

Chewapreecha C, Harris SR, Croucher NJ, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014;46:305–9. https://doi.org/10.1038/ng.2895.

Chin CS, Sorenson J, Harris JB, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med. 2011;364:33–42.

Cohan FM, Perry EB. A systematics for discovering the fundamental units of bacterial diversity. Curr Biol. 2007;17:R373–86. https://doi.org/10.1016/j.cub.2007.03.032.

Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. bioRxiv. 2017. https://doi.org/10.1101/140798.

Comas I, Coscolla M, Luo T, et al. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet. 2013;45:1176–82. https://doi.org/10.1038/ng.2744.

Croucher NJ, Didelot X. The application of genomics to tracing bacterial pathogen transmission. Curr Opin Microbiol. 2015;23:62–7. https://doi.org/10.1016/j.mib.2014.11.004.

Croucher NJ, Harris SRR, Fraser C, et al. Rapid pneumococcal evolution in response to clinical interventions. Science. 2011;331:430–4. https://doi.org/10.1126/science.1198545.

Croucher NJ, Coupland PG, Stevenson AE, et al. Diversification of bacterial genome content through distinct mechanisms over different timescales. Nat Commun. 2014;5:5471. https://doi.org/10.1038/ncomms6471.

Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43:e15. https://doi.org/10.1093/nar/gku1196.

Cui Y, Yu C, Yan Y, et al. Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. Proc Natl Acad Sci U S A. 2013;110:577–82. https://doi.org/10.1073/pnas.1205750110.

Cui Y, Yang X, Didelot X, et al. Epidemic clones, oceanic gene pools and eco-LD in the free living marine pathogen Vibrio parahaemolyticus. Mol Biol Evol. 2015;32:1396–410. https://doi.org/10.1093/molbev/msv009.

Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5:e11147. https://doi.org/10.1371/journal.pone.0011147.

De Maio N, C-H W, O'Reilly KM, Wilson D. New routes to phylogeography: a Bayesian structured coalescent approximation. PLoS Genet. 2015;11:e1005421. https://doi.org/10.1371/journal.pgen.1005421.

De Silva D, Peters J, Cole K, et al. Whole-genome sequencing to determine transmission of Neisseria gonorrhoeae: an observational study. Lancet Infect Dis. 2016;16:1295–303. https://doi.org/10.1016/S1473-3099(16)30157-8.

Dearlove BL, Cody AJ, Pascoe B, et al. Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. ISME J. 2015;10(3):721–9. https://doi.org/10.1038/ismej.2015.149.

Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. Genetics. 2007;175:1251–66. https://doi.org/10.1534/genetics.106.063305.

Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 2015;11:e1004041. https://doi.org/10.1371/journal.pcbi.1004041.

Didelot X, Darling AE, Falush D. Inferring genomic flux in bacteria. Genome Res. 2009a;19:306–17. https://doi.org/10.1101/gr.082263.108.clearly.

Didelot X, Lawson DJ, Falush D. SimMLST: simulation of multi-locus sequence typing data under a neutral model. Bioinformatics. 2009b;25:1442–4. https://doi.org/10.1093/bioinformatics/btp145.

Didelot X, Lawson DJ, Darling AE, Falush D. Inference of homologous recombination in bacteria using whole-genome sequences. Genetics. 2010;186:1435–49. https://doi.org/10.1534/genetics.110.120121.

Didelot X, Eyre DW, Cule M, et al. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Genome Biol. 2012a;13:R118. https://doi.org/10.1186/gb-2012-13-12-r118.

Didelot X, Méric G, Falush D, Darling AE. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. BMC Genomics. 2012b;13:256. https://doi.org/10.1186/1471-2164-13-256.

Didelot X, Pang B, Zhou Z, et al. The role of China in the global spread of the current cholera pandemic. PLoS Genet. 2015;11:e1005072. https://doi.org/10.1371/journal.pgen.1005072.

Didelot X, Walker AS, Peto TE, et al. Within-host evolution of bacterial pathogens. Nat Rev Microbiol. 2016;14:150–62. https://doi.org/10.1038/nrmicro.2015.13.

Dingle KE, Elliott B, Robinson E, et al. Evolutionary history of the clostridium difficile pathogenicity locus. Genome Biol Evol. 2014;6:36–52. https://doi.org/10.1093/gbe/evt204.

Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007;7:214. https://doi.org/10.1186/1471-2148-7-214.

Drummond AJ, Suchard MA. Bayesian random local clocks, or one rate to rule them all. BMC Biol. 2010;8:114. https://doi.org/10.1186/1741-7007-8-114.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol. 2005;22:1185–92. https://doi.org/10.1093/molbev/msi103.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006;4:e88. https://doi.org/10.1371/journal.pbio.0040088.

Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29:1969–73. https://doi.org/10.1093/molbev/mss075.

Earle SG, Wu C, Charlesworth J, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol. 2016;1:16041. https://doi.org/10.1038/nmicrobiol.2016.41.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7. https://doi.org/10.1093/nar/gkh340.

Excoffier L, Foll M. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics. 2011;27:1332–4. https://doi.org/10.1093/bioinformatics/btr124.

Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164:1567–87.

Faria NR, Suchard MA, Rambaut A, et al. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. Philos Trans R Soc Lond Ser B Biol Sci. 2013;368:20120196. https://doi.org/10.1098/rstb.2012.0196.

Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17:368–76. https://doi.org/10.1007/BF01734359.

Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Syst Biol. 1985;39:783–91.

Fitch WM. Toward defining the course of evolution: minimum change for a specific tree topology. Syst Biol. 1971;20:406–16. https://doi.org/10.1093/sysbio/20.4.406.

Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. Prepr arXiv:1207.3907 [q-bio.GN]. 2012; 9.

Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 1997;14:685–95. https://doi.org/10.1093/oxfordjournals.molbev.a025808.

Gire SK, Goba A, Andersen KG, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345:1369–72. https://doi.org/10.1126/science.1259657.

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51. https://doi.org/10.1038/nrg.2016.49.

Grad YH, Kirkcaldy RD, Trees D, et al. Genomic epidemiology of Neisseria gonorrhoeae with reduced susceptibility to cefixime in the USA: a retrospective observational study. Lancet Infect Dis. 2014;14:220–6. https://doi.org/10.1016/S1473-3099(13)70693-5.

Griffiths R, Tavare S. Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc B Biol Sci. 1994;344:403–10.

Guindon S, Dufayard J-F, Lefort V, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21. https://doi.org/10.1093/sysbio/syq010.

Haase JK, Didelot X, Lecuit M, et al. The ubiquitous nature of Listeria monocytogenes clones: a large scale MultiLocus sequence typing study. Environ Microbiol. 2014;16:405–16. https://doi.org/10.1111/1462-2920.12342.

Harris SRR, Feil EJ, Holden MT, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010;327:469–74. https://doi.org/10.1126/science.1182395.

Harris SR, Clarke IN, Seth-Smith HMB, et al. Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing. Nat Genet. 2012;44:413–9. https://doi.org/10.1038/ng.2214.

He M, Miyajima F, Roberts P, et al. Emergence and global spread of epidemic healthcare-associated Clostridium difficile. Nat Genet. 2013;45:109–13. https://doi.org/10.1038/ng.2478.

Hedge J, Wilson J. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. MBio. 2014;5:e02158–14. https://doi.org/10.1128/mBio.02158-14.Editor.

Hellenthal G, Stephens M. msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. Bioinformatics. 2007;23:520–1. https://doi.org/10.1093/bioinformatics/btl622.

Höhna MJ, et al. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst Biol. 2016;65:726–36.

Holt KE, Baker S, Weill F-X, et al. Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. Nat Genet. 2012;44:1056–9. https://doi.org/10.1038/ng.2369.

Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002;18:337–8. https://doi.org/10.1093/bioinformatics/18.2.337.

Hunt DEDE, David LA, Gevers D, et al. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science. 2008;320(5879):1081–5. https://doi.org/10.1126/science.1157890.

Hyatt D, Chen G-L, Locascio PF, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119. https://doi.org/10.1186/1471-2105-11-119.

Ingle DJ, Tauschek M, Edwards DJ, et al. Evolution of atypical enteropathogenic E. coli by repeated acquisition of LEE pathogenicity island variants. Nat Microbiol. 2016;1:15010. https://doi.org/10.1038/nmicrobiol.2015.10.

Jolley KAA, Maiden MCJ. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010;11:595. https://doi.org/10.1186/1471-2105-11-595.

Joseph SJ, Didelot X, Gandhi K, et al. Interplay of recombination and selection in the genomes of Chlamydia trachomatis. Biol Direct. 2011;6:28. https://doi.org/10.1186/1745-6150-6-28.

Joseph SJ, Didelot X, Rothschild J, et al. Population genomics of chlamydia trachomatis: insights on drift, selection, recombination and population structure. Mol Biol Evol. 2012;29:3933–46. https://doi.org/10.1093/molbev/mss198.

Joy JB, Liang RH, Mccloskey RM, et al. Ancestral reconstruction. PLoS Comput Biol. 2016;12: e1004763. https://doi.org/10.1371/journal.pcbi.1004763.

Kennemann L, Didelot X, Aebischer T, et al. Helicobacter pylori genome evolution during human infection. Proc Natl Acad Sci U S A. 2011;108:5033–8. https://doi.org/10.1073/pnas.1018444108.

Kingman JFC. The coalescent. Stoch Process their Appl. 1982;13:235–48. https://doi.org/10.1016/0304-4149(82)90011-4.

Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics. 2011;12:32. https://doi.org/10.1186/1471-2164-12-32.

Kurtz S, Phillippy A, Delcher AL, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12. https://doi.org/10.1186/gb-2004-5-2-r12.

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:R25. https://doi.org/10.1186/gb-2009-10-3-r25.

Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. Trends Genet. 2009;25:107–10. https://doi.org/10.1002/9781118314630.ch15.

Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8:e1002453. https://doi.org/10.1371/journal.pgen.1002453.

Lees JA, Vehkala M, Välimäki N, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun. 2016;7:12797. https://doi.org/10.1101/038463.

Lemey P, Rambaut A, Drummond AJ, Suchard M. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009;5:e1000520. https://doi.org/10.1371/journal.pcbi.1000520.

Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. Mol Biol Evol. 2010;27:1877–85. https://doi.org/10.1093/molbev/msq067.

Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.

Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;165:2213–33. https://doi.org/10.1534/genetics.104.030692.

Li L, Stoeckert CJJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89. https://doi.org/10.1101/gr.1224503.candidates.

Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009;25:2078–9. https://doi.org/10.1093/bioinformatics/btp352.

Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. Nat Rev Microbiol. 2015;13 (12):787–94. https://doi.org/10.1038/nrmicro3565.

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.tions.

Maiden MC, Bygraves JA, Feil EJ, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998;95:3140–5.

Marin JMJ, Pudlo P, Robert CPCP, Ryder R. Approximate Bayesian computational methods. Stat Comput. 2012;22:1167–80.

Martin DP, Murrell B, Golden M, et al. RDP4: detection and analysis of recombination patterns in virus genomes. Virus Evol. 2015;1:vev003. https://doi.org/10.1093/ve/vev003.

Marttinen P, Hanage WP, Croucher NJ, et al. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res. 2012;40:1–12. https://doi.org/10.1093/nar/gkr928.

McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
McNally A, Oren Y, Kelly D, et al. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. PLoS Genet. 2016;12:e1006280. https://doi.org/10.5061/dryad.d7d71.
Medini D, Donati C, Tettelin H, et al. The microbial pan-genome. Curr Opin Genet Dev. 2005;15:589–94. https://doi.org/10.1016/j.gde.2005.09.006.
Milne I, Wright F, Rowe G, et al. TOPALi: software for automatic identification of recombinant sequences within DNA multiple alignments. Bioinformatics. 2004;20:1806–7. https://doi.org/10.1093/bioinformatics/bth155.
Milne I, Lindner D, Bayer M, et al. TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. Bioinformatics. 2009;25:126–7. https://doi.org/10.1093/bioinformatics/btn575.
Mutreja A, Kim DW, Thomson NR, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature. 2011;477:462–5. https://doi.org/10.1038/nature10392.
Nagarajan N, Kingsford C. GiRaF: robust, computational identification of influenza reassortments via graph mining. Nucleic Acids Res. 2011;39:e34. https://doi.org/10.1093/nar/gkq1232.
Nübel U, Dordel J, Kurt K, et al. A timescale for evolution, population expansion, and spatial spread of an emerging clone of methicillin-resistant Staphylococcus aureus. PLoS Pathog. 2010;6:e1000855. https://doi.org/10.1371/journal.ppat.1000855.
Overbeek R, Olson R, Pusch GD, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014;42:206–14. https://doi.org/10.1093/nar/gkt1226.
Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3. https://doi.org/10.1093/bioinformatics/btv421.
Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289–90. https://doi.org/10.1093/bioinformatics/btg412.
Peng Y, et al. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28:1420–8.
Pond SLK, Posada D, Gravenor MB, et al. Sequence analysis GARD: a genetic algorithm for recombination detection. Bioinformatics. 2006;22:3096–8. https://doi.org/10.1093/bioinformatics/btl474.
Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26:1641–50. https://doi.org/10.1093/molbev/msp077.
Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490. https://doi.org/10.1371/journal.pone.0009490.
Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.
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. https://doi.org/10.1086/519795.
Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016;2:vew007. https://doi.org/10.1093/ve/vew007.
Ratmann O, Hodcroft EB, Pickles M, et al. Phylogenetic tools for generalized HIV-1 epidemics: findings from the PANGEA-HIV methods comparison. Mol Biol Evol. 2017;34:185–203. https://doi.org/10.1093/molbev/msw217.
Read TD, Massey RC. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. Genome Med. 2014;6:109. https://doi.org/10.1186/s13073-014-0109-z.

Ronquist F, Teslenko M, van der Mark P, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61:539–42. https://doi.org/10.1093/sysbio/sys029.

Sahl JW, Caporaso JG, Rasko DA, Keim P. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. PeerJ. 2014;2:e332. https://doi.org/10.7717/peerj.332.

Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4:406–25.

Schierup MH, Hein J. Consequences of recombination on traditional phylogenetic analysis. Genetics. 2000;156:879–91.

Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9. https://doi.org/10.1093/bioinformatics/btu153.

Shepheard MA, Fleming VM, Connor TR, et al. Historical zoonoses and other changes in host tropism of staphylococcus aureus, identified by phylogenetic analysis of a population dataset. PLoS One. 2013;8:e62369. https://doi.org/10.1371/journal.pone.0062369.

Sheppard SK, Didelot X, Jolley KA, et al. Progressive genome-wide introgression in agricultural Campylobacter coli. Mol Ecol. 2013a;22:1051–64. https://doi.org/10.1111/mec.12162.

Sheppard SK, Didelot X, Meric G, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci U S A. 2013b;110:11923–7. https://doi.org/10.5061/dryad.28n35.

Smith GJD, Vijaykrishna D, Bahl J, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 2009;459:1122–5. https://doi.org/10.1038/nature08182.

Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22:2688–90. https://doi.org/10.1093/bioinformatics/btl446.

Stoesser N, Sheppard A, Pankhurst L, et al. Evolutionary history of the global emergence of the Escherichia coli epidemic clone ST131. MBio. 2016;7:e02162–15. https://doi.org/10.1128/mBio.02162-15.Invited.

Tang J, Hanage WP, Fraser C, Corander J. Identifying currents in the gene pool for bacterial populations using an integrative approach. PLoS Comput Biol. 2009;5:e1000455. https://doi.org/10.1371/journal.pcbi.1000455.

Tettelin H, Masignani V, Cieslewicz MJ, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 2005;102:13950–5.

Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 2008;12:472–7. https://doi.org/10.1016/j.mib.2008.09.006.

To T-H, Jung M, Lycett S, Gascuel O. Fast dating using least-squares criteria and algorithms. Syst Biol. 2016;65:82–97. https://doi.org/10.1093/sysbio/syv068.

Touchon M, Hoede C, Tenaillon O, et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 2009;5:e1000344. https://doi.org/10.1371/journal.pgen.1000344.

Tritt A, Eisen JA, Facciotti MT, Darling AE. An integrated pipeline for de novo assembly of microbial genomes. PLoS One. 2012;7:e42304. https://doi.org/10.1371/journal.pone.0042304.

Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin Microbiol. 2015;23:148–54. https://doi.org/10.1016/j.mib.2014.11.016.

Visscher PM, Hill WG, Wray NR. Heritability in the genomics era – concepts and misconceptions. Nat Rev Genet. 2008;9:255–66. https://doi.org/10.1038/nrg2322.

Ward MJ, Gibbons CL, McAdam PR, et al. Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of Staphylococcus aureus clonal complex 398. Appl Environ Microbiol. 2014;80:7275–82. https://doi.org/10.1128/AEM.01777-14.

Weinert LA, Chaudhuri RR, Wang J, et al. Genomic signatures of human and animal disease in the zoonotic pathogen Streptococcus suis. Nat Commun. 2015;6:6740. https://doi.org/10.1038/ncomms7740.

Wielgoss S, Didelot X, Chaudhuri RR, et al. A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium Myxococcus xanthus. ISME J. 2016;10:2468–77. https://doi.org/10.1038/ismej.2016.34.

Worobey M, Gemmel M, Teuwen DE, et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature. 2008;455:661–4. https://doi.org/10.1038/nature07390.

Yahara K, Furuta Y, Oshima K, et al. Chromosome painting in silico in a bacterial species reveals fine population structure. Mol Biol Evol. 2013;30:1454–64. https://doi.org/10.1093/molbev/mst055.

Yahara K, Didelot X, Ansari MA, et al. Efficient inference of recombination hot regions in bacterial genomes. Mol Biol Evol. 2014;31:1593–605. https://doi.org/10.1093/molbev/msu082.

Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9. https://doi.org/10.1101/gr.074492.107.

Zhou Z, McCann A, Litrup E, et al. Neutral genomic microevolution of a recently emerged pathogen, Salmonella enterica serovar Agona. PLoS Genet. 2013;9:e1003471. https://doi.org/10.1371/journal.pgen.1003471.

Zinder D, Bedford T, Gupta S, Pascual M. The roles of competition and mutation in shaping antigenic and genetic diversity in influenza. PLoS Pathog. 2013;9:e1003104. https://doi.org/10.1371/journal.ppat.1003104.

What Microbial Population Genomics Has Taught Us About Speciation

B. Jesse Shapiro

Abstract Population genomics has emerged as a valuable tool to define and delimit species and to understand the mechanisms that drive and maintain speciation. Species and speciation have been notoriously difficult to study in microbes owing to their asexual reproduction, promiscuous horizontal gene transfer, and obscure microscopic niches. Over the past few years, whole-genome sequencing of closely related, locally co-occurring populations of microbes, combined with simulations and modelling, has revealed certain general features of microbial speciation: it is usually driven by divergent natural selection between distinct ecological niches (a form of the ecological species concept), and species distinctness is maintained by barriers to gene flow (a form of the biological species concept). In some cases, gene-flow barriers may come about as a natural consequence of ecological specialization. Although these features appear to be quite general, there are exceptions. Trivially, barriers to gene flow cannot be used to delimit clonal populations where there is negligible gene flow. More interestingly, it is unclear whether other barriers to gene flow, such as genetic incompatibilities or differences in phage-host range, are able to drive speciation in the absence of other selective pressures. Here, I discuss the extent to which speciation is driven by natural selection, gene-flow barriers, or a combination of the two, drawing on recent examples from bacterial and archaeal population genomics, experimental evolution, and modelling. I then describe how population genomic data can be used to define and delimit species boundaries, based upon nucleotide identity cutoffs or upon discontinuities in gene flow. Despite important limitations and caveats, delimitation methods provide a useful starting point for more detailed investigation into the genetic and ecological basis of speciation.

Keywords Archaea · Bacteria · Biological species concept · Ecological species concept · ecoSNP · Gene flow · Mosaic sympatry · Niche · Overlapping Habitat Model · Speciation

B. Jesse Shapiro (✉)
Department of Biological Sciences, University of Montreal, Montreal, QC, Canada
e-mail: jesse.shapiro@umontreal.ca

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_10,

1 Introduction

Over 150 years since Darwin published *On the origin of species*, biologists and philosophers are still debating what species are, how they form, and if they really exist (Doolittle and Zhaxybayeva 2009; Doolittle 2012). I have previously argued that species do exist and their origin (the process of speciation) is generally, if not always, driven by natural selection for adaptation to distinct ecological niches (Shapiro et al. 2016). Here, I will critically re-evaluate this argument and discuss alternatives, drawing on the most recent advances from population genomics. Most of the examples will be from bacteria, with some comparisons across other domains of life. Building on the observation that genetically and ecologically coherent units do exist (Caro-Quintero and Konstantinidis 2011; Shapiro and Polz 2014) even if their boundaries may be "fuzzy" (Hanage et al. 2005; Hanage 2013), I will focus on the mechanisms that give rise to these units and keep them distinct. In other words, this chapter is mainly about speciation, not species. However, I will also discuss methods to define and delimit species, which can provide a practical first step toward better understanding of the mechanisms driving speciation.

2 Species Concepts and Definitions

To begin, let us briefly define population genomics and make the distinction between species concepts and definitions. Species *concepts* require at least some notion of mechanism, whereas species *definitions* can be completely operational and agnostic to mechanism but can also be based on a particular species concept (Gevers et al. 2005). I will focus on two popular types of species concepts. The ecological species concept (ESC), favoured by Darwin, posits that speciation is driven by natural selection, with each species adapted to a unique ecological niche (Schluter 2009). The biological species concept (BSC) posits that speciation is driven by barriers to genetic exchange, which is equivalent to reproductive isolation in sexual species (Dobzhansky 1935; Mayr 1942). Strictly speaking, the BSC will never apply to asexual organisms like bacteria. Moreover, bacteria (and other domains of life, including plants and animals) can exchange genes across species boundaries, so barriers to gene flow will always remain somewhat permeable (Shapiro et al. 2016). Therefore, rather than the strict BSC, I will refer mainly to a *BSC-like* concept in which rates of gene flow are higher within than between species, but cross-species gene transfer can still occur. Other species concepts exist, but most are effectively combinations of the ESC and the BSC. For example, the stable ecotype model is essentially the ESC with relatively low rates of genetic exchange (Wiedenbeck and Cohan 2011). Allopatric speciation is a special case of the BSC in which barriers to genetic exchange are initially due to physical isolation, although they can later be reinforced by genetic incompatibilities. Different species concepts predict different and distinctive patterns of genetic variation within and between species (Krause and Whitaker 2015), which can in principle be harnessed to define species.

Population genomics is a valuable tool—perhaps the most valuable tool available—to both inform our concept of species and to precisely define species. The (relatively new) field of population genomics (see the Chapters in this volume on this topic) uses whole-genome information to answer questions posed by the (more mature) field of population genetics—the study of how mutation, selection, and drift change allele frequencies within a population. Populations are generally defined as sets of locally coexisting members of species. If we do not know what species are in the first place, or how to define them, the task of defining species and populations can become circular. Therefore, the application of population genomics to the study of species and speciation usually requires some a priori notion of species or population boundaries, which can then be critically evaluated based on the fit of observed patterns of genomic variation with the predictions of competing species concepts. In some cases, the prior information can include ecological hypotheses, for example, that speciation in marine vibrios is driven by adaptation to either free-living or particle-associated lifestyles (Shapiro et al. 2012). In other cases, a previously named species or genus might be sampled to test whether genome sequence data fits a particular species concept and whether the sampled genomes constitute one or many species (Cadillo-Quiroz et al. 2012; Bobay and Ochman 2017). In general, population genomics requires complete or near-complete genome sequences from several individuals, be they cultured isolated or single-cell genomes. Metagenomic sequencing of bulk DNA from an environment is usually incapable of linking particular genes or mutations back to a specific individual, making it more difficult to test certain species concepts, particularly versions of the BSC that require testing for differences in recombination rates within and between populations. These shortcomings have not prevented researchers from defining "metagenomic species", although such definitions are purely operational and not clearly grounded in any particular concept of species other than the prediction that members of the same species should have correlated abundances over time or across samples (Caro-Quintero and Konstantinidis 2011; Alneberg et al. 2014; Nielsen et al. 2014). Nevertheless, metagenomics can help estimate valuable population genetic parameters such as the nucleotide diversity within a species.

3 Selection, Gene-Flow Barriers, or Both?

Both natural selection and barriers to gene flow can be important in the speciation process, but which is usually the driver that initiates speciation? Certain forms of gene flow, namely, homologous recombination, require a certain degree of sequence identity between donor DNA and the recipient genome (although a few dozen base pairs of identity can be sufficient to initiate the transfer of several kilobases of completely nonhomologous DNA; Mell et al. 2011; Croucher et al. 2012). In principle, the accumulation of mutations could gradually create barriers to homologous recombination, driving speciation in the absence of selection and yielding genetically distinct species fitting the BSC. According to computational modelling,

this is unlikely to occur, unless recombination rates decline unrealistically rapidly with sequence divergence (Fraser et al. 2007). The model suggests that another force—such as divergent natural selection between two niches—is required to drive speciation. A further theoretical argument why selection is required to initiate speciation is based on the competitive exclusion principle (Gause 1934; Tilman 1982). If two species are ecologically equivalent (meaning they are under identical or near-identical regimes of selection), one will inevitably (after some period of time) drive the other to extinction. Only if species are under divergent selection for adaptation to distinct niches will speciation occur.

Beyond these theoretical considerations, what is the population genomic evidence for selection driving speciation? Perhaps the most direct evidence comes from laboratory evolution experiments, combined with whole-genome sequencing. In a long-term evolution experiment starting with a single clone of *E. coli*, a lineage evolved after ~31,000 generations with the ability to metabolize citrate, a previously unused carbon source present in the growth medium (Blount et al. 2008). Sequencing of ancestral CIT- and derived CIT+ genomes revealed the genetic changes required for citrate utilization (Blount et al. 2012). The two ecologically distinct lineages continue to coexist in the experiment, consistent with the ESC. Despite being a clear example of how ecological selection can drive speciation, it is not really a fair test of whether gene-flow barriers can drive speciation because the *E. coli* in the experiment are not competent and do not recombine DNA.

In another evolution experiment using bacteriophage capable of recombination within host cells, Meyer et al. (2016) showed that speciation readily occurred under both allopatric and sympatric conditions, driven by divergent selection for phage to specialize on one of two available bacterial hosts that differed only in their surface phage receptor. In the allopatric experiment, phage were cultured in media containing only one host, and specialization occurred rapidly. In the sympatric experiment, both bacterial hosts were present in the culture media, but specialization still occurred because of the link between ecological preference (one host or the other) and recombination, which only occurs within a host cell. These barriers to gene flow imposed by host preference are analogous to the barriers imposed by particle preference within the marine water column, which appears to be driving sympatric speciation in natural vibrio populations (discussed below). This subtle spatial structure within seemingly homogeneous sympatric environments has been referred to as "mosaic sympatry" (Mallet 2008; Shapiro and Polz 2014) and explains how ecological selection can initiate speciation, which is later reinforced by gene-flow barriers. By sequencing evolved and ancestral phage genomes, Meyer et al. (2016) further showed that several mutations in a single host-recognition gene in the phage genome (*J*) explained host specialization, with different mutations associated with different hosts. The observation of a single gene apparently responsible for speciation is consistent with theoretical predictions that sympatric speciation proceeds more readily when fewer loci are involved in ecological differentiation or reproductive isolations (Kondrashov and Mina 1986; Friedman et al. 2013). Further reducing gene flow between incipient phage species, recombinant *J* alleles encoding combinations of mutations adapted to different hosts were not viable. Therefore, Meyer et al. (2016) appear to have captured a very early stage of sympatric speciation, driven by ecological differentiation and maintained by gene-flow

barriers. A population genomic study of sympatric marine cyanophages suggests the same mechanisms may be at play in natural phage populations, although speciation may be driven by ecological factors other than host identity (Gregory et al. 2016).

Similar patterns have also been observed in recombining natural bacterial populations. For example, we compared the genomes of very closely related *Vibrio cyclitrophicus* isolates (identical 16S and >99% amino acid identity) and concluded that speciation was driven by differential selection for either free-living or particle-associated niches and maintained by the emergence of barriers to gene flow (Shapiro et al. 2012). In other words, the speciation process began with an ESC-like mechanism and was reinforced by a BSC-like mechanism. However, it is difficult to be certain that ecological selection *preceded* the establishment of gene-flow barriers. We found that gene-flow barriers between incipient species are only evident among the most recent detectable recombination events, while older recombination events do not respect species boundaries (Shapiro et al. 2012). I later used an adaptation of the McDonald-Kreitman (MK) test (Vos 2011) to show that the divergence between incipient species involved an unexpected excess of nonsynonymous substitutions, suggesting positive selection driving their divergence (Shapiro 2014). Still, although it is certainly consistent with the "selection first" hypothesis, this does not conclusively prove that ecological selection occurred before the establishment of gene-flow boundaries. Further complicating things, the likely targets of differential selection between free-living and particle-associated habitats—three loci containing >80% of ecoSNPs (the single nucleotide polymorphisms fixed between habitats) and several other genes present in one habitat but not the other—are subject to frequent recombination and were likely acquired from distantly related lineages of *Vibrio*, making it difficult to date their acquisition with certainty. Nevertheless, it is abundantly clear that the two incipient species are ecologically distinct (Yawata et al. 2014) and there is currently no evidence suggesting that gene-flow boundaries emerged before differential selection.

Evidence from several other natural bacterial populations supports the idea that ecological differentiation, due to selection on one or a few "niche-specifying" genes, can occur before any apparent boundaries to gene flow. For example, a population genomic study of *Rhizobium leguminosarum* found that they "form dynamic, diverse populations that are unified by gene flow despite selection acting at one or more loci" (Klinger et al. 2016). Specifically, they found that selection (artificially applied in a 22-year nitrogen fertilization experiment) favoured certain alleles of nitrogen fixation genes, which rose to high frequency in the *R. leguminosarum* population without affecting diversity elsewhere in the genome (Klinger et al. 2016). Such "gene-specific" selective sweeps (Shapiro and Polz 2014) have also been documented in population genomic studies of other bacteria, including *Mesorhizobium* (Porter et al. 2016) and *Streptococcus* (Croucher et al. 2011; Bao et al. 2016). The apparent ease with which natural selection can favour the increase of adaptive genes or alleles in recombining microbial populations suggests that selection could at least plausibly drive speciation, before the establishment of gene-flow boundaries.

Let us now consider the alternative hypothesis that gene-flow barriers directly drive speciation without the need for ecological selection—a version of the BSC without any trace of the ESC. As described above, it is unlikely that gradual mutation accumulation

could cause barriers to homologous recombination. But what about other mechanisms of recombination? Phage-mediated transduction requires the donor and recipient cells of a recombination event to be infected by the same phage. Therefore barriers to phage infection could limit gene flow. Consistent with this idea, a comparative analysis of phage and bacterial genome sequences showed that phage-mediated recombination events are mostly limited to closely related bacterial donors and recipients (Popa et al. 2016). This phage-host specificity could limit genetic exchange to close relatives, providing a natural mechanism for the BSC and leading to more genetic exchange within than between species. In principle, a mutation or recombination event changing a phage receptor could instantaneously create a barrier to gene flow (Rodriguez-Valera et al. 2009; López-Pérez and Rodriguez-Valera 2016), but population genomic evidence of such a BSC-like mechanism driving speciation is still lacking. Large-scale chromosomal rearrangements can play an important role in creating reproductive isolation in yeast (Charron et al. 2014; Leducq et al. 2016), but it is unclear which came first—barriers to gene flow or ecological specialization—or whether both occurred more or less simultaneously to initiate speciation.

4 Models to Interpret Population Genomic Data

Population genomic data can be used to operationally define species and, more importantly, to test competing species concepts. An example of an operational species definition based on genome sequence data is the proposed 95% average nucleotide identity (ANI) threshold (Konstantinidis and Tiedje 2005; Konstantinidis et al. 2006). Pairs of genomes that have below 95% ANI always come from distinct species, according to most species concepts or definitions. However, although a 95% threshold may work well for most species, some recently diverged species might still share 97, 98, or 99% ANI (Doolittle and Zhaxybayeva 2009). For example, the nascent phage (Meyer et al. 2016) and *Vibrio* (Shapiro et al. 2012) species described above would be lumped into a single species using a 95% cutoff. ANI may also vary widely across the genome, leading to "fragmented speciation" in which different parts of the genome effectively speciate at different rates (Retchless and Lawrence 2010). Therefore, a universal ANI-based species definition, while appealing in its simplicity, will likely fail to distinguish "good" species, especially at early stages of speciation. ANI, like other sequence-based thresholds (such as 97% identity in the 16S rRNA gene), is still a useful starting point for a more in-depth testing of species concepts. It has been argued that 97% is a much too inclusive cutoff and that 99% 16S identity or unique sequence types better capture ecologically coherent bacterial species (Acinas et al. 2004; Eren et al. 2013; Koeppel and Wu 2014). No one would argue that genomes sharing less than 95% ANI are part of the same species. However, genomes sharing more than 95% ANI might be divided into two, three, or several species, depending on the choice of species concept.

Testing species concepts requires more than population genomic data. It also requires a model describing the mechanism of speciation, which can then be fit to

population genomic data. One of the first and most influential such models is the stable ecotype model (SEM), which defines species as ecotypes, each inhabiting a distinct ecological niche, such that selective sweeps and neutral drift affect diversity within but not between species (Wiedenbeck and Cohan 2011). In other words, selective sweeps or population bottlenecks that occur within one species (ecotype) do not affect the genetic diversity of other species. Phylogenies based on marker genes often fit well with the predictions of the SEM, namely, that monophyletic groups of closely related bacteria tend to share the same ecological associations (Hunt et al. 2008; Koeppel et al. 2008). However, applied to marker gene sequences from natural populations of *Bacillus*, the SEM fits slightly worse than a neutral model without ecological niches (Fraser et al. 2009), and patterns that appear consistent with the SEM based on marker genes may be inconsistent when genome-wide information is considered (Shapiro et al. 2012).

In the SEM, sweeps or bottlenecks purge genetic diversity genome-wide, because recombination is not strong enough to decouple the evolutionary fates of loci across the genome. Different versions of the SEM can accept increasing levels of recombination (Majewski and Cohan 1999; Wiedenbeck and Cohan 2011), but the SEM always emphasizes strong selection between ecological niches and relatively low rates of gene flow, such that an adaptive allele will always spread by clonal expansion rather than recombination. Such clonal expansions are expected to result in genome-wide selective sweeps, purging genetic diversity across the genome. Although such clonal expansions and genome-wide sweeps likely occur over relatively short time scales [e.g. pathogen outbreaks; (Shapiro 2016)], they appear to be rare in nature, at least among recombining aquatic and soil bacteria studied with genome-wide surveys (Shapiro et al. 2012; Cui et al. 2015; Rosen et al. 2015; Klinger et al. 2016; Porter et al. 2016). For example, of 30 bacterial populations tracked using metagenomics in a lake over a 9-year period, only one appeared to experience a genome-wide purge of diversity (Bendall et al. 2016), although it remains unclear whether the purge was driven by selection or drift (Shapiro 2016). To explain the apparent rarity of genome-wide sweeps in nature, recent models have shown how combinations of negative frequency-dependent selection [e.g. to avoid phage predation; (Takeuchi et al. 2015)] and migration between habitat patches (Niehus et al. 2015) can allow recombination to outpace natural selection, resulting in gene-specific rather than genome-wide selective sweeps. These models help explain population genomic and metagenomic observations consistent with gene-specific sweeps in nature (Coleman and Chisholm 2010; Shapiro et al. 2012; Shapiro and Polz 2014; Klinger et al. 2016; Porter et al. 2016), but did not specifically investigate the process of speciation.

Fraser et al. (2007) used a computational model to investigate the role of homologous recombination in speciation. They confirmed the prediction of the SEM that, in the absence of distinct ecological niches and in the absence of recombination, genetically distinct clusters of bacteria continuously formed and went extinct. Thus, stable species cannot be maintained in a neutral model with only one niche. They went on to show that recombination homogenized the clusters, resulting in a single, stable "cloud" of genetic diversity. When recombination rates declined with genetic

divergence, distinct and stable clusters (reminiscent of species biological species) were maintained—but only using an unrealistically steep rate of decline. In contrast to any parameterization of the Fraser et al. model, real sequence data from the genus *Streptococcus* fall into distinct clusters, despite high rates of recombination. This suggests that a neutral model with or without recombination is not sufficient to explain the formation of stable genetic clusters. For speciation to occur, another ingredient is missing. The missing ingredient could be divergent natural selection between ecological niches or, in special cases of geographic isolation, physical barriers to recombination (Krause and Whitaker 2015).

In the *sym*patric *sim*ulation (*symsim*) model of divergent selection between ecological niches, we found that recombination accelerated the initial rate of niche adaptation but later eroded the distinctness of incipient species, particularly when several (>5) loci are involved in adaptation (Friedman et al. 2013). The model is fully sympatric, meaning that incipient species freely exchange genes despite having completely distinct niches, as might perhaps occur for species inhabiting a well-mixed aquatic environment but specializing on different dissolved nutrients. Qualitatively, the model fit well with the observation of relatively few niche-specifying genes (~3–10) involved in the ecological differentiation of marine vibrios (Shapiro et al. 2012) and suggested that barriers to gene flow (either ecological or physical) might be required to maintain the separateness of species, especially when niche adaptation involves many genes.

Marttinen and Hanage took the next logical step by modelling evolution in two ecological niches with an adjustable level of overlap (Marttinen and Hanage 2017). In this Overlapping Habitat Model (OHM), individuals exchange genes and compete only in their overlapping region of multidimensional niche space (Fig. 1). Unlike *symsim*, which explicitly models the niche-specifying genes, the OHM assumes that niche adaptation is caused by very many loci, such that the recombination of just a few of these loci does not affect niche preference. Using this model, Marttinen and Hanage were able to investigate the rates of genetic divergence under different levels of niche overlap and recombination. Intuitively, with low levels of niche overlap (~20% or less), speciation occurs rapidly due to (implicit) divergent selection between niches and reduced opportunity for genetic exchange (which can only occur in the region of niche overlap). With high niche overlap (~60%), speciation is slow and genetic distances within and between subpopulations (nascent species) continue to overlap significantly, making species difficult to distinguish (as in the case of *V. cyclitrophicus*). Fitting the OHM to real population genomic data, two putative subpopulations of *S. pneumoniae* are predicted to have 41% niche overlap and two putative subpopulations of *C. jejuni* to have 24% overlap. The model further predicts that with fast divergence (no niche overlap), all genes across the genome rapidly accumulate ecoSNPs, similar to the genome-wide divergence predicted by the SEM. With higher niche overlap, ecoSNPs are predicted to accumulate in just a few genes, with most genes containing zero or very few ecoSNPs. This pattern of few dense ecoSNP clusters was observed in both *S. pneumoniae* and *C. jejuni* genomes, suggesting their gradual divergence in the presence of gene flow in partially overlapping niches (Fig. 1). Qualitatively, this also resembles the three dense patches of ecoSNPs in *V. cyclitrophicus* described above, suggesting that the OHM could capture speciation processes in a range

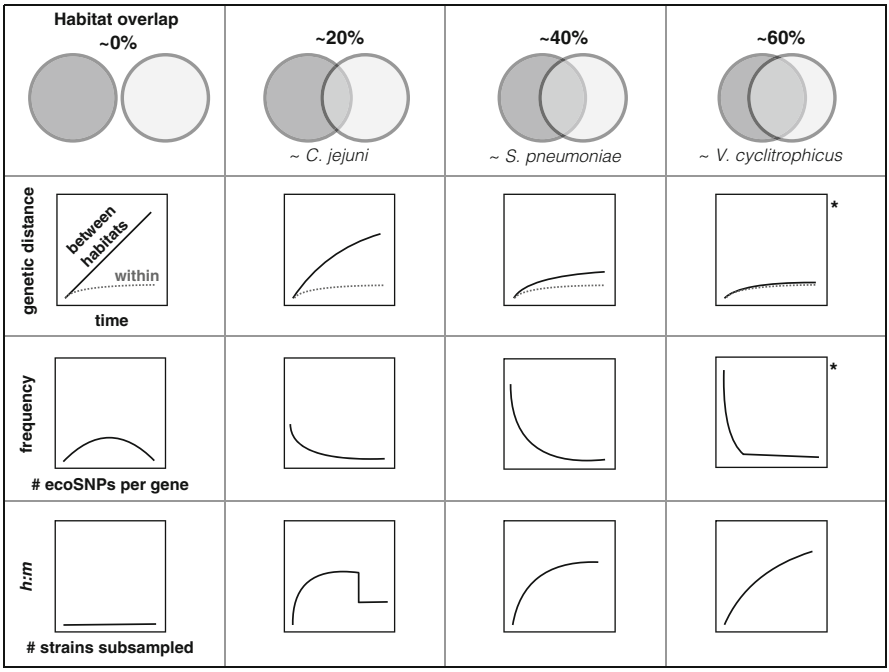

Fig. 1 Population genomic signatures of speciation under the Overlapping Habitat Model (OHM). The first (*top*) row illustrates the extent of habitat overlap between two populations. Populations can live and recombine in their respective habitat or in the region of overlap. Habitats exist in multidimensional niche space. The second row illustrates the genetic distances within and between populations, as predicted by the OHM. When there is little habitat overlap, the two populations diverge rapidly, but as overlap increases, distinct populations become difficult to distinguish from within-species genetic variation. The third row illustrates the predicted distribution of ecoSNPs (fixed single nucleotide differences between the two populations) per gene. The fourth row shows the estimated median homoplasy/mutation (*h:m*) ratio as increasingly large subsamples of genomes are taken from the populations. With ~0% habitat overlap, no recombination is expected between populations; thus the *h:m* ratio will be close to zero, and species are undefinable by the BSC-based method of Bobay and Ochman. In the example of *C. jejuni* (~20% overlap), a discontinuity is observed in the *h:m* ratio, suggesting the existence of two distinct species. The top three panels qualitatively summarize Figs. 2, 4, and 5 from Marttinen and Hanage (2017). Note that the OHM was fit to *C. jejuni* and *S. pneumoniae* datasets, but not *V. cyclitrophicus*. The panels marked with an asterisk are therefore hypothetical, based on the results of Shapiro et al. (2012). The bottom panel qualitatively summarizes portions of Supplementary Fig. 1 from Bobay and Ochman (2017)

of natural bacteria. Because the OHM does not model niche-specifying genes (the genes under divergent selection between niches), it follows that clusters of ecoSNPs in the genome can arise even when these ecoSNPs are not the direct targets of selection. As a consequence, ecoSNP clusters can be either drivers or passengers of the speciation process.

The OHM is appealing for its seamless combination of the ESC and the BSC. Ecology and divergent selection are implicit in the overlap of abstract multidimensional niches.

Barriers to gene flow occur as a consequence of nonoverlapping (or minimally overlapping) niches. The percentage of overlap in multidimensional niche space is a rather abstract concept but provides a point of entry for researchers to determine the main drivers of niche overlap (e.g. physical separation, host preference, nutrient utilization, growth rates, or some combination of these).

5 Species Delimitation Using Population Genomic Data

As discussed above, operational species definitions (such as a 95% ANI threshold) can easily be used to delimit species using population genomic data in the absence of any particular species concept. A more profound use of population genomic data is to detect signals predicted by a specific species concept and define species based on this concept. For example, the BSC predicts higher levels of gene flow within than between species. (Strictly, the BSC predicts zero gene flow between species, a criterion that will never realistically be met in recombining bacteria and archaea; hence only BSC-like concepts are amenable to most microbes and possibly most macrobes; Mallet et al. 2015; Shapiro et al. 2016.) Based on mounting population genomic evidence of higher rates of recombination within than between species or suspected species (Cadillo-Quiroz et al. 2012; Shapiro et al. 2012; Krause and Whitaker 2015; David et al. 2017), a BSC-like concept could plausibly apply to a large variety of microbes. In this BSC-like concept, barriers to gene flow provide a signature of speciation, but the drivers of speciation are not specified.

Bobay and Ochman (2017) recently proposed a way to apply a BSC-like concept to define species based on population genomic data. The method begins with a set of aligned genomes from a putative species (e.g. named species downloaded from NCBI GenBank) and identifies SNPs in the alignment. SNPs are then divided into those that can be placed parsimoniously on the phylogenetic tree, attributed to point mutation, and those that cannot: homoplasies, attributed to recombination. These two classes of SNPs are used to estimate the ratio of recombination to mutation rates (*r:m*) from the ratio of homoplasies to parsimonious mutations (*h:m*). If the alignment includes genomes sampled from just one species, sampling additional genomes will allow the SNP calling procedure to converge on a stable *h:m* ratio. However, if a "contaminant" genome from a second species is added to the alignment, this will cause an abrupt drop in the estimate of *h:m*, because under a BSC-like model, most mutations occurring between species are due to mutation, not recombination. The method therefore accepts "good" species as those that converge on a stable *h:m* estimate and proposes to split species containing "contamination" from other species. Importantly, Bobay and Ochman's method also identifies species that are too clonal (i.e. species with a very low *h:m*) and therefore cannot be classified based on a BSC-like concept.

Studying 105 named species from NCBI GenBank, Bobay and Ochman found that just over half constitute "good" species, about a quarter should be split, and about a quarter are too clonal or lack sufficient numbers of informative SNPs to be

defined (e.g. *Mycobacterium tuberculosis* and *Bacillus anthracis*). Encouragingly, the method identifies a species boundary between familiar animal species such as humans and chimpanzees. The two named species analysed with the OHM model, *S. pneumoniae* and *C. jejuni*, were also included in Bobay and Ochman's analysis, providing an opportunity for comparison (although not exactly the same set of genomes were used). In the *C. jejuni* genomes, a clear discontinuity was identified by Bobay and Ochman (Fig. 1), suggesting that this species should be split in two according to the BSC-like concept. In contrast, *S. pneumoniae* behaves as a single cohesive species (Fig. 1). At face value, this contradicts the OHM model, which predicts that *S. pneumoniae* contains two gradually diverging subpopulations that might be considered distinct species. However, the divergent *S. pneumoniae* subpopulation (SC12) identified by the OHM was not represented in Bobay and Ochman's dataset, highlighting the importance of sampling for any population genomic study of speciation or species delimitation. The two nascent species of *V. cyclitrophicus* (Shapiro et al. 2012) were not identified as distinct species based on the BSC-like criterion, likely because divergence was too recent and barriers to gene flow do not yet extend across the genome. Therefore, very early stages of speciation may be difficult to detect based on a genome-wide gene-flow criterion.

Bobay and Ochman's method is attractive for two main reasons. First, it is not based on any arbitrary threshold of genetic similarity, but rather upon a discontinuity in inferred rates of gene flow. As a result, even if some very early stages of speciation may be missed, the method can delimit species across a range of genetic divergences. Second, it is based on genome sequences, meaning it can be readily and reproducibly applied across a range of different species (including bacteria, archaea, eukaryotes, or even viruses) without "expert" knowledge or complicated phenotypic tests. It also comes with some caveats. For practical reasons, the method tests the coherence of an a priori hypothesized species; it does not define species de novo from a database of all sequenced genomes. More importantly, the method depends strongly on sample size and in fact relies on unbalanced sampling between species for discontinuities in gene flow to be identified. As such, the method is optimized to detect single "contaminant" genomes but will fail to distinguish two species sampled in roughly equal proportions. Like any comparative genomic method, it only measures realized (rather than potential) genetic exchange. Under the strict BSC, individuals that *can* exchange genes are members of the same species, even if in practice they do not (e.g. due to geographic separation and the population structure that results). Determining the potential for genetic exchange requires experiments. All that can be reasonably asked of a comparative genomic method is to assess the realized rates and boundaries of recombination. Therefore, the method provides a useful starting point for further investigation. If a species is split, researchers must go on to ask, was the split due to population structure or ecological differentiation? If the latter, what are the relevant ecological niches?

6 Conclusions

Here I have described how speciation can be initiated by ecological differentiation (an ESC-like species concept) and be maintained by barriers to gene flow (a BSC-like species concept). Population genomic evidence from several groups of bacteria support this "ESC + BSC" paradigm, but there are sure to be exceptions. In effectively nonrecombining bacteria, the BSC does not apply. In some groups of bacteria or archaea, speciation could be driven entirely by barriers to gene flow, but strong examples are still lacking. Even in cases where gene-flow barriers appear to maintain species, it is not clear whether these barriers *initiated* speciation (Cadillo-Quiroz et al. 2012; Krause and Whitaker 2015). Moreover, the distinction between ESC and BSC may be somewhat artificial, because ecological differentiation can create barriers to gene flow, for example, when incipient species favour different hosts or particles (Shapiro et al. 2012; Meyer et al. 2016). This combination of the ESC and BSC is elegantly modelled in the Overlapping Habitat Model, in which gene flow occurs only in the region of niche overlap (Marttinen and Hanage 2017). In many instances, ecological specialization and barriers to gene flow may occur effectively simultaneously, which would explain why the two potential drivers of speciation have proven so difficult to disentangle.

Population genomic and, in some cases, metagenomic data have the potential to delimit species in a standard, reproducible way. For example, genomes that differ at more than 5% of nucleotide sites tend to belong to different species (Konstantinidis and Tiedje 2005; Konstantinidis et al. 2006). While this simple cutoff-based species delimitation may work well in many cases, there are exceptions that are better resolved using concept-based delimitation. For example, *Prochlorococcus marinus* includes genomes that share only 72% average nucleotide identity, but this group still behaves as a coherent gene-flow unit according to a BSC-based species delimitation (Bobay and Ochman 2017). On the other hand, it is well established that there are several, if not hundreds, of genetically and ecologically distinct subclusters within *Prochlorococcus* which appear to stably coexist in the ocean (Rocap et al. 2003; Johnson et al. 2006; Kashtan et al. 2014). It may not matter if there are 1000, 100, or only one species of *Prochlorococcus*—but it is useful to note that *Prochlorococcus* appears to be a relatively homogeneous unit of gene flow, which may contain finer-scale units that go undetected by certain methods (Bobay and Ochman 2017). Similarly, *S. pneumoniae* shows finer genetic substructure within the two major subpopulations, suggesting fine-scale niche partitioning (Marttinen et al. 2015; Marttinen and Hanage 2017). Therefore, although species delimitation methods (Bobay and Ochman 2017) and speciation models (Marttinen and Hanage 2017) can provide impressive fits to the major features of population genomic datasets, these methods and models generally provide only a starting point—a very useful starting point—for more detailed investigations into the ecology, phenotypes, and genetics of the organisms in question.

With the possible exception of experimental evolution experiments, it is effectively impossible to follow a speciation event from start to finish in real time. However, if speciation is indeed common—and it must be if all organisms can be

placed somewhere along a speciation spectrum (Mallet 2008; Shapiro and Polz 2014)—studying diverse microbes at different stages of speciation will allow us to more fully appreciate the order of events driving and maintaining speciation, the general mechanisms involved, and the inevitable exceptional cases.

Acknowledgements I am grateful to the Canada Research Chairs program for funding and to members of my laboratory for useful discussions and comments that improved the manuscript.

Glossary

Niche A specific set of ecological parameters (environments, resources, physical and chemical characteristics, biotic interactions, etc.) to which an organism is adapted. This does not necessarily imply (but does not exclude) physical separation between niches. For the purposes of this chapter, "niche" and "habitat" are used more or less interchangeably, although "habitat" has a more spatial connotation, while niches can be temporal, behavioural, physiological, etc.

Ecological species concept (ESC) A species concept in which speciation is driven by adaptation to distinct habitats or ecological niches, with each species inhabiting a distinct niche.

Biological species concept (BSC) A species concept based on reproductive isolation (in the strict sense) or to barriers to gene flow, resulting in more gene flow within than between species, even if some between-species gene flow still occurs.

Allopatric speciation Speciation driven by physical barriers to gene flow between incipient species, such that speciation may occur in the absence of natural selection.

Sympatric speciation Speciation that occurs in the absence of physical barriers to gene flow, such that speciation must be driven by some combination of natural selection and/or genetic barriers to gene flow.

Mosaic sympatry An intermediate between sympatry and allopatric, in which organisms inhabit different niches (e.g. particles or hosts) within an otherwise well-mixed environment.

Gene flow A general term for exchange of DNA between chromosomes, including both homologous and nonhomologous DNA. In sexual organisms, gene flow occurs during meiosis. In microbes, gene flow can occur by phage-mediated transduction, plasmid-mediated conjugation, or natural competence (uptake of free DNA) followed by homologous or nonhomologous recombination.

Gene-specific selective sweep The process in which an adaptive gene or allele spreads in a population by recombination faster than by clonal expansion. The result is that the adaptive variant is present in more than a single clonal background and that diversity is not purged genome-wide.

Genome-wide selective sweep The process in which an adaptive gene or allele spreads in a population by clonal expansion of the genome that first acquired

it. The result is that diversity is purged genome-wide and that the adaptive variant is linked in the same clonal frame as the rest of the genome.

ecoSNP An ecologically associated single nucleotide polymorphism (SNP) with different nucleotides fixed between two different habitats (e.g. an A allele in habitat 1 and a T allele in habitat 2). Genes under divergent natural selection between niches or habitats ("niche-specifying genes") are expected to contain a large number of ecoSNPs.

References

Acinas SG, Klepac-Ceraj V, Hun DE, Pharino C, Ceraj I, Distel DL, Polz MF. Fine-scale phylogenetic architecture of a complex bacterial community. Nature. 2004;430:551–4.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11:1144–6.

Bao Y-J, Shapiro BJ, Lee SW, Ploplis VA, Castellino FJ, Didelot X, Maiden MCJ, Gevers D, Shapiro BJ, Polz MF, et al. Phenotypic differentiation of streptococcus pyogenes populations is induced by recombination-driven gene-specific sweeps. Sci Rep. 2016;6:36644.

Bendall ML, Stevens SL, Chan L-K, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 2016;10:1589–601.

Blount ZD, Borland CZ, Lenski RE. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proc Natl Acad Sci U S A. 2008;105:7899–906.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE. Genomic analysis of a key innovation in an experimental Escherichia coli population. Nature. 2012;488:513–8.

Bobay L-M, Ochman H. Biological species are universal across life's domains. Genome Biol Evol. 2017;9:491–501.

Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ, Whitaker RJ. Patterns of gene flow define species of thermophilic Archaea. PLoS Biol. 2012;10: e1001265.

Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. Environ Microbiol. 2011;14:347–55.

Charron G, Leducq JB, Landry CR. Chromosomal variation segregates within incipient species and correlates with reproductive isolation. Mol Ecol. 2014;23:4362–72.

Coleman ML, Chisholm SW. Ecosystem-specific selection pressures revealed through comparative population genomics. Proc Natl Acad Sci U S A. 2010;107:18634–9.

Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, et al. Rapid pneumococcal evolution in response to clinical interventions. Science. 2011;331:430–4.

Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD. A high-resolution view of genome-wide pneumococcal transformation. PLoS Pathog. 2012;8:e1002745.

Cui Y, Yang X, Didelot X, Guo C, Li D, Yan Y, Zhang Y, Yuan Y, Yang H, Wang J, et al. Epidemic clones, oceanic gene pools and epigenotypes in the free living marine pathogen Vibrio parahaemolyticus. Mol Biol Evol. 2015;32:1396–410.

David S, Sánchez-Busó L, Harris SR, Marttinen P, Rusniok C, Buchrieser C, Harrison TG, Parkhill J. Dynamics and impact of homologous recombination on the evolution of Legionella pneumophila. PLoS Genet. 2017;13:e1006855.

Dobzhansky T. A critique of the species concept in biology. Philos Sci. 1935;2:344.
Doolittle WF. Population genomics: how bacterial species form and why they don't exist. Curr Biol. 2012;22:R451–3.
Doolittle WF, Zhaxybayeva O. On the origin of prokaryotic species. Genome Res. 2009;19:744–56.
Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML. Methods Ecol Evol. 2013;4:1111–9.
Fraser C, Hanage WP, Spratt BG. Recombination and the nature of bacterial speciation. Science. 2007;315:476–80.
Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The bacterial species challenge: making sense of genetic and ecological diversity. Science. 2009;323:741–6.
Friedman J, Alm EJ, Shapiro BJ. Sympatric speciation: when is it possible in bacteria? PLoS One. 2013;8:e53539.
Gause GF. The struggle for existence. Baltimore: Williams & Williams; 1934.
Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, de Peer YV, Vandamme P, Thompson FL, et al. Opinion: re-evaluating prokaryotic species. Nat Rev Microbiol. 2005;3:733–9.
Gregory AC, Solonenko SA, Ignacio-Espinoza JC, LaButti K, Copeland A, Sudek S, Maitland A, Chittick L, Dos Santos F, Weitz JS, et al. Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. BMC Genomics. 2016;17:930.
Hanage WP. Fuzzy species revisited. BMC Biol. 2013;11:41.
Hanage WP, Fraser C, Spratt BG. Fuzzy species among recombinogenic bacteria. BMC Biol. 2005;3:6.
Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science. 2008;320:1081–5.
Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW. Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. Science. 2006;311:1737–40.
Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science. 2014;344:416–20.
Klinger CR, Lau JA, Heath KD. Ecological genomics of mutualism decline in nitrogen-fixing bacteria. Proc Biol Sci. 2016;283:20152563.
Koeppel AF, Wu M. Species matter: the role of competition in the assembly of congeneric bacteria. ISME J. 2014;8:531–40.
Koeppel AF, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, Rooney AP, Brambilla E, Connor N, Ratcliff RM, et al. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc Natl Acad Sci U S A. 2008;105:2504–9.
Kondrashov AS, Mina MV. Sympatric speciation: when is it possible? Biol J Linn Soc Lond. 1986;27:201–23.
Konstantinidis KT, Tiedje JM. Towards a genome-based taxonomy for prokaryotes. J Bacteriol. 2005;187:6258–64.
Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci. 2006;361:1929–40.
Krause DJ, Whitaker RJ. Inferring speciation processes from patterns of natural variation in microbial genomes. Syst Biol. 2015;64:926–35.
Leducq J-B, Nielly-Thibault L, Charron G, Eberlein C, Verta J-P, Samani P, Sylvester K, Hittinger CT, Bell G, Landry CR. Speciation driven by hybridization and chromosomal plasticity in a wild yeast. Nat Microbiol. 2016;1:15003.
López-Pérez M, Rodriguez-Valera F. Pangenome evolution in the marine bacterium Alteromonas. Genome Biol Evol. 2016;8:1556–70.

Majewski J, Cohan FM. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. Genetics. 1999;152:1459–74.
Mallet J. Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. Philos Trans R Soc Lond B Biol Sci. 2008;363:2971–86.
Mallet J, Besansky N, Hahn MW. How reticulated are species? Bioessays. 2015;38:140–9.
Marttinen P, Hanage WP. Speciation trajectories in recombining bacterial species. PLoS Comput Biol. 2017;13:e1005640.
Marttinen P, Croucher NJ, Gutmann MU, Corander J, Hanage WP. Recombination produces coherent bacterial species clusters in both core and accessory genomes. Microb Genom. 2015;1:e000038.
Mayr E. Systematics and the origin of species. New York: Columbia University Press; 1942.
Mell JC, Shumilina S, Hall IM, Redfield RJ. Transformation of natural genetic variation into Haemophilus influenzae genomes. PLoS Pathog. 2011;7:e1002151.
Meyer JR, Dobias DT, Medina SJ, Servilio L, Gupta A, Lenski RE. Ecological speciation of bacteriophage lambda in allopatry and sympatry. Science. 2016;354:1301–4.
Niehus R, Mitri S, Fletcher AG, Foster KR. Microbial genomes into multiple niches. Nat Commun. 2015;6:1–9.
Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32:822–8.
Popa O, Landan G, Dagan T. Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. ISME J. 2016;11:543–554.
Porter SS, Chang PL, Conow CA, Dunham JP, Friesen ML. Association mapping reveals novel serpentine adaptation gene clusters in a population of symbiotic Mesorhizobium. ISME J. 2016;11:248–62.
Retchless AC, Lawrence JG. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. Proc Natl Acad Sci U S A. 2010;107:11453–8.
Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, et al. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature. 2003;424:1042–7.
Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F, Mira A. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7:828–36.
Rosen MJ, Davison M, Bhaya D, Fisher DS. Microbial diversity. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science. 2015;348:1019–23.
Schluter D. Evidence for ecological speciation and its alternative. Science. 2009;323:737–41.
Shapiro BJ. Signatures of natural selection and ecological differentiation in microbial genomes. Adv Exp Med Biol. 2014;781:339–59.
Shapiro BJ. How clonal are bacteria over time? Curr Opin Microbiol. 2016;31:116–23.
Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 2014;22:235–47.
Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ. Population genomics of early events in the ecological differentiation of bacteria. Science. 2012;336:48–51.
Shapiro BJ, Leducq JB, Mallet J. What is speciation ? PLoS Genet. 2016;12:e1005860.
Takeuchi N, Cordero OX, Koonin EV, Kaneko K. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. BMC Biol. 2015;13:1–11.
Tilman D. Resource competition and community structure. Princeton: Princeton University Press; 1982.
Vos M. A species concept for bacteria based on adaptive divergence. Trends Microbiol. 2011;19:1–7.

Wiedenbeck J, Cohan FM. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev. 2011;35:957–76.

Yawata Y, Cordero OX, Menolascina F, Hehemann J-H, Polz MF, Stocker R. A competition-dispersal tradeoff ecologically differentiates recently speciated marine bacterioplankton populations. Proc Natl Acad Sci U S A. 2014;111:5622–7.

Peering into the Genetic Makeup of Natural Microbial Populations Using Metagenomics

Vincent J. Denef

Abstract This chapter focuses on how metagenomic data are applied to examine the genomic heterogeneity of natural microbial populations. It highlights the opportunities and challenges inherent to the approach and describes recently developed methods to maximally leverage the potential of these datasets while tackling some of the challenges. We describe how performing population genomic analyses using metagenomic data allows (1) resolution of ecologically and genetically cohesive populations in the environment, (2) tracking of evolutionary processes within them, and (3) application of metatranscriptomic and metaproteomic analyses to determine the in situ physiology of distinct populations. While challenges remain that are inherent to the approach, the current wave of new bioinformatic tools is starting to realize the theoretical potential of metagenomics to peer into the spatiotemporal dynamics of the genetic structure of natural populations.

Keywords Bacteria · Bioinformatics · Gene content variation · Metagenomics · Natural populations · Recombination · Selection · Sequence variation · Strain-resolved

1 Introduction

1.1 Scope of This Chapter

This chapter explores the advances that have been made in the area of population genomics through studies that make use of metagenomic data. It covers methodological advances and challenges and biological insights that have been gathered. Metagenomics [also called environmental or community genomics (Handelsman

V. J. Denef (✉)
Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA
e-mail: vdenef@umich.edu

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_14,

2004; Tyson et al. 2004)] is the analysis of data produced through randomly sequencing fragments from a DNA pool extracted from environmental samples (Fig. 1). In general, metagenomics finds its application in the study of the composition and functional potential of microbial communities in their native environment. However, a growing number of studies are leveraging these datasets to gain unprecedented insights into the genetic composition of natural populations, i.e., groups of individuals belonging to the same species that co-occur in space and time. In the context of metagenomics, the term population genomics was initially—and continues to be—used as a synonym for metagenomics, a possible cause of confusion. This has been particularly the case when referring to the reconstruction of a consensus genome of a population through curated assembly of metagenomic data (DeLong 2004, 2005; Handelsman 2004; Tyson et al. 2004). In the context of this book chapter, the strictest definition of metagenomic-based population genomics refers to the analysis of genome-wide heterogeneity existing between individuals belonging to the same species/ecotype (Whitaker and Banfield 2006).

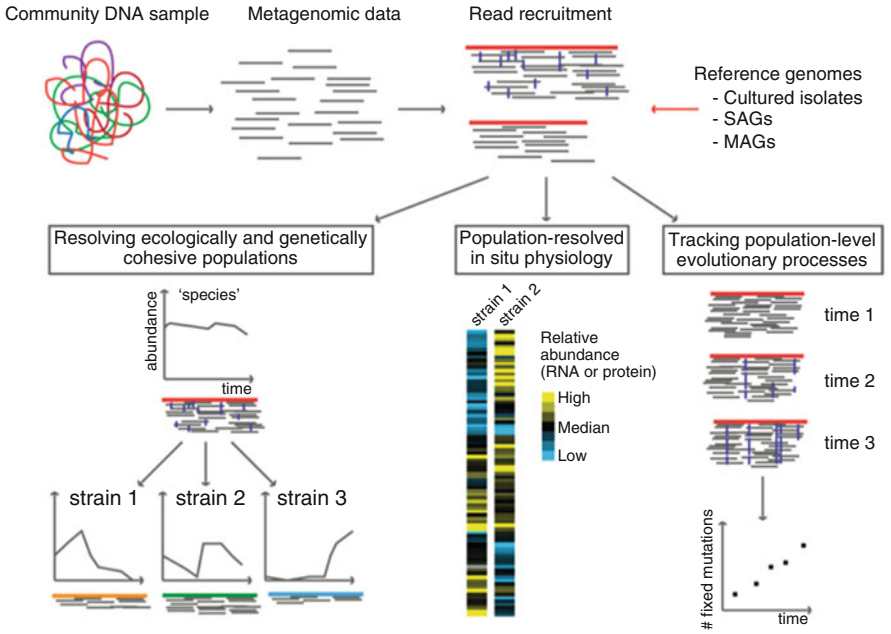

Fig. 1 Overview of applications of metagenomic-based population genomics discussed in this chapter. After extracting, fragmenting, and generating sequencing reads to create a metagenomic dataset, reads are typically aligned to a reference sequence, obtained from microbial isolates, single cells (SAGs), or assembled from metagenomic data (MAGs). These data can be used to (left) resolve ecologically distinct populations by identifying genetically similar reads originating from distinct strains, (middle) determine in situ gene expression (using metatranscriptomic or metaproteomic data), or (right) track evolutionary processes, for example, by identifying polymorphic sites where specific nucleotides rise to fixation over time

While more theoretically motivated definition of populations is the focus of another chapter in this book (Shapiro 2017), it is important to highlight how population genomics using metagenomic data has contributed to our efforts to recognize and delineate ecologically and evolutionary cohesive populations. Most prominent is the demonstration of sequence-discrete populations in environmental samples [Fig. 2; sensu (Caro-Quintero and Konstantinidis 2012)]. Conceptually, a sequence-discrete population was defined by Caro-Quintero and Konstantinidis as "the natural entity present in a community/sample that comprises genotypes, which are clearly distinguishable from their closest co-occurring relatives (if any) based on their high genetic relatedness and comparable relative abundance *in situ*." Technically, such genotypic clusters are identified by comparing the nucleotide identity of the short reads gathered in a metagenomic survey to a reference genome by means of a process called read recruitment. Whereas similar observations of sequence-discrete populations based on a single or multiple marker genes sequenced from bacterial isolates had been made before (Hanage et al. 2006; Hunt et al. 2008; Rocap et al. 2003), the advent of metagenomic methodology allowed for a genome-wide assessment of genetic relatedness of randomly sampled cells present in environmental samples. These insights from isolate and metagenomic studies have helped move forward the discussion regarding the existence of microbial species and specifically how to define a microbial population. While genotypic variation within a defined species can be large at a regional or global scale, thus complicating our ability to define clear species boundaries, it is important to stress the inherent property of a population to contain only individuals that are occurring in the same place at the same time, i.e., that they are sympatric (Shapiro and Polz 2014; Cordero and Polz

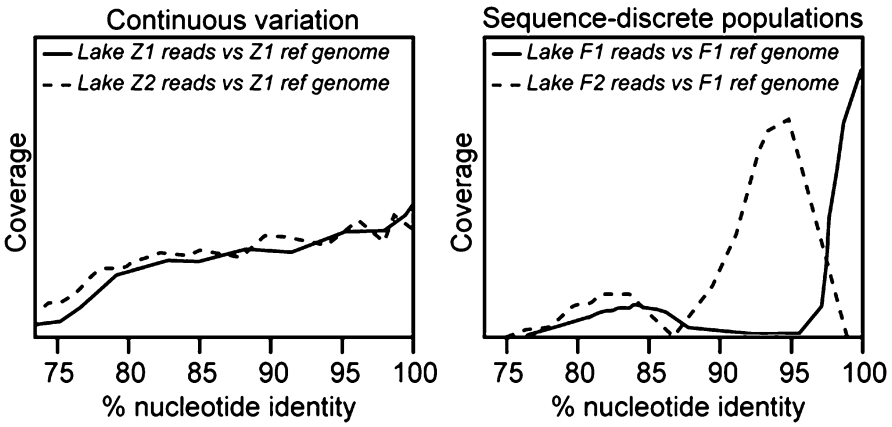

Fig. 2 Identification of populations using metagenomic data. Comparisons of sequences generated from community DNA samples from two lakes to a reference genome of a bacterial isolate from one of these two lakes. If discrete populations would not exist, the patterns on the left could be obtained, while if ecologically and evolutionary cohesive populations that are distinct from the reference population exist, the pattern on the right would be expected. The patterns on the right are the most commonly observed [e.g., (Bendall et al. 2016, Caro-Quintero and Konstantinidis 2012)]

2014). While implied by the definition of a population, early challenges to the idea that discrete populations exist based on sequenced isolates from disparate locations did not respect this condition of sympatry [e.g., (Welch et al. 2002)]. Further discussion of insights into the microbial species concept derived from bacterial population genomics is covered in other chapters (Shapiro and Polz 2015; Shapiro 2017) and will not be discussed in detail here.

1.2 Approaches Included in This Chapter

Examining population-level variability within natural populations requires the availability of a reference sequence in most metagenomic approaches to population genomics. There are multiple ways to obtain reference genomes: (1) from isolates, preferably originating from the same environment/sample the metagenomic data are derived from; (2) from genomes assembled from metagenomic data (metagenomic assembled genome, MAG), which tends to be a composite sequence not representative of a single cell in the population; or (3) from a single cell genomic dataset (single amplified genome, SAG). While our ability to generate MAGs was initially limited to low complexity communities (Tyson et al. 2004), we can now reconstruct 100 s to 1000 s of genomes from metagenomic datasets. Few of these MAGs are complete, and although contamination from other populations cannot be completely excluded, obtaining >90% completeness with limited (<5%) contamination is commonly achieved (Anantharaman et al. 2016; Delmont et al. 2017; Parks et al. 2017). Tools to refine sequence bins and estimate their completeness and purity are continuously being developed (Broeksema et al. 2017; Eren et al. 2015; Parks et al. 2015) and will continue to improve genome reconstruction and bin refinement, so as to provide a sounder basis for downstream population genomic analyses. More uncertainty will remain in genomes assembled from metagenomes compared to those from isolates.

Independent of the approach used to generate the reference genome, the central tenet of the analysis is the comparison of randomly sampled and sequenced DNA fragments with this reference genome to assess sequence content and compositional variation of populations within and between environmental samples. A variety of approaches to generate metagenomic DNA fragments can be applied. Most straightforward is to randomly generate sequences from DNA extracted from environmental samples. Other approaches first reduce the diversity of the community, e.g., by passing the sample through a series of filters with decreasing pore sizes (Baker et al. 2010) or through (in situ) enrichments (Delmont et al. 2015). This allows for the enrichment of specific populations of interest, therefore increasing sequencing depth for population genomic analyses. Finally, instead of focusing on the entire genome, one can target a series of sites across the genome. An approach recently implemented, which may see much broader application, is to extract multi-locus sequencing typing (MLST) genes from complex metagenomic datasets using reference sequences (Berry et al. 2017; Zolfo et al. 2017).

Although the work discussed in this chapter relates to studies that use random sequencing of DNA extracted from complex natural communities, occasionally single cell genomics is categorized as a metagenomic approach. Single cell sequencing approaches sort a single cell from an environmental sample using dilution, flow cytometry, or microfluidic approaches and subsequently perform DNA amplification and genome sequencing (Blainey 2013). Genome sequencing analyses of single cells representing the same naturally occurring population is similar to metagenomics, as both eliminate culturing biases. Other than that aspect, population genomics based on single cell sequencing [e.g., (Kashtan et al. 2017; Malmstrom et al. 2013; Zaremba-Niedzwiedzka et al. 2013)] is conceptually and methodologically similar to the analysis of representative isolate genomes from natural populations [e.g., (Hunt et al. 2008)], discussed elsewhere in this book. Testament to the power of this approach, Kashtan and colleagues used single cell genomic data of *Prochlorococcus* cells from ocean water to demonstrate the existence of hundreds of co-occurring populations. These populations were shown to differ from each other in terms of genome content, such as the presence of small genomic islands that most likely conferred predation resistance and phage recognition), as well as genome-wide sequence composition (Kashtan et al. 2014). These insights were gained by characterizing >1,000 cells by sequencing of the rRNA ITS region, while a subset of 69 cells was sequenced to >70% estimated genome completeness. As mentioned above, single cell genome sequencing is also commonly used to generate a reference sequence, after which sequencing reads from metagenomic surveys from the same or different environments can be aligned to this genome to evaluate population-level heterogeneity, which is of relevance to this chapter [e.g., (Thrash et al. 2014)].

2 Opportunities and Challenges

2.1 Opportunities

The advantage of metagenomic approaches compared to single isolate approaches is the ability to sample a very high number of individuals without culturing bias. While the number of individual cells that can be analyzed is rapidly increasing for single cell approaches, the number of cells sampled and typical reconstructed genome completeness remains higher in metagenomic approaches. This offers the unprecedented ability to peer deeply into the genetic structure of natural populations and has revealed the extraordinary genetic diversity that exists among groups of closely related microbes. The extent of this diversity and the correlation between the abundance of genetic subclusters with distinct environmental conditions can result in the division of previously named taxonomic units into ecologically distinct populations (Bhaya et al. 2007; Denef et al. 2010a). A hallmark of such newly defined populations is extensive diversity in gene content and sequence composition (Denef et al. 2010a; Simmons et al. 2008). Initially, there were doubts that environments beyond reduced complexity systems such as the acid mine drainage system, in

which pioneering studies of population genomics using metagenomics were conducted, could be tackled (Denef et al. 2010b). However, recent work has expanded the approach to systems ranging from the human microbiome to aquatic environments (Bendall et al. 2016; Nayfach et al. 2016; Olm et al. 2017). Importantly, the use of metagenomic approaches allows us to access 10 s–1000 s of populations at the same time (Anantharaman et al. 2016; Parks et al. 2017). The tremendous growth of publicly available metagenomic datasets, as well as reference genomes from microbial isolates or single cell genomics projects, is another major opportunity to tackle new population genomic questions without additional sequencing efforts. This was demonstrated in recent studies that have leveraged thousands of human microbiome metagenomic datasets to uncover strain-level dynamics and infer mode of transmission and biogeographical patterns among hundreds of bacterial populations simultaneously (Nayfach et al. 2016; Truong et al. 2017).

2.2 Challenges

However, many challenges remain (Table 1). The first and possibly most important major challenge is the lack of linkage between variant sites (nucleotide substitutions, insertions and deletions, rearrangements), i.e., using metagenomic data we are unable to determine which alleles across the genome are present in one lineage versus another. As a consequence, most environmental population genomic approaches have relied on isolate or single cell-derived sequences (Kashtan et al. 2014; Krause and Whitaker 2015; Shapiro and Polz 2014). Multiple factors can be responsible for lack of linkage in metagenomic data: (1) the number of variant loci across the genome is typically too low compared to sequencing read length (generally 100–150 nucleotides) or sequencing library fragment size (up to several hundred nt) to enable linkage across more than a few hundred nucleotides, and (2) in metagenomic datasets, each sequencing read typically originates from a different individual. As a result, we are limited to identifying which sites are polymorphic in the population or which sites are divergent between coexisting closely related populations. Determining which mutations occur across the genomes of a single lineage remains possible only by using isolate or single cell genomic data unless population structure is very simple (Denef and Banfield 2012), although new approaches based on statistical inferences may change this (e.g., DESMAN, see below). Considering these challenges due to read length of the most commonly used next-generation sequencing platforms, it is thus not surprising that some of the most thorough metagenomic population genomic work has been carried out using longer sequences, such as those obtained by Sanger sequencing. These studies were largely successful because long sequence reads and library insert size enable linkage across several kilobases at a time (Allen et al. 2007; Eppley et al. 2007; Konstantinidis and DeLong 2008; Simmons et al. 2008). Technological innovation is ongoing, and newer sequencing platforms (e.g., PacBio, Oxford Nanopore) may resolve the read length issues, as long as they continue to offer high sampling depth and limited

Table 1 Fundamental challenges for metagenomic-based population genomic analyses

Challenge	Approaches to address challenge	Remaining issues	Example studies
Linking SNPs co-occurring in the same individual	Link by relative abundance	Only works for low within-population genetic diversity or requires high number of samples	Denef and Banfield (2012); Quince et al. (2017)
	Focus on overall patterns of polymorphisms across genome	Limited to broad interpretations regarding genome-wide vs gene-specific selective sweeps	Bendall et al. (2016)
	Long-read sequencing technologies	High error rates, or lower sequence coverage, or not broadly available	Sharon et al. (2015)
Differentiating SNPs vs errors	Tool for identifying true sequence variants from sequencing error (Varcap)	Only helps resolve SNPs present in >2% of population	Zojer et al. (2017)
	Tools for removing error-based bias in population genetic parameter calculations	Platform-specific, unclear if it removes sequence-library-dependent bias	Johnson and Slatkin (2006); Johnson and Slatkin (2008); Johnson and Slatkin (2009); Nielsen et al. (2011)
Obtaining sufficient sequence coverage	Physical- or affinity-based enrichment of target populations	Affinity-based methods often technically challenging, physical methods restricted to populations with outlier cell size	Baker et al. (2006); Hatzenpichler et al. (2016); Pernthaler et al. (2008)
Tracking gene gain/loss	Reference-free sequencing read dataset comparisons	Untested for population genomic analyses	Nijkamp et al. (2013)
	Sample-by-sample genome reconstruction	Restricted to populations that are abundant across time series	Bendall et al. (2016)

sequencing errors (see below), which is typically not yet the case [but see (Sharon et al. 2015)]. Alternatively, methods that use cross-linking of DNA within the cell may allow the connection of physically linked variants (Marbouty et al. 2014), but this method has not been applied to population genomic studies.

The second major challenge is the issue of sequencing error, which results in false positive polymorphic sites. These errors usually occur as low-frequency "mutations" that are often observed only once (Schirmer et al. 2016). As several population genetic parameters require knowledge of all polymorphic sites, including those occurring at low frequency, errors will bias their metagenomics-based estimates. Case in point is Watterson's theta (Watterson 1975), which estimates the genetic diversity present in a population based on which we can estimate mutation rates and/or effective population sizes and which requires even the knowledge of the frequency of singleton variant sites. As new sequencing platforms have emerged,

each with their specific error spectrum, a variety of tools have been developed to differentiate between true variants and sequencing errors. Some tools assign confidence levels to observed variants, with the goal of reducing false positive variants [e.g., VarCap (Zojer et al. 2017)] focuses on improving reliability of identifying variants >2% of the population), or to remove bias in population genetic parameter estimates (Johnson and Slatkin 2006, 2008, 2009; Nielsen et al. 2011). In addition to sequencing platform-specific error profiles (Schirmer et al. 2016), a complication arises from the observation that errors can be sequencing library preparation protocol dependent. This issue was highlighted recently in a reanalysis of the preterm fecal microbiome metagenomic data that is frequently used as a benchmark dataset for new bioinformatic tools (Sharon et al. 2013). It could be shown that the day-by-day alternation between two SNP patterns in a bacterial population was caused by different library preparation methods used on even and odd days [http://merenlab.org/2016/12/14/coverage-variation/ comment on (Eren et al. 2015) based on data from (Sharon et al. 2013)].

The third challenge is obtaining sufficient sampling depth, which is the number of sequences that cover a particular region of the genome. Sufficient sequence coverage is needed to accurately estimate allele frequencies, and as such most analyses are currently limited to the most abundant populations in environmental samples. Yet, we have progressed far beyond what the research community envisioned just a few years ago with respect to the number of near-complete genomes that we can reconstruct from environmental samples. Such genomes can subsequently be used to examine the genetic structure of the corresponding populations. In part, this is due to the development of physical (e.g., size-selective filtration)- or affinity (e.g., based on use of fluorescent in situ hybridization and cell sorting)-based methods (Baker et al. 2006; Hatzenpichler et al. 2016; Pernthaler et al. 2008) through which populations of interest (e.g., based on taxonomic identity or metabolic activity) can be enriched, facilitating population genomic analysis from the corresponding metagenomic data (Deng et al. 2014).

Finally, the identification of gene content differences within a population can be challenging when using metagenomic data. Unless extensive manual curation of an assembly is performed [e.g., (Simmons et al. 2008)], genomic regions (islands) carried only by low-abundance subpopulations (i.e., part of a population's "flexible genome") will generally not be binned in the consensus genome of the population of interest (i.e., the "core genome"). This is because these genomic regions (a) diverge in their k-mer (specific stretches of nucleotides, e.g., tetramers ATGC, AATG, etc.) composition [used in most binning applications that seek to group fragments of contiguous sequence (contigs) originating from the same genome such as VizBin (Laczny et al. 2015), CONCOCT (Alneberg et al. 2014), and TETRA-ESOM (Dick et al. 2009)] compared to core genome contigs, (b) have differential coverage patterns that diverge from the core genome of the population [also commonly used in binning applications (e.g., CONCOCT (Alneberg et al. 2014), GroopM (Imelfort et al. 2014), metagenome (Albertsen et al. 2013))], and/or (c) will often fail to assemble into large enough contigs to allow accurate binning due to the low abundance of the subpopulations they derive from. Similarly, when using

metagenomic sequences to identify variants across datasets by mapping sequence reads to reference genomes, we can only track changes in frequency among regions shared by these populations and cannot identify the addition of new genomic regions due to horizontal gene transfer [see (DeLong 2012) in commentary on (Denef and Banfield 2012)]. Yet, we know that such differences constitute a significant fraction of population-level genomic heterogeneity. Evidence regarding the physiological importance of these unique regions is mixed (Denef et al. 2010a; Frias-Lopez et al. 2008; Gogarten and Townsend 2005; Kuo and Ochman 2009; Thompson et al. 2011; Hehemann et al. 2016). Nonetheless, it is important to try to include these regions in metagenomic-enabled population genomic analyses as gene frequencies at either intermediate or low levels result from frequency-dependent selective pressures by social and ecological interactions and thus suggests adaptive roles for flexible genome content (Coleman et al. 2006; Cordero et al. 2012; Cordero and Polz 2014; Kashtan et al. 2014; Rodriguez-Valera et al. 2016).

3 Current Applications

We present a series of recently developed tools to facilitate population genomic analyses using metagenomic data and explore three types of applications of these methods (Fig. 1; Table 2). First, we provide an overview of how these methods are being used to resolve ecologically and genetically distinct populations that would previously have been considered as a single operational taxonomic unit (OTU). Most commonly OTUs are defined based on 16S rRNA gene sequence identity, but these can similarly be defined based on multiple housekeeping genes or complete genomes. Second, we show how these approaches can be used to infer the physiology of distinct populations. Third, we summarize applications of these methods to gather insights into evolutionary processes occurring in natural microbial populations.

3.1 Methods

Read mapping to a reference sequence is a key step in most population genomic approaches. Over time a wide array of read alignment tools have become available, each with their own user-specified tunable parameters. Naturally, this flexibility may affect our ability to accurately perform population genomic analyses. In a recent comparative analysis, popular tools such as bwa (Li and Durbin 2010) and bowtie2 (Langmead and Salzberg 2012) resulted in similar and more accurate results than some other tools when all were run using default parameter settings (http://merenlab.org/2015/06/23/comparing-different-mapping-software/). In more recent years, reference-free methods have been developed that avoid some drawbacks of the reference-based approach, particularly the inability to detect parts of the population's

Table 2 Goals, approaches, and challenges for metagenomic-based population genomic analyses

Goal		Approaches	Challenges	Example studies
Resolving ecologically and genetically cohesive populations	Identifying sequence-discrete populations	Read recruitment (e.g., bowtie2, bwa) + custom scripts for data plotting	Determine relevant sampling scales to capture sympatric individuals (e.g., bulk water vs size-fractionated samples)	Caro-Quintero and Konstantinidis (2012)
	Distinguish diverging within-species ecological dynamics	Growing suite of automated tools such as Constrains, MetaMLST, MIDAS, PanPhlAn, StrainPhlAn, DESMAN, and LSA	Database dependency of many tools limits us to species with extensive reference genome availability	Luo et al. (2015); Zolfo et al. (2017); Nayfach et al. (2016); Asnicar et al. (2017); Ward et al. (2016); Quince et al. (2017); Cleary et al. (2015)
	Identify strain-specific gene content and SNPs		Most approaches need a large number of samples to be effective	
Determining physiology of ecologically and genetically cohesive populations	Identify in situ differences in gene expression between co-occurring strains	Custom scripts/manual as well as automated tools to resolve metatranscriptomic or metaproteomic data (e.g., PanPhlAn)	Relationship between expression levels and process rates rarely known	Wilmes et al. (2008); Denef et al. (2010a); Brooks et al. (2015); Asnicar et al. (2017)
	Estimate in situ growth rates	iRep, based on metagenomic coverage patterns	Not benchmarked against measured growth rates thus far	Olm et al. (2017)
Tracking evolutionary processes within ecologically and genetically cohesive populations	Homologous recombination vs mutation	Manual tools (e.g., Strainer) to visually identify and quantify recombination sites or automated tools to determine recombination vs mutation rates	Manual work is low throughput	Eppley et al. (2007); Johnson and Slatkin (2009)
	Gene gain/loss	Custom scripts to determine gene content differences between MAGs representing same population across time series samples	Can we discriminate gene gain/loss in population vs strain replacement?	Bendall et al. (2016)
	Natural selection	Custom scripts/manual approach to determine mutation rates and/or gene-specific vs genome-wide selective sweeps	Challenging in "open" systems	Denef and Banfield (2012); Roux et al. (2014); Bendall et al. (2016)

flexible genome. One such tool is able to detect gene frequency patterns across samples of regions of the genome that are not represented in reference sequences [e.g., MARYGOLD (Nijkamp et al. 2013)]. Historically, population genomic analysis of metagenomic data relied on manual analysis, either through existing assembly visualization/curation software (Morowitz et al. 2011; Simmons et al. 2008), through generic graphing software (e.g., excel, R) to visualize the distribution of sequence similarities among reads mapping to a population's contigs (Fig. 2; [Bendall et al. 2016; Caro-Quintero and Konstantinidis 2012; Oh et al. 2011)], or through software developed specifically for the resolution of bacterial strains in metagenomic data [e.g., Strainer (Eppley et al. 2007)]. While these approaches worked, a drawback of these methods is the labor-intensiveness and difficulty to reproduce similar results by independent users with different levels of expertise.

More recently, a variety of tools have been developed to (1) remove bias due to sequencing error [VarCap (Zojer et al. 2017)], (2) extract population genomic metrics (e.g., Watterson's theta) from next-generation sequencing data (Haubold et al. 2010; Johnson and Slatkin 2006), (3) visualize SNP patterns across sample series in assembled contigs [e.g., using Anvi'o (Eren et al. 2015)], and (4) resolve closely related strains from metagenomic datasets [e.g., ConStrains (Luo et al. 2015), MIDAS (Nayfach et al. 2016), DESMAN (Quince et al. 2017), StrainPhlAn (Truong et al. 2017), PanPhlAn (Scholz et al. 2016), and LSA (Cleary et al. 2015)]. The development of the latter set of tools is particularly exciting, as it promises to greatly facilitate the resolution of strain dynamics and the coupling of gene content and sequence composition data with dynamics in population abundance across environmental or temporal gradients as has been performed manually previously (Denef et al. 2010a; Morowitz et al. 2011).

Most of the strain resolution tools rely heavily on whole genome reference databases, which are reasonably representative for some microbial systems such as the human microbiome, but much less so for other systems such as terrestrial and aquatic biomes. The reliance on reference genomes limits the ability for strain resolution in these other environments at this point (Nayfach et al. 2016). All of the reference-based tools are able to analyze thousands of metagenomic datasets at the same time while extracting strain dynamics for many species at the same time, e.g., 135 in the case of the study by Truong et al. (2017). Based on their own benchmark study, StrainPhlAn appears to reduce the per-nucleotide nucleotide variant identification error to less than 0.1%, granting more accurate strain identification than tools such as MIDAS and ConStrains. PanPhlAn is similar in approach to StrainPhlAn, but is focused on identifying strain-specific gene content, rather than nucleotide substitutions (Scholz et al. 2016).

In contrast, strain resolution tools such as DESMAN and LSA take reference sequence-independent approaches and, in the case of LSA, even an assembly-independent approach. DESMAN identifies strains, including both genotype-specific nucleotide substitutions and gene content variation, from metagenomic data generated from a sample collection. After validating their approach with a mock dataset, they applied their method to examine abundance patterns of different strains within a large set of ocean metagenomic data (TARA Oceans). While

DESMAN allows the identification of novel strains, it is highly dependent on the quality of the assembly and binning steps and requires a relatively large number of samples to be effective (Quince et al. 2017). Many researchers currently rely on automatic binning approaches to generate their metagenomic sequence bins, but these can be highly inaccurate, depending on community composition including the extent of co-occurring closely related populations and the extent of community turnover in the temporal or spatial sample series. While DESMAN has the ability to further resolve multi-strain bins, careful manual curation, aided by tools such as Anvi'o (Delmont et al. 2017) or ICoVeR (Broeksema et al. 2017), may be necessary for downstream population genomic analyses. The second assembly-independent approach discussed here, latent genome analysis (LSA; Cleary et al. 2015), separates sequencing reads prior to assembly by calculating unobserved variables call "eigengenomes" that reflect covariance in k-mer abundances across a sample series. This method allowed for the separation and downstream assembly of specific genomic regions of strains sharing less than 99.5% average nucleotide identity, while regions of the genome highly conserved between strains were grouped together as sequencing reads from which conserved core genome could be assembled.

Beyond whole genome approaches, several approaches have been developed to extract population-specific sequences for a set of core genes. The concept of multi-locus sequence typing (MLST) (Maiden et al. 1998) used for population genetic analysis of isolates has been implemented in metagenomic data analysis either through a series of custom bioinformatic scripts (Berry et al. 2017) or through more streamlined packages such as MetaMLST (Zolfo et al. 2017), ConStrains (Luo et al. 2015), and MetaPhlAn2 (Truong et al. 2015). Finally, to resolve true sequence variants from sequencing errors, tools such as oligotyping (Eren et al. 2013) can be applied to sequence reads covering marker genes, though we are not aware of applications to metagenomic data thus far.

3.2 Resolving Ecologically and Genetically Cohesive Populations

The motivation to develop these new metagenomic tools originated from the realization that studies using single marker genes clustered at a fixed identity level (i.e., OTUs) likely miss key community dynamics since multiple ecologically distinct populations were clustered together in a single OTU (Acinas et al. 2004; Denef et al. 2010a; Eckburg et al. 2005; Fraser et al. 2009; Fuhrman and Campbell 1998; Giovannoni et al. 1990; Hahn et al. 2016; Hunt et al. 2008; Larkin and Martiny 2017; Morowitz et al. 2011; Rocap et al. 2003; Shapiro et al. 2012; Shapiro and Polz 2014; Sharon et al. 2013; Wilmes et al. 2008). The ability to resolve strain-level

differences in microbial communities and detect the dynamics of highly related genotypes will likely lead to rapid advances in our ability to study microbial ecology at the appropriate resolution. We present here some of the most recent examples of how streamlined strain-resolved analyses are leading to previously unrecognized ecological patterns.

Using MIDAS, researchers were able to identify strains in metagenomic datasets and this revealed dynamics that could not be observed at a higher taxonomic level (e.g., species) (Nayfach et al. 2016). Conceptually, the finding that important ecological dynamics are masked by clustering distinct populations into higher taxonomic levels is similar to previous findings. Particularly, a study by Rodriguez-Brito and coauthors showed that hidden underneath the observed stability at coarser genetic resolution ("species" level) were strongly fluctuating abundances of ecologically distinct "strains" grouped at the species level (Rodriguez-Brito et al. 2010). The study by Nayfach and coauthors revealed that mothers pass on a large percentage of bacterial allele variants to their children in the early days after birth. In the subsequent postnatal months, even as the number of species shared between mother and child increases, the strain composition gradually diverges (Fig. 3a, b), indicating increasing importance of colonization from other sources (Nayfach et al. 2016). These findings were confirmed in a similar study using PanPhlAn and StrainPhlAn (Asnicar et al. 2017). At a larger spatial scale, links between geographic distance and strain correspondence have been found in human populations using StrainPhlAn as well and indicate limited overlap in strains between geographically distinct populations (Truong et al. 2017).

Several of the recently developed methods allow us to pinpoint the specific gene content and SNP variation that differentiates closely related but ecologically distinct populations from each other to attempt to explain their distinct population dynamics. For example, resolution of strains and identifying strain-specific gene content has allowed for the identification of specific strains involved in diseases where traditional approaches failed to do so. Using PanPhlAn, Ward and coauthors identified strain-specific gene content of *Escherichia coli* using 166 infant microbiomes and identified strains associated with infant risk for necrotizing enterocolitis to be enriched in genes involved in iron acquisition and specific energy and amino acid metabolism functions (Ward et al. 2016). In another study, an analysis of regional strain-level variability identified regionally distinct horizontally transferred genes, in large part glycosyl transferase family proteins likely reflecting dietary differences at both large and small spatial scales (Brito et al. 2016) (Fig. 3c). While these studies did not aim to resolve co-occurring closely related populations, the same approach could be applied to identify genes differentiating sympatric populations.

The field of epidemiology is also embracing metagenomic tools to better understand disease outbreaks. As MLST is a common method used in epidemiological studies using isolates, tools adapted to metagenomic data, such as MetaMLST, have been used to identify strains in disease outbreaks (Zolfo et al. 2017). In addition, the large number of reference sequences available for pathogenic bacteria in

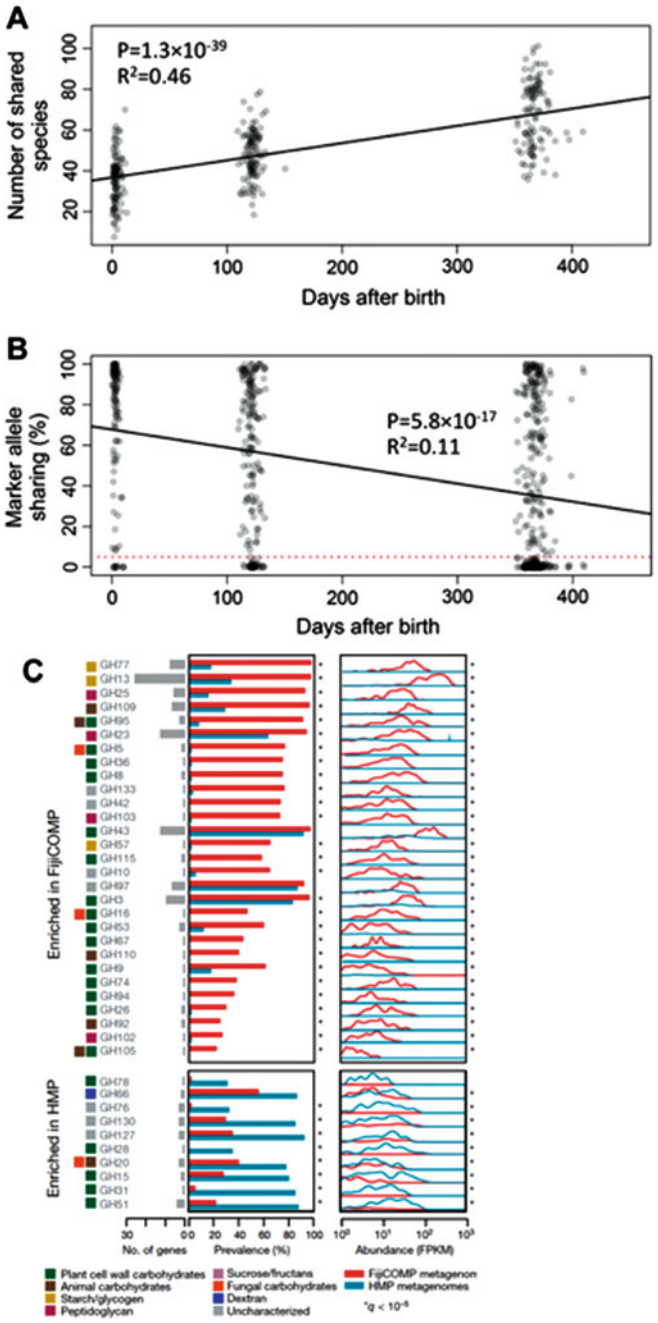

Fig. 3 Examples of the automated resolution of strains in metagenomic data. (**a**, **b**) Comparison of metagenomic data in mother-infant pairs using MIDAS indicated that while the number of shared species increases with time after birth (**a**), vertical transmission of strains is particularly important

combination with automated tools greatly facilitates the use of metagenomic data to perform epidemiological studies. This allows us to expand on the studies that were thus far limited to isolate sequence data and enables insights into strain transmission, retention, and tissue specificity within the human body in the absence of any culturing bias (Donati et al. 2016).

Outside of the human microbiome, we are currently limited to analyzing a handful of lineages that have an adequate representation in the databases, although the generation of novel genomes reconstructed from metagenomic data and single cell genomics is rapidly increasing the number and taxonomic coverage of available references. For two well-represented taxa, the marine *Pelagibacter* and *Prochlorococcus*, MIDAS has been used to determine differences between populations in different oceanic regions by evaluating gene content overlap (Nayfach et al. 2016). Conventional approaches failed to detect these phenomena (Sunagawa et al. 2015). Whether these patterns were due to dispersal limitation or due to environmental selection according to conditions that differ between oceanic regions and that correlated with distance could not be resolved.

Expanding population-resolved analyses beyond taxa currently well represented in genomic databases, Garcia and coauthors generated their own system-specific (*in casu*, a specific lake) database of 33 reference genomes using single cell genomics and did read recruitment using metagenomic data from a 5-year sample time series from the same lake. They revealed distinct patterns for several abundant lineages. Some lineages could be resolved into distinct genotypes with clearly distinguished ecological dynamics that likely represented separate populations (e.g., *Actinobacteria* acI lineages). Other lineages (e.g., *Alphaproteobacteria* LD12, the freshwater sister group to marine *Pelagibacter*) did not have sequence-discrete nor ecologically distinct within-group dynamics (Garcia et al. 2016); thus distinct populations could not be resolved, or all sampled cells belonged to a single population. The availability of more nonhuman microbiome reference sequences, in combination with the recently developed automated tools to deconvolute strain patterns and identify alleles and gene content differences associated with these strains, is promising.

Fig. 3 (continued) early in life and decreases as time goes on, based on the % of shared alleles in core genome marker genes (**b**). (**c**) Resolution of strain-specific differences due to divergence in mobile element gene content showed that the type and abundance of specific glycoside hydrolase gene families diverged significantly between a cohort from Fiji (FijiCOMP) and North America (HMP). Prevalence indicates the % of fecal samples in the cohort that the protein family was identified in. Abundance, expressed in fragments per kilobase of protein coding sequence per million mapped reads (FPKM), presents the relative abundance spectrum across all samples in each cohort. Asterisks indicate significant differences in prevalence and abundance. Figure adapted from Nayfach et al. (2016) Fig. 3 and Brito et al. (2016) Fig. 1

3.3 Determining Physiology of Ecologically and Genetically Cohesive Populations

In contrast to tracking population dynamics of closely related genotypes, only limited exploration of their physiological similarities and differences in the environment has been performed. When cultured isolates are available, it has been shown that closely related strains can adopt widely divergent physiologies, e.g., based on light spectrum preferences (Moore and Chisholm 1999) or temperature (Yung et al. 2015). Similar to metagenomics, a culture-independent approach can be taken to determine physiology of strains directly in the environment. This could theoretically be done by a combination of in situ hybridization [e.g., targeting genes sufficiently divergent to enable strain-specific hybridization using fluorescent in situ hybridization (Barrero-Canosa et al. 2017)] with assays gathering insights on physiology such as Raman spectrometry (Huang et al. 2007) or nano-SIMS (Behrens et al. 2008) that determine the ability for specific substrate uptake and/or metabolism.

Thus far, however, inferences about physiological differences between closely related but ecologically/genetically distinct populations have been made primarily by determining differences in transcript or protein abundances using metatranscriptomic and metaproteomic approaches. While translating gene expression to process rates remains challenging, recent studies integrating in situ expression and process measurements indicate the possibility to use gene expression data for process rate predictions (Wilson et al. 2017). Resolving expression patterns between closely related populations is particularly insightful when they are sympatric as these data can provide clues to the genetic differences that underlie ecological differences between these populations. Examples include the use of strain-resolved proteomics to show strain-level differences in biological phosphorus removal bioreactor communities (Wilmes et al. 2008), to identify pathways underpinning r- vs K-strategy ecotypes in biofilm development (Denef et al. 2010a), and to show physiological differences in chemotaxis and motility between closely related strains with distinct successional dynamics during preterm infant gut colonization (Brooks et al. 2015). All of these studies relied on a relatively labor-intensive manual effort to resolve strain-specific protein abundance levels and typically are focused on a single "species"-level group. More recently, automated strain-resolved metagenomic methods have also been used at a metatranscriptomic level [PanPhlAn; (Scholz et al. 2016)]. The method has been focused mostly on confirming activity of organisms in situ at strain-level resolution, for example, to show that strains vertically transmitted from mother to child were active in both the mother and child's gut environments (Asnicar et al. 2017).

Innovative tools have also been developed to gain insights into in situ growth rates of natural populations. When recruiting sequencing reads to assembled contigs from metagenomic data, it becomes apparent that replicating bacterial populations generate distinct coverage trends during bidirectional genome replication. Coverage is higher at the origin of replication and decreases toward the terminus. iRep is a tool that exploits this pattern to estimate an index of replication, which can be interpreted

as the fraction of the population that is actively making one genome copy at the time of sampling (Brown et al. 2016). The iRep estimate is a population-average value, and the existence of multiple replication forks during genome replication can bias this index (i.e., values >2 can be achieved). Olm and coauthors used iRep to track growth rates of strains across different body sites of preterm infants. First, they determined that identical strains could be found on multiple body parts. However, using iRep, they found that the replication rates of each strain differed depending on body site (Olm et al. 2017).

3.4 *Tracking Evolutionary Processes Within Ecologically and Genetically Cohesive Populations*

As stated at the start of the chapter, the analysis of metagenomic data allows us to resolve the genetic structure of natural populations. We discuss here findings related to the role of homologous recombination relative to mutation, variability in the flexible genome, and using metagenomic data to study natural selection.

Metagenomic analyses of the genetic structure of natural populations have led to new insights regarding the importance of homologous recombination within and between natural populations. Manual inspection of some of the first genomic datasets reconstructed from metagenomic data from an acid mine drainage system revealed the coexistence of multiple *Ferroplasma* populations that were inferred to be mosaic genomes originating from homologous recombination between at least three parent populations (Tyson et al. 2004). These findings were confirmed when comparing environmental metagenomic data to the genome of an isolate of the same species (Allen et al. 2007). A more quantitative approach was applied by Eppley and coauthors who found that the recombination rate within a *Ferroplasma* population was higher than the recombination rate between *Ferroplasma* populations. This suggested the presence of a species boundary based on genetic distance and within-species genetic cohesion mediated by homologous recombination (Eppley et al. 2007). Nonetheless, recombination still occurred between the two *Ferroplasma* populations, at rates proportional to varying sequence similarity across the genome. The continuation of homologous recombination in more conserved regions of the genome, while more divergent regions being already genetically more isolated, is in line with the model of temporally fragmented speciation proposed by Retchless and Lawrence (2007).

While all the studies mentioned in the previous paragraph focused on the same populations in acid mine drainage systems, they inspired new research on the importance of recombination in other systems and the development of automated methods to estimate recombination rates while controlling for sequencing errors (Johnson and Slatkin 2009). Subsequent studies found recombination to be common in marine populations, though at rates roughly four times lower than those observed in the acid mine drainage system archaeal populations (Konstantinidis and DeLong

2008). In thermophilic cyanobacteria, recombination rates have been shown to be similar to mutation rates observed through comparing metagenomic data with isolate genome sequence data (Rosen et al. 2015). In contrast, very low recombination rates relative to mutation rates were observed when comparing single cell genomes of LD12, the freshwater sister lineage of the abundant marine group *Pelagibacter* (Zaremba-Niedzwiedzka et al. 2013). These results indicate that recombination rates can be highly population-specific, and no generalization regarding the importance of recombination relative to mutation should be made. At the same time, it has to be noted that these rate comparisons generally do not control explicitly for differences in genetic distance between the sequences (and corresponding strains) considered.

Metagenomic data has also been used to study recombination within viral populations. A particular focus has been put on the CRISPR locus, which primarily functions as an adaptive defense system against viruses and is composed of an array of repeats interspersed with unique DNA segments called spacers. These CRISPR spacers most likely originate from the DNA of viruses infecting the microbial host that carries the CRISPR array in its DNA. Sequence reads that contained a sequence identical to a spacer sequence but no CRISPR repeats were identified as belonging to the targeted viruses and subsequently used to reconstruct viral genomic datasets. These reconstructions indicated the ability of some viruses to escape the microbial host's CRISPR viral defense system by homologous recombination. Erosion of linkage between viral genome variant positions at sequence lengths similar to the size of the CRISPR spacers leads to evasion of the CRISPR defense system by the viruses (Andersson and Banfield 2008). Similarly, by introducing multiple phage genotypes in a phage-bacterial coevolution experiment, recombination was shown to be an important mechanism to overcome CRISPR-based immunity (Paez-Espino et al. 2015).

Gene content differences between and within ecologically cohesive populations are observed commonly in studies using isolates. The analysis of metagenomic data has made it abundantly clear that genomic heterogeneity at the level of gene content is a hallmark of natural populations due to rapid gene gain and loss (Wilmes et al. 2009). The benefit of metagenomic data is that is has allowed for a quantitative assessment of the differential abundance of particular genomic islands between divergent environments (Coleman and Chisholm 2010) and over time (Bendall et al. 2016). The evolutionary origin of these gene content differences has been hypothesized to lie in a variety of ecological interactions (Cordero and Polz 2014) including viral predation (Rodriguez-Valera et al. 2009). From the enrichment of nutrient uptake genes under nutrient limitation (Coleman and Chisholm 2010), to the extensive gene flux in the mobile gene pool within and across species boundaries (Boucher et al. 2011), gene content differences are commonly observed to differentiate populations across space or time, despite overall cohesion of the rest of the genome.

Finally, efforts have been focused on identifying the effects of selection, which has been reviewed previously (Wilmes et al. 2009). Since that review, deep sampling of natural populations with metagenomic data generated from time series from a

relatively closed system (acid mine drainage) has been used to determine nucleotide fixation rates in the environment (Denef and Banfield 2012). The estimated rate was similar to findings in laboratory experimental evolution experiments (Barrick et al. 2009). Also, the loci affected by fixed non-synonymous mutations were biased toward regulatory genes in both the laboratory and environmental studies (Barrick et al. 2009), pointing to the importance of gene expression evolution in the early stages of evolutionary and ecological differentiation.

Despite challenges posed by dispersal in more open systems, a recent application of time-series metagenomics in a freshwater lake was able to show that both gene-specific and genome-wide selective sweeps occur in natural populations (Bendall et al. 2016) (Fig. 4). Other studies using isolates have indicated the possibility of gene-specific selective sweeps as well (Shapiro et al. 2012), and a previous metagenomic study has shown that orthologous regions differentiating coexisting organisms based on nucleotide substitutions did not show evidence of positive selection, contrary to predictions from the ecotype model (Simmons et al. 2008). Thus, population genomic studies using metagenomic data have added support for the importance of both gene-specific selective sweeps and genome-wide selective sweeps. The latter are in support of the ecotype model, i.e., that all diversity in an ecologically and evolutionary cohesive cluster of cells is regularly purged by selection of one specific adaptive genotype within the cluster, while the former indicates that recombination rates can be sufficiently high to undo the effects of selection. As argued by Shapiro and Polz (2015), there likely is no single model of speciation, but rather a spectrum determined by the contributions from gene flow and selection.

Time series metagenomic analyses have also suggested genome-wide selective sweeps in viral populations (Roux et al. 2014). Moreover, the dynamic interplay between viral and bacterial evolution has attracted the attention of researchers applying metagenomic tools, with a particular focus on dynamic changes occurring as a result of selection at CRISPR viral defense system loci. These analyses have given us insight into individual cell lineages' exposure history to viruses and have shown that CRISPR loci can be a population genome's most highly diverse loci (Tyson and Banfield 2008). Time series analyses of CRISPR sites have been used to (1) determine the retention of spacers and changes occurring in both CRISPR spacers and targeted viral genome loci (Sun et al. 2016), (2) model the evolutionary benefits of conservation of trailer-end (i.e., older) CRISPR spacers (Weinberger et al. 2012), and (3) identify molecular mechanisms such as incomplete immunity based on a single CRISPR spacer that may explain coevolutionary dynamics that deviate from those predicted by basic CRISPR immunity phage-bacteria population models (Levin et al. 2013). Similar datasets could be used to test recently proposed models of the dynamic coevolution between hosts and viruses based on CRISPR immunity (Childs et al. 2012).

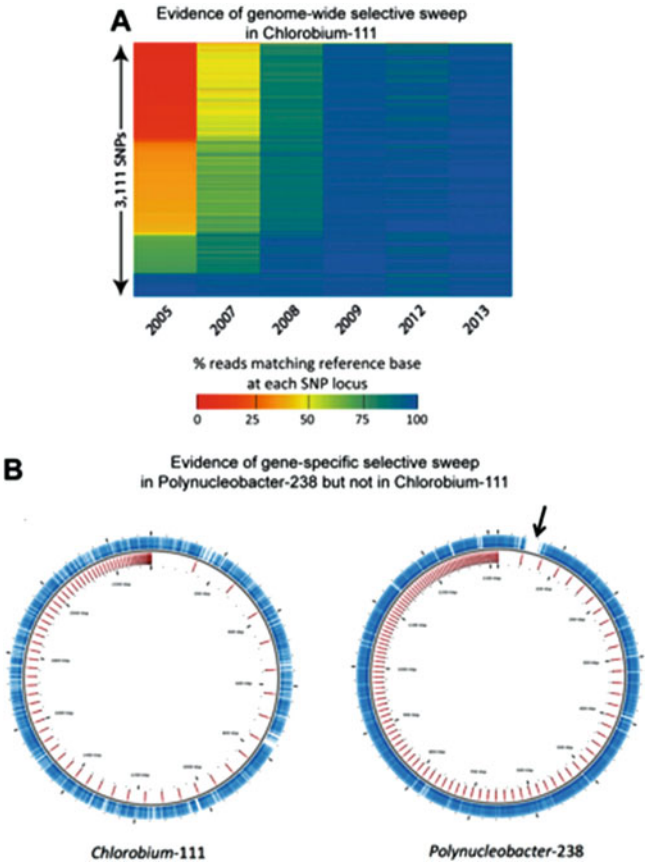

Fig. 4 Identifying selection events using metagenomic-based population genomic analyses. (**a**) Read recruitment of metagenomic data generated from samples collected from the same lake over an 8-year period to a MAG of *Chlorobium*-111 indicated gradual purging of diversity at all polymorphic sites, i.e., a genome-wide selective sweep. The "reference base" is the base most commonly observed in the final sample in 2013. Data from samples from each year were combined for read recruitment. (**b**) Comparison of distribution SNPs (blue bars) detected in the metagenomic data across the MAGs of *Chlorobium*-111 and *Polynucleobacter*-238. Contigs breaks are indicated by red lines. The *Polynucleobacter* genome shows a 21 kbp region with no SNPs (black arrow), which the authors interpreted as evidence of a gene-specific selective sweep preceding the first sample time point. Adapted from Bendall et al. (2016) Fig. 4 and Supplementary Figure S5

4 Outlook

Despite tremendous insights into natural population genomic heterogeneity gained from metagenomic approaches, some of the key limitations of metagenomic data have kept metagenomics from replacing isolate or single cell genomic approaches to perform population genomic analyses. This is particularly true for the estimation of

key population genetic parameters. While future advances in read length and base calling accuracies may facilitate the use of metagenomic data for population genomic analyses sensu stricto (Koren and Phillippy 2015), recently developed tools discussed in this chapter are allowing us to mine current metagenomic data to identify and track strains across space and time (Asnicar et al. 2017; Nayfach et al. 2016; Quince et al. 2017). As discussed above, such approaches are most powerful in the context of extensive reference genomic databases, making them currently most useful in human microbiome research. Yet, the ability to readily obtain (partial) genomic sequences from 100 s to 1000 s of single cells per sample (Kashtan et al. 2014, 2017) or directly from metagenomic data (Anantharaman et al. 2016; Delmont et al. 2017) is opening avenues to apply these tools to all microbial systems. We will likely also see further integration of population genomic analyses with metatranscriptomic or metaproteomic data or even high-throughput measurements of phenotypic features (Props et al. 2016) to gain insights into both the role of within- and between-population genomic heterogeneity and phenotypic plasticity. Improving our ability to see changes in population genetic structure of microbial populations across space and time will improve our understanding of both the evolutionary and the ecological processes that shape microbial populations (Cohan 2016; Dudaniec and Tesson 2016; Shapiro and Polz 2015). These insights are paramount in our efforts to understand how microbial populations and the communities they are part of change in composition and functioning in light of change, particularly disturbances caused by human activities.

Acknowledgments I thank Ruben Props (Ghent University) and Prof. Martin Polz (MIT) for constructive comments to help improve this chapter.

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF. Fine-scale phylogenetic architecture of a complex bacterial community. Nature. 2004;430:551–4.

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31:533–8.

Allen EE, Tyson GW, Whitaker RJ, Detter JC, Richardson PM, Banfield JF. Genome dynamics in a natural archaeal population. Proc Natl Acad Sci U S A. 2007;104:1883–8.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11:1144–6.

Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nature Commun. 2016;7:13219.

Andersson AF, Banfield JF. Virus population dynamics and acquired virus resistance in natural microbial communities. Science. 2008;320:1047–50.

Asnicar F, Manara S, Zolfo M, Truong DT, Scholz M, Armanini F, Ferretti P, Gorfer V, Pedrotti A, Tett A, Segata N. Studying vertical microbiome transmission from mothers to infants by strain-

level metagenomic profiling mSystems. 2017;2(1). pii: e00164-16. doi: https://doi.org/10.1128/mSystems.00164-16.
Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, Allen EE, Banfield JF. Lineages of acidophilic archaea revealed by community genomic analysis. Science. 2006;314:1933–5.
Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD, Land ML, Verberkmoes NC, Hettich RL, Banfield JF. Enigmatic, ultrasmall, uncultivated Archaea. Proc Natl Acad Sci U S A. 2010;107:8806–11.
Barrero-Canosa J, Moraru C, Zeugner L, Fuchs BM, Amann R. Direct-geneFISH: a simplified protocol for the simultaneous detection and quantification of genes and rRNA in microorganisms. Environ Microbiol. 2017;19:70–82.
Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature. 2009;461:1243–7.
Behrens S, Lösekann T, Pett-Ridge J, Weber PK, Ng WO, Stevenson BS, Hutcheon ID, Relman DA, Spormann AM. Linking microbial phylogeny to metabolic activity at the single-cell level by using enhanced element labeling-catalyzed reporter deposition fluorescence in situ hybridization (EL-FISH) and NanoSIMS. Appl Environ Microbiol. 2008;74:3143–50.
Bendall ML, Stevens SL, Chan LK, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, Froula J, Kang D, Tringe SG, Bertilsson S, Moran MA, Shade A, Newton RJ, McMahon KD, Malmstrom RR. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 2016;10:1589–601.
Berry MA, White JD, Davis TW, Jain S, Johengen TH, Dick GJ, Sarnelle O, Denef VJ. Are oligotypes meaningful ecological and phylogenetic units? A case study of Microcystis in freshwater lakes. Front Microbiol. 2017;8:365.
Bhaya D, Grossman AR, Steunou AS, Khuri N, Cohan FM, Hamamura N, Melendrez MC, Bateson MM, Ward DM, Heidelberg JF. Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. ISME J. 2007;1:703–13.
Blainey PC. The future is now: single-cell genomics of bacteria and archaea. FEMS Microbiol Rev. 2013;37:407–27.
Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Bapteste E, Lopez P, Tarr CL, Polz MF. Local mobile gene pools rapidly cross species boundaries to create endemicity within global Vibrio cholerae populations. MBio 2011;2(2). pii: e00335-10. doi: https://doi.org/10.1128/mBio.00335-10
Brito IL, Yilmaz S, Huang K, Xu L, Jupiter SD, Jenkins AP, Naisilisili W, Tamminen M, Smillie CS, Wortman JR, Birren BW, Xavier RJ, Blainey PC, Singh AK, Gevers D, Alm EJ. Mobile genes in the human microbiome are structured from global to individual scales. Nature. 2016;535:435–9.
Broeksema B, Calusinska M, McGee F, Winter K, Bongiovanni F, Goux X, Wilmes P, Delfosse P, Ghoniem M. ICoVeR—an interactive visualization tool for verification and refinement of metagenomic bins. BMC Bioinformatics. 2017;18:233.
Brooks B, Mueller RS, Young JC, Morowitz MJ, Hettich RL, Banfield JF. Strain-resolved microbial community proteomics reveals simultaneous aerobic and anaerobic function during gastrointestinal tract colonization of a preterm infant. Front Microbiol. 2015;6:654.
Brown CT, Olm MR, Thomas BC, Banfield JF. Measurement of bacterial replication rates in microbial communities. Nat Biotechnol. 2016;34:1256–63.
Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. Environ Microbiol. 2012;14:347–55.
Childs LM, Held NL, Young MJ, Whitaker RJ, Weitz JS. Multiscale model of CRISPR-induced coevolutionary dynamics: diversification at the interface of Lamarck and Darwin. Evolution. 2012;66:2015–29.
Cleary B, Brito IL, Huang K, Gevers D, Shea T, Young S, Alm EJ. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nat Biotechnol. 2015;33:1053–60.

Cohan FM. Bacterial speciation: genetic sweeps in bacterial species. Curr Biol. 2016;26:R112–5.
Coleman ML, Chisholm SW. Ecosystem-specific selection pressures revealed through comparative population genomics. Proc Natl Acad Sci U S A. 2010;107:18634–9.
Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, Chisholm SW. Genomic islands and the ecology and evolution of Prochlorococcus. Science. 2006;311:1768–70.
Cordero OX, Polz MF. Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol. 2014;12:263–73.
Cordero OX, Ventouras LA, DeLong EF, Polz MF. Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. Proc Natl Acad Sci U S A. 2012;109:20059–64.
Delmont TO, Eren AM, Maccario L, Prestat E, Esen ÖC, Pelletier E, Le Paslier D, Simonet P, Vogel TM. Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. Front Microbiol. 2015;6:358.
Delmont TO, Quince C, Shaiber A, Esen OC, Lee ST, Lucker S, Eren AM. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in the surface ocean. bioRxiv. 2017:129791.
DeLong EF. Microbial population genomics and ecology: the road ahead. Environ Microbiol. 2004;6:875–8.
DeLong EF. Microbial community genomics in the ocean. Nat Rev Microbiol. 2005;3:459–69.
DeLong EF. Microbial evolution in the wild. Science. 2012;336:422–4.
Denef VJ, Banfield JF. In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. Science. 2012;336:462–6.
Denef VJ, Kalnejais LH, Mueller RS, Wilmes P, Baker BJ, Thomas BC, VerBerkmoes NC, Hettich RL, Banfield JF. Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. Proc Natl Acad Sci U S A. 2010a;107:2383–90.
Denef VJ, Mueller RS, Banfield JF. AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. ISME J. 2010b;4:599–610.
Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P, Sullivan MB. Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. Nature. 2014;513:242–5.
Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF. - Community-wide analysis of microbial genome sequence signatures. Genome Biol. 2009;10: R85.
Donati C, Zolfo M, Albanese D, Tin Truong D, Asnicar F, Iebba V, Cavalieri D, Jousson O, De Filippo C, Huttenhower C, Segata N. Uncovering oral Neisseria tropism and persistence using metagenomic sequencing. Nat Microbiol. 2016;1:16070.
Dudaniec RY, Tesson SV. Applying landscape genetics to the microbial world. Mol Ecol. 2016;25:3266–75.
Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. Diversity of the human intestinal microbial flora. Science. 2005;308:1635–8.
Eppley JM, Tyson GW, Getz WM, Banfield JF. Genetic exchange across a species boundary in the archaeal genus ferroplasma. Genetics. 2007;177:407–16.
Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol. 2013;4
Eren AM, Esen C, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. Anvio: an advanced analysis and visualization platform for omics data. PeerJ. 2015;3:e1319.
Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The bacterial species challenge: making sense of genetic and ecological diversity. Science. 2009;323:741–6.
Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, DeLong EF. Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A. 2008;105:3805–10.
Fuhrman JA, Campbell L. Marine ecology: microbial microdiversity. Nature. 1998;393:410–1.

Garcia SL, Stevens SL, Crary B, Martinez-Garcia M, Stepanauskas R, Woyke T, Tringe SG, Andersson S, Bertilsson S, Malmstrom RR. Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations. bioRxiv. 2016. https://doi.org/10.1101/080168.

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. Genetic diversity in Sargasso Sea bacterioplankton. Nature. 1990;345:60.

Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol. 2005;3:679–87.

Hahn MW, Jezberová J, Koll U, Saueressig-Beck T, Schmidt J. Complete ecological isolation and cryptic diversity in Polynucleobacter bacteria not resolved by 16S rRNA gene sequences. ISME J. 2016;10:1642–55.

Hanage WP, Fraser C, Spratt BG. Sequences, sequence clusters and bacterial species. Phil Trans R Soc B. 2006;361:1917–27.

Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev. 2004;68:669–85.

Hatzenpichler R, Connon SA, Goudeau D, Malmstrom RR, Woyke T, Orphan VJ. Visualizing in situ translational activity for identifying and sorting slow-growing archaeal-bacterial consortia. Proc Natl Acad Sci U S A. 2016;113:E4069–78.

Haubold B, Pfaffelhuber P, Lynch M. mlRho—a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. Mol Ecol. 2010;19(Suppl 1):277–84.

Hehemann JH, Arevalo P, Datta MS, Yu X, Corzett CH, Henschel A, Preheim SP, Timberlake S, Alm EJ, Polz MF. Adaptive radiation by waves of gene transfer leads to fine-scale resource partitioning in marine microbes. Nat Commun. 2016;7:12860.

Huang WE, Stoecker K, Griffiths R, Newbold L, Daims H, Whiteley AS, Wagner M. Raman-FISH: combining stable-isotope Raman spectroscopy and fluorescence in situ hybridization for the single cell analysis of identity and function. Environ Microbiol. 2007;9:1878–89.

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science. 2008;320:1081–5.

Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ. 2014;2:e603.

Johnson PL, Slatkin M. Inference of population genetic parameters in metagenomics: a clean look at messy data. Genome Res. 2006;16:1320–7.

Johnson PL, Slatkin M. Accounting for bias from sequencing error in population genetic estimates. Mol Biol Evol. 2008;25:199–206.

Johnson PL, Slatkin M. Inference of microbial recombination rates from metagenomic data. PLoS Genet. 2009;5:e1000674.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science. 2014;344:416–20.

Kashtan N, Roggensack SE, Berta-Thompson JW, Grinberg M, Stepanauskas R, Chisholm SW. Fundamental differences in diversity and genomic population structure between Atlantic and Pacific Prochlorococcus. ISME J. 2017;11(9):1997–2011.

Konstantinidis KT, DeLong EF. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J. 2008;2:1052–65.

Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol. 2015;23:110–20.

Krause DJ, Whitaker RJ. Inferring speciation processes from patterns of natural variation in microbial genomes. Syst Biol. 2015;64:926–35.

Kuo C-H, Ochman H. The fate of new bacterial genes. FEMS Microbiol Rev. 2009;33:38–43.

Laczny CC, Sternal T, Plugaru V, Gawron P, Atashpendar A, Margossian HH, Coronado S, Lv d M, Vlassis N, Wilmes P. VizBin—an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome. 2015;3(1):1.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

Larkin AA, Martiny AC. Microdiversity shapes the traits, niche space, and biogeography of microbial taxa. Environ Microbiol Rep. 2017;9:55–70.

Levin BR, Moineau S, Bushman M, Barrangou R. The population and evolutionary dynamics of phage and bacteria with CRISPR-mediated immunity. PLoS Genet. 2013;9:e1003312.

Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.

Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol. 2015;33:1045–52.

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998;95:3140–5.

Malmstrom RR, Rodrigue S, Huang KH, Kelly L, Kern SE, Thompson A, Roggensack S, Berube PM, Henn MR, Chisholm SW. Ecology of uncultured Prochlorococcus clades revealed through single-cell genomics and biogeographic analysis. ISME J. 2013;7:184–98.

Marbouty M, Cournac A, Flot JF, Marie-Nelly H, Mozziconacci J, Koszul R. Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. Elife. 2014;3:e03318.

Moore LR, Chisholm SW. Photophysiology of the marine cyanobacterium Prochlorococcus: Ecotypic differences among cultured isolates. Limnol Oceanogr. 1999;44:628–38.

Morowitz MJ, Denef VJ, Costello EK, Thomas BC, Poroyko V, Relman DA, Banfield JF. Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. Proc Natl Acad Sci U S A. 2011;108:1128–33.

Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. 2016;26:1612–25.

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

Nijkamp JF, Pop M, Reinders MJ, de Ridder D. Exploring variation-aware contig graphs for (comparative) metagenomics using MaryGold. Bioinformatics. 2013;29:2826–34.

Oh S, Caro-Quintero A, Tsementzi D, DeLeon-Rodriguez N, Luo C, Poretsky R, Konstantinidis KT. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. Appl Environ Microbiol. 2011;77:6000–11.

Olm MR, Brown CT, Brooks B, Firek B, Baker R, Burstein D, Soenjoyo K, Thomas BC, Morowitz M, Banfield JF. Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. Genome Res. 2017;27:601–12.

Paez-Espino D, Sharon I, Morovic W, Stahl B, Thomas BC, Barrangou R, Banfield JF. CRISPR immunity drives rapid phage genome evolution in Streptococcus thermophilus. MBio. 2015;6.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.

Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017. https://doi.org/10.1038/s41564-017-0012-7.

Pernthaler A, Dekas AE, Brown CT, Goffredi SK, Embaye T, Orphan VJ. Diverse syntrophic partnerships from deep-sea methane vents revealed by direct cell capture and metagenomics. Proc Natl Acad Sci U S A. 2008;105:7052–7.

Props R, Monsieurs P, Mysara M, Clement L, Boon N. Measuring the biodiversity of microbial communities by flow cytometry. Meth Ecol Evol. 2016;7:1376–85.
Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, Eren AM. DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol. 2017;18:181.
Retchless AC, Lawrence JG. Temporal fragmentation of speciation in bacteria. Science. 2007;317:1093–6.
Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature. 2003;424:1042–7.
Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M, Buchanan J, Desnues C, Dinsdale E, Edwards R, Felts B, Haynes M, Liu H, Lipson D, Mahaffy J, Martin-Cuadrado AB, Mira A, Nulton J, Pasić L, Rayhawk S, Rodriguez-Mueller J, Rodriguez-Valera F, Salamon P, Srinagesh S, Thingstad TF, Tran T, Thurber RV, Willner D, Youle M, Rohwer F. Viral and microbial community dynamics in four aquatic environments. ISME J. 2010;4:739–51.
Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F, Mira A. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7:828–36.
Rodriguez-Valera F, Martin-Cuadrado AB, López-Pérez M. Flexible genomic islands as drivers of genome evolution. Curr Opin Microbiol. 2016;31:154–60.
Rosen MJ, Davison M, Bhaya D, Fisher DS. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science. 2015;348:1019–23.
Roux S, Hawley AK, Beltran MT, Scofield M, Schwientek P, Stepanauskas R, Woyke T, Hallam SJ, Sullivan MB. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell-and meta-genomics. Elife. 2014;3:e03125.
Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics. 2016;17:125.
Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods. 2016;13:435–8.
Shapiro BJ. What microbial population genomics has taught us about speciation. Popul Genom. 2017. https://doi.org/10.1007/13836_2018_10.
Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 2014;22:235–47.
Shapiro BJ, Polz MF. Microbial speciation. Cold Spring Harb Perspect Biol. 2015;7(10):a018143. https://doi.org/10.1101/cshperspect.a0181.
Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, Alm EJ. Population genomics of early events in the ecological differentiation of bacteria. Science. 2012;336:48–51.
Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. 2013;23:111–20.
Sharon I, Kertesz M, Hug LA, Pushkarev D, Blauwkamp TA, Castelle CJ, Amirebrahimi M, Thomas BC, Burstein D, Tringe SG, Williams KH, Banfield JF. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res. 2015;25:534–43.
Simmons SL, Dibartolo G, Denef VJ, Goltsman DS, Thelen MP, Banfield JF. Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. PLoS Biol. 2008;6:e177.
Sun CL, Thomas BC, Barrangou R, Banfield JF. Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. ISME J. 2016;10:858–70.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A. Structure and function of the global ocean microbiome. Science. 2015;348:1261359.

Thompson AW, Huang K, Saito MA, Chisholm SW. Transcriptome response of high-and low-light-adapted Prochlorococcus strains to changing iron availability. ISME J. 2011;5:1580–94.

Thrash JC, Temperton B, Swan BK, Landry ZC, Woyke T, DeLong EF, Stepanauskas R, Giovannoni SJ. Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. ISME J. 2014;8:1440–51.

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12:902–3.

Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 2017;27:626–38.

Tyson GW, Banfield JF. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol. 2008;10:200–7.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004;428:37–43.

Ward DV, Scholz M, Zolfo M, Taft DH, Schibler KR, Tett A, Segata N, Morrow AL. Metagenomic sequencing with strain-level resolution implicates uropathogenic E. coli in necrotizing enterocolitis and mortality in preterm infants. Cell Rep. 2016;14:2912–24.

Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975;7:256–76.

Weinberger AD, Sun CL, Pluciński MM, Denef VJ, Thomas BC, Horvath P, Barrangou R, Gilmore MS, Getz WM, Banfield JF. Persisting viral sequences shape microbial CRISPR-based immunity. PLoS Comput Biol. 2012;8:e1002475.

Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A. 2002;99:17020–4.

Whitaker RJ, Banfield JF. Population genomics in natural microbial communities. Trends Ecol Evol. 2006;21:508–16.

Wilmes P, Andersson AF, Lefsrud MG, Wexler M, Shah M, Zhang B, Hettich RL, Bond PL, VerBerkmoes NC, Banfield JF. Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. ISME J. 2008;2:853–64.

Wilmes P, Simmons SL, Denef VJ, Banfield JF. The dynamic genetic repertoire of microbial communities. FEMS Microbiol Rev. 2009;33:109–32.

Wilson ST, Aylward FO, Ribalet F, Barone B, Casey JR, Connell PE, Eppley JM, Ferrón S, Fitzsimmons JN, Hayes CT, Romano AE, Turk-Kubo KA, Vislova A, Armbrust EV, Caron DA, Church MJ, Zehr JP, Karl DM, DeLong EF. Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium *Crocosphaera*. Nat Microbiol. 2017;2:17118.

Yung CM, Vereen MK, Herbert A, Davis KM, Yang J, Kantorowska A, Ward CS, Wernegreen JJ, Johnson ZI, Hunt DE. Thermally adaptive tradeoffs in closely related marine bacterial strains. Environ Microbiol. 2015;17:2421–9.

Zaremba-Niedzwiedzka K, Viklund J, Zhao W, Ast J, Sczyrba A, Woyke T, McMahon K, Bertilsson S, Stepanauskas R, Andersson SG. Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. Genome Biol. 2013;14:R130.

Zojer M, Schuster LN, Schulz F, Pfundner A, Horn M, Rattei T. Variant profiling of evolving prokaryotic populations. PeerJ. 2017;5:e2997.

Zolfo M, Tett A, Jousson O, Donati C, Segata N. MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. Nucleic Acids Res. 2017;45:e7.

A Reverse Ecology Framework for Bacteria and Archaea

Philip Arevalo, David VanInsberghe, and Martin F. Polz

Abstract Advances in sequencing technologies have led to a rapid increase in available bacterial and archaeal genome information, but for much of this diversity, little ecological information is available. Reverse ecology provides a potential path forward by using genomic information to gain insight into the ecological associations and niche spaces of organisms. A crucial first step is to predict population structure, which provides the basis for analyzing genomes for evidence of ecological differentiation. Although delineation of bacterial and archaeal populations remains difficult, we outline how gene flow information can be used to identify populations as genetic units, which also are ecological units because adaptations can spread through them in a specific manner. This approach is particularly powerful when closely related populations are analyzed for signatures of differential selection that indicate recent ecological differentiation. Genome-wide association studies can also help identify mutations and genes underlying ecologically relevant traits. Albeit still in their infancy, reverse ecology approaches have the potential to order microbial diversity into genetically and ecologically cohesive units and hence provide the opportunity to test hypotheses about the evolutionary mechanisms creating and maintaining diversity within and between populations.

Keywords Archaea · Bacteria · Genomics · Population structure · Reverse ecology · Speciation

P. Arevalo · D. VanInsberghe · M. F. Polz (✉)
Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: mpolz@mit.edu

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_46,

1 Introduction

The rapid growth in research in microbial ecology in the past few years has been spurred by the recognition of the critical role microbes play in global environmental processes and in human health. Progress has, to a large extent, been enabled by the increasingly low cost of sequencing, which was critical in overcoming the formidable challenge of characterizing the vast co-existing genetic diversity in most of the Earth's habitats (Goodwin et al. 2016). The growth in available microbial genomes and metagenomes has helped finding associations of microbial groups with environmental variables, other organisms, or disease states (Balding 2006; Hughes-Martiny et al. 2006; Knight et al. 2012). Yet it is also true that the rapid generation of genomic sequence data has outpaced our ability to gain specific knowledge about the ecology of microbes, certainly at the level of detail required to understand the creation and maintenance of genomic diversity in the wild.

Linking ecology to genomic diversity is not a trivial enterprise due to several problems of both theoretical and practical nature. Most importantly, it has remained difficult to define the appropriate units of diversity at which to measure microbial associations, and sampling happens, more often than not, on far too large spatial and temporal scale to directly measure microbial interactions (Polz et al. 2006). For sexually reproducing eukaryotes, which form reproductive or gene-flow units (Coyne and Orr 2004; Mayr 1942), ecologically relevant units of diversity are locally co-existing members of species (populations). This congruence arises because adaptive mutations can spread within populations in an exclusive manner. It is therefore at least theoretically straightforward to analyze how physiological, behavioral, or metabolic adaptations differentiate one population from another and how selection and demography affect genetic diversity. Bacteria and archaea, however, engage in promiscuous recombination, which can, in principle, lead to the incorporation of genetic material from any other organism (Babteste and Boucher 2008; Doolittle and Papke 2006; Doolittle and Zhaxybayeva 2009). This insight has contributed to the continued largely operational (and arbitrary) definition of species (Thompson et al. 2015). However, recent work suggests that in spite of the lack of formal species definitions in the microbial world, microbial diversity is clustered and that, if appropriately chosen, sequence clusters represent ecologically distinct units (Polz et al. 2013; Shapiro and Polz 2014). This development provides hope that a biologically informed species concept will be possible in the future.

These advances in identifying ecological distinct units among microbes are exciting as they also suggest the possibility of a reverse ecology approach in which genomic features can be used to infer the ecology of a group of organisms (Fig. 1). The term reverse ecology was originally introduced in the context of eukaryotic population genomics to identify genome regions under divergent selection as evidence of recent ecological differentiation of two populations (Li et al. 2018). More recently, reverse ecology has also been used to describe an approach where systems biological modeling of metabolic features of a microbe is used to refine understanding of ecological interactions (Levy and Borenstein 2012). While

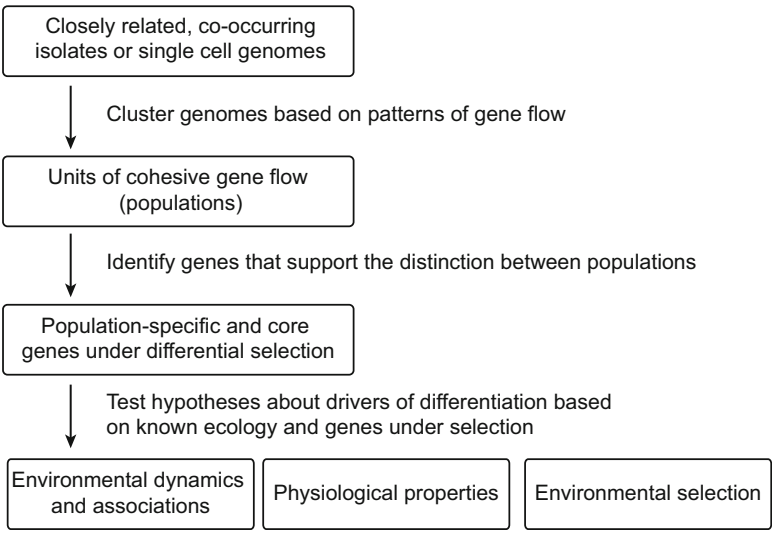

Fig. 1 Schematic of potential workflow in reverse ecology approaches

the latter is, strictly speaking, not dependent on defining population structure, it still benefits from it. In the following, we first review the application of reverse ecology in the context of population genomics of eukaryotes and then outline how reverse ecology can be applied to bacteria and archaea without requiring a formal species definition. We do this by discussing how, in spite of problematic species definitions for bacteria and archaea, progress has been made toward identifying genetic and ecological units and how these can be used for reverse ecology approaches.

2 Reverse Ecology in Eukaryotes

One of the first studies that used whole-genome sequences in a reverse ecology framework examined genomic regions under selection in two morphologically identical populations of the mosquito *Anopheles gambiae* (Lawniczak et al. 2010). These populations have overlapping geographical distributions with differences in their preferred larval environment, but no other significant behavioral or phenotypic differences are evident. There is, however, evidence that gene flow is restricted between the two populations (della Torre et al. 2005). By identifying genomic regions with low intrapopulation divergence, Lawniczak and colleagues inferred soft sweeps and hence differential positive selection. These included regions containing genes related to insecticide resistance, suggesting that anthropogenic influences may be a key driver of population differentiation. Interestingly, the researchers also discovered that these two populations were further advanced in the speciation process than previously thought and posited that this may have limited their ability to detect genes that are “instrumental” as opposed to “incidental” to

population differentiation (Lawniczak et al. 2010). In other words, factors important in initiating population differentiation are most easily observed in recently diverged populations.

The population structure of organisms of interest is, however, not always known a priori. In these cases, a robust conceptual understanding of eukaryotic speciation allows the prediction of populations from genomic data alone. This was the approach taken by Ellison and colleagues in studying the cosmopolitan yeast species *Neurospora crassa* (Ellison et al. 2011). Phylogenetic analyses as well as models of admixture and demographic history supported the existence of two distinct *N. crassa* populations, which were geographically separated (localized to Louisiana and the Caribbean, respectively) but not differentiated in any other obvious manner. Two genomic regions undergoing differential selection were identified by measuring three different metrics of nucleotide diversity and linkage disequilibrium in transcriptomic data from globally collected isolates. These regions contained genes related to temperature adaptation and circadian regulation. Importantly, this led to a specific ecological hypothesis: because the Louisiana population experiences colder winters than the Caribbean population, the Louisiana population would be better adapted to cold stress. Direct testing of growth rates at different temperatures confirmed this hypothesis, demonstrating the potential of reverse ecology to inform a targeted experimental approach.

Insights into population-specific adaptation have also been gained using reverse ecology in several other fungal groups including *Penicillium* molds used in the manufacturing of blue cheeses (Ropars et al. 2017) and the North American mushroom species *Suillus brevipes* (Branco et al. 2017). While climatic factors were found to structure *S. brevipes* populations similarly to *N. crassa*, an antifungal growth inhibition gene appears to play a key role in the differentiation of *Penicillium* populations (Ropars et al. 2015). The genomic region containing this gene was confirmed to inhibit the growth of other *Penicillium* strains when grown on a cheese-based medium, highlighting that reverse ecology can identify both abiotic and biotic determinants of population structure.

Taken together, these examples from eukaryotic species provide a broad outline for how reverse ecology can successfully be applied (Fig. 1). First, recently diverged populations differentiated by unknown factors must be identified. Next, whole-genome-based methods can be used to identify candidate genes under differential selection in the identified populations. Finally, the predicted function of these candidate genes can be used to form specific hypotheses about differentiation, which can be tested using experimental or observational methods.

3 The Bacterial and Archaeal Species Problem

If reverse ecology has been successful in eukaryotic systems for which genomic data are difficult to obtain, the dearth of such studies in bacteria and archaea may seem surprising given the relative abundance of available genomic sequences (as of July 2018, there are ~24 times more prokaryotic genomes than eukaryotic

genomes available on GenBank). However, the lack of a coherent microbial species definition represents an impediment to a reverse ecology approach because patterns of selection can only be interpreted in the context of population structure. While many attempts have been made at organizing the diversity of microbial life (Achtman and Wagner 2008; Bobay and Ochman 2017; Cohan 2002; Fraser et al. 2009; Vos 2011, #4446; Gevers et al. 2005; Konstantinidis et al. 2006; Konstantinidis and Tiedje 2005; Rosselló-Mora and Amann 2001), there is still debate how to best incorporate ecological information. Is it possible to delineate cohesive evolutionary and ecological units? The following section will give a brief overview of these attempts and their limitations.

Systematically organizing the diversity of microbial life has been a contentious task from its beginning. Early attempts took their inspiration from traditional classification schemes and categorized microbes according to a suite of phenotypic and morphological characteristics including motility, cell wall composition, and resource usage (Thompson et al. 2015; Vandamme et al. 1996). With the development of more complex methods in molecular biology, genetic measures such as DNA-DNA hybridization (Wayne et al. 1987) and restriction fragment length polymorphism (RFLP) analysis (Lee et al. 1998) were quickly added to the growing list of relevant traits, ushering in the age of so-called polyphasic taxonomy (Vandamme et al. 1996). This is essentially a "more is better" approach to taxonomy and holds that all available information is potentially relevant for classification, without an overarching theoretical framework (Gevers et al. 2005; Vandamme et al. 1996).

Within the last two decades, however, the limitations of such a system have become apparent (Thompson et al. 2015). In a polyphasic system, divisions are driven by what human observers are able to measure or have a particular interest in, regardless of the relevance of such measures to the ecology and evolution of microbes themselves. Without an a priori theory-based concept of what a population is, how can we ensure that a polyphasic system delineates ecologically and evolutionarily cohesive units?

With the rapid increase in the amount of available sequence data, comparison of similarity of genes initially seemed to provide a way to add rigor to taxonomy by defining species based on sequence identity cutoffs (Gevers et al. 2005; Keswani and Whitman 2001; Konstantinidis and Tiedje 2005; Stackebrandt and Goebel 1994). This meant that microbes could be separated into discrete, unambiguous units. Perhaps the most widely used method for delineating microbial groups centers around an operational taxonomic unit (OTU) that defines bacterial and archaeal species as groups of individuals that are >97% similar in their 16S ribosomal RNA sequence. This figure was based primarily on the observation that DNA-DNA hybridization values of 70% correspond with previously defined bacterial species and in turn that 97% rRNA identity roughly corresponded with this 70% cutoff (Stackebrandt and Goebel 1994) but was later revised to 99% rRNA identity cutoff (Keswani and Whitman 2001). However, OTUs do not group microbes into genetically or ecologically cohesive units. Close examination of groups of organisms clustered in this way reveals that significant substructure exists within these clusters

(Jaspers and Overmann 2004; Koeppel and Wu 2013). Nor is there a reason to assume that a uniform sequence cutoff reflects the evolutionary process of differentiation across highly diverse organisms. Indeed, even organisms with identical 16S rRNA have been found to inhabit different ecological niches (e.g., Hunt et al. 2008; Jaspers and Overmann 2004; Kashtan et al. 2014), highlighting that OTUs based on single-gene identity lack resolution to detect fine-scale ecological differentiation.

The problem of low resolution could potentially be solved by incorporating more genetic information, either by considering multiple genes or the entire genome. This idea guides the delineation of microbial units based on phylogenetic multilocus sequence analysis (MLSA) (Gevers et al. 2005) and genome-wide average nucleotide identity (ANI) (Goris et al. 2007; Varghese et al. 2015). In MLSA-based approaches, phylogenetic clusters are often observed, but where to draw the line delineating clusters is unclear. Indeed, such clusters among bacteria and archaea can appear "fuzzy," without a clear boundary separating them (Hanage et al. 2005; Papke et al. 2007). Similarly, the problem of arbitrary boundaries limits ANI analyses. Since identity cutoffs are based on pre-existing, taxonomic species definitions, units delineated by ANI may not correspond to ecologically cohesive populations.

4 The Nature of Gene Flow in Bacteria and Archaea

In addressing whether we can identify genetically and ecologically congruent units, we need to first consider how bacteria and archaea differ in their mode of evolution from eukaryotes. Firstly, incorporation of new genetic material is always unidirectional and leads either to gene conversion by homologous recombination or gene addition by nonhomologous recombination. While a small fraction of taxa follows a highly clonal mode of evolution (~15%), most engage in recombination (Bobay and Ochman 2017; Vos and Didelot 2009). The rates of homologous recombination can, however, differ by several orders of magnitude (Vos and Didelot 2009), and the rates of nonhomologous recombination remain poorly constrained, but there is some evidence that there are two classes of genes that differ in their turnover rate (Baumdicker 2014; Baumdicker et al. 2012). In fact, there is mounting evidence that gene addition and loss frequently happen by homologous recombination of the flanking regions so that these may turn over with similar rates to gene conversion (Cordero and Polz 2014; Cordero et al. 2012; Croucher et al. 2016). The difficulties these different modes of gene flow represent for bacterial and archaeal population genetics have recently been reviewed (Rocha 2018); what is of concern here is their effect on genotypic integrity and ecological adaptation.

In particular, horizontal gene transfer (HGT) among distantly related organisms can create genotypes that vary in properties of ecological relevance by acquiring functions, such as antibiotic resistance or nitrogen fixation, that distinguish them from otherwise closely related genotypes (Doolittle and Papke 2006; Syvanen 2012). At the same time, the recipient genotype has also become ecologically similar, in at least one niche dimension, to the organism from which it acquired

the novel pathway. In fact, such functional differentiation is observed among closely related environmental isolates (e.g., Hahn and Pockl 2005; Hehemann et al. 2016) and, in combination with high gene turnover, has been taken as evidence that gene acquisition and loss is so high as to quickly erode any niche association of lineages (Doolittle and Papke 2006). By extension, the very notion of a lineage has been questioned on the same grounds – with the consequence that nearly each genotype might represent its own independent ecological unit (Doolittle and Zhaxybayeva 2009) that can only be recognized by the functional genes it carries (Wiedenbeck and Cohan 2011). The idea that genotypic clusters should be rapidly eroded by HGT might in part be an artifact of early comparative studies of quite anciently diverged genomes. In these, only a fraction of genes in the core genome showed phylogenetic congruence, and the flexible genome seemed to be completely unrelated (Doolittle and Papke 2006; Welch et al. 2002).

In contrast to this radical view of uncoupling of lineage and function, analysis of environmental isolates and metagenomes has demonstrated that microbial communities consist of genotypic clusters of closely related organisms despite also showing evidence for extensive gene flow (e.g., Bendall et al. 2016; Bobay and Ochman 2017; Denef et al. 2010; Gevers et al. 2005; Hanage et al. 2005; Konstantinidis and DeLong 2008; Luo et al. 2011; Oh et al. 2011). Moreover, cohesive ecological dynamics and associations have been demonstrated for a growing number of cases, including for vibrios (Hunt et al. 2008) and cyanobacteria (Kashtan et al. 2014), as well as for organisms represented in several marine, freshwater, and acid-mine drainage community metagenomes (Bendall et al. 2016; Caro-Quintero and Konstantinidis 2012; Denef et al. 2010; Konstantinidis and DeLong 2008; Whitaker and Banfield 2006). These observations suggest congruence of genotypic and ecological units and are, in principle, consistent with the notion of populations as locally co-existing members of a species. The challenge is then to develop an understanding of how genotypic clusters originate and are maintained and whether they are selectively optimized to occupy sufficiently different niches to co-exist with other clusters. Importantly, any such attempt needs to take into account the considerable genotypic diversity encountered in environmental populations, which often consist of genomes differing by a considerable fraction of their gene content and displaying large allelic diversity even if most of their genes suggest close relationships (Cordero and Polz 2014).

5 Evolution of Genetic and Ecological Units

Ecological units, in the most basic sense, denote groups of organisms with common ecological functions. It is obvious that this definition represents an abstraction by the observer and is hence subject to individual preferences of how finely one wishes to demarcate units (Jax 2006). For example, does the acquisition of an antibiotic-resistant gene generate a new ecological unit or simply a variant within an existing unit? Do all sulfate-reducing bacteria represent one ecological unit since they all

carry out a common, highly relevant environmental function? In other words, is an ecotype (defined here as ecologically completely equivalent genotypes) the right unit, or should we define ecological units more broadly, and is it possible to define them based on natural processes?

To address the problem of defining natural ecological units, we return to the observation that genetic information is clustered within communities and focuses on the mechanisms capable of clustering genetic diversity, namely, migration, mutation, recombination, and selection. Although one principal way of clustering is genetic differentiation due to allopatric speciation (Denef et al. 2010; Konstantinidis and Tiedje 2005; Nemergut et al. 2011; Whitaker 2006), sympatric speciation is thought to be more common in microbial populations (Vos 2011). For sympatric speciation, the key questions are whether selection is required for cluster formation and, once clusters are formed, whether they act as genetic units within which adaptive mutations can spread. If so, clusters can be regarded as selectionally optimized, natural ecological units akin (but not identical) to populations in sexually reproducing eukaryotes.

There is, in fact, mounting evidence from modeling and empirical studies that formation of genotypic clusters (speciation) requires selection. We refer the reader for a detailed description to the chapter by Shapiro (2018); here, we focus on the information required to identify clusters and carry out reverse ecology studies and describe only briefly how the interplay of recombination and selection can have two possible outcomes. The first outcome, popularized by Frederic Cohan's ecotype model (Cohan 2002; Koeppel et al. 2008), is that under low recombination regimes, adaptive alleles or genes may spread by genome-wide selective sweeps. However, when recombination is high enough, a different outcome may result: genes or genomic regions may sweep within a population independent of the rest of the genome (Polz et al. 2013). Such gene-specific sweeps are most likely possible because both ecological (Boucher et al. 2011; Smillie et al. 2011) and genetic similarity (Fraser et al. 2007; Majewski 2001) allow for greater gene flow, so that taken together genotypically and ecologically similar populations have higher probability of both gene transfer and recombination (Shapiro and Polz 2014). In both outcomes, adaptive mutations may thus spread within a cluster, but speciation may lead to formation of new clusters if the adaptive mutation triggers to spatial or temporal niche or habitat separation (Polz et al. 2013). In order for such separation to arise, the adaptation must be accompanied with a trade-off so that the nascent population possesses a disadvantage in the ancestral but an advantage in the new niche or habitat (Wiedenbeck and Cohan 2011).

Finally, clustering is a natural consequence of speciation for both outcomes of selective sweeps (Fig. 2). Because of the high linkage, during genome-wide selective sweeps, the genotype carrying the fitness-conferring mutation expands within the niche, outcompeting all other genotypes, effectively setting diversity within the population to zero (Cohan 2002). Diversification of the winner genotype leads to formation of a cluster, which can be reinforced by recurrent sweeps (periodic selection). However, clusters can also arise under the second speciation scenario involving gene-specific sweeps (Polz et al. 2013). Because the sister populations will

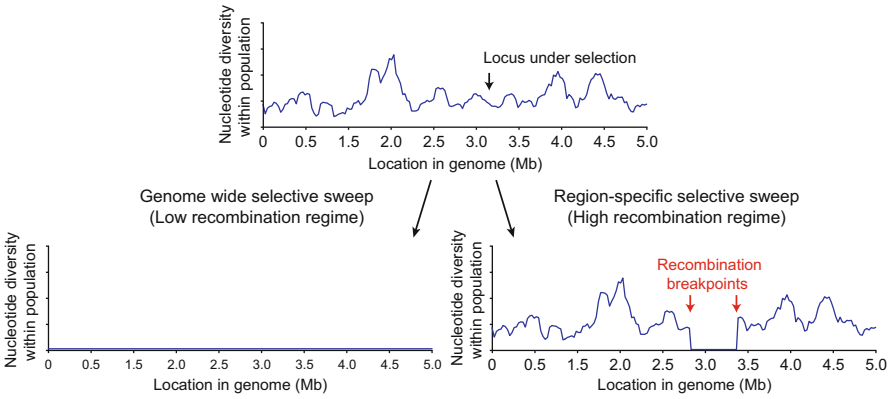

Fig. 2 The two outcomes of selective sweeps in microbial populations under high- and low-frequency recombination regimes. Under low-recombination regimes, selective sweeps will reduce overall genome-wide diversity within a population essentially to zero. Under high recombination, alleles become unlinked, and selection can operate on individual loci without affecting the diversity of alleles across the rest of the genome

occupy a different habitat than the ancestral population due to the requirement of a trade-off outlined above, gene flow between the two populations will be initially depressed because of lower encounter rates. Subsequently, the combined action of population-specific accumulation of mutation and recombination will further genetically isolate the nascent from the ancestral population, eventually leading to the formation of a distinct cluster (Polz et al. 2013).

There are examples for both genome-wide and gene-specific sweeps in environmental bacteria and archaea (Bao et al. 2016; Bendall et al. 2016; Cadillo-Quiroz et al. 2012; Dutilh et al. 2014; Rosen et al. 2015; Shapiro et al. 2012; Toro et al. 2017), although there is only a single well-documented case of the first. This genome-wide sweep was revealed by metagenomics of a 9-year time series analyzing bacterioplankton in a small lake where a population of green sulfur bacteria showed a purge of nearly all single-nucleotide polymorphism over the observation period (Bendall et al. 2016). In contrast, several other populations co-existing in the same lake showed evidence for gene-specific sweeps, i.e., progressively reduced diversity in specific genome regions, while the rest of the genome remained diverse. The first evidence for gene-specific sweeps was revealed by analysis of 20 *Vibrio cyclitrophicus* genomes with identical 16S rRNA genes and >99% amino acid identity genome wide (Shapiro et al. 2012). Despite being so genetically similar, two separate groups with distinct ecological preferences were hypothesized based on differential distribution in ocean water: while some isolates were obtained from organic particles, others occurred free-living (Hunt et al. 2008). These distinct lifestyles are made possible by the patchy distribution of resources in the ocean (Polz et al. 2006), which might promote a form of mosaic sympatry. Nearly at the same time, a second investigation targeted 12 genomes of the archaeon *Sulfolobus islandicus* from a hot spring in Kamchatka, Russia (Cadillo-Quiroz et al. 2012). These studies were similar in that both sampled closely related isolates from the

same geographic location, without apparent barriers to genetic exchange (either a single hot spring or a single bucket of seawater), and from groups of bacteria or archaea with relatively high rates of recombination (Vos and Didelot 2009; Whitaker et al. 2005). The studies differed in that the first study showed clear evidence for gene-specific sweeps, while the Cadillo-Quiroz et al. study detected continents of differentiation, perhaps indicating that the speciation process was further advanced. Moreover, while the *Vibrio* study had an a priori notion of ecological association for the two populations due to the sequencing of a gene under potential environmental selection (Shapiro et al. 2012), the *Sulfolobus* study took a purer reverse ecology approach, identifying two populations based on overall genomic similarity, then investigating recombination rates within and between groups, and characterizing phenotypic differences between them.

6 Gene Content Variation and Population Structure

One vexing issue in bacterial and archaeal population genetics is the observation that even very closely related genomes can vary considerably in gene content, leading to the categorization of the core (shared by all) and flexible (unique or shared by few) genome. How to interpret the flexible genome, the sum of which is often called the pan-genome, remains a particularly difficult problem since it is unclear how genes occurring at low frequency within populations affect the ecology of the organisms.

Part of the issue undoubtedly stems from the fact that it is very difficult to identify correct population boundaries (Rocha 2018). Because of the species issue discussed above, the unit of comparison is almost always too broad so that genes judged to be part of the flexible genome may actually be part of the core genome of more narrowly defined populations. Moreover, this issue is exacerbated by sparse sampling of closely related genomes that represent the diversity within populations. For example, in the marine cyanobacterium *Prochlorococcus*, populations in the Atlantic contain genes responsible for efficient phosphorus acquisition that are absent from populations in the Pacific, consistent with prevailing environmental concentrations of this essential nutrient (Coleman and Chisholm 2010). Hence these genes are part of the core genome of Atlantic populations but would be judged flexible genes if closely related isolates were compared from both ocean regions. The recent discovery of even finer population differentiation in sympatry by large-scale sampling of closely related genomes suggests that an even higher portion of previously flexible genes should be recategorized as core (Kashtan et al. 2014, 2017). Another example is *Campylobacter jejuni* strains that were isolated from both cattle and chickens, but the genome-wide phylogeny provided little evidence for host preference (Sheppard et al. 2013). In other words, host switching is relatively rapid, and long-term host preferences have not been established. However, a gene cluster involved in vitamin B_5 biosynthesis is universally present in cattle isolates but mostly absent in chicken isolates. This gene cluster appears to provide a selective advantage in B_5-depleted environments, which might include the cattle gut (Sheppard et al. 2013). This example stresses the importance of defining populations

in order to be able to interpret which genes are part of the core genome and hence likely essential for survival of the population.

But even if population boundaries are narrowly defined, the flexible genome remains substantial enough that cohesive populations may nevertheless contain high levels of genotypic (and to some extent, phenotypic) diversity within them. How can this be explained? First, as discussed earlier, niche-specifying variants (genes or alleles) may come with a fitness trade-off, such that they are adaptive in one niche but not in another. In a genetically cohesive population that spans two niches, different niche-specifying variants will be maintained in each niche, leading to variation at the level of the entire population (Marttinen and Hanage 2017; Niehus et al. 2015). Second, frequency-dependent selection might maintain diversity in a subset of genes involved in niche complementarity, social interactions, vaccine resistance, and predator-prey interactions (Cordero and Polz 2014; Corander et al. 2017). A relatively high proportion of genes in the flexible genome may be involved in such interactions. It has been argued previously that many genes occurring at intermediate to low frequency within genomes are involved in predation evasion by varying surface antigenicity (Cordero and Polz 2014; Rodriguez-Valera et al. 2009). Moreover, intermediate-frequency genes may be involved in frequency-dependent interactions such as public good production and cheating as well as niche-complementation (Cordero and Polz 2014). This may also explain some phenotypic variation frequently observed among closely related genotypes. For example, any secreted compound, such as enzymes, antibiotics, or signaling molecules, can become a public good that may invite cheating given sufficiently stable population structure. Indeed, broadcast exoenzymes involved in polysaccharide degradation as well as secreted siderophores for iron acquisition have been shown to occur at relatively low frequency within populations, and there is evidence that cheater populations have evolved (Cordero et al. 2012; Hehemann et al. 2016). Lastly, we should not forget that many genes, typically localized in genomic islands of high variation, appear to have such high turnover within populations that a high fraction might be (nearly) neutral to bacterial fitness (Baumdicker 2014; Baumdicker et al. 2012; Berg and Kurland 2002; Haegeman and Weitz 2012; Thompson et al. 2005). Similarly, if genome-wide selective sweeps do not periodically reduce diversity, substantial allelic diversity will be preserved through speciation. In other words, allelic diversity will be much older than the population itself (Castillo-Ramirez et al. 2011). Importantly, interpretation of such microevolutionary changes in the context of selection and population dynamics requires that sympatric genomes (i.e., from the same population) are sampled.

7 Toward Reverse Ecology in Bacteria and Archaea

As suggested above, the analysis of speciation processes can be instrumental in building hypotheses of ecological differentiation (see also Shapiro and Polz 2014). We stress that such exercise is most informative when very closely related genomes are analyzed (e.g., identical or nearly identical in 16S rRNA marker genes) since

detection of ecological differentiation becomes increasingly trivial if more divergent genomes are compared. Two kinds of hypotheses can be formed in the context of reverse ecology studies. First, the identification of nascent clusters can provide general hypotheses of genetically differentiated units that can then be investigated broadly for genetic, ecological, or physiological differentiation. This was the case in the *Sulfolobus* example mentioned above where gene flow analysis suggested recently diverged populations and screening of representative isolates from the two populations revealed growth differences in standard media (Cadillo-Quiroz et al. 2012). In general, hypothesizing population structure followed by searches for ecological or physiological differences is the only possible analysis in this speciation scenario if genome-wide selective sweeps occur since the high linkage erases signatures of selection by resetting diversity to low levels across the entire genome. In contrast, gene-specific sweeps afford building of more nuanced hypotheses of how selection has differentially affected two diverging populations by looking for genes and genome regions that show reduced nucleotide diversity (Fig. 2). This was the case in *V. cyclitrophicus*, which represents, to our knowledge, the most direct reverse ecology analysis in bacteria to date and which we describe in some detail below before outlining a general approach.

The hypothesis of two recently diverged populations originally arose from a study assessing to what extent bacteria of the family Vibrionaceae co-occurring in the same seawater samples are partitioned into ecologically differentiated populations. Over 1,000 isolates were obtained from size-fractionated water representing different potential lifestyles such as free-living or attachment to various organic particles or larger organisms (Hunt et al. 2008). Fine-scale phylogenetic relationships were subsequently assessed using multilocus sequencing, and hypotheses of population structure were generated using a mathematical model (AdaptML) that establishes the evolutionary history of ecological differentiation, which in this case is differential association with the various size fractions. Importantly, this analysis revealed several genotypic clusters that appeared nearly identical in multilocus genes but nonetheless differentially associated. The most closely related cluster containing two predicted populations, representing the aforementioned *V. cyclitrophicus*, was subsequently chosen for genomic analysis to gain further insights into the processes involved in evolution of differential association (Shapiro et al. 2012).

The analysis of genomes sampled from the two populations revealed that in spite of very close relationships (average amino acid identity across the genome ~99%) and a history of extensive recombination, there was evidence for population-specific sweeps of genome regions (Shapiro et al. 2012). Moreover, the most recent homologous recombination was population specific, and flexible gene content showed clustering consistent with population prediction, indicating that the gene pools had begun to separate. Importantly, annotation of sweep regions provided specific hypotheses of functional differentiation among the populations. Several genes in particular pointed to a differential adaptation for free-living and attached lifestyles. These comprised the *sypC* and *sypG* genes, which are important in biofilm formation and were the only locus affected by a sweep in both populations and the genes encoding the MSHA pilus, which is important in surface attachment and was present

in all genomes from the attached but absent from the free-living population (Shapiro et al. 2012). Based on these findings, a series of experiments were designed to test the prediction of behavioral underpinnings of the differential association (Yawata et al. 2014).

Detailed microfluidic experiments testing a variety of behavioral properties of members of the two nascent *V. cyclitrophicus* populations revealed that the observed habitat separation appears to be associated with an ecological trade-off, known from macroecology as competition-dispersal trade-off (Yawata et al. 2014). While one population specializes in organic particle exploitation through strong attachment and growth in biofilms, the other population only rarely attaches yet is specialized for dispersal by rapidly detecting and swimming toward new resources, implying that it can better exploit short-lived nutrient patches consisting of locally available dissolved organic matter (Yawata et al. 2014). Hence the experiments indicated that the evolution of fine-scale behavioral adaptations may have been responsible for the onset of ecological differentiation between strains of the ancestral population. This is because differential specialization for particulate and dissolved resources is associated with decreased encounter rates and hence may serve to initiate the gene pool separation required for speciation as discussed above.

8 Guidelines for Reverse Ecology Approaches in Bacteria and Archaea

Based on the findings outlined above, it is possible to propose a more general reverse ecology approach for bacteria and archaea (Shapiro and Polz 2014). We reiterate that the goal is to determine whether a sample of closely related, sympatric genome sequences constitute one or more genotypic units and to test how these units might differ in their ecology either by mapping of these clusters onto environmental gradients or patches or by laboratory tests.

In most cases, some notion of the ecology, metabolism, or phylogenetic relatedness of the target group will exist and can influence the sampling scheme. For example, a collection of closely related isolates could be chosen from two or more hypothesized niches (or microhabitats) or phenotypic groups in order to test whether these groups behave as separate genotypic units and to uncover the genes or mutations that might contribute to their ecological differences. Isolates should, however, be sampled from the same geographic location in order to reduce the effects of allopatric divergence and focus on the effects of local selection and recombination. Some a priori information – perhaps from a previous phylogenetic or metagenomic survey – may also be required in order to select a subset of closely related populations from the community.

It is possible to either choose a genomic or metagenomic approach to assess the diversity within populations. Whole-genome sequencing of cultured isolates or isolated (but uncultured) single cells is preferable because it reveals information about how genes and mutations are linked within genomes, facilitating inferences

about recombination events among genomes. Metagenomic sequencing has the advantage of sampling more individuals within an environment than are generally possible to isolate, but linkage information will be limited by the sequencing read length and quality of the assembly (Denaf 2018). Most importantly, unbiased metagenomic sequencing will only provide an appropriate population genomic dataset for populations that are relatively abundant in the sampled environment. The power of metagenomic data can be boosted significantly if they are gathered as a time series. Although such datasets are currently rare and potentially challenging to collect, they can follow the speciation process in real time and potentially catch selective sweeps and niche-specifying events in action (Bendall et al. 2016). Fine-grained time series or highly resolved spatial sampling might also follow shifting ecological conditions over time and space, revealing independent behaviors of different clusters (Martin-Platero et al. 2018).

Assembly and alignment of genomes follow general methods (reviewed in Denaf 2018; Didelot 2017) and are only briefly outlined here. Complete genome sequences are more readily assembled from isolates, but assembly can also be attempted on metagenomic data, taking care to guard against or account for different individuals being co-assembled into a single genome. The main goal of alignment of genomes is to define core and flexible components. Here, particular care must be taken to only define these categories for organisms that co-occur and hence have the potential to be connected by contemporary gene flow and be subjected to consistent environmental selection.

The next step is to evaluate phylogenetic signals in SNPs found in the core genome. Standard phylogenetic methods can be used to build a core genome-wide phylogeny, and the average impact of recombination can be measured by assessing linkage disequilibrium between SNPs. Specific recombination events and breakpoints can then be identified using methods such as BratNextGen (Marttinen et al. 2012), ClonalFrame/ClonalOrigin (Didelot et al. 2010), and STARRInIGHTS (Shapiro et al. 2012), which are also reviewed in Didelot (2017). These analyses will reveal the number of major genotypic units (well-supported monophyletic groups) and whether these units are supported genome wide (consistent with mostly clonal evolution) or in "islands" or "continents" of the genome (Cadillo-Quiroz et al. 2012; Shapiro et al. 2012). Importantly, linkage can be high, depending on the frequency of recombination and size of incorporated DNA, so that considerable hitchhiking with mutations under selection may occur (Rocha 2018).

If populations were hypothesized a priori based on an ecological axis of interest, as in the *V. cyclitrophicus* example above, it is possible to assess whether these presumed populations correspond to genotypic clusters or not. If genome-wide diversity is clustered according to ecology, this suggests that stable clusters have formed, i.e., the speciation process is further advanced. If there is a preference for recombination within rather than between ecological groups, the single population might be on a trajectory toward speciation. If there is little or no phylogenetic clustering according to ecology, the hypothesized populations likely constitute a single, phenotypically diverse population. In this case, certain (flexible) genes or

(core) mutations that associate with ecology might be identified by genome-wide association study (GWAS) for which a number of techniques are now available that are robust toward the modes of bacterial and archaeal evolution (Chen and Shapiro 2015; Falush and Bowden 2006; Sheppard et al. 2013; Yahara et al. 2017).

If populations were not hypothesized a priori (a "purer" reverse ecology approach), a first step is to assess how many phylogenetic groups were identified. If phylogenetic groupings are supported genome wide, this suggests stable differentiation, the ecological basis of which remains unknown but can be tested by phenotypically characterizing representative isolates from each group and/or mapping genotypic clusters onto environmental samples. If groupings are not supported genome wide, genomic regions containing the bulk of the phylogenetic signal, or signals of positive selection (Shapiro 2014; Shapiro et al. 2009), frequent recombination, or dense polymorphism can be functionally annotated to generate hypotheses about their possible ecological roles.

9 Future Perspectives

Reverse ecology is in its infancy, but the potential for revolutionizing our understanding of how microbial diversity is organized in the environment is clear. It is already evident from the handful of studies that have looked at ecological partitioning of closely related groups that ecological differentiation can happen on very fine scales of genotypic relatedness, certainly below the resolution afforded by 16S rRNA marker genes. Importantly, when genotypic clusters can be detected, they should be ecologically differentiated from other such clusters. Although this does not preclude some level of ecological diversity within these clusters due to the potential of acquisition of novel, niche-specifying genes, such diversity should be relatively minor because selection can only maintain a limited number of ecologically divergent loci within the same, genetically mixed population (Friedman et al. 2013). Hence a reverse ecology strategy, in which genotypic clusters among co-existing microbes are identified as a first step toward identifying ecologically cohesive populations, is potentially easier than the forward approach, which is to map marker genes onto many environmental samples in the hopes of finding significant ecological associations.

As genome sequencing becomes more and more broadly accessible because of decreased cost and increased throughput, it will become feasible to sequence sufficient numbers of closely related genomes from the same environmental samples, either in the form of isolates or single-cell genomes. Moreover, improved coverage and assembly techniques will also allow increased identification of genotypic clusters from metagenomic samples. Once these genomes are available, they can serve two purposes. First, they can be used to delineate clusters, and second, they can help build hypotheses of environmental differentiation by searching for genes of potential ecological relevance. In that way, some guess as to the population's niche can be made before engaging in the exercise of mapping the cluster onto environmental

samples and identifying correlations with biotic and abiotic environmental metadata. We stress that this exercise must consider the fine structure of the environment since microbial habitats and interactions often occur at small spatial (micro- to millimeters) and temporal scales (minute to days) (David et al. 2014; Martin-Platero et al. 2018; Polz et al. 2006). Given sufficient environmental and genomic sampling, GWAS can provide valuable further insights as to the causes of allele and gene diversity within and between populations. Enabled by our still-evolving knowledge of microbial diversity, the combined toolkits of population genomics, reverse ecology, and GWAS will no doubt continue to enrich and expand our understanding as we move from descriptive to more mechanistic models of the ecological and genetic structure of microbes in the wild.

References

Achtman M, Wagner M. Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol. 2008;6:431–40.

Babteste E, Boucher Y. Lateral gene transfer challenges principles of microbial systematics. Trends Microbiol. 2008;16:200–2007.

Balding D. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006;7:781–91.

Bao YJ, Shapiro BJ, Lee SW, Ploplis VA, Castellino FJ. Phenotypic differentiation of Streptococcus pyogenes populations is induced by recombination-driven gene-specific sweeps. Sci Rep. 2016;6:36644.

Baumdicker F. The site frequency spectrum of dispensable genes. Theor Popul Biol. 2014;100C:13–25.

Baumdicker F, Hess WR, Pfaffelhuber P. The infinitely many genes model for the distributed genome of bacteria. Genome Biol Evol. 2012;4:443–56.

Bendall ML, Stevens SL, Chan LK, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 2016;10:1589–601.

Berg OG, Kurland CG. Evolution of microbial genomes: sequence acquisition and loss. Mol Biol Evol. 2002;19:2265–76.

Bobay LM, Ochman H. Biological species are universal across life's domains. Genome Biol Evol. 2017; https://doi.org/10.1093/gbe/evx026.

Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Bapteste E, Lopez P, Tarr CL, Polz MF. Local mobile gene pools rapidly cross species boundaries to create endemicity within global Vibrio cholerae populations. MBio. 2011;2:e00335–10.

Branco S, Bi K, Liao HL, Gladieux P, Badouin H, Ellison CE, Nguyen NH, Vilgalys R, Peay KG, Taylor JW, et al. Continental-level population differentiation and environmental adaptation in the mushroom Suillus brevipes. Mol Ecol. 2017;26:2063–76.

Cadillo-Quiroz H, Didelot X, Heid NL, Herrara A, Darling A, Reno ML, Krause DJ, Whitaker RJ. Patterns of gene flow define species of thermophilic Archaea. PLoS Biol. 2012;10:e1001265.

Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. Environ Microbiol. 2012;14:347–55.

Castillo-Ramirez S, Harris SR, Holden MT, He M, Parkhill J, Bentley SD, Feil EJ. The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS Pathog. 2011;7:e1002129.

Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015;25:17–24.

Cohan FM. What are bacterial species. Annu Rev Microbiol. 2002;56:457–87.
Coleman ML, Chisholm SW. Ecosystem-specific selection pressures revealed through comparative population genomics. Proc Natl Acad Sci U S A. 2010;107:18634–9.
Corander J, Fraser C, Gutmann MU, Arnold B, Hanage WP, Bentley SD, Lipsitch M, Croucher NJ. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. Nat Ecol Evol. 2017;1:1950–60.
Cordero OX, Polz MF. Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol. 2014;12:263–73.
Cordero OX, Ventouras LA, Delong EF, Polz MF. Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. Proc Natl Acad Sci U S A. 2012;109:20059–64.
Coyne JA, Orr HA. Speciation. Sunderland: Sinauer Associates; 2004.
Croucher NJ, Mostowy R, Wymant C, Turner P, Bentley SD, Fraser C. Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict. PLoS Biol. 2016;14:e1002394.
David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, Erdman SE, Alm EJ. Host lifestyle affects human microbiota on daily timescales. Genome Biol. 2014;15:R89.
della Torre A, Tu Z, Petrarca V. On the distribution and genetic differentiation of Anopheles gambiae s.s. molecular forms. Insect Biochem Mol Biol. 2005;35:755–69.
Denef VJ. Peering into the genetic makeup of natural microbial populations using metagenomics. In: Population genomics. Cham: Springer; 2018. https://doi.org/10.1007/13836_2018_14.
Denef VJ, Mueller RS, Banfield JF. AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. ISME J. 2010;4:599–610.
Didelot X. Computational methods in microbial population genomics. In: Population genomics. Cham: Springer; 2017. https://doi.org/10.1007/13836_2017_3.
Didelot X, Lawson D, Darling A, Falush D. Inference of homologous recombination in bacteria using whole-genome sequences. Genetics. 2010;186:1435–49.
Doolittle WF, Papke RT. Genomics and the bacterial species problem. Genome Biol. 2006;7:116.
Doolittle WF, Zhaxybayeva O. On the origin of prokaryotic species. Genome Res. 2009;19:744–56.
Dutilh BE, Thompson CC, Vicente AC, Marin MA, Lee C, Silva GG, Schmieder R, Andrade BG, Chimetto L, Cuevas D, et al. Comparative genomics of 274 Vibrio cholerae genomes reveals mobile functions structuring three niche dimensions. BMC Genomics. 2014;15:654.
Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, Taylor JW. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proc Natl Acad Sci. 2011;108:2831–6.
Falush D, Bowden R. Genome-wide association mapping in bacteria? Trends Microbiol. 2006;14:353–5.
Fraser C, Hanage WP, Spratt BG. Recombination and the nature of bacterial speciation. Science. 2007;315:476–80.
Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The bacterial species challenge: making sense of genetic and ecological diversity. Science. 2009;232:741–6.
Friedman J, Alm EJ, Shapiro BJ. Sympatric speciation: when is it possible in bacteria? PLoS One. 2013;8:e53539.
Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, de Peer YV, Vandamme P, Thompson FL, et al. Re-evaluating prokaryotic species. Nat Rev Microbiol. 2005;3:733–9.
Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.
Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007;57:81–91.
Haegeman B, Weitz JS. A neutral theory of genome evolution and the frequency distribution of genes. BMC Genomics. 2012;13:196.

Hahn MW, Pockl M. Ecotypes of planktonic actinobacteria with identical 16S rRNA genes adapted to thermal niches in temperate, subtropical, and tropical freshwater habitats. Appl Environ Microbiol. 2005;71:766–73.

Hanage WP, Fraser C, Spratt BG. Fuzzy species among recombinogenic bacteria. BMC Biol. 2005;3:6. https://doi.org/10.1186/1741-7007-1183-1186.

Hehemann JH, Arevalo P, Datta MS, Yu X, Corzett CH, Henschel A, Preheim SP, Timberlake S, Alm EJ, Polz MF. Adaptive radiation by waves of gene transfer leads to fine-scale resource partitioning in marine microbes. Nat Commun. 2016;7:12860.

Hughes-Martiny JB, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Divine MC, Kane M, Krumins JA, Kuske CR, et al. Microbial biogeography: putting microorganisms on the map. Nat Rev Microbiol. 2006;4:102–12.

Hunt DE, David LD, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science. 2008;320:1081–5.

Jaspers E, Overmann J. Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. Appl Environ Microbiol. 2004;70:4831–9.

Jax K. Ecological units: definitions and application. Q Rev Biol. 2006;81:237–58.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science. 2014;344:416–20.

Kashtan N, Roggensack SE, Berta-Thompson JW, Grinberg M, Stepanauskas R, Chisholm SW. Fundamental differences in diversity and genomic population structure between Atlantic and Pacific Prochlorococcus. ISME J. 2017;11:1997–2011.

Keswani J, Whitman WB. Relationship of 16S rRNA sequence similarity to DNA hybridization in prokaryotes. Int J Syst Evol Microbiol. 2001;51:667–78.

Knight R, Jansson J, Field D, Fierer N, Desai N, Fuhrman JA, Hugenholtz P, van der Lelie D, Meyer F, Stevens R, et al. Unlocking the potential of metagenomics through replicated experimental design. Nat Biotechnol. 2012;30:513–20.

Koeppel AF, Wu M. Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. Nucleic Acids Res. 2013;41:5175–88.

Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, Rooney AP, Brambilla E, Connor N, Ratcliff RM, et al. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc Natl Acad Sci U S A. 2008;105:2504–9.

Konstantinidis KT, DeLong EF. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J. 2008;10:1052–65.

Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. 2005;102:2567–72.

Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomics area. Philos Trans R Soc Lond Ser B Biol Sci. 2006;361:1929–40.

Lawniczak MKN, Emrich SJ, Holloway AK, Regier AP, Olson M, White B, Redmond S, Fulton L, Appelbaum E, Godfrey J, et al. Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences. Science. 2010;330:512–4.

Lee IM, Gundersen-Rindal DE, Davis RE, Bartoszyk IM. Revised classification scheme of phytoplasmas based on RFLP analyses of 16S rRNA and ribosomal protein gene sequences. Int J Syst Bacteriol. 1998;48:1153–69.

Levy R, Borenstein E. Reverse ecology: from systems to environments and back. Adv Exp Med Biol. 2012;751:329–45.

Li QL, Yi SC, Li DZ, Nie XP, Li SQ, Wang MQ, Zhou AM. Optimization of reverse chemical ecology method: false positive binding of Aenasius bambawalei odorant binding protein 1 caused by uncertain binding mechanism. Insect Mol Biol. 2018;27:305–18.

Luo CW, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. Proc Natl Acad Sci U S A. 2011;108:7200–5.
Majewski J. Sexual isolation in bacteria. FEMS Microbiol Lett. 2001;199:161–9.
Martin-Platero AM, Cleary B, Kauffman K, Preheim SP, McGillicuddy DJ, Alm EJ, Polz MF. High resolution time series reveals cohesive but short-lived communities in coastal plankton. Nat Commun. 2018;9:266.
Marttinen P, Hanage WP. Speciation trajectories in recombining bacterial species. PLoS Comput Biol. 2017;13:e1005640.
Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res. 2012;40:e6.
Mayr E. Systematics and the origin of species. New York: Columbia University Press; 1942.
Nemergut DR, Costello EK, Hamady M, Lozupone C, Jiang L, Schmidt SK, Fierer N, Townsend AR, Cleveland CC, Stanish L, et al. Global patterns in the biogeography of bacterial taxa. Environ Microbiol. 2011;13:135–44.
Niehus R, Mitri S, Fletcher AG, Foster KR. Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nat Commun. 2015;6:8924.
Oh S, Caro-Quintero A, Tsementzi D, DeLeon-Rodriguez N, Luo C, Poretsky R, Konstantinidis KT. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. Appl Environ Microbiol. 2011;77:6000–11.
Papke RT, Zhaxybayeva O, Feil EJ, Sommerfeld K, Muise D, Doolittle WF. Searching for species in haloarchaea. Proc Natl Acad Sci U S A. 2007;104:14092–7.
Polz MF, Hunt DE, Preheim SP, Weinreich DM. Patterns and mechanisms of genetic and phenotypic differentiation in marine microbes. Philos Trans R Soc Lond B. 2006;361:2009–21.
Polz MF, Alm EJ, Hanage WP. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet. 2013;29:170–5.
Rocha EPC. Neutral theory, microbial practice: challenges in bacterial population genetics. Mol Biol Evol. 2018;35:1338–47.
Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F, Mira A. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7:828–36.
Ropars J, Rodríguez de la Vega RC, López-Villavicencio M, Gouzy J, Sallet E, Dumas É, Lacoste S, Debuchy R, Dupont J, Branca A, et al. Adaptive horizontal gene transfers between multiple cheese-associated fungi. Curr Biol. 2015;25:2562–9.
Ropars J, López-Villavicencio M, Snirc A, Lacoste S, Giraud T. Blue cheese-making has shaped the population genetic structure of the mould Penicillium roqueforti. PLoS One. 2017;12:e0171387.
Rosen MJ, Davison M, Bhaya D, Fisher DS. Microbial diversity. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science. 2015;348:1019–23.
Rosselló-Mora R, Amann R. The species concept for prokaryotes. FEMS Microbiol Rev. 2001;25:39–67.
Shapiro BJ. Signatures of natural selection and ecological differentiation in microbial genomes. Adv Exp Med Biol. 2014;781:339–59.
Shapiro BJ. What microbial population genomics has taught us about speciation. In: Population genomics. Cham: Springer; 2018. https://doi.org/10.1007/13836_2018_10.
Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 2014;22:235–47.
Shapiro BJ, David LA, Friedman J, Alm EJ. Looking for Darwin's footprints in the microbial world. Trends Microbiol. 2009;17:196–204.
Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ. Population genomics of early events in the ecological differentiation of bacteria. Science. 2012;336:48–51.

Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MC, Parkhill J, Falush D. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci U S A. 2013;110:11923–7.

Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. Ecology drives a global network of gene exchange connecting the human microbiome. Nature. 2011;480:241–4.

Stackebrandt E, Goebel BM. Taxonomic note: a place for DNA-DNA reassociation kinetics and sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol. 1994;44:846–9.

Syvanen M. Evolutionary implications of horizontal gene transfer. Annu Rev Genet. 2012;46:341–58.

Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, Sarma-Rupavtarm R, Distel DL, Polz MF. Genotypic diversity within a natural coastal bacterioplankton population. Science. 2005;307:1311–3.

Thompson CC, Amaral GR, Campeao M, Edwards RA, Polz MF, Dutilh BE, Ussery DW, Sawabe T, Swings J, Thompson FL. Microbial taxonomy in the post-genomic era: rebuilding from scratch? Arch Microbiol. 2015;197:359–70.

Toro N, Villadas PJ, Molina-Sanchez MD, Navarro-Gomez P, Vinardell JM, Cuesta-Berrio L, Rodriguez-Carvajal MA. The underlying process of early ecological and genetic differentiation in a facultative mutualistic Sinorhizobium meliloti population. Sci Rep. 2017;7:675.

Vandamme P, Pot B, Gillis M, De Vos P, Kersters K, Swings J. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev. 1996;60:407–38.

Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati A. Microbial species delineation using whole genome sequences. Nucleic Acids Res. 2015;43:6761–71.

Vos M. A species concept for bacteria based on adaptive divergence. Trends Microbiol. 2011;19:1–7.

Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 2009;3:199–208.

Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, Moore LH, Moore WEC, Murray RGE, Stackebrandt E, et al. Report of the Ad Hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Evol Microbiol. 1987;37:463–4.

Welch RA, Burland V, Plunkett G III, Redford P, Roesch P, Rasko D, Buckles EL, Liou S-R, Boutin A, Hackett J, et al. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci U S A. 2002;99:17020–4.

Whitaker RJ. Allopatric origins of microbial species. Philos Trans R Soc Lond Ser B Biol Sci. 2006;361:1975–84.

Whitaker RJ, Banfield JF. Population genomics in natural microbial communities. Trends Ecol Evol. 2006;21:508–16.

Whitaker RJ, Grogan DW, Taylor JW. Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus. Mol Biol Evol. 2005;22:2354–61.

Wiedenbeck J, Cohan FM. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev. 2011;35:957–76.

Yahara K, Meric G, Taylor AJ, de Vries SP, Murray S, Pascoe B, Mageiros L, Torralbo A, Vidal A, Ridley A, et al. Genome-wide association of functional traits linked with Campylobacter jejuni survival from farm to fork. Environ Microbiol. 2017;19:361–80.

Yawata Y, Cordero OX, Menolascina F, Hehemann JH, Polz MF, Stocker R. Competition-dispersal tradeoff ecologically differentiates recently speciated marine bacterioplankton populations. Proc Natl Acad Sci U S A. 2014;111:5622–7.

Part II
Population Genomics of Bacteria and Archaea

What Is a *Pseudomonas syringae* Population?

David A. Baltrus

Abstract Although they are often best known as causative agents of agricultural disease, many phytopathogen lineages, like *Pseudomonas syringae*, have been sampled across a wide range of environmental contexts. These may be frequently isolated as epiphytes on disease-free plants as well as from sources associated with the water cycle like rivers, lakes, rain, snow, and clouds. The ability of these bacteria to persist across such diverse environments poses a great challenge for understanding population dynamics because adaptation likely occurs across numerous distinct niches and evolutionary parameters and will likely differ widely depending on specific contexts. Within the literature, there is an intrinsic tendency to treat all strains within these lineages the same, but such a treatment likely obscures interesting and important nuances between isolates. In this chapter, I will focus on *P. syringae* and explore what is known about the evolutionary dynamics of this group at the levels of genomes, phylogroups, and (broadly defined) species. I will highlight many ways in which populations could differ and will touch upon what is known and has been learned from numerous genome sequencing efforts, which hopefully shine a light toward a path forward to resolve numerous nomenclatural challenges. I will point toward the generality of what is known about *P. syringae* and how this may apply to other environmental systems. While there remains much to learn, the ever-increasing rate of accumulation of genomic data from diverse sources has certainly helped our ability to at least frame the evolutionarily important questions. Building from these, an impending wave of future data promises to be a powerful tool for resolving some of these discussions.

Keywords Environmental survival · Epidemiology · Phytopathogen evolution · Population structure · *Pseudomonas syringae*

D. A. Baltrus (✉)
School of Plant Sciences, University of Arizona, Tucson, AZ, USA

School of Animal and Comparative Biomedical Sciences, University of Arizona, Tucson, AZ, USA
e-mail: baltrus@email.arizona.edu

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_25,

1 Introduction

Step outdoors and there is a reasonable chance you'll find bacteria that fall under the taxonomic umbrella of *Pseudomonas syringae*. Although nomenclature can be a bit muddled (Baltrus 2016), lineages belonging to the *P. syringae* species complex are best known as important phytopathogens of numerous crops around the world and as a model system for understanding the molecular basis of plant pathogenicity (Hirano and Upper 2000; Baltrus et al. 2017; O'Brien et al. 2011; Mansfield et al. 2012; Xin and He 2013). However, close relatives of these pathogenic strains (and in some cases even known pathogens) can readily be found as epiphytes of plants, in leaf litter, and from a variety of sources that constitute important parts of the water cycle such as rivers, lakes, rainwater, snow, and clouds (Morris et al. 2008, 2013). Strains of *P. syringae* are even capable of being vectored by insects, albeit under laboratory conditions, and are considered potential entomopathogens (Stavrinides et al. 2009; Hendry et al. 2014). In reality, the *P. syringae* species complex represents a geographically and environmentally ubiquitous group of bacteria, and these very characteristics create a variety of problems for discussing evolutionary dynamics within their populations. Diversity of roles and environments breeds tension across ecological and evolutionary descriptions for this group, as well as many other environmentally ubiquitous taxa, because the lineages don't occupy a simple niche that can easily be classified. This chapter will not come close to resolving any of these challenges, but I will build on previous treatments of this topic (Vinatzer and Monteil 2014; Vinatzer et al. 2014; Baltrus et al. 2017) to bridge the gaps between how *P. syringae* is seen from the perspective of ecological, evolutionary, and phytopathogenic viewpoints and highlight different ways to think about this bacterium at the population level.

Since this discussion relies heavily on how to interpret evolutionary trends and diversity throughout *P. syringae*, it will inherently broach questions concerning how to define the term "population." It is my hope that these discussions will provide a foothold to organize thoughts concerning *P. syringae* because, although specific conditions and transmission capabilities could be vastly divergent across strains, questions concerning the structures of interactions are relevant regardless. My intent is that this discussion can be viewed broadly beyond a focus on just one type of bacteria and that that same lines of thought could apply across other systems involving phytopathogens or symbionts found widely throughout the environment (like *Erwinia*, *Burkholderia*, and *Streptomyces*). In the end, evolutionary dynamics are governed by a core set of parameters, and with sufficient understanding of these parameters, we should be able to gauge how strains evolve. The problem with many phytopathogens is that we lack basic data about these parameters, and thus a first step is to identify, as famously coined in a different context, the known and unknown "unknowns."

2 *Pseudomonas syringae*: A Model Phytopathogen

The *P. syringae* species complex is perhaps best described as a ragtag group of misfits, with fairly broad metabolic capabilities and the ability to survive across a variety of environments (Baltrus 2016; Hirano and Upper 2000; Morris et al. 2013). *P. syringae* doesn't quite get the attention of its human pathogenic cousin, *P. aeruginosa*, and appears to be slightly more uniform than other species, such as the well-known physiological powerhouse *P. putida*, and plant growth promoters like *P. fluorescens*. Crop diseases caused by strains of *P. syringae* can manifest as blights, spots, and streaks on leaves and as cankers or galls on woody plants with specific symptoms dependent on particular combinations of strains and hosts (O'Brien et al. 2011; Mansfield et al. 2012; Baltrus et al. 2017). Virulence in *P. syringae* is thought to be critically dependent on a structure called a type three secretion system (T3SS) and effector proteins which are translocated into plant cells to disrupt immune responses (Collmer et al. 2000). Almost all strains that clearly fall within the *P. syringae* species complex appear to harbor at least one T3SS, the one known exception being a strain from pathovar *actinidiae* that has naturally deleted this structure, and each strain is thought to contain between approximately five and 40 effectors (Baltrus et al. 2017; O'Brien et al. 2011; McCann et al. 2017; Dillon et al. 2017). T3SS effector proteins are produced by the bacteria but are translocated into host cells where they can manipulate multiple plant pathways in a variety of ways (Lindeberg et al. 2012). Pathogenic strains may additionally produce a handful of secreted secondary metabolites which clearly affect plant physiology, notably toxins and plant hormone mimics such as syringolin, syringomycin, mangotoxin, syringopeptin, coronatine, tabtoxin, and phaseolotoxin (Bender et al. 1999; Carrión et al. 2012; Schellenberg et al. 2008). As with many other pathogen-host systems, disease is the manifestation of a complicated dance between multiple virulence pathways that differentially interact with host immune and metabolic pathways.

There has been plenty already written about the role of *P. syringae* as an agricultural pest and about the molecular basis of pathogenicity for this species. There also exist many valid questions concerning diversity in how disease symptoms from *P. syringae* are manifested, in how host range has traditionally been considered, and in how "virulence" is defined. As it's not my goal to delve into these topics too much here, I'll point the reader to a variety of reviews on these subjects (Baltrus et al. 2017; Lindeberg et al. 2012; O'Brien et al. 2011; Hirano and Upper 2000; Ichinose et al. 2013). However, it is worth pointing out that the traditional strict focus on phytopathogenicity and related properties has obscured the idea that *P. syringae* can frequently be found naturally covering leaves of many plants as epiphytes in the absence of disease (Hirano and Upper 2000; Morris et al. 2017). These strains seemingly live in happy commensalism with plants and yet still reach decently high census sizes, 10^4 or 10^5 cells per g of plant tissue (Morris et al. 2008). Some strains found in the environment are closely related to those known to cause disease on crop plants and are virtually indistinguishable from pathogens at a genomic level (Monteil et al. 2013, 2016). It is possible that known virulence pathways also play a

role in *P. syringae* survival in these alternate contexts and that secondary hosts could enable a large pool that significantly contributes to widespread dispersal of strains, but to date data addressing these hypotheses is limited.

3 *Pseudomonas syringae*: A Burgeoning Ecological Model

Recent years have seen a growing appreciation for the ecology of *P. syringae* outside of the contexts of plant disease (Morris et al. 2008, 2010, 2013). Strains have been isolated from various water sources like streams and lakes but also from clouds as well as freshly fallen snow and rain. It is clear that *P. syringae*, and other phytopathogen relatives from genera like *Erwinia*, can be lifted up into the atmosphere and dispersed over wide areas through precipitation (Failor et al. 2017). For the purposes of evolutionary arguments, it's unclear whether strains actively grow during periods of atmospheric transit or if there are significant effects of natural selection due to differential death under these conditions (Morris et al. 2010, 2017). Also unclear are the relative propensities for certain lineages and phylogroups to be dispersed through the atmosphere or survive in environmental conditions *ex planta*. As with plant epiphytes, many of the strains that can be isolated from environmental sources are nearly indistinguishable from known phytopathogen lineages, which adds to questions concerning the mixing of strains from agricultural and environmental sources (Bartoli et al. 2015; Monteil et al. 2016).

There are many questions and a few guesses as to characteristics that enable strains to be carried up into the atmosphere, and actually testing for the genetic basis of these traits is challenging for many reasons (Morris et al. 2013). One characteristic that has received much attention because of striking visual and intuitive phenotypic effects, and which impacts both environmental and phytopathogenic aspects of the *P. syringae* life cycle, is the ability of some strains to produce ice-nucleation proteins (Pietsch et al. 2017; Christner et al. 2008b; Morris et al. 2008). While it's true that many other proteins can act to promote ice nucleation, proteins encoded by the *inaZ* gene in *P. syringae* do so at higher temperatures than other potential nucleating molecules (Pandey et al. 2016). A very good demonstration of this effect can be found in a video created by Dr. Mark Martin and found here: https://youtu.be/SenJud3cHLc. So efficient is nucleation from InaZ that this protein has been co-opted by the skiing industry to produce artificial snow and which is marketed as Snowmax®. Ice nucleation is thought to promote plant pathogenicity by breaking open cells to allow access to nutrients and entry into leaves after brief freezes, and thus *P. syringae* can lead to crop damage due to frost at higher than expected temperatures (Lindow et al. 1982). Conversely, Ice− strains that are high-quality plant colonizers have been intentionally released to the environment and do appear to lower disease symptoms through competitive exclusion of Ice+ strains (Lindow 1992). However, ice nucleation can also promote precipitation of bacteria from the atmosphere and thus might also act as a dispersal agent (Morris et al. 2008, 2014; Christner et al. 2008a). It is unclear whether the main selective force on the Ice+

phenotype is through interactions with plants or environmental dispersal (or both), but this intriguing property of some strains certainly changes how dispersal shapes population and community structures (Morris et al. 2008).

4 Are There Differences Between Phytopathogenic and Environmental Strains of *Pseudomonas syringae?*

Many previous discussions about population dynamics within *P. syringae* have been heavily skewed by sampling biases. For somewhat obvious reasons, many well-characterized strains were originally isolated from disease outbreaks on crop species. Other related lineages that might have been present within the same field or present on different host plants were simply discarded or left unsampled. As such, the foundation for understanding genotypic and phenotypic diversity throughout *P. syringae* was (and largely still is) based on known pathogenic strains despite recent efforts to categorize all possible lineages. In light of broader sampling initiatives, numerous questions remain about how sampling location might alter inferences about evolutionary and ecological patterns (Morris et al. 2010; Demba Diallo et al. 2012; Monteil et al. 2013, 2016). However, we have now sequenced upward of a thousand different and diverse strains, and genome characteristics are roughly similar in a set of key parameters. There is usually one main chromosome, approximately 5.5–6.5 Mb, with a GC content that hovers around 60% and with a core of around ~2,500 proteins shared across strains of the species (Baltrus et al. 2017; O'Brien et al. 2011; Nowell et al. 2014; Dillon et al. 2017). This size is relatively large when considering common bacterial workhorses such as *Escherichia coli* and human pathogens like *Streptococcus pneumoniae* but is fairly standard (if even a little small) when placed up against other terrestrial proteobacteria that are culturable, such as *Burkholderia* and *Ensifer* (Land et al. 2015).

While there have been numerous ways to classify strains within *P. syringae* throughout the years, lately the most prevalent system has invoked "phylogroups" as a standard reference for defining clades or groups of related strains (Baltrus 2016; Berge et al. 2014). Phylogroups are somewhat subjective groups of *P. syringae* that form monophyletic clades within phylogenetic trees of the species. Currently, there are 13 recognized phylogroups, with genetic distance between strains within a group being $<5\%$ and between groups $>5\%$ [(Berge et al. 2014) and Fig. 1]. Given this overall level of divergence, somewhat akin to that between *Escherichia coli* and *Salmonella*, one could make the case that this group is composed of a variety of different species (Vinatzer and Monteil 2014; Vinatzer et al. 2014; Baltrus et al. 2017; Dillon et al. 2017). However, due largely to the momentum of history, all of these lineages are still subjectively considered to fall within the group of *P. syringae* sensu lato. Originally, phylogroups were defined based on phylogenies inferred using either four or seven loci conserved throughout *P. syringae* as parts of multilocus sequence analysis (MLSA) or multilocus sequence analysis typing

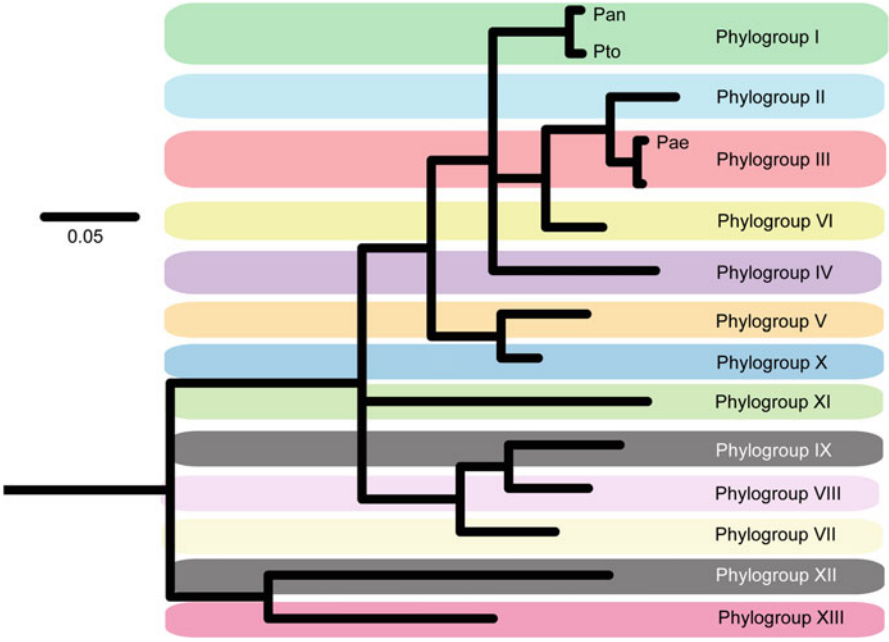

Fig. 1 *Pseudomonas syringae* phylogroups. A Bayesian phylogeny of *P. syringae* strains from all 13 phylogroups was inferred using a fragment of the *cts* gene as per (Berge et al. 2014), with *P. rhizosphaerae* as an outgroup, and using Mr. Bayes 3.2. The analysis was run for 1,000,000 generations with a burn-in period of 250,000 generations. All phylogroups with colored boxes include strains that have been found as pathogens of plants, or which have demonstrated pathogenic potential under laboratory conditions. Phylogroups with gray boxes and white letters indicate that no member of that group has been demonstrated to be a phytopathogen at this time. Three pathogenic lineages discussed in the text (*Pae*, *Pto*, *Pan*) are also highlighted

(MLST) schemes (Hwang et al. 2005; Sarkar and Guttman 2004). More recent whole genome analyses have largely confirmed phylogroup relationships established on the basis of MLSA schemes, although such convergence of evolutionary inferences across methods has not always been the case for more closely related strains (Baltrus 2016; Baltrus et al. 2014, 2017; Monteil et al. 2016).

A thorough investigation analyzing sequences from 198 different strains demonstrated that gene content doesn't differ in a systematic way between strains isolated as agricultural pathogens or from environmental sources. There are a handful of loci (notably the type III effectors *hopQ* and *hopD*) that are significantly correlated with status as a pathogen, but it seems definitive overall that there is no great distinction to be made based on whether *P. syringae* strains are isolated from infected plants or environmental sources (Monteil et al. 2016). Furthermore, although strains of *P. syringae* aren't well known to be competent for natural transformation (but see this report involving episomal elements in Lovell et al. 2009), genome analysis indicates that recombination within phylogroups can occur at relatively high rates (ρ/θ between 1 and 6 depending on the strain combinations examined), although

mutation appears to be the dominant way that genetic variation is introduced between phylogroups (Monteil et al. 2016; Sarkar and Guttman 2004; Yan et al. 2008; Cai et al. 2011a, b; McCann et al. 2017). A more thorough recent analysis analyzing whole genome sequences supports the idea that intra-phylogroup recombination occurs most frequently, but that intra-phylogroup recombination (and specifically involving genes that provide ecological coherence) does play an important evolutionary role across strains (Dillon et al. 2017). Moreover, this whole genome analysis further supports the idea that there are two main evolutionary clusters within the species currently known as *P. syringae* and that recombination between strains does provide structure to these groups (Dillon et al. 2017). That signals of recombination have been seen in housekeeping genes argues against a simple and clean scenario whereby plasmids and phage facilitate gene exchange, and although homologous recombination of chromosomal regions into plasmids/phage could potentially explain gene movement, a clear mechanism behind such natural recombination events remains undefined (but see Swingle et al. 2010; Bao et al. 2012). As a further demonstration of overlap between environmental strains and those known to be pathogens, there is apparently no barrier for recombination between strains found in the environment and on crops, suggesting that evolutionarily relevant interactions occur between strains found in both environments (Monteil et al. 2016).

Upon a wider analysis over multiple studies, phylogroup membership is more indicative of phenotypic characteristics of strains than sampling location per se, and some phylogroups and subclades appear to have distinctly different niches (Berge et al. 2014). As of recent publications, a small minority of phylogroups have been shown to contain only strains isolated from the environment or from asymptomatic plants (Berge et al. 2014). Most phylogroups contain at least one strain that's canonically considered a pathogen. Many groups contain crop pathogens together with a variety of isolates from the environment where there is little to no published information concerning their potential pathogenicity (Berge et al. 2014; Demba Diallo et al. 2012; Monteil et al. 2016) or where strains are often isolated from asymptomatic plants (Kniskern et al. 2011; Clarke et al. 2010; Demba Diallo et al. 2012). Phylogroup XIII strains also appear to be frequently found in the environment and have only recently shown potential as disease-causing agents (Berge et al. 2014; Busquets et al. 2017). Additionally, there is a subclade within phylogroup III (containing the model strain *Pph*1448a) that has never been sampled during environmental sampling and which appears to contain lower metabolic capabilities than other clades (Morris et al. 2010; Monteil et al. 2016; Rico and Preston 2008; Berge et al. 2014). This could reflect a shift in ecological niche to more specifically infect plants that is correlated with lower survival rates under environmental conditions.

One large-scale trend that's quite apparent is that phylogroup II strains are known as etiological agents of disease across a relatively broad range of hosts compared to other phylogroups and are also sampled more frequently from the environment (Baltrus et al. 2017; Hwang et al. 2005; Demba Diallo et al. 2012; Morris et al. 2010; Berge et al. 2014). Moreover, strains from this phylogroup contain a lower number of type III effectors on average than other phylogroups, a trend that

correlates with presence of a suite of toxins like mangotoxin, syringopeptin, syringomycin, and syringolin (Baltrus et al. 2011, 2017). This trade-off between number of effectors and toxin presence is supported by at least two distinct and independent pieces of evolutionary data. One group of strains in phylogroup II (pea pathogens of pathovar *pisi*) has lost the toxins and regained many effectors compared to its likely ancestor (Baltrus et al. 2011). Likewise, there is a group of phylogroup X strains that also appear to contain a lower number of effectors and have independently gained syringomycin (Hockett et al. 2014). The relationship between phylogroups II and X is even more intriguing because gene content for genomes from these strains appears to have been independently shaped by convergent forces (Dillon et al. 2017). As such, even across known pathogenic strains, there likely exists at least two distinct ways to be a pathogen, and these differences correlate with subtle changes in life history (toxin-producing phylogroup II strains seem more environmentally hardy than other strains). There also exists a subclade within phylogroup II that has swapped out the traditional and canonical tripartite PAI containing structural genes for the T3SS and has replaced this locus with an alternate yet related T3SS at a different region of the chromosome (Demba Diallo et al. 2012; Clarke et al. 2010). While the functions of this atypical system under natural conditions remain unknown, it is regulated differently than the canonical system even though it is still competent to deliver effectors (Clarke et al. 2010). To this point, it is also worth noting again that phylogroup II strains often have broad host ranges and often contain highly active ice-nucleation capabilities, which could help to explain their environmental prevalence and worldwide persistence (Pietsch et al. 2017; Morris et al. 2010; Berge et al. 2014; Demba Diallo et al. 2012).

5 What Is a *Pseudomonas syringae* Population?

P. syringae lineages are frequently found across environments, with seemingly little differentiation between strains isolated as pathogens and from nonpathogenic conditions. Moreover, there currently exists little consensus as to relevant scales for sampling that could provide definitive views into what constitutes a niche for each lineage. Even so, as an entrance into a discussion, I'd like to propose simple thought experiment to demonstrate the challenges of understanding *P. syringae* populations. Imagine you're taking that same stroll outside as in the introductory paragraph but this time remembered to bring plates to sample bacteria. After a bit of work and a bit of sequencing, you come to realize that you have sampled nearly identical strains repeatedly from the diseased leaves of one plant species, from healthy-looking leaves of a different plant, from forest leaf litter, as well as from a nearby stream. Imagine also that you have a friend on the other side of the world that samples this nearly identical strain genotype from the bark of a tree in their backyard. As often happens, sampling from each of these environments will also yield a wide collection of *P. syringae* strains that are genotypically different than those mentioned immediately above. Are all of these closely related isolates members of the same

population, from multiple connected populations, from independently evolving separate populations that form a metapopulation? Although divergence in the niches of isolation may suggest that these are different lineages, each responding to selective pressures inherent in their given environments, it's possible that these strains inhabit environments at separate points in time but form one large metapopulation connected by dispersal. If these do represent different populations, since there is almost no variation throughout their whole genomes, they must be evolving quite slowly, dispersing quite rapidly, or experiencing a combination of both. Monteil et al. (2016) have provided perhaps the closest genomic peek at this scenario to date and suggest that a combination of both rapid dispersal and relatively slow evolutionary rates can likely explain observed levels of genomic diversity, but even then it is difficult to extrapolate given biases in sampling. In the absence of more intensive whole genome sampling schemes across hosts, environments, as well as spatial and temporal scales, it's nearly impossible to come to a clear decision.

As the thought experiment above hopefully suggested, despite extensive knowledge concerning mechanisms of pathogenesis and a growing body of literature focused on natural ecology of *P. syringae*, the critical evolutionary question of "How do we describe the relevant scale for *P. syringae* populations?" remains unanswered. Part of this challenge stems from the reality that the name "*P. syringae*" characterizes a hugely diverse group of bacterial isolates, as a result of a somewhat tortured taxonomic history (Vinatzer and Monteil 2014; Baltrus 2016) with a range of genomic average nucleotide identity values that nearly spans the difference from *Salmonella* to *Escherichia* (Vinatzer et al. 2017). We know that multiple lineages of *P. syringae* can be found in the same ecosystems, in the same plants, and even in the same leaves, but it is often difficult to measure whether these lineages interact in an evolutionary relevant way (Morris et al. 2008, 2010, 2017; Humphrey et al. 2014). It is possible that these lineages fill the same niche and therefore directly compete, but it's also possible that they are merely spatially similar but ecologically disparate. Given that an Ice− strain can successfully outcompete pathogens, and therefore lessen disease, we know that competitive exclusion between strains does at least exist under natural conditions (Lindow 1992). The true answer may lie in understanding that spectra of interaction exist, and for each type of genetic variant, the relevant population parameters might differ. What are the dominant evolutionary forces that will shape newly arising genetic variation within *P. syringae* lineages? How much variation arises over the course of an outbreak within a single field, and how much of this variation makes it through transmission bottlenecks? Piece by piece, we can step through what we understand and compare this to data arising from studies of outbreaks (like Cai et al. 2011b; McCann et al. 2017) and work out from there to at least narrow the list of possible answers.

Newly introduced genetic variation is the currency for understanding population structure, and we can learn a significant amount by simply following the evolutionary fates of new variants (Cordero and Polz 2014; Shapiro and Polz 2014; Shapiro et al. 2009; Choudoir et al. 2012). Genetic variation is introduced to strains through mutation or horizontal gene transfer, and the fate of this new variation is then subject to population-level forces such as genetic drift and natural selection. The strength of

genetic drift is inherently correlated with effective population size (N_e), and our understanding of drift is therefore intrinsically linked to defining populations (Nei and Tajima 1981). It's even possible and highly relevant for this discussion, albeit with some assumptions, to back calculate from the rate of mutation fixation in bacteria to effective population size (Sung et al. 2016). Although the strength of selection is traditionally understood to rely on environmental context and competition between clones and species, competition can also occur between variant gene sites within a clone (Gerrish and Lenski 1998). Therefore, our understanding of how selection shapes *P. syringae* will also depend on how many circulating genetic variants are competing within a population. As highlighted in other parts of this book, we know much about relevant population-level parameters for some well-studied bacterial systems across environments. Laboratory studies of clonal microbial evolution have shown that, under relatively simple and controlled conditions, adaptation within microbial populations can be approximated using a small number of parameters including N_e and mutation rate (u) (Wiser et al. 2013; Sniegowski and Gerrish 2010). N_e in fluctuating populations, as one would likely find in *P. syringae*, can be calculated by taking the harmonic mean of population sizes throughout the course of population existence (Sjödin et al. 2005). As such, calculation of harmonic mean is dominated by the smallest numbers sampled. Even though you may have 10^{10} cells within a flask as a census size, the level of variation seen by evolution could be driven by the bottleneck that this population went through at the beginning of growth. Furthermore, while it's quite easy to measure u, it's also quite likely that u will physiologically change during population growth and may even genetically change over the course of bacterial adaptation (Kivisaar 2010; Lynch 2010; Denamur and Matic 2006). So, even under the simplifying routines of daily transfer and growth in a minimally complex environment, it can be difficult to get a handle on microbial population dynamics.

Extrapolating from what we've learned under very simple conditions to facultative phytopathogens, the challenge of modern-day microbial population genetics becomes apparent. While census population sizes may be huge even on single leaves (10^{10}), N_e for these very same leaves will likely be driven by bottlenecks that occur during colonization and transmission and could be much lower than census size. If microbes are obligate pathogens and grow only in one host, there will still likely be changes in the origin of subpopulations due to differential niches and spatial structuring. Mutation rates will likely change with physiology of infection and add to that that host defenses themselves can be mutagenic to resident bacteria. Population sizes for *P. syringae* will not be constant across hosts or environmental reservoirs and will differ with type of plant and between genetic variants within a host. Furthermore, different environmental contexts (leaf litter, rainwater, clouds, etc.) will have different carrying capacities for *P. syringae.* Differential population sizes between environments as well as the amount of growth possible within environments will skew evolutionary forces on *P. syringae* populations that experience multiple environments. N_e may be quite large on some hosts for a fraction of time, but since it is calculated from harmonic means, it may be bottlenecks between hosts and long periods of time in low-carrying capacity environments that drive population-level evolutionary phenomena.

P. syringae genomes display relatively high rates of plasticity and are quite diverse in terms of gene content and functional capabilities (Nowell et al. 2014). For many facultative pathogens, and certainly for *P. syringae*, only a relatively small portion of their genome is devoted to virulence in hosts, and this abundance of "other" genes likely indicates that much of their life cycle is underappreciated. It wouldn't be shocking to see disparate, yet closely related, lineages of *P. syringae* occupy different niches on the same leaf in proximity to one another. Perhaps one lineage specializes as a pathogen attacking the plant, while the other is a secondary colonizer and feasts on breakdown products of the primary infection. It's also likely that lineages only compete for short amounts of time in certain plant hosts but then spend the majority of their life cycle specialized toward completely independent environments with selective forces shaped by different parameters. Largely disparate lineages may be part of the same proximate population or community for brief periods but not over the majority of time. Alternatively, there could be enough spatial structure, even within a single leaf, for multiple competing lineages to actually coexist in small microcolonies or huddled together to take advantage of nutrient oases (Lindow and Brandl 2003). Interactions within and between lineages of *P. syringae* that are in contact, and the relevant differences in growth and survival of these lineages, ultimately will determine which genetic variants reach high frequencies, and thus an examination of these parameters is quite worthwhile for understanding evolutionary dynamics.

6 What Can We Learn from *Pseudomonas syringae* Epidemics?

Assuming that *P. syringae* epidemic strains behave similarly to other strains under natural conditions, it's an instructive exercise to compare the overall diversity of *P. syringae* mentioned above to genetic variation that is present across multiple genome sequences of strains sampled from three independent epidemics of *P. syringae*. Indeed, temporal sampling over such outbreaks provides an unprecedented viewpoint into population-level differentiation. Furthermore, since these outbreaks occurred on three different types of host plant (horse chestnut, tomato, and kiwi) that represent a spectrum of domestication times and agricultural practices, commonalities in pathogen evolutionary dynamics could provide a more general view of the species as a whole.

The first genomic analyses of *P. syringae* from an epidemic focused on *P. syringae* pv. *aesculi* (*Pae*) causing bleeding canker disease on horse chestnut across Northwest Europe (Green et al. 2010; de Keijzer et al. 2012). Diseased trees were first noticed around 2002–2003, with reports from Belgium, Germany, the Netherlands, and the UK. Horse chestnut is a long-lived, woody host and has not been domesticated. Although there is little agricultural value in horse chestnut, the prominence of this tree as a landscape staple coupled with the unique opportunity to

study bacterial epidemics with the potential to wipe out long-lived hosts motivated relatively extensive research resources to study this pathogen. Two epidemic strains were isolated in 2006, from Scotland (Glasgow) and England (Surrey), with a third analyzed strain isolated from Scotland (Pitlochry) in 2008 (Green et al. 2010). At the time of the analysis, the closest sequenced strain to the outbreak lineage was the type strain of *Pae* isolated from a diseased tree in India in 1969, ~40 years before the European strains, but which was genotypically identical over seven loci used for MLSA studies. This comparison is interesting because *Pae* strains had only ever previously been isolated as a foliar disease and had only ever been isolated in India. Therefore, to the extent that conclusions can be drawn from this set of samples, the Northwest European epidemic lines appear to have subtly switched hosts, from Indian to European horse chestnut, and had likely developed the ability to cause bleeding cankers in woody tissue over the span of a few decades. Comparison of four *Pae* genomes showed that the European strains differed from the Indian type strain by 1,613 single-nucleotide polymorphisms (SNPs) over 3 Mb (out of approximately a 5.5 Mb genome). The European strains were not identical, but there were only a total of three SNPs across this same 3 Mb across these three strains. Roughly 5% of the genome differed in presence/absence between the Indian and European strains, with the European strains possessing 245 additional proteins compared to their Indian cousin. Much of the genetic variation found between the European strains appears to already be present between strains isolated from different parts of the UK in 2006 rather than having accumulated between the two Scottish strains from 2006 to 2008. Notably, gene gains in the European strains likely occurred through plasmid acquisition, but none of the European isolates maintained the same plasmid suite. From these data points, the overall message seems to be that relatively little nucleotide variation arose in the years since the start of the epidemic in Northern Europe but that acquisition and loss of plasmids contribute largely to the rapid creation of intra-epidemic variation.

The second examined epidemic (in effect at least) provides an orthogonal viewpoint to the pv. *aesculi* strains mentioned above because temporally isolated strains causing speck of tomato were analyzed for the accumulation of genetic variation (Cai et al. 2011b). In this case, there are three clonal lineages of *P. syringae* pv. *tomato* (*Pto*) responsible for disease worldwide: T1, JL1065, and DC3000. MLSA analysis of *Pto* strains isolated between 1942 and 2009 from a worldwide collection demonstrated that T1 is currently the dominant disease-causing lineage, but that the dominance of this clone only occurred after replacement of both JL1065 and DC3000 somewhere around 1960. However, both JL1065 and DC3000 clones are extant and are occasionally isolated as causative agents of speck disease outbreaks. Analysis of five T1 clones isolated in Europe and North America from 1961 to 2005 showed any two pairs of these strains only differed by between 53 and 183 SNPs over ~3.5 Mb of their core genome. Even given this reasonably dense sample of temporal isolates, divergence times between these strains were estimated to be between 1,000 and 1,000,000 years, demonstrating the challenge of estimating clone divergence with only estimates of crucial parameters like mutation rate. Denser sampling of known SNPs over all the catalogued isolates showed there

was intra-clone evolution as well, with specific genotypes replacing each other as the main disease-causing lineage within the T1 lineage. Furthermore, there was a bit of a geographic signal in *Pto* clone identity and evolutionary dynamics, with the JL1065 persisting in developing countries and the appearance of biased dispersal of genotypes between Europe and North America and between South America and Africa. There were also hints of independent evolutionary events for *Pto* occurring in different geographic regions, suggesting that (at least) evolutionary relevant subpopulations do exist on the worldwide scale for this pathogen.

Recent outbreaks of disease on kiwi have driven *P. syringae* pv. *actinidiae* (*Psa*), the causative agent of bleeding canker of kiwi, to be the best characterized of any lineage within *P. syringae* (McCann et al. 2013, 2017; Fujikawa and Sawada 2016; Wilstermann et al. 2017; Everett et al. 2011). It's an interesting pathosystem to investigate because kiwi is native to Asia and has only been domesticated within the last 100 years. These outbreaks therefore represent a particularly interesting window to view adaptation of phytopathogens in new crop species, because (unlike crops such as tomato that have been shaped by human selection for thousands of years) pathogens haven't had much time to adapt to the specific agricultural lineages (McCann et al. 2013). In the early 1980s, farmers in Japan and China began to notice a *P. syringae* disease causing cankers, leaf spots, and flower damage infecting their orchards. Observations of this pathogen increased through the next couple of decades, but 2007/2008 was the beginning of a pandemic where *Psa* severely threatened the kiwi supplies (McCann et al. 2017). This outbreak has provided researchers the chance to follow phytopathogenic adaptation to a new crop in real time through dense sampling of diseased and asymptomatic hosts.

Psa can be split into four phylogenetically coherent clades (McCann et al. 2017). One clade is represented by strains that caused outbreaks in Japan and Italy in the 1980s, another clade caused a limited outbreak in Korea in 2007–2008, and a third clade is represented by a small pocket of strains in Japan. The most recent pandemic and destructive clone, *Psa*-3, appears to have originated in China and spread worldwide from there, even though the center of diversity and likely origin of *Psa* itself likely lie in Korea or Japan. Each of these four main *Psa* clades is separated from each other by between one and 4,000 SNPs (McCann et al. 2017). However, there also appears to be a relatively high level of gene exchange between these four *Psa* clones, with SNPs being introduced by recombination between clones at a slightly higher rate than through mutation. While recombination occurs within *Psa* clonal complexes themselves, within clone evolution is driven much more by mutation (7×) than recombination. This pattern suggests that there is a diverse reservoir where all extant clones can come in contact, that occasionally a particularly pathogenic type emerges to cause a pandemic, and that the three previous outbreaks have been caused by phylogenetically distinct lineages of *Psa*. Strains related to those implicated in the New Zealand outbreak, clone *Psa*-3, have been particularly well sampled over temporal and geographic diversity and provide a rich epidemic snapshot for how diversity emerges throughout *P. syringae*. China harbors a few different subclades of strains from within *Psa*-3, with the most divergent strains possessing hundreds of SNPs on average compared to each other. Epidemiological

genomics suggests two independent escapes from this source population to Italy and New Zealand. Transmission events to Italy and New Zealand are clearly different because all New Zealand *Psa*-3 strains share four informative SNPs compared to other lineages. In addition to these four shared SNPs, New Zealand strains themselves differ from each other on average by about four SNPs each. Subsequently, this *Psa*-3 clone was introduced to Japan either from New Zealand or from the same source population in China, as the Japanese isolates of *Psa*-3 share the four conserved SNPs found in New Zealand isolates. Genome sequences of the New Zealand and Japanese strains, along with extrapolation as to divergence times, suggest a relatively slow rate of accumulation for nucleotide substitutions throughout *Psa*-3 during the epidemic (McCann et al. 2017). In contrast, gene gain and loss is much more prevalent than the generation of SNPs across all lineages of *Psa*, with horizontal gene transfer of ICE elements between strains representing a particularly significant event for pathovar *actinidiae* (McCann et al. 2013). Overall, the story from *Psa* is one of a somewhat diverse and interacting group of related clones at the source population where every once in a while a particularly good pathogen emerges at the right time and right place and is dispersed throughout the world. After dispersal, it appears as though SNPs do accumulate but do so at a lower rate than other well-sampled animal pathogens.

There is a question as to how well we can extrapolate from these outbreaks to *P. syringae* as a whole. It's possible that the same geographic and evolutionary patterns reflected in the outbreak strains hold true across all of *P. syringae*. It's also possible that these outbreaks represent rare outliers in the overall spectrum of diversity and population dynamics. Outbreak strains may reflect lineages that capture the right genotypic information at the right time, persist until a better strain comes along, and may be sampled in a biased way due to human interest. For agricultural pathogens, there may be distinctly separate evolutionary dynamics for outbreak strains compared to closely related environmental isolates because outbreak strains potentially undergo different selection regimes and experience a different landscape for host diversity (Stukenbrock and McDonald 2008). A relatively high percentage of possible hosts in small geographic areas, as one would see in a monoculture field, could enable a positive feedback loop of increases in effective population size, rapid growth, introduction of new diversity through mutation, and transmission to new monoculture fields. This feedback loop may lead to increasing specialization on agriculturally relevant crops for the outbreak strains compared to those found in the environment. Strains present under more "natural" conditions may not experience such widespread availability of hosts in proximity, or may not grow to as high a level inside of these hosts. Even with these differences in mind, multiple papers have demonstrated that environmental isolates can infect plants at high rates and that there is substantial genomic overlap between known phytopathogen lineages and those from other sources (Cai et al. 2011a; Monteil et al. 2013, 2016). However, clearer answers to these questions will likely arise with increasing sequence data from environmental strains, but it is at least worth considering that context for which we have the most information about *P. syringae* population dynamics (or for any other relevant agricultural pathogen) may not be representative

of the majority of strains. As usually happens, the answer probably lies somewhere in the middle of the spectrum, where outbreak strains are unique in some of their population dynamics because of particular host distributions but reflect overall qualities in terms of dispersal, gene gain and loss, and mutation that are in line for the expectation for *P. syringae* as a whole.

7 Why Does the Delineation of Bacterial Populations Matter?

Defining population structures and evolutionary parameters is difficult, even for well-studied bacteria like *E. coli* (Shapiro et al. 2009). However, these estimates are important for modeling real-time scenarios for evolution (for instance, in predicting the spread of antibiotic resistance) as well as teasing apart epidemiological histories of outbreaks (McCann et al. 2013, 2017; Cai et al. 2011b). Basic questions about how bacterial populations are structured are also critical for understanding fundamental rules and constraints governing evolution of environmental bacteria under natural conditions (Choudoir et al. 2012). While some parameters can be estimated from widespread genomic data (i.e., N_e in Sung et al. 2016), at least enough to provide ballpark estimates, the best case scenarios have temporally and geographically stratified samples that can be used to calibrate rates of evolution. Even with such data, extrapolating backward still relies on an implicit definition for understanding how frequently two genetic variants are going to be competing against one another for fixation.

Questions about species, lineages, and populations take on a heightened importance when considering agricultural pathogens. Typically, transport of and research on known phytopathogens are strictly regulated at the level of localities as well as countries. What strains are called and how evolutionary lineages are defined matters, because it is this very nomenclature that controls how strains are regulated (Baltrus 2016). Exacerbating these underlying problems of confusion across strain names, in many cases, the resources dedicated to classification of agricultural pathogens are less than those for human and animal pathogens, and thus changes to regulatory definitions occur more slowly than the research to define strains. It's ironic to an extent, because transport of pathogenic strains may be highly restricted to regions where very similar (or identical) strains are already endemic. I, and I'm assuming numerous others, have dealt with the frustrating scenario where a particular bacterial strain has been requested but where the recipient doesn't have the correct permits for strain import. However, closely related strains would likely be available naturally on plants growing in the recipients region if not right outside their door. Our lack of understanding of how to interpret evolutionary patterns within phytopathogens hampers our abilities to accurately and efficiently develop bureaucracies to minimize risks of these pathogens. Aside from providing data to inform basic evolutionary questions, a better understanding of how populations of *P. syringae* evolve under

natural conditions could significantly help to sort out the Rube Goldbergian morass of nomenclature for these strains that has arisen over time (Baltrus 2016).

8 Extrapolating from *Pseudomonas syringae*

I have focused on *P. syringae* for this chapter, partly because it's a species I've come to know quite well but also because it's a great example of a bacterium where extensive knowledge about one of its potential habitats completely overshadows our understanding of other important factors in its existence. *P. syringae* is not an outlier in this regard though, and the same kinds of evolutionary and population-level questions can apply to nearly all other bacteria found in the phyllosphere and rhizosphere. Strains of *Erwinia* and *Pantoea* have been sampled from a variety of host-associated contexts in both the presence and absence of disease and without much geographic provenance but are also abundantly sampled as parts of the water cycle and are ice nucleators (Kado 2006; Starr and Chatterjee 1972; Walterson and Stavrinides 2015; Christner et al. 2008b). Strains of *Ensifer*, *Rhizobium*, and *Burkholderia* are well known as symbionts of nodulating plants but are often found living freely within the soil or in phytobiomes. Although there is great potential for local adaptation at microscales within these species, overall patterns are somewhat equivocal when it comes to evidence for geographic or local adaptation (Van Cauwenberghe et al. 2014; Lemaire et al. 2016; Harrison et al. 2017; Stopnisek et al. 2014). Perhaps the most interesting of this bunch are members of the genus *Streptomyces* which are widely found throughout soils across the world but can also be both pathogens and endophytes of plants. In contrast to other bacteria mentioned above, isolates of *Streptomyces* appear to follow some geographic pattern, which raises new questions both about the extent of isolation between strains and about patterns of local adaptation (Choudoir et al. 2016; Andam et al. 2016). They also have the ability to form spores, which, although not as durable as endospores, may shift both their abilities for dispersal and survival (Lennon and Jones 2011).

What these species all have in common is that they are found associated with plants widely throughout the planet and are fairly easy to culture, but to date there are relatively few studies that have been carried out with enough power to truly understand the effects of evolution at the level of populations. Even across these species the relevant parameters of dispersal or growth may differ, but we stand at a crossroads where we know much about their evolutionary potential but relatively little about their evolutionary reality in nature. After the initial wave of genome sequencing of interesting isolates, we can begin to fill in these gaps by focusing on identifying geographic patterns at whole genome levels and with sequentially sampling of the same sites throughout time. With those data in hand, we can begin to address questions about relative dispersal capabilities between strains and genera and to investigate more specific hypotheses about topics such as how spore forming shapes population dynamics under natural conditions. Ultimately, it may be possible

to begin to match the precision and efficiency with which marine microbiologists have described ecological niches and selective pressures (Follows et al. 2007). No doubt that each system will have its own strengths and unique weaknesses, but overall the picture will begin to come together where our understanding of how genome level diversity changes throughout time and space can guide delineation of populations of microbes evolving independently from each other.

9 Conclusions

The ubiquity of *P. syringae* across environmental contexts, from various types of freshwater habitats to plants to (sometimes) even insects, creates interesting questions about population dynamics but also challenges in defining basic evolutionary parameters. The likelihood is that, across the broadly defined species *P. syringae* sensu lato, there will be clades and subclades of related strains that behave the same but also significant diversity throughout the entire group in terms of niches and population dynamics. I would suggest not getting too hung up on the taxonomic challenges correlated with these questions, but instead basing analyses on relative phylogenetic relationships between strains. Phylogroups have shown to do a pretty good job of approximating relevant evolutionary groups of *P. syringae*, in so much as recombination seems to occur more frequently within phylogroups than between them. Even so, strictly basing evolutionary analyses at the level of phylogroups will undoubtedly miss important nuances because membership in phylogroups (at this point at least) is somewhat subjectively determined (Berge et al. 2014). It is worth noting that there exists a burgeoning movement to base relative classification schemes on whole genome sequences for *P. syringae* and extending across bacterial taxa (Vinatzer et al. 2017). Once bacterial taxonomists have fully embraced such approaches, emergent evolutionary trends and patterns may arise over different scales that are currently obscured by nomenclatural confusion, clutter, and rot.

Stepping back to the question of how *P. syringae* populations evolve under natural conditions, it appears as though there are a few clear takeaway messages from larger-scale genomic studies that can inform answers going forward. Every *P. syringae* clone is not everywhere as there does exist a signal of geographic provenance across multiple clades. That studies can epidemiologically trace dispersal and transmission routes for clones themselves speaks loudly to geographic differences. Whatever the relevant population metric is within and across phylogroups of *P. syringae*, there is also likely going to be a center of diversity somewhere in the world from which new clones emerge and are dispersed. These centers of diversity and the modes of dispersal need not be the same for every phylogroup, so that interesting patterns will likely emerge as an increasing number of strains are sequenced from across the globe. Recombination is likely going to occur more frequently within phylogroups than between, and therefore one clear prediction is that centers of diversity for a relevant phylogenetic clade may be hotspots for interstrain recombination by currently unknown mechanisms. Within

a "population" of *P. syringae*, regardless of how that group is defined, it's also quite clear that gene acquisition and loss are going to be the dominant evolutionary forces generating diversity over short periods of time, especially when genes of interest are found on prophage and plasmids. Single-nucleotide variants certainly matter when it comes to evolutionary dynamics, especially when these variants impact host immune recognition, but the frequency of such SNPs will likely be dwarfed by gene gain and loss.

I don't have a clear answer to the thought experiment proposed earlier, but given all of the data, I'd like to make some informed guesses that could certainly change in the future. As of this moment, I would feel confident saying that two nearly genotypically identical strains of *P. syringae* found in the same proximate area regardless of sampling source have a pretty good chance of being members of the same population. The likelihood of population membership will fall with geographic distance of sampling, although we don't know much about the rate at this moment in time, therefore I can't easily rule out that two nearly genotypically identical strains found on opposite sides of the world are from the same population. If two divergent *P. syringae* strains are isolated from the same plant, I would say at this moment that they are likely from within the same population in an evolutionarily relevant sense if they are members of the same phylogroup, but if they are members of different phylogroups, they are probably from two different populations that are either coexisting peacefully or in direct competition.

10 Future Perspectives

There will no doubt be many future efforts to isolate, identify, and sequence *P. syringae* strains. Even during the course of writing this chapter, multiple groups have shared an incredible amount of data that can inform questions I've highlighted using *P. syringae* (Dillon et al. 2017; Straub et al. 2017; Karasov et al. 2018). This future accumulation of data will provide increasingly refined estimates of important evolutionary parameters such as estimates of mutation rate, measurements of nucleotide diversity, size of core genomes shared across strains, and the amounts of recombination and/or horizontal gene transfer that occurs between groups and species. The publication of such information will suggest wonderful new avenues for evolutionary research and could transform our understanding of the forces that shape natural populations of environmentally ubiquitous microbes.

However, I also posit that despite this inevitable flood of nucleotide sequence, we will eventually run into the similar challenges as we do now (albeit with more information at our disposal) in terms of identifying evolutionarily relevant lineages unless serious efforts are made to incorporate geographic and temporal sampling on a wider scale. It's my belief that the true transformative moment will come when the power of genomic analyses are combined with time series across different hosts and geographies. Then, we will firmly know just how much diversity can arise within a field in between transmission bottlenecks during a disease outbreak, we will be able

to differentiate between specialist and generalist populations across hosts, and we will know just how capable recombination, selection, and drift can be in shaping species-wide diversity patterns. To these points, Karasov et al. (2018) provide a unique viewpoint into community-level diversity of pseudomonads (using 16S rRNA sequences) across levels of single plants scattered over geographic scales and can be used as a framework for scaffolding studies to evaluate whole genome diversity across hosts.

While we are currently awash in genomic sequences from this species, the coming years promise much more in terms of sampling and sequencing. Genome sequences and relevant information about characteristics such as mutation spectrum will accumulate, as will isolates from diverse habitats and geographic regions. There are currently no clear answers to many of the questions raised in this chapter, but I have it on good authority that multiple groups are working to gather data pertinent to these questions so that we'll likely have a better idea sooner rather than later. Even as this data accumulates, *P. syringae* has proven to be a fascinating organism, and there are no doubt many other mysteries will continue to confound and challenge us for decades.

Acknowledgments I would like to thank numerous individuals that helped improve this chapter by reading earlier versions, especially Brians Smith and Kvitko. I would especially like to thank Boris Vinatzer for his thoughtful and careful critique. D.A.B. is supported by the National Science Foundation (NSF) IOS-1354219 and US Department of Agriculture (USDA) 2016-67014-24805.

References

Andam CP, et al. A latitudinal diversity gradient in terrestrial bacteria of the genus Streptomyces. mBio. 2016;7(2):e02200–15.

Bao Z, Cartinhour S, Swingle B. Substrate and target sequence length influence RecTE(Psy) recombineering efficiency in *Pseudomonas syringae*. PLoS One. 2012;7(11):e50617.

Bartoli C, et al. A framework to gauge the epidemic potential of plant pathogens in environmental reservoirs: the example of kiwifruit canker. Mol Plant Pathol. 2015;16(2):137–49.

Bender CL, Alarcón-Chaidez F, Gross DC. *Pseudomonas syringae* phytotoxins: mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. Microbiol Mol Biol Rev. 1999;63(2):266–92.

Berge O, et al. A user's guide to a data base of the diversity of *Pseudomonas syringae* and its application to classifying strains in this phylogenetic complex. PLoS One. 2014;9(9):e105547.

Baltrus DA. Divorcing strain classification from species names. Trends Microbiol. 2016;24 (6):431–9.

Baltrus DA, et al. Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 Pseudomonas syringae isolates. PLoS Pathog. 2011;7(7):e1002132.

Baltrus DA, et al. Incongruence between multi-locus sequence analysis (MLSA) and whole-genome-based phylogenies: *Pseudomonas syringae* pathovar *pisi* as a cautionary tale. Mol Plant Pathol. 2014;15(5):461–5.

Baltrus DA, McCann HC, Guttman DS. Evolution, genomics and epidemiology of *Pseudomonas syringae*: challenges in bacterial molecular plant pathology. Mol Plant Pathol. 2017;18 (1):152–68.

Busquets A, et al. *Pseudomonas caspiana* sp. nov., a citrus pathogen in the *Pseudomonas syringae* phylogenetic group. Sys Appl Microbiol. 2017;40(5):266–73.
Cai R, et al. Reconstructing host range evolution of bacterial plant pathogens using *Pseudomonas syringae* pv. *Tomato* and its close relatives as a model. Infect Genet Evol. 2011a;11(7):1738–51.
Cai R, et al. The plant pathogen *Pseudomonas syringae* pv. *tomato* is genetically monomorphic and under strong selection to evade tomato immunity. PLoS Pathog. 2011b;7(8):e1002130.
Carrión VJ, et al. The mbo operon is specific and essential for biosynthesis of mangotoxin in *Pseudomonas syringae*. PLoS One. 2012;7(5):e36709.
Choudoir MJ, Campbell AN, Buckley DH. Grappling with Proteus: population level approaches to understanding microbial diversity. Front Microbiol. 2012;3:336.
Choudoir MJ, Doroghazi JR, Buckley DH. Latitude delineates patterns of biogeography in terrestrial Streptomyces. Environ Microbiol. 2016;18(12):4931–45.
Christner BC, Cai R, et al. Geographic, seasonal, and precipitation chemistry influence on the abundance and activity of biological ice nucleators in rain and snow. Proc Nat Acad Sci USA. 2008a;105(48):18854–9.
Christner BC, et al. Ubiquity of biological ice nucleators in snowfall. Science. 2008b;319 (5867):1214.
Collmer A, et al. *Pseudomonas syringae* Hrp type III secretion system and effector proteins. Proc Natl Acad Sci U S A. 2000;97(16):8770–7.
Cordero OX, Polz MF. Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol. 2014;12(4):263–73.
Clarke CR, et al. *Pseudomonas syringae* strains naturally lacking the classical *P. syringae* hrp/hrc locus are common leaf colonizers equipped with an atypical type III secretion system. Mol Plant Microbe Interact. 2010;23(2):198–210.
de Keijzer J, et al. Histological examination of horse chestnut infection by *Pseudomonas syringae* pv. *aesculi* and non-destructive heat treatment to stop disease progression. PLoS One. 2012;7 (7):e39604.
Demba Diallo M, et al. *Pseudomonas syringae* naturally lacking the canonical type III secretion system are ubiquitous in nonagricultural habitats, are phylogenetically diverse and can be pathogenic. ISMEJ. 2012;6(7):1325–35.
Denamur E, Matic I. Evolution of mutation rates in bacteria. Mol Microbiol. 2006;60(4):820–7.
Dillon MM, et al. Recombination of ecologically and evolutionarily significant loci maintains genetic cohesion in the Pseudomonas syringae species complex. bioRxiv. 2017. https://doi.org/10.1101/227413.
Everett KR, et al. First report of *Pseudomonas syringae* pv. *actinidiae* causing kiwifruit bacterial canker in New Zealand. Aust Plant Dis Notes. 2011;6(1):67–71.
Failor KC, et al. Ice nucleation active bacteria in precipitation are genetically diverse and nucleate ice by employing different mechanisms. ISMEJ. 2017. https://doi.org/10.1038/ismej.2017.124.
Follows MJ, et al. Emergent biogeography of microbial communities in a model ocean. Science. 2007;315(5820):1843–6.
Fujikawa T, Sawada H. Genome analysis of the kiwifruit canker pathogen *Pseudomonas syringae* pv. *actinidiae* biovar 5. Sci Rep. 2016;6:21399.
Gerrish PJ, Lenski RE. The fate of competing beneficial mutations in an asexual population. Genetica. 1998;102–103(1–6):127–44.
Green S, et al. Comparative genome analysis provides insights into the evolution and adaptation of *Pseudomonas syringae* pv. *aesculi* on Aesculus hippocastanum. PLoS One. 2010;5(4):e10224.
Harrison TL, et al. No evidence for adaptation to local rhizobial mutualists in the legume Medicago lupulina. Ecol Evol. 2017;7(12):4367–76.
Hendry TA, Hunter MS, Baltrus DA. The facultative symbiont rickettsia protects an invasive whitefly against entomopathogenic *Pseudomonas syringae* strains. Appl Environ Microbiol. 2014;80(23):7161–8.
Hirano SS, Upper CD. Bacteria in the leaf ecosystem with emphasis on Pseudomonas syringae – a pathogen, ice nucleus, and epiphyte. Microbiol Mol Biol Rev. 2000;64(3):624–53.

Hockett KL, et al. *Pseudomonas syringae* CC1557: a highly virulent strain with an unusually small type III effector repertoire that includes a novel effector. Mol Plant Microbe Interact. 2014;27(9):923–32.
Humphrey PT, et al. Diversity and abundance of phyllosphere bacteria are linked to insect herbivory. Mol Ecol. 2014;23(6):1497–515.
Hwang MSH, et al. Phylogenetic characterization of virulence and resistance phenotypes of *Pseudomonas syringae*. Appl Environ Microbiol. 2005;71(9):5182–91.
Ichinose Y, Taguchi F, Mukaihara T. Pathogenicity and virulence factors of *Pseudomonas syringae*. J Gen Plant Pathol. 2013;79(5):285–96.
Kado CI. Erwinia and related genera. In: Dworkin M, et al., editors. The prokaryotes. New York, NY: Springer; 2006. p. 443–50.
Karasov TL, et al. *Arabidopsis thaliana* populations support long-term maintenance and parallel expansions of related *Pseudomonas* pathogens. bioRxiv. 2018. https://doi.org/10.1101/241760.
Kivisaar M. Mechanisms of stationary-phase mutagenesis in bacteria: mutational processes in pseudomonads. FEMS Microbiol Lett. 2010;312(1):1–14.
Kniskern JM, Barrett LG, Bergelson J. Maladaptation in wild populations of the generalist plant pathogen *Pseudomonas syringae*. Evolution. 2011;65(3):818–30.
Land M, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;15(2):141–61.
Lindeberg M, Cunnac S, Collmer A. *Pseudomonas syringae* type III effector repertoires: last words in endless arguments. Trends Microbiol. 2012;20(4):199–208.
Lindow SE. Ice− strains of *Pseudomonas syringae* introduced to control ice nucleation active strains on potato. In: Biological control of plant diseases, NATO ASI series. Boston, MA: Springer; 1992. p. 169–74.
Lindow SE, Brandl MT. Microbiology of the phyllosphere. Appl Environ Microbiol. 2003;69(4):1875–83.
Lindow SE, Arny DC, Upper CD. Bacterial ice nucleation: a factor in frost injury to plants. Plant Physiol. 1982;70(4):1084–9.
Lemaire B, et al. Biogeographical patterns of legume-nodulating Burkholderia spp.: from African Fynbos to continental scales. Appl Environ Microbiol. 2016;82(17):5099–115.
Lennon JT, Jones SE. Microbial seed banks: the ecological and evolutionary implications of dormancy. Nat Rev Microbiol. 2011;9(2):119–30.
Lovell HC, et al. Bacterial evolution by genomic island transfer occurs via DNA transformation in plants. Curr Biol. 2009;19(18):1586–90.
Lynch M. Evolution of the mutation rate. Trends Genet. 2010;26(8):345–52.
Mansfield J, et al. Top 10 plant pathogenic bacteria in molecular plant pathology. Mol Plant Pathol. 2012;13(6):614–29.
McCann HC, et al. Genomic analysis of the kiwifruit pathogen *Pseudomonas syringae* pv. *actinidiae* provides insight into the origins of an emergent plant disease. PLoS Pathog. 2013;9(7):e1003503.
McCann HC, et al. Origin and evolution of the kiwifruit canker pandemic. Genome Biol Evol. 2017;9(4):932–44.
Monteil CL, et al. Nonagricultural reservoirs contribute to emergence and evolution of *Pseudomonas syringae* crop pathogens. New Phytol. 2013;199(3):800–11.
Monteil CL, et al. Population-genomic insights into emergence, crop adaptation and dissemination of *Pseudomonas syringae* pathogens. Microb Genom. 2016;2(10):e000089.
Morris CE, et al. The life history of the plant pathogen *Pseudomonas syringae* is linked to the water cycle. ISMEJ. 2008;2(3):321–34.
Morris CE, et al. Inferring the evolutionary history of the plant pathogen *Pseudomonas syringae* from its biogeography in headwaters of rivers in North America, Europe, and New Zealand. mBio. 2010;1(3). https://doi.org/10.1128/mBio.00107-10.
Morris CE, Monteil CL, Berge O. The life history of *Pseudomonas syringae*: linking agriculture to earth system processes. Annu Rev Phytopathol. 2013;51:85–104.

Morris CE, et al. Bioprecipitation: a feedback cycle linking earth history, ecosystem dynamics and land use through biological ice nucleators in the atmosphere. Glob Chang Biol. 2014;20 (2):341–51.
Morris CE, et al. Frontiers for research on the ecology of plant-pathogenic bacteria: fundamentals for sustainability: challenges in bacterial molecular plant pathology. Mol Plant Pathol. 2017;18 (2):308–19.
Nei M, Tajima F. Genetic drift and estimation of effective population size. Genetics. 1981;98 (3):625–40.
Nowell RW, et al. The extent of genome flux and its role in the differentiation of bacterial lineages. Genome Biol Evol. 2014;6(6):1514–29.
O'Brien HE, Thakur S, Guttman DS. Evolution of plant pathogenesis in *Pseudomonas syringae*: a genomics perspective. Annu Rev Phytopathol. 2011;49:269–89.
Pandey R, et al. Ice-nucleating bacteria control the order and dynamics of interfacial water. Sci Adv. 2016;2(4):e1501630.
Pietsch RB, Vinatzer BA, Schmale DG 3rd. Diversity and abundance of ice nucleating strains of Pseudomonas syringae in a freshwater lake in Virginia, USA. Front Microbiol. 2017;8:318.
Rico A, Preston GM. *Pseudomonas syringae* pv. *tomato* DC3000 uses constitutive and apoplast-induced nutrient assimilation pathways to catabolize nutrients that are abundant in the tomato apoplast. Mol Plant Microbe Interact. 2008;21(2):269–82.
Sarkar SF, Guttman DS. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. Appl Environ Microbiol. 2004;70(4):1999–2012.
Schellenberg B, Ramel C, Dudler R. Syringolin A: action on plants, regulation of biosynthesis and phylogenetic occurrence of structurally related compounds. In: Pseudomonas syringae pathovars and related pathogens – identification, epidemiology and genomics. Dordrecht: Springer; 2008. p. 249–57.
Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 2014;22(5):235–47.
Shapiro BJ, et al. Looking for Darwin's footprints in the microbial world. Trends Microbiol. 2009;17(5):196–204.
Sjödin P, et al. On the meaning and existence of an effective population size. Genetics. 2005;169 (2):1061–70.
Sniegowski PD, Gerrish PJ. Beneficial mutations and the dynamics of adaptation in asexual populations. Philos Trans R Soc Lond Ser B Biol Sci. 2010;365(1544):1255–63.
Starr MP, Chatterjee AK. The genus Erwinia: enterobacteria pathogenic to plants and animals. Annu Rev Microbiol. 1972;26:389–426.
Stavrinides J, McCloskey JK, Ochman H. Pea aphid as both host and vector for the phytopathogenic bacterium Pseudomonas syringae. Appl Environ Microbiol. 2009;75(7):2230–5.
Stopnisek N, et al. Genus-wide acid tolerance accounts for the biogeographical distribution of soil Burkholderia populations. Environ Microbiol. 2014;16(6):1503–12.
Straub C, et al. The ecological genetics of Pseudomonas syringae residing on the kiwifruit leaf surface. bioRxiv. 2017. https://doi.org/10.1101/235853.
Stukenbrock EH, McDonald BA. The origins of plant pathogens in agro-ecosystems. Annu Rev Phytopathol. 2008;46:75–100.
Sung W, et al. Evolution of the insertion-deletion mutation rate across the tree of life. G3 (Bethesda). 2016;6(8):2583–91.
Swingle B, et al. Recombineering using RecTE from *Pseudomonas syringae*. Appl Environ Microbiol. 2010;76(15):4960–8.
Van Cauwenberghe J, et al. Population structure of root nodulating Rhizobium leguminosarum in Vicia cracca populations at local to regional geographic scales. Syst Appl Microbiol. 2014;37 (8):613–21.
Vinatzer BA, Monteil CL. *Pseudomonas syringae* genomics: from comparative genomics of individual crop pathogen strains toward population genomics. In: Gross DC, Lichens-Park A, Kole C, editors. Genomics of plant-associated bacteria. Berlin: Springer-Verlag; 2014. p. 79–98.

Vinatzer BA, Monteil CL, Clarke CR. Harnessing population genomics to understand how bacterial pathogens emerge, adapt to crop hosts, and disseminate. Annu Rev Phytopathol. 2014;52:19–43.

Vinatzer BA, Tian L, Heath LS. A proposal for a portal to make earth's microbial diversity easily accessible and searchable. Antonie Van Leeuwenhoek. 2017;110(10):1271–9.

Wilstermann A, et al. Potential spread of kiwifruit bacterial canker (*Pseudomonas syringae* pv. *actinidiae*) in Europe. EPPO Bulletin. 2017;47(2):255–62.

Wiser MJ, Ribeck N, Lenski RE. Long-term dynamics of adaptation in asexual populations. Science. 2013;342(6164):1364–7.

Walterson AM, Stavrinides J. Pantoea: insights into a highly versatile and diverse genus within the Enterobacteriaceae. FEMS Microbiol Rev. 2015;39(6):968–84.

Xin X-F, He SY. *Pseudomonas syringae* pv. *tomato* DC3000: a model pathogen for probing disease susceptibility and hormone signaling in plants. Annu Rev Phytopathol. 2013;51:473–98.

Yan S, et al. Role of recombination in the evolution of the model plant pathogen *Pseudomonas syringae* pv. *tomato* DC3000, a very atypical tomato strain. Appl Environ Microbiol. 2008;74 (10):3171–81.

An Introductory Narrative to the Population Genomics of Pathogenic Bacteria, Exemplified by *Neisseria meningitidis*

Kanny Diallo and Martin C. J. Maiden

Abstract The ability to study populations of bacteria, rather than individual isolates from cases of disease, represented a step change in our understanding of the bacterial pathogenesis. The last few decades of the twentieth century and the first two of the twenty-first century saw the development of conceptual, technical, and analytical approaches that enabled the development of first bacterial population genetics and then bacterial population genomics, with the study of pathogens in the forefront of this development. These investigations have enabled the diversity of bacterial pathogen lifestyles to be revealed, including details of their ecology and evolution. Studies of the pathogenic *Neisseria* and specifically *Neisseria meningitidis* were in the forefront of these developments, driven in part because of the complexities of the pathobiology of this organism. In addition to insights into the biology of the meningococcus, these studies have provided insights into bacterial population genomics generally, provided a number of broadly applicable techniques, and had major impacts on understanding and controlling meningococcal disease with vaccination.

Keywords Epidemiology · Evolution · Meningococcus · MLST · Vaccination

1 Introduction

Population genomics is the combination of genome-wide analyses of nucleotide sequence variation with the concepts of population genetics: the genetic analysis of representative samples from biological populations. It has a wide variety of applications including understanding the phylogenetic relationships of members of a

K. Diallo
Department of Zoology, University of Oxford, Oxford, UK

Centre pour les Vaccins en Développement, Bamako, Mali

M. C. J. Maiden (✉)
Department of Zoology, University of Oxford, Oxford, UK
e-mail: martin.maiden@zoo.ox.ac.uk

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_37,

given biological population and elucidating the impact of evolutionary processes and functional variation on their biology (Luikart et al. 2003). The term was initially applied in 1998 in the study of human genomes, but the approach has been much more widely applied, including in the study of bacterial pathogens (Gulcher and Stefansson 1998). Population genomics involves the differentiation between evolutionary processes that influence individual loci, i.e. mutation, selection, and recombination, and those which lead to bacterial adaptation from those that act genome-wide, such as genetic drift or population bottlenecks (Black et al. 2001).

In common with their application to other organisms, population genomic analyses of bacterial pathogens aim to (1) characterise evolutionary forces (mutation, gene flow, recombination, and genetic drift), (2) determine bacterial population structures, (3) elucidate mechanisms of pathogen evolution, and, more pragmatically, (4) improve isolate characterisation and identification (typing) (Robinson et al. 2010). Because of their intrinsic interest, the relatively small size of bacterial pathogen genomes, and the very large numbers of bacterial isolates available from cases of human and animal diseases, very many population genomic studies have been undertaken on a variety of bacterial pathogens including the *Neisseria* (Maiden 2008). However, notwithstanding the advantages that have promoted these studies, they are not without their difficulties, due in large part to the fact that bacterial pathogens are very diverse, having evolved multiple times from non-pathogenic ancestors to exploit different hosts. These difficulties include:

1. Population sampling. Depending on the ecological relationship of the pathogen to the host, a collection of disease isolates may be a more-or-less representative sample of the natural population of the pathogen.
2. Genetic diversity. Some pathogens are genetically highly diverse, while others are essentially monomorphic.
3. Mechanisms of evolution. While all bacteria are fundamentally asexual and clonal, in that they reproduce by binary fission, the impact of the parasexual processes of horizontal gene transfer (HGT) varies widely from non-existent or negligible to extensive and highly disruptive of clonality.
4. Analysis. Due to reasons (1)–(3), it is not straightforward to analyse population genomic data for bacterial pathogens. As discussed elsewhere in this book, population genomic models remain a rapidly developing area of study, and it is usually necessary to adapt the analysis approaches available as well as to develop new ones.

Given the impossibility of comprehensively reviewing the wide variety of different bacterial pathogens for which genomic data are available within a limited space, this chapter will concentrate on the knowledge that has been gained from the population genomic analysis of the encapsulated bacterial pathogen *N. meningitidis* (the meningococcus), with reference to a number of key studies in other organisms. The meningococcus was among the first bacterial pathogens to be investigated using these approaches, and meningococcal biology has a number of features that illustrate the difficulties outlined above, which are broadly applicable to the field (Maiden 2008).

The meningococcus is a major cause of meningitis and septicaemia worldwide. It is especially notorious and feared because of its propensity to cause a rapidly developing, very severe, and frequently fatal disease (invasive meningococcal disease, IMD) in infants, children, and young adults, although it can affect any age group (Rodrigues and Maiden 2018). Notwithstanding this fearsome global reputation as an aggressive pathogen, the meningococcus is normally an asymptomatic coloniser of the human nasopharynx, which relies on transmission among healthy hosts for its survival. Invasive disease therefore represents a dysfunctional bacterium-host relationship that is detrimental, and indeed potentially fatal, to both (Caugant and Maiden 2009). Interestingly, this feature of being an unintentional or 'accidental' pathogen is shared with two other principal causes of meningitis, *Streptococcus pneumoniae* and *Haemophilus influenzae*. Another intriguing feature of the meningococcus, shared with these two otherwise unrelated organisms, is the high diversity of its populations, combined with its competence for DNA uptake throughout its life cycle, which means that HGT has played an important role in its evolution and population structure.

2 Population Sampling

An ideal population genomic study will employ a random, unbiased sample of the population in question; however, such a sample is almost always impossible to obtain due to sampling constraints, including limited number of samples, access to the natural population, and other practical limitations. It is possible to reduce the impact of these issues by aiming to collect an appropriate representation of population diversity, based on an understanding of the natural history of the pathogen (Maiden 2006). Such problems can also be ameliorated by taking care to frame questions appropriate to the samples that have been collected. For obligate pathogens, i.e. those for which disease is an integral part of transmission such as *Mycobacterium tuberculosis*, sampling bacteria from cases of invasive disease may be sufficient to get a good representation of population diversity, at least of those organisms that are transmitted frequently in the host population. It is more complicated for bacteria that have more than one host or that are found in the animal reservoirs and/or the environment, such as the zoonotic pathogens *Campylobacter* spp. and *Salmonella* spp., where sampling the population requires sampling the appropriate reservoir (Young et al. 2007).

In the case of the meningococcus, a particular problem is that culture collections are dominated by organisms isolated from cases of IMD, which represent only a small portion of the population compared to those meningococci that are asymptomatically carried (Caugant and Maiden 2009). Indeed, as they have caused disease, isolates from IMD are, by definition, *unrepresentative* of the natural population of asymptomatically carried organisms. The focus on sequencing isolates from cases of invasive disease can introduce bias into population genomic studies (Caugant and Maiden 2009). Consequently, although population genomic studies of IMD isolates

can and have been undertaken to investigate the spread of particular invasive meningococcal genotypes, they have to be interpreted with care when using them to understand the broader features of meningococcal biology.

At the time of writing, most population genomic studies generated bacterial genome sequence data from cultured isolates; however, it is known that culture methods are only effective for a small number of bacterial taxa and often underestimate the diversity of bacterial populations. With improvements in methodologies and reductions in cost, it is now possible to consider metagenomic approaches that identify all taxa present in a given sample, independent of culture (Bilen et al. 2018). Discussed elsewhere in this volume, these new metagenomic methods will enable a more complete analysis of bacterial populations, as they will permit the composition of communities and populations to be studied; however, there remain difficulties in the prosecution of such investigations, especially for locations such as the nasopharynx, where obtaining sufficient samples of the microbiota free of human DNA can be more difficult than some other locations (Goldberg et al. 2015).

3 Genetic Diversity

The diversity of a given biological population is primarily a consequence of its age, with mutation accumulating over time, but this diversity is also influenced by the selection pressures that the population experiences and the occurrence of HGT (Achtman and Wagner 2008). Thus, the asexual pathogens mentioned below (*M. tuberculosis*, *M. leprae*, *B. anthracis*, and *Y. pestis*) have very limited genetic variation, i.e. they are genetically monomorphic, having evolved relatively recently with limited or no HGT (Achtman 2008). Genomes also experience different types of selection: stabilising (negative) selection, diversifying (positive) selection, and drift (neutral) variation. These selection pressures are not evenly distributed around the genome. In bacteria, which have small compact genomes, most genes are under strong stabilising selection for conservation of function, but intriguingly even such 'housekeeping' genes in bacteria such as the meningococcus can harbour high levels of genetic variation. Other genes, especially those that encode surface components that are recognised by the host immune system, may be under strong diversifying selection. An interesting example of where the selection pressures acting on a gene can change is the penicillin-binding protein (*pbp*) genes of the meningococcus. These were under stabilising selection for the conservation of function until the introduction of penicillin, when diversifying selection pressures became more predominant (Spratt et al. 1989). The identification of the differing selection pressures experienced by different genes by population genomic analyses provides important insight into gene function and is a major contribution of this area of study to the unravelling of complex bacterial phenotypes.

4 Mechanisms of Evolution

Models of bacterial evolution and population structure have been revolutionised by the availability of nucleotide sequencing data (Achtman 2004). It has long been known that bacteria divide asexually by binary fission, which results in two identical daughter cells. Initially, it was thought that genetic exchange played only a marginal role in bacterial evolution and the generation of population structure. Population genomic studies have shown to be the case in certain high-profile pathogenic bacteria, including the monomorphic pathogens *M. tuberculosis*, *Mycobacterium leprae*, *Yersinia pestis*, and *Bacillus anthracis*. As nucleotide sequence data became increasingly available, however, it became clear that HGT is far more significant than originally thought playing a major role in the evolution of most bacteria, including most pathogenic organisms (Yahara et al. 2016). The meningococcus and its close relative *Neisseria gonorrhoeae* (the gonococcus) were important paradigms in establishing this in the late 1990s. The impact of HGT varies among organisms: it is unusually extensive in the gastric pathogen *Helicobacter pylori*, but more limited in most organisms, including the meningococcus (Vos and Didelot 2009).

4.1 Analysis Approaches

The variable impact of HGT in different bacterial populations complicates the analysis of population genetic data from bacteria in general and pathogens in particular (Didelot and Wilson 2015). A detailed description of these problems is beyond the scope of the present chapter, being discussed elsewhere in this volume, but will be mentioned briefly. In the absence of HGT, bacterial diversification follows a phylogenetic model: each division event leads to an identical daughter cell, with variation introduced by occasional mutations. Except in the very unusual case of mutation occurring twice or a reverse mutation, these mutations are passed on exclusively to the descendants of the cells in which they occurred. Thus, as population growth occurs, accompanied by the inevitable accumulation of mutations and diversity reduction events caused by periodic selection and bottlenecking, evolution follows a branching process. This is relatively easily modelled as a phylogenetic tree, and if data on mutation rate can be estimated reliably, major events can even be dated with reasonable certainty. This is the case with the genetically monomorphic pathogens mentioned above, *M. tuberculosis*, *M. leprae*, *Y. pestis*, and *B. anthracis*. Conversely, in sexually reproducing organisms, or bacteria where HGT is extensive and reassorts variation genome-wide, such as in *H. pylori*, classical principles of population genetics, developed in the early twentieth century for sexually reproducing organisms, can be applied. As discussed above, however, neither of these extremes applies to most bacteria and bacterial pathogens, including the meningococcus. Consequently, a pragmatic combination of different analysis

approaches, phylogenetic and population genetic, is essential for studies of pathogen population genomics. The best and most developed approaches attempt to model evolution directed by both processes, but this is complicated and computationally intensive (Didelot and Wilson 2015; Didelot and Falush 2007). Here we shall illustrate this with several studies, including early pre- and 'first-generation' population genomic studies of the meningococcus.

5 Phenotypic Methods for Bacterial Characterisation and Diversity Studies

Before the availability of nucleotide sequencing methods, which expanded at the end of the twentieth century and the beginning of the twenty-first century, phenotype-based methods were essential for bacterial characterisation. They remain an important part of clinical microbiology and are often a prerequisite for population analysis. Different tests have been elaborated to study the diversity of bacteria by exploring the observed phenotypic diversity of the organisms, which ultimately represent the diversity encoded by the genome. Bacterial culture is still widely used for bacterial identification, but decisions on appropriate culture methods require an extensive amount of a priori knowledge of the bacterium of interest. Optimal growth conditions have yet to be identified for most bacteria, which remains a major challenge for studying its diversity using culture-based methods; however, many human pathogenic bacteria have been well-characterised in this respect. Serological assays have also played an important role in improving the understanding of the epidemiology of meningitis by facilitating the identification and typing of pathogenic bacteria like *N. meningitis*, *S. pneumoniae*, *Escherichia coli*, *Salmonella* spp., and *H. influenzae* which are known to harbour different serological types, with only few causing most cases of disease.

One of the first applications of bacterial population genetic approaches to bacterial diversity was multilocus enzyme electrophoresis (MLEE). This method, first used in the study of drosophila (Lewontin and Hubby 1966) and human (Harris 1976) population genetics, was extended to bacteria in the 1980s (Selander et al. 1986). MLEE allowed the assessment of genetic discrimination within bacterial population samples by assigning electrophoretic types (ETs), based on the differences in the electrophoretic mobility of housekeeping enzyme variants in starch gel electrophoresis (Caugant et al. 1986). This was the first approach to indicate the large spectrum of population structures present among different bacteria, ranging from the predominantly clonal organisms such as *Bordetella* spp. (Musser et al. 1987), encapsulated *H. influenzae* (Musser et al. 1988), and *Salmonella* spp. (Selander et al. 1990) to less strongly clonal bacteria such as *N. meningitidis* (Caugant et al. 1987a), *Staphylococcus aureus* (Musser and Kapur 1992), and *S. pneumoniae* (Hall et al. 1996).

Early MLEE analysis of *N. meningitidis* indexed the sequence diversity of 9 enzymes among 152 *N. meningitidis* isolates of diverse serogroups, but with a majority of serogroup B meningococci (*Nm*B), and identified 55 distinct ETs. No substantive genetic differences were observed between carried and invasive isolates with this method; however, one ET, ET-5, was associated with IMD, as it represented 58% of invasive isolates in the samples but only 18% of the carried isolates, demonstrating that distinct *N. meningitidis* genotypes were differentially distributed among isolates from disease and carriage, with some meningococci exhibiting a 'hyperinvasive' phenotype. The analysis of the allele distributions among the different genotypes was also the first indication of extensive HGT, mediated by homologous recombination in meningococcal populations (Caugant et al. 1986).

MLEE was also used to study the diversity of invasive and carried *N. meningitidis* (Caugant et al. 1987a, b, 1988; Olyhoek et al. 1987). A similar approach, with the additional analysis of the diversity of OMPs, was also used to investigate the diversity of serogroup A (*NmA*) epidemic isolates, suggesting a predominantly clonal population structure (Olyhoek et al. 1987). Twelve different enzymes were used in the MLEE analyses of 655 *E. coli* isolates, and 60 different ETs were identified (Caugant et al. 1984); similarly, MLEE was used to study the diversity among 261 *E. coli* isolates sharing the same serological antigens O, H, and K and found that there was almost as much diversity among the isolates sharing a single antigen as between randomly chosen isolates (Caugant et al. 1985). The study of 242 isolates of *H. influenzae*, including 65 nontypable and 177 type b isolates, characterised at 15 enzymes, identified 94 ETs. This study confirmed the genetic distinction between the type b isolates which were associated with specific ETs and the nontypable ones, which had distinct ET for each 65 isolates tested. From these results, it was hypothesised that the ancestors of *H. influenzae* were most likely encapsulated (Musser et al. 1986a). In contrast the MLEE analysis of 60 strains of *Bordetella* spp., using 15 enzymes, yielded 14 ETs that were very similar even among different species like *Bordetella parapertussis* and *Bordetella bronchiseptica* (Musser et al. 1986b). MLEE established bacterial population genetics as a discipline, as it assessed variation at loci and could be correlated directly to allelic changes in the enzyme-encoding DNA sequences. It could therefore be used for phylogenetic analysis; however, the method was difficult to reproduce among different laboratories, and the full potential of population genomics that was indicated by these pioneering studies was realised by the advent of nucleotide sequence-based methods.

6 Sequence-Based Methods for Bacterial Characterisation and Diversity Studies

Even though it was possible to sequence the whole genome of a bacterium by the end of the 1990s (Tettelin et al. 2000), it remained very expensive, technically difficult, and mostly undertaken only by specialist genome centres on single, or at most a few,

isolates of given bacterial pathogens. Bacterial population genetic studies therefore relied on a series of sequence-based molecular methods focusing on a particular gene or set of genes within a pathogen population. These methods were all at the interface between population genetics and population genomics as it has subsequently developed.

What might be termed 'first-generation' bacterial population genomics, which linked bacterial population genetics with what is currently thought of as population genomics, began in the late 1990s and consisted of the sequencing of a limited number of loci across a large number of isolates using dideoxy sequencing, originally established by Fred Sanger in 1977 (Sanger et al. 1977) and subsequently developed into a high-throughput approach (Prober et al. 1987). By the late 1990s, the Sanger method provided sequence reads, of up to about 1,000 base pairs (bp) with reliability on about 400–500 bp, within a single experiment, with sufficiently high accuracy that a nucleotide sequence could be reliably established by sequencing a given gene once on each DNA strand. The first-generation automated sequencing instruments allowed high-throughput sequencing of individual loci, making them suitable for multilocus studies and the sequencing of single complete genomes (Prober et al. 1987).

One of the most common applications of Sanger sequencing in bacterial population genomics was multilocus sequence typing (MLST), which at the time of writing remained an important paradigm for bacterial typing (Maiden et al. 1998). MLST schemes have been developed for many different bacteria, and 97 bacterial MLST schemes were hosted on the PubMLST database (https://www.pubmlst.org) at the time of writing (Jolley and Maiden 2014). MLST adapted the principles exploited by MLEE but assessed the variation in the sequences of small number (usually seven) of fragments of genes under stabilising selection for conservation of metabolic function (housekeeping genes) (Maiden et al. 1998; Achtman et al. 2012; Meats et al. 2003; Enright and Spratt 1998). MLST indexes allelic variation by assigning an arbitrary allele number to each unique sequence described and combining the resulting seven numbers obtained for each isolate into an allelic profile or sequence types (ST), analogous to the ET of MLEE. The method led to the discovery that for many bacteria, STs could be grouped into groups of related STs called clonal complexes (cc), in a similar way that ETs had been grouped previously. For many organisms including the meningococcus, ccs persist in bacterial populations through time and geographical spread and are surrogates for genetic lineages (Bratcher et al. 2012). In the case of *N. meningitidis*, this allowed the identification of hyperinvasive lineages, equivalent to those observed by MLEE (Maiden et al. 1998; Yazdankhah et al. 2004); however, unlike MLEE, MLST was scalable, easily reproduced among laboratories, and amenable to dissemination electronically, leading to the establishment of global databases containing MLST data. In 2009 MLST analyses had established that the following ccs corresponded to the hyperinvasive lineages found to be responsible for IMD globally: cc1, cc5, cc8, cc11, cc18, cc23, cc32, cc41/44, cc103, cc162, cc269, and cc334. Other clonal complexes were found principally in samples from asymptomatic carriage (Caugant and Maiden 2009; Bratcher et al. 2012).

Despite its many advantages over MLEE, MLST remained relatively labour-intensive, required expensive equipment and some level of computer literacy, and was consequently challenging to implement in many routine settings. MLST schemes remain still very relevant to epidemiological studies of many bacteria, but in the era of next-generation sequencing (NGS) technologies (Loman et al. 2012), it is increasingly more cost-effective to determine MLST loci from whole genome sequence (WGS) data rather than sequencing them individually. The MLST approach is, however, highly scalable in terms of the number of loci used (Maiden et al. 2013).

Phylogenetic inference made by these sequence analyses confirmed the importance of HGT, as suggested by MLEE, in the evolution of bacteria (Didelot and Maiden 2010). Different bacterial recombination mechanisms have been defined: transduction, the introduction of genetic variation through incorporation of DNA from a viral phage; conjugation, the transfer of DNA sequences via a direct cell to cell contact between two bacterial cells; and transformation, involving DNA uptake from the environment (Goodman and Scocca 1988). Despite the recognition of the occurrence of HGT, the clonal paradigm of bacterial population was considered predominant, and studies of invasive serogroup A meningococci suggested a clonal model of evolution of these bacteria (Olyhoek et al. 1987; Wang et al. 1992; Nicolas et al. 2001); however, the inclusion of more carriage isolates in the sequence analysis has shown that the meningococcus was actually a highly recombinogenic bacterium with high levels of HGT, as suggested by the different phylogenetic relationships inferred from the trees reconstructed from sequences of different genes (Feil et al. 1996; Zhou et al. 1997; Salvatore et al. 2002). Recombination rates calculated from MLST data using the ClonalFrame algorithm (Vos and Didelot 2009) have been described to be about 30 times those of mutation, for *Salmonella enterica*, 23 times in the case of *S. pneumoniae* (Hanage et al. 2005), 7 times for *N. meningitidis* (Jolley et al. 2005), 4 times for *H. influenzae* (Meats et al. 2003), 0.3 times for *Klebsiella pneumoniae* (Diancourt et al. 2005), and 0.1 times in the case of *Staphylococcus aureus* (Enright et al. 2000).

7 The Genome Era: The Population Genomic Approach Comes of Age

The improvement in the methods and decrease in cost of WGS technologies made the sequencing of large numbers of genomes feasible, replacing and complementing single-gene sequencing. The first bacterium to be whole genome sequenced was *H. influenzae* in 1995 using a 'shotgun' sequencing approach (Hood et al. 1996). This was achieved with Sanger sequencing methods with individual reads computationally assembled into a single contiguous sequence, or 'contig', of 1,830,137 bp. This took about a year to complete and was a proof of concept that such method could work with bacterial genomes, being the first of many pathogen genomes

sequenced in this way (Fleischmann et al. 1995). The first meningococcal genomes to be sequenced were the serogroup A isolate Z2491 (Parkhill et al. 2000) and the serogroup B isolate MC58 (Tettelin et al. 2000) both published in 2000. It was not until 2010 that the genome sequence of a non-pathogenic *Neisseria*, *Neisseria lactamica*, became available (Bennett et al. 2010); indeed, despite the possibility of WGS, most projects continued to rely on Sanger sequencing methods to study the population genomics of particular medically relevant organisms up until a significant drop in prices of the different technologies to affordable ranges. Not only have the prices decreased, but the length of time to generate a sequence has also considerably reduced to a few days, and user-friendly analysis tools have been developed making the genome assembly and analysis more accessible to microbiologists, without the necessity of an extensive training in bioinformatics. Robust population genomic studies of bacterial pathogens have been made possible through those technical improvements (Gardy and Loman 2018).

7.1 Impact of Next-Generation Sequencing Technologies

Once the population of interest has been identified and appropriately sampled, different sequencing approaches can be used to analyse the isolates, depending on the question(s) to be addressed. DNA extraction must be adequately prepared as they represent the starting point of any WGS analysis. Following the success of first-generation sequencing, new methods commonly called 'next-generation sequencing' (NGS) were developed and became available from 2005 (Junemann et al. 2013). The first of such NGS platforms consisted of higher-throughput sequencing systems, referred to in this chapter as second-generation sequencing (SGS). These increased throughputs by sequencing large number of DNA molecules in parallel. Generally, these approaches generated shorter, less accurate sequences compared to dideoxy sequencing, but the extremely high sequencing capacity of these instruments, combined with computational developments, enabled these disadvantages to be overcome by high levels of sequence coverage (Loman et al. 2012). The SGS allowed sequencing of hundreds or thousands of bacterial genomes to 'high-quality draft' status (Chain et al. 2009), that is, not complete finished genomes, but the great majority of the genome assembled into a number of sequence contigs at high accuracy. At the time of writing, for example, the *Neisseria* PubMLST database contained more than 43,172 isolates with genetic and epidemiological data and 13,384 whole genome sequences from various geographical regions.

7.2 Evolution and Population Structure

Genomic analyses of large numbers of *N. meningitidis* isolates have elucidated many aspects of meningococcal evolution and population structure, for example, enabling an improved understanding of the role and mechanisms of HGT in this organism.

For example, it has been shown that HGT was favoured by the presence of multiple copies of DNA uptake sequences (DUS) throughout the genome, which are especially concentrated in and around conserved genes with essential metabolic functions. This led to the suggestion that HGT plays an important conservative role in meningococcal evolution and is not solely, or even principally, a means of generating variation that can be acted on by diversifying selection (Davidsen et al. 2004; Treangen et al. 2008). Analysis of the first *N. meningitidis* genome sequences identified almost 1,900 DUS copies (Davidsen et al. 2004; Treangen et al. 2008). These DUS have also been identified in other non-pathogenic *Neisseria* species (Marri et al. 2010), which are known to exchange DNA with *N. meningitidis* (Bennett et al. 2009), and an increase rate of HGT is observed between meningococci sharing similar DUS (Frye et al. 2013). Other repetitive elements present in both pathogenic and non-pathogenic *Neisseria* include Correia elements, which are mobile sequences (Buisine et al. 2002) flanked by 26 bp inverted Correia repeats (Correia et al. 1986), and dRS3 elements which are a family of 20 bp repeat sequences abundant in *N. meningitidis* genomes and involved in recombination (Parkhill et al. 2000); both have been shown to be involved in gene regulation and sequence variation in pathogenic *Neisseria* (Bentley et al. 2007).

Meningococcal repetitive genome elements facilitate another important phenomenon, phase and antigenic variation, a mechanism which allows meningococci to turn the expression of some of their OMPs *on* or *off*, allowing for immune evasion when they are *off*. The changes in the length of these repeated sequences in these protein-encoding genes are the results of recombination events altering the coding sequences (Tan et al. 2016; Seib et al. 2015). Phase variation mechanisms are found in genes that are important for bacterial adaptation to different environments, some of which play a major role in invasion and virulence (Moxon et al. 1994). Although phase variable genes are also present in non-pathogenic *Neisseria*, they tend to not have the repeating elements and are probably subject to less phase switching (Marri et al. 2010) than their homologs in pathogenic species.

We now know that most HGT events among meningococci are mediated by homologous recombination between very closely related organisms. This makes the process difficult to distinguish from genetic variation arising by mutation, as the bacteria exchange sequences that are often similar with few nucleotide changes. The limited number of genomes from carried isolates, only 751 carried *N. meningitidis* were recorded in PubMLST at the time of writing, has been a limiting step in our understanding of the evolution of the meningococci.

The study of *N. meningitidis* population structures has benefited from the higher-resolution characterisation of the isolates. A meningococcal core genome has been established, comprising 1,605 loci present in 95% of their 108 representative meningococcal isolates and distinguishing 10 lineages in agreement with the major invasive clonal complexes characterised by seven-locus MLST (Bratcher et al. 2014). This core genome has been subsequently used in the analysis of different isolate collections, such as the Meningitis Research Foundation Meningococcus Genome Library (MRF-MGL), a database of all the confirmed cases of *N. meningitidis* infection in England and Wales since 2010; the analysis of the

899 isolates collected in the epidemiological years 2010–2011 and 2011–2012 identified more than 20 distinct lineages and a high level of recombination (Hill et al. 2015). Lineages or genogroup-specific population structure studies can be conducted at genomic levels of resolution, as exemplified by the analysis of the global cc11 isolates, which identified a South American/UK serogroup W strain, distinct from those identified in other regions of the world (Lucidarme et al. 2015) and the study of the invasive serogroup Y meningococci in Sweden (Toros et al. 2015). The analysis of the 92 cc11 serogroup W meningococci collected between 1994 and 2012 from the African meningitis belt (AMB) revealed a phylogeographic clustering of the isolates, as their phylogeny reflected their geographical origin (Retchless et al. 2016), while the analysis of 81 serogroup C isolates collected during the 2015 epidemic in Niger showed that they were distinct from those circulating globally (Kretz et al. 2016).

WGS studies have also focused on the identification of genomic determinants of virulence, following on from studies undertaken with MLST or antigen gene sequences (Callaghan et al. 2008; Climent et al. 2010). For example, the genomes of disease and carriage isolates have been compared to identify genetic differences between these groups of isolates, for both serogroup Y (Oldfield et al. 2016) and serogroup A meningococci (Diallo et al. 2017); however, to date no genomic study had identified a single virulent factor that differentiates between these phenotypes, although the meningococcal disease-associated (MDA) island (Bille et al. 2005) was shown to be involved in the ability of the bacteria to cause disease especially in young adult (Bille et al. 2008). Nonetheless, it was found that the bacterial population circulating during the Chadian serogroup A epidemic in 2011 was not homogeneous, as would have been suggested by genogrouping or MLST analysis (Fig. 1), and was separated into distinct clusters associated with the age of the individuals sampled, indicating that external factors, host-related or environmental, could play a role in the microevolution of those bacteria (Fig. 2) (Diallo et al. 2017).

7.3 Disease Surveillance

At the start of the twenty-first century, there was a decline in the incidence of IMD caused by the disease-associated meningococcal lineages that had characterised global epidemiology over the latter half of the twentieth century (Hill et al. 2015). This was due, at least in part, to the highly effective immunisation campaigns using the capsular polysaccharide conjugate vaccines, with substantial reductions of IMD caused by serogroup C and serogroup A organisms in those regions where the vaccines had been deployed (Maiden 2013). Similar reductions had occurred in the incidence of *H. influenzae* type b (Hib) disease, where the Hib-conjugate polysaccharide vaccine had been widely used. However, continuing problems with serogroup B IMD, where no conjugate polysaccharide vaccine was available, the recrudescence of serogroup Y and W IMD (where vaccines were available), and the issues with vaccine escape in *S. pneumoniae*, all indicated that continued

Fig. 1 rMLST neighbour-joining tree of 23 *NmA* from Chad and publicly available cc5 *Nm*A. The relationship from the concatenated nucleotide sequences of the ribosomal genes between the *Nm*A isolates from Chad ($n = 23$) and other publicly available cc5 *Nm*A genomes from the PubMLST database is represented in this tree. The label on each node indicates the PubMLST ID number, the country, the date, and the ribosomal sequence type (rST) for each isolate represented. A total of 141 other cc5 *Nm*A isolates were found in PubMLST, but only 1 representative of each unique strain (defined as isolates sharing the same alleles at all 53 ribosomal loci) was included in the tree alongside all the Chadian *Nm*A from the 2011 meningitis epidemic. The seven-locus MLST profiles of the isolates are indicated by different coloured boxes. The position of the reference genome used in this study (WUE 2594) is represented by a black star. This figure was published by Diallo et al. (2017)

epidemiological surveillance was necessary to maintain and improve control of IMD (Elberse et al. 2016).

At the time of writing, WGS approaches were increasingly the method of choice for meningococcal isolate characterisation in high-income countries. In these settings, they were already the most cost-effective means of collecting multilocus data, including those required for MLST and antigen typing. In addition, WGS provided

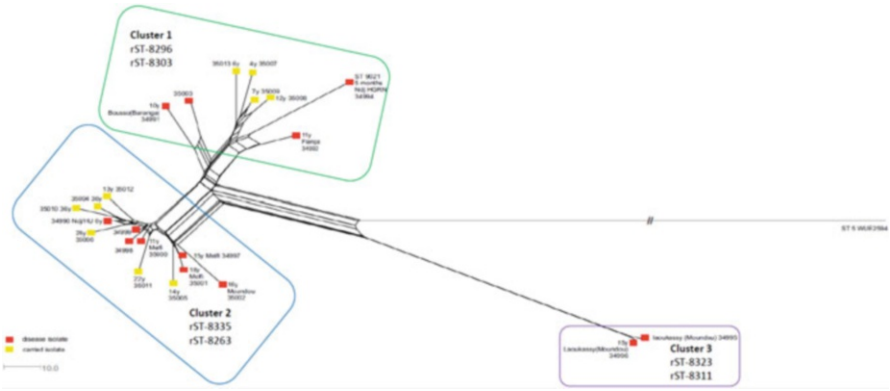

Fig. 2 wgMLST neighbour-net tree of the 23 *NmA* from the 2011 meningitis epidemic in Chad. The genomic relationship based on wgMLST between the Chad *Nm*A isolates is depicted in relation to the reference genome WUE2594. Three clusters are observed and labelled on the tree. The invasive isolates are depicted in red and the carried ones in yellow. The rSTs contained in each cluster are also indicated as well as the age and region of the patient/healthy volunteer are indicated when available; ND corresponds to the absence of any epidemiological information for the specific isolate. The tree was produced based on a comparison in terms of $n = 2{,}070$ loci defined in the reference genome. This figure was published by Diallo et al. (2017)

additional information from the remainder of the genome that was useful for high-resolution examination of disease outbreaks and surveillance. The first routine application of WGS for meningococci for disease surveillance, and one of the first for any bacterium, was the establishment of the Meningitis Research Foundation Meningococcus Genome Library (MRF-MGL), hosted within the pubMLST.org/neisseria database (Jolley and Maiden 2010, 2014). The MRF-MGL contains high-quality draft genome sequences for all the meningococci isolates received at the UK national reference laboratories since 2010 (Hill et al. 2015). This allows automatic typing of isolates, rapid sharing of data via the Internet, and the possibility of additional genomic analysis and comparisons with other genomes, from other countries, present in the database.

As more European countries systematically sequence their isolates and deposit them in web-accessible databases (Bratcher et al. 2018), a regional genomic surveillance system will emerge, enabling high-resolution monitoring of the molecular epidemiology of IMD; however, a complete picture of the molecular diversity of meningococcal populations requires the simultaneous collection and characterisation of carried isolates. The effort to sequence invasive isolates has not been limited to *N. meningitidis*; indeed, in the UK Public Health England has adopted WGS as the routine typing method for all the *Salmonella* isolates received, thereby replacing the serotyping assays and creating a large pool of sequence data (Ashton et al. 2016).

In 2017, however, WGS remained very far from being accessible in LMIC settings, even in well-established research centres and certainly not in national reference laboratories. The national surveillance systems still relied mostly on

clinical diagnosis and classical laboratory methods, although more national reference laboratories were equipped to perform molecular tests such as PCR confirmation or occasionally MLST. The regional surveillance supported by the World Health Organization (WHO) provides a platform to implement a genomic surveillance of meningococcal disease or any other bacterial disease of interest using a web-based interface which would facilitate the sharing of results and encourage a community data analysis.

Examples of the potential of this approach include the use of WGS-based surveillance to identify (1) aggressive serogroup W meningococcal lineages in the UK and South America that were distinct from the serogroup W 'Hajj clone' associated with a previous outbreak (Lucidarme et al. 2015), indicating the evolution of the global cc11 serogroup W meningococci (Mustapha et al. 2015), and (2) serogroup C meningococci circulating in the AMB that have caused large epidemics (Kretz et al. 2016). Other WGS studies defined the meningococci responsible for distinct waves of serogroup A:cc5 meningococci (ST7 and ST2859) in the AMB: one study suggested this might be due to herd immunity evasion via homologous recombination affecting noncapsular exposed antigens (Lamelas et al. 2014), but an alternative explanation was that this could be due to changes in genes involved in metabolic functions affecting transmission (Watkins and Maiden 2017). Such analyses improved the understanding of the evolution of the disease isolates and would not have been possible at lower resolution.

WGS analyses have been crucial in meningitis outbreak investigation in several regions of the world, allowing the identification of new variants that were indistinguishable when characterised by MLST (Diallo et al. 2017; Lavezzo et al. 2013; Mulhall et al. 2016). WGS data have also been used for identification of outbreak strains and their discrimination from unrelated cases (Jolley et al. 2012). If coupled with epidemiological data, there is potential for the study of transmission dynamics, as previously undertaken for other bacterial pathogens, for example, in the investigation of multidrug-resistant *Staphylococcus aureus* (MRSA) outbreaks (Millar et al. 2017).

8 Future Perspectives

The advent of high-throughput, cost-effective WGS technologies in the first decades of the twenty-first century provided a powerful impetus to the analysis of populations of pathogenic bacteria. It became possible to analyse the complete, or very new near-complete (high-quality draft), genomes for hundreds or even thousands of bacterial isolates. The sequence data were combined with a number of analysis approaches, both conventional and newly developed, to establish bacterial population genetics as a major paradigm in the analysis of bacterial biology. This was first widely exploited for the analysis of bacterial pathogens of humans but is a broadly applicable approach with wide application.

At the time of writing, population genomics has been exploited to investigate a number of questions, ranging from high-resolution epidemiology (trees combined with maps) through bacterial evolution and population structure to host-pathogen interactions and virulence. There are also major opportunities for the investigation of antibiotic resistance, the establishment of global bacterial disease surveillance databases, and the development of new vaccines; however, the accessibility to these technologies is still skewed to high-income countries, despite the highest burden of infectious disease being localised in low- and middle-income countries (LMICs), especially sub-Saharan Africa. Consequently, there remains a need to translate population genomics to the LMIC setting. In principle, there are many advantages to using sequencing in resource-poor settings: although the set-up costs are high, this is also true for more conventional microbiological analyses, with molecular and sequencing analyses presenting opportunities for a 'technology jump' where novel laboratory capacity is set up a priori with molecular methods that are broadly applicable with a wide range of pathogens. Some early examples of this were the deployment of the Oxford Nanopore and Ion Torrent sequencing platforms for molecular epidemiology in the Ebola epidemic in West Africa in 2015 (Quick et al. 2016).

A further challenge is the establishment of representative sampling frames for specimens from across the globe. The under-sampling of particular regions of the world and oversampling of others increase the risk of biases and can lead to an inaccurate picture of the diversity and evolution of a given bacterial population. It is therefore important that sustainable systems are put in place for improved access to appropriate sampling technology, combined with adequate training of infectious disease scientists, and the development of improved surveillance systems in LMIC settings. Similarly, the emphasis on bacterial isolates obtained from cases of invasive disease can also give a distorted view of the bacterial population. The development and deployment of effective metagenomic analyses and increased sampling of asymptomatic infection and environmental reservoirs will have important practical and conceptual implications (Bilen et al. 2018).

When it comes to translation of these data and ideas to public health, it is important that the different analysis methods and bioinformatics pipelines are standardised, with clear and reproducible procedures that have been validated with appropriate publicly accessible trials. Regulatory and advisory bodies will need to generate guidelines for the certification of genomic methods and agree on protocols and algorithms to be implemented. Furthermore, studies are required to establish the practical impact of implementation of these approaches in hospital settings for patient care management and improved disease outcomes. Such studies are also needed to determine the level at which genomic analysis should most appropriately be performed (hospital, regional, national, or international) and when real-time genomic data is clinically relevant.

The increase in data availability poses challenges in data storage and handling and also generates a requirement for increased computational power and bioinformatics expertise. Population genomics has not answered all the questions raised by the intriguing life cycles of bacteria such as the meningococcus. For example, it

remains unclear why hyperinvasive variants of this accidental pathogen persist; however, the integration of genomics with the other 'omics' approaches, transcriptomics, proteomics, and epigenomics, will generate new insights into such questions.

In conclusion, population genomics continues to revolutionise the study of pathogenic bacteria, providing tools to address a wide range of questions but also providing information and approaches that are directly applicable to public health. As technologies and analysis approaches continue to develop and be deployed, their utility and dissemination will increase in all settings, but, perhaps most importantly, they provide the prospect of the 'technology jump' in LMIC settings, where it is possible to envisage a move directly from first-generation technology, involving culture and serotyping, to near-patient sequence analysis, with a consequent revolution in clinical microbiology and public health.

References

Achtman M. Population structure of pathogenic bacteria revisited. Int J Med Microbiol. 2004;294(2–3):67–73.

Achtman M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu Rev Microbiol. 2008;62:53–70.

Achtman M, Wagner M. Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol. 2008;6(6):431–40.

Achtman M, et al. Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. PLoS Pathog. 2012;8(6):e1002776.

Ashton PM, et al. Identification of Salmonella for public health surveillance using whole genome sequencing. PeerJ. 2016;4:e1752.

Bennett JS, Thompson EA, Kriz P, Jolley KA, Maiden MC. A common gene pool for the *Neisseria* FetA antigen. Int J Med Microbiol. 2009;299(2):133–9.

Bennett JS, et al. Independent evolution of the core and accessory gene sets in the genus *Neisseria*: insights gained from the genome of *Neisseria lactamica* isolate 020-06. BMC Genomics. 2010;11:652.

Bentley SD, et al. Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. PLoS Genet. 2007;3(2):e23.

Bilen M, et al. The contribution of culturomics to the repertoire of isolated human bacterial and archaeal species. Microbiome. 2018;6(1):94.

Bille E, et al. A chromosomally integrated bacteriophage in invasive meningococci. J Exp Med. 2005;201(12):1905–13.

Bille E, et al. Association of a bacteriophage with meningococcal disease in young adults. PLoS One. 2008;3(12):e3885.

Black WC, Baer CF, Antolin MF, DuTeau NM. Population genomics: genome-wide sampling of insect populations. Annu Rev Entomol. 2001;46:441–69.

Bratcher HB, Bennett JS, Maiden MCJ. Evolutionary and genomic insights into meningococcal biology. Future Microbiol. 2012;7(7):873–85.

Bratcher HB, Corton C, Jolley KA, Parkhill J, Maiden MC. A gene-by-gene population genomics platform: *de novo* assembly, annotation and genealogical analysis of 108 representative *Neisseria meningitidis* genomes. BMC Genomics. 2014;15:1138.

Bratcher HB, et al. Establishment of the European meningococcal strain collection genome library (EMSC-GL) for the 2011 to 2012 epidemiological year. Euro Surveill. 2018;23(20).

Buisine N, Tang CM, Chalmers R. Transposon-like Correia elements: structure, distribution and genetic exchange between pathogenic *Neisseria* sp. FEBS Lett. 2002;522(1–3):52–8.

Callaghan MJ, et al. Opa protein repertoires of disease-causing and carried meningococci. J Clin Microbiol. 2008;46(9):3033–41.

Caugant DA, Maiden MC. Meningococcal carriage and disease – population biology and evolution. Vaccine. 2009;27(Suppl 2):B64–70.

Caugant DA, Levin BR, Selander RK. Distribution of multilocus genotypes of *Escherichia coli* within and between host families. J Hyg (Lond). 1984;92(3):377–84.

Caugant DA, et al. Genetic diversity in relation to serotype in *Escherichia coli*. Infect Immun. 1985;49(2):407–13.

Caugant DA, et al. Multilocus genotypes determined by enzyme electrophoresis of *Neisseria meningitidis* isolated from patients with systemic disease and from healthy carriers. J Gen Microbiol. 1986;132:641–52.

Caugant DA, et al. Genetic relationships and clonal population structure of serotype 2 strains of *Neisseria meningitidis*. Infect Immun. 1987a;55(6):1503–13.

Caugant DA, et al. Intercontinental spread of *Neisseria meningitidis* clones of the ET-5 complex. Antonie Van Leeuwenhoek. 1987b;53(6):389–94.

Caugant DA, Kristiansen BE, Frøholm LO, Bovre K, Selander RK. Clonal diversity of *Neisseria meningitidis* from a population of asymptomatic carriers. Infect Immun. 1988;56(8):2060–8.

Chain PS, et al. Genome project standards in a new era of sequencing. Science. 2009;326(5950):236–7.

Climent Y, et al. Clonal distribution of disease-associated and healthy carrier isolates of *Neisseria meningitidis* between 1983 and 2005 in Cuba. J Clin Microbiol. 2010;48(3):802–10.

Correia FF, Inouye S, Inouye M. A 26-base-pair repetitive sequence specific for *Neisseria gonorrhoeae* and *Neisseria meningitidis* genomic DNA. J Bacteriol. 1986;167(3):1009–15.

Davidsen T, et al. Biased distribution of DNA uptake sequences towards genome maintenance genes. Nucleic Acids Res. 2004;32(3):1050–8.

Diallo K, et al. Hierarchical genomic analysis of carried and invasive serogroup A *Neisseria meningitidis* during the 2011 epidemic in Chad. BMC Genomics. 2017;18(1):398.

Diancourt L, Passet V, Verhoef J, Grimont PAD, Brisse S. Multilocus sequence typing of *Klebsiella pneumoniae* nosocomial isolates. J Clin Microbiol. 2005;43(8):4178–82.

Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. Genetics. 2007;175(3):1251–66.

Didelot X, Maiden MC. Impact of recombination on bacterial evolution. Trends Microbiol. 2010;18(7):315–22.

Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 2015;11(2):e1004041.

Elberse KE, et al. Pneumococcal population in the era of vaccination: changes in composition and the relation to clinical outcomes. Future Microbiol. 2016;11(1):31–41.

Enright MC, Spratt BG. A multilocus sequence typing scheme for *Streptococcus pneumoniae:* identification of clones associated with serious invasive disease. Microbiology. 1998;144(11):3049–60.

Enright MC, Day NPJ, Davies CE, Peacock SJ, Spratt BG. Multilocus sequence typing for the characterization of methicillin-resistant (MRSA) and methicillin-susceptible (MSSA) clones of *Staphylococcus aureus*. J Clin Microbiol. 2000;38:1008–15.

Feil E, Zhou J, Maynard Smith J, Spratt BG. A comparison of the nucleotide sequences of the *adk* and *recA* genes of pathogenic and commensal *Neisseria* species: evidence for extensive interspecies recombination within *adk*. J Mol Evol. 1996;43(6):631–40.

Fleischmann RD, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* RD. Science. 1995;269:496–512.

Frye SA, Nilsen M, Tonjum T, Ambur H. Dialects of the DNA uptake sequence in Neisseriaceae. PLoS Genet. 2013;9(4):e1003458.

Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet. 2018;19:9–20.
Goldberg B, Sichtig H, Geyer C, Ledeboer N, Weinstock GM. Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. MBio. 2015;6(6):e01888.
Goodman SD, Scocca JJ. Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. Proc Natl Acad Sci U S A. 1988;85:6982–6.
Gulcher J, Stefansson K. Population genomics: laying the groundwork for genetic disease modeling and targeting. Clin Chem Lab Med. 1998;36(8):523–7.
Hall LM, Whiley RA, Duke B, George RC, Efstratiou A. Genetic relatedness within and between serotypes of *Streptococcus pneumoniae* from the United Kingdom: analysis of multilocus enzyme electrophoresis, pulsed-field gel electrophoresis, and antimicrobial resistance patterns. J Clin Microbiol. 1996;34(4):853–9.
Hanage WP, et al. Using multilocus sequence data to define the pneumococcus. J Bacteriol. 2005;187(17):6223–30.
Harris H. Enzyme variants in human-populations. Johns Hopkins Med J. 1976;138(6):245–52.
Hill DMC, et al. Genomic epidemiology of age-associated meningococcal lineages in national surveillance: an observational cohort study. Lancet Infect Dis. 2015;15(12):1420–8.
Hood DW, et al. Use of the complete genome sequence information of *Haemophilus influenzae* strain Rd to investigate lipopolysaccharide biosynthesis. Mol Microbiol. 1996;22(5):951–65.
Jolley KA, Maiden MC. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010;11(1):595.
Jolley KA, Maiden MC. Using MLST to study bacterial variation: prospects in the genomic era. Future Microbiol. 2014;9:623–30.
Jolley KA, Wilson DJ, Kriz P, McVean G, Maiden MC. The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. Mol Biol Evol. 2005;22(3):562–9.
Jolley KA, et al. Resolution of a meningococcal disease outbreak from whole genome sequence data with rapid web-based analysis methods. J Clin Microbiol. 2012;50(9):3046–53.
Junemann S, et al. Updating benchtop sequencing performance comparison. Nat Biotechnol. 2013;31(4):294–6.
Kretz CB, et al. Whole-genome characterization of epidemic Neisseria meningitidis serogroup C and resurgence of serogroup W, Niger, 2015. Emerg Infect Dis. 2016;22(10):1762–8.
Lamelas A, et al. Emergence of a new epidemic *Neisseria meningitidis* serogroup A clone in the African meningitis belt: high-resolution picture of genomic changes that mediate immune evasion. MBio. 2014;5(5):e01974–14.
Lavezzo E, et al. Genomic comparative analysis and gene function prediction in infectious diseases: application to the investigation of a meningitis outbreak. BMC Infect Dis. 2013;13:554.
Lewontin RC, Hubby JL. A molecular approach to study of genic heterozygosity in natural populations. 2. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics. 1966;54(2):595–609.
Loman NJ, et al. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol. 2012;10(9):599–606.
Lucidarme J, et al. Genomic resolution of an aggressive, widespread, diverse and expanding meningococcal serogroup B, C and W lineage. J Infect. 2015;71(5):544–52.
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(12):981–94.
Maiden MC. Multilocus sequence typing of bacteria. Annu Rev Microbiol. 2006;60:561–88.
Maiden MC. Population genomics: diversity and virulence in the *Neisseria*. Curr Opin Microbiol. 2008;11(5):467–71.
Maiden MC. The impact of protein-conjugate polysaccharide vaccines: an endgame for meningitis? Philos Trans R Soc Lond Ser B Biol Sci. 2013;368(1623):20120147.

Maiden MCJ, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998;95(6):3140–5.

Maiden MC, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013;11(10):728–36.

Marri PR, et al. Genome sequencing reveals widespread virulence gene exchange among human *Neisseria* species. PLoS One. 2010;5(7):e11835.

Meats E, et al. Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. J Clin Microbiol. 2003;41(4):1623–36.

Millar EV, et al. Genomic characterization of USA300 methicillin-resistant *Staphylococcus aureus* (MRSA) to evaluate intraclass transmission and recurrence of skin and soft tissue infection (SSTI) among high-risk military trainees. Clin Infect Dis. 2017;65(3):461–8.

Moxon ER, Rainey PB, Nowak MA, Lenski RE. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr Biol. 1994;4(1):24–33.

Mulhall RM, et al. Resolution of a protracted serogroup B meningococcal outbreak with whole genome sequencing shows inter species genetic transfer. J Clin Microbiol. 2016;54(12):2891–9.

Musser JM, Kapur V. Clonal analysis of methicillin-resistant *Staphylococcus aureus* strains from intercontinental sources – Association of the Mec Gene with divergent phylogenetic lineages implies dissemination by horizontal transfer and recombination. J Clin Microbiol. 1992;30(8):2058–63.

Musser JM, Barenkamp SJ, Granoff DM, Selander RK. Genetic relationships of serologically nontypable and serotype b strains of *Haemophilus influenzae*. Infect Immun. 1986a;52(1):183–91.

Musser JM, Hewlett EL, Peppler MS, Selander RK. Genetic diversity and relationships in populations of *Bordetella* spp. J Bacteriol. 1986b;166(1):230–7.

Musser JM, Bemis DA, Ishikawa H, Selander RK. Clonal diversity and host distribution in *Bordetella bronchiseptica*. J Bacteriol. 1987;169(6):2793–803.

Musser JM, Kroll JS, Moxon ER, Selander RK. Evolutionary genetics of the encapsulated strains of *Haemophilus influenzae*. Proc Natl Acad Sci U S A. 1988;85(20):7758–62.

Mustapha MM, et al. Genomic epidemiology of hypervirulent serogroup W, ST-11 *Neisseria meningitidis*. EBioMedicine. 2015;2(10):1447–55.

Nicolas P, et al. Clonal expansion of sequence type (ST-)5 and emergence of ST-7 in serogroup A meningococci, Africa. Emerg Infect Dis. 2001;7(5):849–54.

Oldfield NJ, et al. Genomic analysis of serogroup Y *Neisseria meningitidis* isolates reveals extensive similarities between carriage-associated and disease-associated organisms. J Infect Dis. 2016;213(11):1777–85.

Olyhoek T, Crowe BA, Achtman M. Clonal population structure of *Neisseria meningitidis* serogroup A isolated from epidemics and pandemics between 1915 and 1983. Rev Infect Dis. 1987;9:665–82.

Parkhill J, et al. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature. 2000;404(6777):502–6.

Prober JM, et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science. 1987;238(4825):336–41.

Quick J, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530(7589):228–32.

Retchless AC, et al. The establishment and diversification of epidemic-associated serogroup W Meningococcus in the African Meningitis Belt, 1994 to 2012. mSphere. 2016;1(6).

Robinson DA, Falush D, Feil EJ. Bacterial population genetics in infectious disease. Hoboken: Wiley; 2010. p. 420.

Rodrigues CMC, Maiden MCJ. A world without bacterial meningitis: how genomic epidemiology can inform vaccination strategy. F1000Res. 2018;7:401.

Salvatore P, et al. Phenotypes of a naturally defective *recB* allele in *Neisseria meningitidis* clinical isolates. Infect Immun. 2002;70(8):4185–95.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–7.

Seib KL, et al. Specificity of the ModA11, ModA12 and ModD1 epigenetic regulator N(6)-adenine DNA methyltransferases of *Neisseria meningitidis*. Nucleic Acids Res. 2015;43:4150–62.

Selander RK, et al. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl Environ Microbiol. 1986;51:837–84.

Selander RK, et al. Evolutionary genetic relationships of clones of *Salmonella* serovars that cause human typhoid and other enteric fevers. Infect Immun. 1990;58(7):2262–75.

Spratt BG, et al. Recruitment of a penicillin-binding protein gene from *Neisseria flavescens* during the emergence of penicillin resistance in *Neisseria meningitidis*. Proc Natl Acad Sci U S A. 1989;86:8988–92.

Tan A, et al. Distribution of the type III DNA methyltransferases *modA*, *modB* and *modD* among *Neisseria meningitidis* genotypes: implications for gene regulation and virulence. Sci Rep. 2016;6:21015.

Tettelin H, et al. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science. 2000;287(5459):1809–15.

Toros B, et al. Genome-based characterization of emergent invasive *Neisseria meningitidis* serogroup Y isolates in Sweden from 1995 to 2012. J Clin Microbiol. 2015;53(7):2154–62.

Treangen TJ, Ambur OH, Tonjum T, Rocha EP. The impact of the neisserial DNA uptake sequences on genome evolution and stability. Genome Biol. 2008;9(3):R60.

Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 2009;3(2):199–208.

Wang JF, et al. Clonal and antigenic analysis of serogroup A *Neisseria meningitidis* with particular reference to epidemiological features of epidemic meningitis in China. Infect Immun. 1992;60:5267–82.

Watkins ER, Maiden MCJ. Metabolic shift in the emergence of hyperinvasive pandemic meningococcal lineages. Sci Rep. 2017;7:41126.

(WHO/IST) IcST-WA. Meningitis weekly reports. In: WHO, editor. Epidemiological information-meningitis. Ouagadougou: WHO-Multi-Disease Surveillance Centre, Regional Meningitis Surveillance. http://www.who.int/emergencies/diseases/meningitis/epidemiological/en/.

Yahara K, et al. The landscape of realized homologous recombination in pathogenic bacteria. Mol Biol Evol. 2016;33(2):456–71.

Yazdankhah SP, et al. Distribution of serogroups and genotypes among disease-associated and carried isolates of *Neisseria meningitidis* from the Czech Republic, Greece, and Norway. J Clin Microbiol. 2004;42(11):5146–53.

Young KT, Davis LM, Dirita VJ. *Campylobacter jejuni:* molecular biology and pathogenesis. Nat Rev Microbiol. 2007;5(9):665–79.

Zhou J, Bowler LD, Spratt BG. Interspecies recombination, and phylogenetic distortions, within the glutamine synthetase and shikimate dehydrogenase genes of *Neisseria meningitidis* and commensal *Neisseria* species. Mol Microbiol. 1997;23(4):799–812.

Population Genomics of Archaea: Signatures of Archaeal Biology from Natural Populations

David J. Krause and Rachel J. Whitaker

Abstract Since the advance of high-throughput whole-genome sequencing, microbial population biology has been providing insight into the processes that generate and maintain genomic diversity and shedding light on the complex lives of microorganisms in the natural environment. The domain archaea harbors a wealth of diverse populations useful for studying microbial population biology in highly varied environments, and their deep divergence from Bacteria creates a distinct, independent field of study despite superficial similarities. Today, much of the knowledge derived from archaeal population genomics is the result of culturing individuals from the environment and sequencing isolates, which has enabled the study of population biology for several archaeal species, including mutation rates, recombination rates, and the influence of environmental selection. With a constantly increasing volume of metagenomic data and advancing technology for single-cell genomics, population genomics is making its way into the uncultured majority that has otherwise evaded previous population genomic techniques, and the unique biology of the archaea is poised to enhance our understanding of microbial population biology.

Keywords Archaea · Genome architecture · Population structure · Recombination · Selection · Species

D. J. Krause
Laboratory of Genetics, Genome Center of Wisconsin, Wisconsin Energy Institute, J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, WI, USA

R. J. Whitaker (✉)
Department of Microbiology, Carl R. Woese Institute for Genomic Biology, University of Illinois Urbana-Champaign, Champaign, IL, USA
e-mail: rwhitaker@life.illinois.edu

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_49,

1 Introduction

The discovery of archaea as a third domain of life in the 1970s transformed the way we think about the origin of life, the scope of natural diversity, and the relationships among organisms (Woese and Fox 1977; Woese et al. 1990). Archaea are single-celled microorganisms that share many superficial similarities with the Bacteria, despite being more closely related to Eukarya. However, a historic bias toward bacteria, especially pathogenic strains, and eukaryotes has led to archaea being the least sampled of the three domains of life (Schloss et al. 2016). For decades, environmental 16S rRNA gene sequencing projects used primers biased toward bacterial diversity, but often missing swaths of archaeal diversity (Klindworth et al. 2013), and despite the early sequencing of *Methanococcus jannaschii* (Bult et al. 1996), bacterial genomes quickly outnumbered those of their archaeal counterparts, as a current search of NCBI genome yields 1,400 archaeal genomes, but an excess of 21,000 bacterial genomes.

The sequencing of individual genomes enabled the study of microbial populations, as the highly similar 16S rDNA sequence clusters within populations could be interrogated for further diversity using multi-locus sequence analysis (MLSA) (Maiden 2006; Maiden et al. 1998). Further advances in sequencing technology have enabled researchers to sequence many genomes of organisms within a population (Mau et al. 2006). The most common methodology for this approach is to isolate many individuals from a natural microbial population in the laboratory followed by sequencing individual genomes. The essential component of this method is the ability to interrogate individual level variation across the genome. The specificity of culturing systems has enabled the research of archaeal populations, often due to the ability to isolate archaeal organisms that harbor unique traits. For example, methanogenic archaea can be isolated anaerobically using CO_2 and H_2 or various C1 or C2 compounds (Wolfe 2011), thermoacidophilic *Sulfolobus* spp. can be isolated on low pH media at high (>80°C) temperature (Brock et al. 1972), halophilic archaea can be isolated using excess of 1M salt concentrations (Torreblanca et al. 1986), and ammonia-oxidizing archaea can be isolated via selecting for chemoautotrophic growth on ammonia (Könneke et al. 2005).

However, the appreciation for the difficulty, or impossibility, of culturing the vast diversity of microbes led to the practice of metagenomic sequencing of environmental DNA samples to access the genomes of organisms without the need for culturing (Gilbert and Dupont 2011; Venter et al. 2004). Again, the progress of sequencing technology has now enabled read lengths and coverage depths that make it possible to assemble individual members of microbial communities from shotgun metagenomic data (Parks et al. 2017; Tully et al. 2018). This is a fascinating and encouraging approach to microbial population genomics; however, the difficulty in linking variable regions of the genome into individual genotypes prevents many of the population genetic methods described below. Also, in many environments archaea are rare and masked by dominating bacteria. Further advances are now enabling the sequencing of individual cells taken from natural environments,

facilitating identification of individual genomes while maintaining the independence from culture-based techniques (Gawad et al. 2016; Marcy et al. 2007; Woyke et al. 2009).

In all three domains, population genomics uses the patterns of natural variation in the genomes of closely related organisms to infer key parameters of evolution, including mutation, selection, gene flow through migration, recombination and horizontal gene transfer, genetic drift, and how these forces interact. Understanding these forces in natural populations can elucidate the basis of diversity, the formation of species, and identify particular genes under selection in natural populations. While some forces (selection, genetic drift, migration) are likely universal across the domains, the underlying molecular mechanisms that establish mutation rates, recombination, and horizontal gene transfer in archaea are likely different than those in bacteria and may be similar to eukaryotes. This difference will shape the way these forces interact in the architecture of archaeal genomes. The exciting and powerful potential of studying the archaeal domain is evident in how their unique molecular mechanisms are present in patterns of natural diversity.

Here we will summarize relevant progress in the field of archaeal population genomics using distinct technical approaches: culturing members of archaeal populations from nature and sequencing the genomes of individuals and assembling the genomes of archaeal populations from larger metagenomic datasets.

2 Population Genomics of Cultured Archaeal Isolates

Sequencing the genomes of cultured isolates can be straightforward. Because individual genomes are sequenced in isolation, SNPs can be called with high confidence and information about linkage between SNPs within an organism is absolute. The approach is limited to organisms that can be cultured in the laboratory; meaning it is inaccessible to uncultured taxa or even restricted to only the easily cultured members of generally culturable taxa. However, several archaeal systems have been investigated using population genomics of cultured isolates, and these are currently the most in-depth studies of all the technological approaches currently available.

3 The Thermoacidophilic *Sulfolobus islandicus*

Originally isolated from acidic hot springs in Iceland, *S. islandicus* has also been found in thermal areas in Russia and North America. This species can be cultured on complex media at low pH and high temperature, allowing for efficient selection of this species from hot spring water samples. Prior to the accessibility of high-throughput genome sequencing, multi-locus sequence studies had identified that

populations of *S. islandicus* were isolated by large geographic distance (Whitaker et al. 2003), and within endemic populations, recombination between strains was common (Whitaker et al. 2005). As genomes became more prevalent, comparative genomics approaches for strains from various populations further validated those previous conclusions from MLST data, as well as identified polymorphism patterns and variation in gene content that varied with respect to physical distance from the multiple origins of replication (Flynn et al. 2010; Reno et al. 2009).

The first genomics study to leverage a single *S. islandicus* population focused on the genome sequences of 12 strains isolated from a single hot spring in Kamchatka, Russia (Cadillo-Quiroz et al. 2012). MLSA had identified two divergent subpopulations within the larger Mutnovsky population, and the genomic approach solidified this observation with genome-wide data. By modelling recombination events within and between the subpopulation using ClonalOrigin (Didelot et al. 2010), the authors found that genetic exchange was more frequent within than between the subpopulations, a feature consistent with the biological species concept for diverging species (Cadillo-Quiroz et al. 2012). The authors calculated F_{ST}, a metric for fixed differences between the two subpopulations, to find that fixation was highest near the origins of replication, in the lower diversity regions. This view of "continents" of fixation was an alternative to other studies of speciation that have identified more targeted selection driving species apart in "islands" (Shapiro and Polz 2014). Further analysis of the population genomic dataset using phylogenetic inferences revealed that recombination rates varied around the chromosome and the reduced recombination around origins of replication could explain the decreased nucleotide diversity in these regions due to background selection. Further, the high F_{ST} values calculated in these regions may have also been the result of reduced interpopulation recombination in these regions relative to other regions (Krause et al. 2014).

S. islandicus and the related *Sulfolobus* spp. are prime systems for interrogating population biology of thermophilic archaea in light of several of their novel features. With a habitat restricted to acidic hot springs, the confounding influence of migration is limited when studying isolated populations. *S. islandicus*, like other *Sulfolobus* spp. and many other archaea but unlike most bacteria, have multiple origins of replication (Lundgren et al. 2004; Robinson et al. 2004; Wu et al. 2014). Although *S. islandicus* reproduces asexually, genetic recombination can be inferred from genome sequences as well as observed in the laboratory (Cadillo-Quiroz et al. 2012; Krause et al. 2014; Whitaker et al. 2005; Zhang et al. 2013). Recombination rate calculation in *S. islandicus* estimated a *r/m* value of 1.2, similar to another archaeal taxon *Halorubrum* sp. at 2.1 (Vos and Didelot 2009).

Elucidation of the unique ways that *Sulfolobus* spp. exchange genetic material is still in progress (Ajon et al. 2011; Schleper et al. 1995; Stedman et al. 2000; van Wolferen et al. 2016).

4 The Methanogenic *Methanosarcina mazei*

The methanogen *M. mazei* is a euryarchaeota with a wide metabolic diversity, being capable of growth via anaerobic conversion of acetate, methylamines, and methanol to methane, carbon dioxide, and ammonia in the case of methylamine (Deppenmeier et al. 2002). The species can be isolated anaerobically on these substrates from various anaerobic habitats. A recent study sampled various locations within the Columbia River Estuary, isolating *M. mazei* from these samples and sequencing their genomes (Youngblut et al. 2015). Genome analysis revealed the presence of two major subpopulations within the isolate set. The authors identified several genes characterized by high levels of fixation and divergence between the populations, measured by F_{ST} and sequence identity, respectively. These genes included three molybdate transport genes and three molybdenum-containing formylmethanofuran dehydrogenase subunits, as well as other genes including several encoding hypothetical proteins. Some of these genes were also concentrated in a highly fixed and divergent region of the chromosome, raising the possibility of genomic architecture playing a role. The subpopulations also differed in gene presence/absence patterns, and although the majority of these genes were annotated as "hypothetical proteins" as is frequent in archaeal genomes, several gene annotations represented multiple instances of nearly fixed gene content, including CRISPR-associated proteins, glycosyltransferases, and restriction-modification proteins. Further, the authors measured phenotypic variation between the subpopulations, as assayed by rates of methane production from various substrates. They identified significant differences between the subpopulations in methane production from trimethylamine.

The unique metabolism of methanogenic archaea allows them to fulfill a very specific niche in anaerobic habitats. More population studies that continue to go beyond single markers and into whole genomes will elucidate features of their population structure and how metabolic diversity plays a role in the differentiation of populations within species.

5 The Halophilic *Halorubrum* spp.

Halophilic archaea can be isolated on complex media from medium to high-salinity environments, ranging from the Dead Sea, to salt flats, to even Antarctica (Anderson et al. 2016; Mormile et al. 2003; Mullakhanbhai and Larsen 1975). A recent study in the halophilic archaeal genus *Halorubrum* sequenced the genomes of 17 isolates from a single saline lake in Aran-Bidgol, Iran (Fullmer et al. 2014). Genera such as *Halorubrum* and other halophilic archaea are commonly found in saline lakes based on both isolates and metagenomic data (Naghoni et al. 2017). Previous work performed on isolates of *Halorubrum* from a set of saline ponds in Spain had identified high levels of recombination using MLSA (Papke et al. 2004). In Fullmer et al. (2014), the authors identified four major subpopulations within their isolate

set based on maximum-likelihood gene trees. Two of these subpopulations were considered cohesive groups based on high bootstrap support in the individual and concatenated gene analyses, average nucleotide identity (ANI), and tetramer frequency analysis. These groups differed in their presence and absence of CRISPR systems as well as intein elements, protein segments that excise themselves from the larger protein in which they are contained. Given the frequent observation of transmission of inteins within recombining populations, intein divergence between the subpopulations provided evidence that recombination barriers existed between the phylogroups, laying the groundwork for future potential studies of speciation between these groups.

As with *Sulfolobus*, the feature of recombination in asexually reproducing organisms can help to reveal patterns in the population structure of the organisms, such as barriers to recombination and migration. The novel feature of intein sequences in *Halorubrum* further allowed for understanding how genetic elements outside of plasmids and viruses can play a role in population structure as well.

6 Metagenomics and Single-Cell Genomics: Interrogating the Uncultured Majority

Given that most of microbial diversity is not accessible to culture approaches, this precludes the ability to easily interrogate individual microbes from a population. The rise in popularity of metagenomics, combined with deeper sequencing, longer reads, and computational pipeline advances, has led to the ability to assemble individual genomes from larger community read datasets. There are inherent problems with this method. While SNPs can be confidently identified throughout genomic assemblies, it is not possible to assemble true individuals. SNP frequencies can be interrogated at genomic positions, but they cannot be confidently linked to one another beyond the length of an individual read pair. Despite these limitations, assembling individuals from larger metagenomic datasets has revealed information about populations of uncultured archaea in the natural environment.

7 The Acidophilic *Ferroplasma*

Even before the shift from MLST to whole-genome sequencing for isolates taken from archaeal populations, the first archaeal population was being studied at a whole-genome level using high-throughput sequence data, but without laboratory isolation of the organisms. Shotgun metagenomic data from a pink biofilm in acid mine drainage near Redding, California, was assembled, leading to scaffolds that corresponded to two different "types" of the archaeal genus *Ferroplasma* (Eppley et al. 2007). By analyzing individual read pairs for patterns of variation that

distinguished between the two types, the authors were able to identify strain-level variants within the two *Ferroplasma* populations. Using the information from these strain-level variant, the authors inferred recombination events among strains within an individual population, and even among strains across the two populations. They also observed a decrease in calculated recombination rates as sequence divergence between strains increased, a pattern commonly found in bacterial and eukaryotic systems (Datta et al. 1997; Fraser et al. 2007).

The uniqueness of the acid mine drainage site lies in its high acidity (pH = 0.7), which enables the oxidation of ferrous iron as an energy source for *Ferroplasma* (Golyshina et al. 2000). Again, recombination in archaeal populations is a major feature underlying population structure, as would be confirmed later using genomes from other archaeal species. The low diversity of the acid mine drainage microbial community also facilitated the assembly of these populations, as more diverse environments can make assembling individuals from populations far more challenging.

8 Assembling Genomes from Massive Metagenomic Datasets

Especially when species are not in high abundance within a microbial community, populations are difficult to interrogate from assembling genomes and mapping reads. Progress is being made toward assembling genomes from massive metagenomic datasets, which can in the future facilitate more targeted population-level read mapping for population genomics. A recent assembly effort using 1,550 metagenomic datasets assembled 43 TACK-group archaea, 41 DPANN-group archaea, and 538 Euryarchaeota (Parks et al. 2017). Metagenomic approaches in the human gut microbiome have further identified features of uncultured methanogenic archaea, possibly utilizing trimethylamine produced within the gut (Borrel et al. 2017). Metagenomics has also been able to identify archaeal viruses in DNA samples, and possibly even identify new ones (Bolduc et al. 2012; Gudbergsdóttir et al. 2016).

9 Sequencing Single-Cell Archaeal Genomes from Natural Environments

Single-cell genomics has not yet fully matured to the point of probing many members of archaeal populations. Single-cell amplification and sequencing of 32 individual cells of the uncultured archaeal lineage MSBL1 from several Brine Pools in the Red Sea yielded the ability to construct a likely metabolic network for these yet uncultured organisms, but did not analyze any population-level

metrics among the sequenced strains (Mwirichia et al. 2016). Other approaches have combined the use of single-celled amplified genomes (SAGs) with metagenomics, to pull out metagenomic reads corresponding to a particular lineage. This was performed for a population of the NAG1 archaeal lineage inhabiting the Great Boiling Spring in Nevada (Becraft et al. 2017; Rinke et al. 2013). Again, the analysis was capable of constructing likely metabolic networks for uncultured organisms for which metabolic reconstruction would be impossible without genomic or culture data, but no analyses comparing individual members of the populations were performed.

10 Future Perspectives

Population genomics has yielded valuable insight into the population structures of various archaeal organisms. Studying the genomes of closely related cultured isolates has allowed for the identification and study of speciation events, barriers to recombination between populations, and phenotypic differentiation between populations. Future work with cultured isolates will be invaluable for directly correlating phenotypic variation observed in the laboratory with inferences made from genomic data. Currently, cultured isolates are still the most efficient way to interrogate many individuals within a population while maintaining linkage among sites and the potential to interrogate other features of individual strains. While metagenomic sequencing has been in favor for directly interrogating microbial communities for more than a decade, technological advances are only recently allowing for the interrogation of individual strains within larger metagenomic datasets. Single-cell genomics is continuing to increase in popularity and make technological advances; however, population genomics for archaea at this scale has not yet been performed. Combining single-cell draft genomes with metagenomics may be a strong intermediate between the two, especially for the uncultured majority.

The archaea harbor many novelties that may influence various aspects of their population biology. Multiple origins of replication may underlie larger chromosomal architectures that influence how different parts of the genome evolve. The extreme environments in which many archaea can be found may also underlie aspects of population structure, such as limited migration. While many of the archaea described here originate from environments with extreme temperatures, pH, or salinity, mesophilic archaea can also be found in non-extreme habitats (Könneke et al. 2005). Future work should aim to highlight the novelty of archaea in understanding what factors shape archaeal populations, including the role of recombination and recombination barriers in speciation, how ecological or phenotypic differentiation arises and is maintained between lineages, and how genomes evolve in regions and as a whole.

References

Ajon M, Fröls S, van Wolferen M, Stoecker K, Teichmann D, Driessen AJM, Grogan DW, Albers S-V, Schleper C. UV-inducible DNA exchange in hyperthermophilic archaea mediated by type IV pili. Mol Microbiol. 2011;82:807–17.

Anderson IJ, DasSarma P, Lucas S, Copeland A, Lapidus A, Del Rio TG, Tice H, Dalin E, Bruce DC, Goodwin L, et al. Complete genome sequence of the Antarctic Halorubrum lacusprofundi type strain ACAM 34. Stand Genomic Sci. 2016;11:70.

Becraft ED, Dodsworth JA, Murugapiran SK, Thomas SC, Ohlsson JI, Stepanauskas R, Hedlund BP, Swingley WD. Genomic comparison of two family-level groups of the uncultivated NAG1 archaeal lineage from chemically and geographically disparate hot springs. Front Microbiol. 2017;8:2082.

Bolduc B, Shaughnessy DP, Wolf YI, Koonin EV, Roberto FF, Young M. Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs. J Virol. 2012;86:5562–73.

Borrel G, McCann A, Deane J, Neto MC, Lynch DB, Brugère J-F, O'Toole PW. Genomics and metagenomics of trimethylamine-utilizing archaea in the human gut microbiome. ISME J. 2017;11:2059–74.

Brock TD, Brock KM, Belly RT, Weiss RL. Sulfolobus: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature. Arch Mikrobiol. 1972;84:54–68.

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, et al. Complete genome sequence of the Methanogenic archaeon, Methanococcus jannaschii. Science. 1996;273:1058–73.

Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ, Whitaker RJ. Patterns of gene flow define species of thermophilic archaea. PLoS Biol. 2012;10:e1001265.

Datta A, Hendrix M, Lipsitch M, Jinks-Robertson S. Dual roles for DNA sequence identity and the mismatch repair system in the regulation of mitotic crossing-over in yeast. Proc Natl Acad Sci U S A. 1997;94:9757–62.

Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R, Henne A, Wiezer A, Bäumer S, Jacobi C, et al. The genome of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea. J Mol Microbiol Biotechnol. 2002;4:453–61.

Didelot X, Lawson D, Darling A, Falush D. Inference of homologous recombination in Bacteria using whole-genome sequences. Genetics. 2010;186:1435–49.

Eppley JM, Tyson GW, Getz WM, Banfield JF. Genetic exchange across a species boundary in the archaeal genus ferroplasma. Genetics. 2007;177:407–16.

Flynn KM, Vohr SH, Hatcher PJ, Cooper VS. Evolutionary rates and gene dispensability associate with replication timing in the archaeon Sulfolobus islandicus. Genome Biol Evol. 2010;2:859–69.

Fraser C, Hanage WP, Spratt BG. Recombination and the nature of bacterial speciation. Science. 2007;315:476–80.

Fullmer MS, Soucy SM, Swithers KS, Makkay AM, Wheeler R, Ventosa A, Gogarten JP, Papke RT. Population and genomic analysis of the genus Halorubrum. Front Microbiol. 2014;5:140.

Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016;17:175–88.

Gilbert JA, Dupont CL. Microbial metagenomics: beyond the genome. Annu Rev Mar Sci. 2011;3:347–71.

Golyshina OV, Pivovarova TA, Karavaiko GI, Kondratéva TF, Moore ER, Abraham WR, Lünsdorf H, Timmis KN, Yakimov MM, Golyshin PN. Ferroplasma acidiphilum gen. nov., sp. nov., an acidophilic, autotrophic, ferrous-iron-oxidizing, cell-wall-lacking, mesophilic member of the Ferroplasmaceae fam. nov., comprising a distinct lineage of the archaea. Int J Syst Evol Microbiol. 2000;50(Pt 3):997–1006.

Gudbergsdóttir SR, Menzel P, Krogh A, Young M, Peng X. Novel viral genomes identified from six metagenomes reveal wide distribution of archaeal viruses and high viral diversity in terrestrial hot springs. Environ Microbiol. 2016;18:863–74.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013;41:e1.

Könneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA. Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature. 2005;437:543–6.

Krause DJ, Didelot X, Cadillo-Quiroz H, Whitaker RJ. Recombination shapes genome architecture in an organism from the archaeal domain. Genome Biol Evol. 2014;6:170–8.

Lundgren M, Andersson A, Chen L, Nilsson P, Bernander R. Three replication origins in Sulfolobus species: synchronous initiation of chromosome replication and asynchronous termination. Proc Natl Acad Sci U S A. 2004;101:7046–51.

Maiden MCJ. Multilocus sequence typing of bacteria. Annu Rev Microbiol. 2006;60:561–88.

Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998;95:3140–5.

Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA, et al. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci U S A. 2007;104:11889–94.

Mau B, Glasner JD, Darling AE, Perna NT. Genome-wide detection and analysis of homologous recombination among sequenced strains of Escherichia coli. Genome Biol. 2006;7:R44.

Mormile MR, Biesen MA, Gutierrez MC, Ventosa A, Pavlovich JB, Onstott TC, Fredrickson JK. Isolation of Halobacterium salinarum retrieved directly from halite brine inclusions. Environ Microbiol. 2003;5:1094–102.

Mullakhanbhai MF, Larsen H. Halobacterium volcanii spec. nov., a Dead Sea halobacterium with a moderate salt requirement. Arch Microbiol. 1975;104:207–14.

Mwirichia R, Alam I, Rashid M, Vinu M, Ba-Alawi W, Anthony Kamau A, Kamanda Ngugi D, Göker M, Klenk H-P, Bajic V, et al. Metabolic traits of an uncultured archaeal lineage – MSBL1 – from brine pools of the Red Sea. Sci Rep. 2016;6:19181.

Naghoni A, Emtiazi G, Amoozegar MA, Cretoiu MS, Stal LJ, Etemadifar Z, Fazeli SAS, Bolhuis H. Microbial diversity in the hypersaline Lake Meyghan, Iran. Sci Rep. 2017;7:11522.

Papke RT, Koenig JE, Rodríguez-Valera F, Doolittle WF. Frequent recombination in a saltern population of Halorubrum. Science. 2004;306:1928–9.

Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1533–42.

Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ. Biogeography of the Sulfolobus islandicus pan-genome. Proc Natl Acad Sci U S A. 2009;106:8605–10.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499:431–7.

Robinson NP, Dionne I, Lundgren M, Marsh VL, Bernander R, Bell SD. Identification of two origins of replication in the single chromosome of the archaeon Sulfolobus solfataricus. Cell. 2004;116:25–38.

Schleper C, Holz I, Janekovic D, Murphy J, Zillig W. A multicopy plasmid of the extremely thermophilic archaeon Sulfolobus effects its transfer to recipients by mating. J Bacteriol. 1995;177:4417–26.

Schloss PD, Girard RA, Martin T, Edwards J, Thrash JC. Status of the archaeal and bacterial census: an update. MBio. 2016;7:e00201–16.

Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 2014;22:235–47.
Stedman KM, She Q, Phan H, Holz I, Singh H, Prangishvili D, Garrett R, Zillig W. pING family of conjugative plasmids from the extremely thermophilic archaeon Sulfolobus islandicus: insights into recombination and conjugation in crenarchaeota. J Bacteriol. 2000;182:7014–20.
Torreblanca M, Rodriguez-Valera F, Juez G, Ventosa A, Kamekura M, Kates M. Classification of non-alkaliphilic halobacteria based on numerical taxonomy and polar lipid composition, and description of Haloarcula gen. nov. and Haloferax gen. nov. Syst Appl Microbiol. 1986;8:89–99.
Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data. 2018;5:170203.
van Wolferen M, Wagner A, van der Does C, Albers S-V. The archaeal Ced system imports DNA. Proc Natl Acad Sci. 2016;113:2496–501.
Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004;304:66–74.
Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 2009;3:199–208.
Whitaker RJ, Grogan DW, Taylor JW. Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science. 2003;301:976–8.
Whitaker RJ, Grogan DW, Taylor JW. Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus. Mol Biol Evol. 2005;22:2354–61.
Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A. 1977;74:5088–90.
Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. Proc Natl Acad Sci U S A. 1990;87:4576–9.
Wolfe RS. Techniques for cultivating methanogens. Methods Enzymol. 2011;494:1–22. https://doi.org/10.1016/B978-0-12-385112-3.00001-9.
Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H, Saw JH, Senin P, Yang C, Chatterji S, et al. Assembling the marine metagenome, one cell at a time. PLoS One. 2009;4:e5299.
Wu Z, Liu J, Yang H, Xiang H. DNA replication origins in archaea. Front Microbiol. 2014;5:179.
Youngblut ND, Wirth JS, Henriksen JR, Smith M, Simon H, Metcalf WW, Whitaker RJ. Genomic and phenotypic differentiation among Methanosarcina mazei populations from Columbia River sediment. ISME J. 2015;9:2191–205.
Zhang C, Krause DJ, Whitaker RJ. Sulfolobus islandicus: a model system for evolutionary genomics. Biochem Soc Trans. 2013;41:458–62.

Part III
Population Genomics of Fungi

Advances in Genomics of Human Fungal Pathogens

Daniel Raymond Kollath, Marcus de Melo Teixeira, and Bridget Marie Barker

Abstract Fungi are responsible for 1.5 million deaths every year, and one-third of the human population has experienced a fungal infection. The increasing numbers of immunocompromised people are associated with the increased incidence of human mycosis, either from medical interventions such as cancer therapy or transplantation or due to other underlying diseases such as HIV/AIDS or diabetes. Additionally, climate change has been implicated in widening distributions of endemic fungi, potentially expanding beyond previously restricted ranges. In this chapter, we will address two main classes of fungal pathogens: first, the globally distributed fungi such as *Candida*, *Aspergillus*, and *Cryptococcus*, followed by a discussion of endemic fungal pathogens and their relatives *Paracoccidioides*, *Histoplasma*, *Coccidioides*, and *Emmonsia*. In the past, virulence and pathogenesis studies were limited to few infection models and biomarkers, but these studies have progressed significantly with the advances of DNA sequencing and genetic tools. Newly sequenced structural (DNA) and functional (RNA and protein) genomes provide a scaffold to understand gene gain and loss that might be associated with mammalian infection and disease progression. During infection, these pathogens express a wide range of genes that are associated with either establishment of infection or escaping recognition by host immune cells. Moreover, population genomic studies reveal that pathogen complexes exhibit different strategies to generate genetic diversity either via sexual or parasexual recombination, and this phenomenon may be implicated in altered virulence, disease presentation, and antifungal resistance. The literature of genomic studies of the abovementioned pathogenic fungal genera are summarized, and molecular taxonomy and population structure are explored, as well as a survey of the main genomic characteristics, chromosomal variation, gene content, and expression. Comparative genomics between pathogenic and nonpathogenic close-related species provides evidence of both convergent and unique adaptation of those fungal lineages to mammalian hosts. A better understanding of patterns of gene flow

D. R. Kollath · M. de Melo Teixeira · B. M. Barker (✉)
The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA
e-mail: drk87@nau.edu; marcus.teixeira@unb.br; bridget.barker@nau.edu

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_41,

among species (hybridization), adaptation and evolutionary potential, fully closed reference genomes, and general improvement of gene annotation and function is needed.

Keywords Comparative genomics · Fungal infection · Fungal pathogen genomics · Population genetics of fungi

1 Introduction

The fungal kingdom is a diverse group of eukaryotes, harboring several million species, which likely originated between 500 and 1,000 mybp (Hawksworth and Lucking 2017). Although fungi are predominately saprobic, a number of fungi are symbiotic, parasitic, or predatory species that feed on living organisms and sometimes kill the host (Taylor 2014). Worldwide, fungi kill more people each year than malaria, tuberculosis, or breast cancer, and one-third of the human population has experienced a fungal infection (Brown et al. 2012; Bongomin et al. 2017). The frequency of invasive fungal infections is increasing and is often associated with climate change, increased numbers of immunocompromised people who are undergoing cancer therapy, transplantation, as well as those with underlying diseases, like granulomatous disease or HIV/AIDS (Park et al. 2005; Oladele et al. 2018; Chastain et al. 2017; Brown et al. 2012). Moreover, there are several fungal pathogens that infect otherwise healthy people, and the immune condition of the host is not always directly linked to disease pathology (Kohler et al. 2017). As outlined in several of the papers cited above, the number of invasive fungal infections in the human population has dramatically increased, and the fungi responsible cover a broad range of taxa. *Candida* ssp. (*Saccharomycetes*), *Aspergillus* ssp. (*Eurotiales*), and *Cryptococcus neoformans* (*Tremellales*) represent fungal infections where the impaired immune status of an individual is directly associated with the establishment of the disease. Additionally, certain pathogens affect both immunocompetent and immunocompromised people, and the immune condition of the host is not necessarily associated with the infection and disease progression. These infections are caused by fungal pathogens such as *Cryptococcus gattii* (*Tremellales*), *Histoplasma capsulatum*, *Blastomyces* ssp., *Paracoccidioides* spp., *Coccidioides* spp. (*Onygenales*), *Talaromyces marneffei* (former *Penicillium* – *Eurotiales*), *Sporothrix* sp. (*Ophiostomatales*), and black yeast-like fungi (*Chaetothyriales*) which are responsible for thousands of severe infections, especially in tropical and subtropical countries around the globe (Queiroz-Telles et al. 2017).

Fungal infections represent a major challenge to public health, as fungi are ubiquitous in the environment and rarely transmitted from human to human, or transmitted by a vector, making exposure impossible to completely avoid (Barberan et al. 2015). Furthermore, the distribution, exposure rates, and genetic basis of either virulence or host escape mechanisms of these pathogens represent a major knowledge gap in the medical mycology field. Thus, fungal epidemics and outbreaks are challenging to predict and prevent, and the consequences are severe if not correctly

addressed. The human fungal pathogens are polyphyletically distributed across different phyla in the fungal kingdom, and different evolutionary trajectories and adaptive strategies have led the emergence of different mechanisms of virulence and adaptation to living hosts across these lineages (Fisher et al. 2012).

It is estimated that five million different species of fungi exist in the world, and new fungi are discovered frequently (Dukik et al. 2017; Hawksworth and Lucking 2017). How do we determine if a novel organism is truly a new species? A common definition for a microbial species is the evolutionary species concept (ESC), which is summarized as a single lineage of ancestor-descendent populations which maintain distinct identity from other such lineages and has its own evolutionary tendencies and historical fate (Wiley 1978). In addition, morphological (MSC), biological (BSC), and phylogenetic (PSC) species concepts all have specific criteria that have been used to describe species boundaries in fungi (Hawksworth 2006; Taylor et al. 2000). The hallmark publication of Hawksworth describes the identification of 70,000 fungal species using morphological or phenotypic characters (Hawksworth and Rossman 1997). The taxonomical classification of human fungal pathogens was traditionally based on those morphological characters that are responsible for the formation of the sexual structures (i.e., fruiting bodies) and asexual structures (i.e., conidia/conidiophores and phialides) or by characterizing the pathogenic phase derived from animal hosts (i.e., yeast-like forms) often resulting in incorrect taxonomic assignments (Bowman et al. 1992b).

The main challenges of defining pathogenic fungal species by morphological characters or ability to form viable progeny are (1) not all fungal pathogens are cultivable in artificial media; (2) fungal pathogens exhibit a high polyphyletic morphological profile within species complexes or even between different genera; (3) some fungal pathogens may take up to 6 weeks to be precisely characterized due to slow growth rate under laboratory conditions; (4) not all fungal pathogens produce fruiting bodies, or induction of these structures under laboratory conditions are laborious and time-consuming; and (5) some fungi are homothallic (self-fertile), and in this case the meiospores will not display a detectable inheritance pattern. In these fungi, the presence of meiospores is not sufficient to determine if mating takes place, necessitating genetic markers to ensure that the progeny has two distinct parents (Taylor et al. 1999, 2000, 2006; Taylor and Ellison 2010; Bowman et al. 1992a). With the advances in molecular biology tools, the typing methods have significantly increased in precision of defining species limits as well as the detection of fungal pathogens. Prior to sequencing-based methods, molecular techniques, such as restriction fragment length polymorphisms (Botstein et al. 1980), random amplified polymorphic DNAs (Williams et al. 1990), amplified fragment length polymorphisms (Vos et al. 1995), and simple sequence repeats (Tautz 1989), have been used to determine species. Limitations include cost, rapidity, technical proficiency required, the use of radioactive materials, and confidence in the frequency of specific polymorphisms within species and populations.

In fungi, direct analysis of nucleotide sequence is most consistent with the evolutionary species concept. Molecular phylogenetics has become a common approach to define species, because changes in nucleotide sequences and allele

frequencies at different loci can be detected prior to changes in morphology or mating patterns (Taylor 2014). The hallmark publication of Taylor, Bruns, Lee, and White proposed the idea of fungal barcoding (White et al. 1990). Universal PCR primers targeting the ribosomal DNA of fungi were developed, which took advantage of emerging Sanger DNA sequencing methodology, and became a powerful tool to discriminate fungal species. Additional advances in Sanger sequencing and cloning methods led to systematic functional genomic characterization of conserved genes in fungi and eventually the first fungal genomes. The baker's yeast *Saccharomyces cerevisiae* was the first eukaryotic genome to be sequenced and paved the way not only for fungal pathogen genomics, such as *Candida albicans* and *Aspergillus fumigatus*, but genomics as a whole (Goffeau et al. 1996; Jones et al. 2004; Nierman et al. 2005).

The abundance of Sanger-derived DNA sequence data from fungi allowed researchers to develop targeted sequencing and analysis of homologous genes among fungal pathogens. In light of this, the genealogical concordance for phylogenetic species recognition (GCPSR) was proposed as a method for species assignment among fungi and is widely used in medical mycology (Taylor et al. 2000). This technique offers many advantages over the BSC and MSC, particularly in morphologically homogeneous fungi or when there is insufficient knowledge of sexual reproduction. According to the GCPSR method, each locus will produce slightly different gene genealogies within the same species due to the existence of recombination. However, when comparing two different species, the genealogies for the different loci will be concordant due to accumulation of genetic differences due to genetic drift or selection. It is expected that within species there may be a conflict between the branches generated due to recombination and gene flow. The detection of common branches shared among different gene trees is the key to identification of phylogenetic species.

Many fungal pathogens exist as either haploid or diploid mycelia and are propagated by mitotic (clonal) divisions. In a clonal population, a single genotype is maintained in a population, reducing diversity and maintaining an adaptive phenotype (Taylor et al. 2015). However, environmental stress, exposure to the host, or antimicrobial drug exposure may induce some fungal pathogens to undergo sexual or parasexual (the parasexual cycle involves plasmogamy, karyogamy, and meiosis taking place at non-specified times in the fungal life cycle) reproduction to generate recombinant progeny and genetic diversity that might adapt to the novel environment (Heitman 2010). Recombination has been identified by evolutionary analysis in natural populations of human fungal pathogens, where divergent clades emerge due to natural selection and speciation, which may lead to phenotypic changes among these species, such as disease variation, adaptation strategies, virulence, and pathogenicity, among others. The consequences of sexual and parasexual recombination include gene duplications and losses, mutations, chromosomal rearrangements, and loss of heterozygosity, which, coupled with horizontal gene transfer and de novo gene formation, are the main forces of phenotypic variation in fungi.

The sexual cycle in fungi can occur via a heterothallic system (self-sterile) or a homothallic system (self-fertile) mechanism. Among heterothallic species, mating occurs between two sexually compatible individuals, which are morphologically identical but are genetically determined by the mating-type (*MAT*) locus. Homothallic species contain two complementary sexual loci in the same haploid genome, and thus each individual can self-fertilize (Ni et al. 2011). The gene content and order of the genes in the *MAT* locus vary according to each phylum. In *Ascomycetes*, the *MAT1-1* locus encodes for an α-box gene, while the *MAT1-2* locus encodes for an HMG-type gene, and the locus usually spans 10 kb. In basidiomycete yeasts, the MAT locus (a or α) is composed of homeodomain genes (e.g., SXI1α and SXI2a in *Cryptococcus* ssp.) and spans 100 kb (Yan et al. 2007). These transcription factors regulate the downstream expression of pheromones and receptors in the MAPK mating signaling pathway, pheromone-forming enzymes, and other transcription factors. Same-sex mating has also been suggested in both *Candida* and *Cryptococcus* species, and this may generate genetic and phenotypic plasticity (Hirakawa et al. 2017; Fu et al. 2015).

The ability to sequence entire genomes provides insight into biological and metabolic diversity within the fungal kingdom and has advanced research in medical science, agricultural science, bioremediation, biotechnology, ecology, and many other disciplines (Sharma 2016). As more genomes are sequenced and made available for public reference, we gain a greater understanding of the biological diversity within the fungal kingdom. The first fungal genome to be entirely sequenced (*Saccharomyces cerevisiae*) provided insight into the evolution and natural history of yeasts. It also served as a model organism for eukaryotic biology, giving rise to novel advancement in cancer biology and genetics. Fungi also serve as important model organisms for biomedical and infectious disease research. The first full genome sequence of a fungal pathogen was *Candida albicans*, as well as one of the first eukaryotic pathogens (Jones et al. 2004). Sequencing genomes of pathogens provides information about the evolution of pathogenicity and virulence factors as well as potential targets for treatment (Sharma 2016). *Aspergillus fumigatus* is one of the most prevalent causes of mycoses worldwide, but the basic biology of the organism was not understood until comparative genomics was carried out. These analyses revealed species-specific horizontally acquired genes that lead to extremely rapid environmental adaptations (Fedorova et al. 2008). Genomic analysis was used to help understand the recent emergence of *Cryptococcus gattii* in the Pacific Northwest of North America. This geographical expansion of a previously described endemic tropical and subtropical genotypes revealed that naturally acquired genomic adaptations lead to the emergence of the pathogen in this region as well as a recent microevolution caused by different selection pressures in the new environment (Engelthaler et al. 2014). When doing comparative genomic analysis on the *C. albicans* genome, allelic differences were identified that lead to resistance to certain antifungals. These examples show how genomic sequencing of fungal pathogens is an important tool to understand factors that may have led to the emergence or outbreak of disease as well as acquired drug resistance (Cuomo 2017).

In recent years many fungal genomic databases were created and made publicly available for reference. The *Saccharomyces* database allows non-bioinformaticians to access the full genome of strain S288C and provides extensive information on mapping, sequence information, protein domains, expression data, and much more (Sharma 2016). There are also databases that are dedicated to a specific genera of fungi such as candidagenome.org that provides functional information about genes and proteins of *Candida* spp. or aspergillus-genomes.org.uk, which offers web-based tools to analyze genomic features. FungiDB (fungidb.org) is a database that provides an interface that allows researchers to compare genomic data of multiple species of fungi (Basenko et al. 2018). These databases allow researchers with diverse skill sets to access a huge quantity of genomic data from a wide range of species almost immediately. The growth of these existing databases and the creation of new ones can increase the quality and speed for new research in the medical mycology field (Cuomo 2017).

The following sections will look at the genomics and biology of several fungal pathogens but are certainly not exhaustive. The use of population, evolutionary, and functional genomics can give great insights into pathogenicity and other clinically relevant traits by examining molecular variation within the same populations and across different populations and species. We strive to summarize a large body of knowledge to guide the reader to recent developments in human fungal pathogen genomics and encourage the reader to further explore and engage with these developing stories.

2 Human Fungal Pathogens from the Genus *Candida*

Candida is polyphyletic genus of fungi nested within the *Saccharomycotina* subphylum of the *Ascomycota* and contains many pathogenic yeast species (Boekhout et al. 2009; Turner and Butler 2014). *Candida* species are the most prevalent cause of opportunistic fungal infections and are responsible for high rates of morbidity and mortality worldwide (Eggimann et al. 2003). Whole genome sequencing has led to a number of new findings, including novel codon usage distribution, recombination patterns, gene gain/loss, and antifungal resistance mechanisms.

Disturbances in the normal microbial flora, particularly in patients that are immunocompromised or have a severe primary disease, present an opportunity for *Candida* spp. to cause disease. Disease states range from superficial infections of skin and mucosal tissues (e.g., oral thrush, vaginal yeast infections) to systemic infections that can affect a wide range of organs and tissues (Jackson et al. 2009). Most *Candida* species belong to a single clade that is characterized by the unique translation of CUG codons and harbor haploid (*C. lusitaniae*) and diploid species (*C. albicans*, *C. dubliniensis*, *C. tropicalis*, and *C. parapsilosis*). The *C. parapsilosis* complex is composed of at least two additional species revealed by multilocus analysis: *C. metapsilosis* and *C. orthopsilosis*. *C. glabrata*, *C. nivariensis*, and *C. bracarensis* are haploids and descendants from a common ancestor that

underwent a whole genome duplication process (WGD) (Butler et al. 2009; Gabaldon et al. 2013, 2016). Recently, the three *Candida* WGD species were placed into the *Nakaseomyces* genus, along with three environmental species: *Nakaseomyces delphensis*, *Nakaseomyces bacillisporus*, and *Nakaseomyces castellii* (Gabaldon et al. 2013).

Even though there is tremendous variation in phenotype and genome size among the *Candida* species, the total number of protein-coding genes is similar (Butler et al. 2009). Genome size varies broadly among the *Candida* species, ranging from 10.6 to 15.5 megabases, and haploid species have a smaller genome than the diploid species, and the overall frequency and distribution of single nucleotide polymorphisms (SNPs) vary between diploid species (Turner and Butler 2014; Butler et al. 2009; Jones et al. 2004). The subphylum *Saccharomycotina* consists of pathogenic (genus *Candida*) and nonpathogenic yeasts. Within the genus *Candida* are opportunistic pathogens that are assigned to this genus because they are pathogens but belong to distinct lineages that are comprised of both pathogens and free-living yeasts, suggesting that the ability to infect humans has evolved independently several different times within the subphylum *Saccharomycotina* (Gabaldon et al. 2016). Comparing *Candida* genomes reveals virulence factors and utilization of the CUG codon for serine instead of leucine, which appears to increase diversity of surface proteins to evade the host immune system by changing the recognition patterns of immune cells (Gabaldon et al. 2016; Miranda et al. 2013).

All members of the *Candida* genus, with the exception of *C. glabrata* and *C. krusei*, have a unique way of translating CUG codons. While CUG codons are normally translated to the amino acid leucine, the *Candida* clade translates CUG to serine (Butler et al. 2009). There have been many studies that attempted to understand the evolutionary advantage of this genetic code change. Using *Saccharomyces cerevisiae*, a close relative of *C. albicans*, there is evidence that selection drives a molecular mechanism that requires CUG to have ambiguity. Codon ambiguity usually shows decreased fitness, which means there has to be some positive evolutionary response to this negative impact. It has been shown that CUG ambiguity induces increased expression of a novel set of stress proteins that triggers the general stress response that gives the fungus a competitive edge in stressful conditions (Santos et al. 1999). This genetic code change could give *Candida* species an evolutionary advantage that could allow for the occupation of new ecological niches by outcompeting other organisms under stress conditions. This also may be an explanation as to why *Candida* species are exceptional opportunistic pathogens.

It is still unclear whether or not all *Candida* pathogens can undergo meiosis. *C. albicans*, for example, was considered a strict diploid asexual fungus. Pathogenic *Candida* species have a wide variety of strategies when it comes to sexual reproduction (Bennett 2010). There seem to be differences between haploid species and diploid species, as well as differences within the ploidy groups. Diploid species have members that reproduce sexually via parasexual cycles, where there is a mating of compatible diploid cells followed by mitosis and chromosome loss rather than

meiosis, homothallism (self-mating), and no observed outcrossing (Bennett 2010). Haploid species have been observed to complete sexual cycles via heterothallic mating (outcross mating) and homothallic mating cycles (Butler 2010; Butler et al. 2009). Idiomorphs at the mating-type locus (MTL) seem to determine mating type in *Candida* species (Butler et al. 2009). There was an effort to find orthologs of meiotic genes that have been observed in *S. cerevisiae* as well as in *Candida* pathogens. Genetic analysis and mating crossings showed that sexual reproduction in *C. albicans* occurs between two different and compatible mating-type cells (Butler et al. 2009). Moreover, the ability to undergo the sexual cycle is dependent on a phenotypic switch from white asexual stage to the opaque mating-efficient stage that is controlled by the transcription factor Wor1 (Cain et al. 2012). Each of the a and α diploid counterparts carries a single MTL, and mating occurs in vitro using specific media conditions or by using in vivo mammalian models to generate a/α tetraploid progeny. Likewise, *C. tropicalis* also undergoes a cryptic sexual cycle that is also dependent on the white to opaque switch (Porman et al. 2011). Recently, *C. albicans* was shown to regulate mating not only by the MTL but also by epigenetic phenotypic shifts (Bennett 2015). Cells switch from a white (round) state to an opaque (elongated) state with the help of the transcription factor Wor1. Sexual mating between opaque cells was shown to be a million times more efficient than mating between white cells (Bennett 2015). This suggests that environment plays an important role in the expression of a sexual versus asexual cycle and that mating is linked with traits associated with pathogenicity and drug resistance (Hirakawa et al. 2017).

To date, for both *C. parapsilosis* complex and *C. glabrata*, no functional sexual mechanism has been found, despite discovery that these "non-mating" species have the MTL as well as genes involved in the pheromone response pathway, indicating that cryptic sex may exist (Butler et al. 2009). In addition, genomic analysis showed that IME1 meiotic regulators are absent in all *Candida* species and the DMCL-dependent pathway (involved in meiotic recombination) was absent in heterothallic species (Butler 2010). Surprisingly, non-mating species showed conserved pheromone response pathways indicating that these meiotic pathways are used for other purposes such as virulence factors (Butler et al. 2009). According to population genomic analysis, recombination due to sexual reproduction is low in *Candida* species compared to other fungi, but there may be other novel mechanism to maintain genetic variation and adaptation mechanisms of these pathogens (Salazar et al. 2018; Holland et al. 2014; Turner and Butler 2014). Genetic studies suggest that mitotic recombination followed by loss-of-heterozygosity events is one main source of phenotypic variation in *C. albicans* (Gomez-Raja et al. 2008). Chromosome break-induced replication, loss, and segmental deletions can result in the expression of recessive genes and reveal adaptive alleles important for virulence, mating competence (Magee and Magee 2000), auxotrophy (Gomez-Raja et al. 2008), and antifungal drug resistance (Coste et al. 2007).

Twenty-one gene families are significantly enriched among pathogenic *Candida*, including genes that encode for lipases, oligopeptide transporters, and adhesins (Butler et al. 2009). These gene families are all known to be associated with

pathogenicity and virulence, especially adhesins that facilitate adherence of the pathogen to host cells or surfaces. Hwp1, ala1p, als5p, als1p, and epa1p are proteins in the class glycosylphosphatidylinositol-dependent cell wall proteins (GPI-CWP), which play a role in adherence to host endothelial and epithelial cells (Ariyachet et al. 2013; Liu and Filler 2011; Zhu and Filler 2010; Almeida et al. 2008). The *Candida* clade also harbors more gene families associated with extracellular enzymes and transmembrane transporter proteins. Cell-surface transporters (oligopeptide transporters and amino acid permeases) are also enriched in those pathogens, providing more evidence of the importance of extracellular activities for pathogens to be successful (Butler et al. 2009). Additionally, gene families enriched for cell wall, hyphal, pseudohyphal, and biofilm growth have been identified in pathogenic *Candida* species.

Many genes specific to *C. albicans* are associated with virulence, which may provide an explanation as to why this member of the *Candida* clade is such a successful pathogen. To investigate this, the genomes of *C. albicans* and *C. dubliniensis* (a less virulent and closely related fungus) were compared (Jackson et al. 2009). The genomes of the two species are virtually colinear with a few exceptions. The secreted aspartic proteinase (SAP) gene family controls hydrolytic responses of host tissues and is a crucial virulence factor for *C. albicans* (Naglik et al. 2003a). These hydrolytic enzymes destroy host cell membranes, invade cells, and degrade host immune defense molecules, such as lactoferrin, secretory IgA, and macrophage proteins, to resist antimicrobial activity (Zakikhany et al. 2007). There are four SAP loci on chromosome 6 of *C. albicans*; however, *C. dubliniensis* is missing SAP4 and SAP5 (Jackson et al. 2009). This suggests an inversion, insertion-deletion, or transposition event has led to *C. dubliniensis*-specific differences from *C. albicans.* In addition, genes that encode for common fungal virulence-associated cell wall proteins (e.g., Hyr/Iff proteins and Als adhesins) are found in *C. albicans* genomes. Als adhesins are associated with adhesion to host surfaces, acquiring iron from host, and invasion of host cells. These Als gene families are enriched in *C. albicans* but are absent in nonpathogenic *Saccharomyces* fungi (Sheppard et al. 2004; Loza et al. 2004; Maguire et al. 2013). Patterns of gene enrichment suggest that *C. albicans* has specifically evolved and adapted to occupy the host niche.

Candida glabrata is another important human pathogen in the *Candida* clade. This pathogen accounts for a high number of mucosal systemic infections and is the second most common yeast human pathogen behind *C. albicans* (Rodrigues et al. 2014). Both fungi show significant similarities in their genomic content for virulence factors, but there are also important differences. Whereas *C. albicans* primarily employs hyphal formation and proteinase secretion as virulence factors, *C. glabrata* relies on the production of lectins (Kaur et al. 2005). The genome shows a loss of genes involved in nitrogen, phosphate, and sulfur metabolism and simple sugar utilization, and this gene loss may be a result of a close association and evolution with mammalian hosts.

Although *C. albicans* utilizes secreted proteinases for infection, *C. glabrata* does not produce a significant amount of extracellular proteinases (Turner and Butler 2014). The two yeasts both produce phospholipase B (PLB) enzymes that seem to be

related to virulence (Naglik et al. 2003b). In *C. albicans* this is associated with gastrointestinal tract colonization. PLB protein genes were found to be highly expressed during *C. glabrata* vaginal infections, indicating that they play a role in the metabolism of phospholipids and cause damage to the vaginal epithelium. As compared to hyphal formation in *C. albicans*, *C. glabrata* initiates a signaling cascade for the formation of pseudohyphae via mitogen-activated protein kinases under depleted nitrogen conditions (Kaur et al. 2005). The transcription factor ste12 is required for pseudohyphae formation in low nitrogen conditions and is highly conserved among fungi. The deletion of *ste12* in *C. glabrata* results in lower pathogenicity, which suggests that pseudohyphal formation plays a role in infection, but the mechanism is not yet understood (Calcagno et al. 2003). Adherence to host epithelial cells is controlled by different mechanisms in *C. glabrata* as compared to *C. albicans.* HWP1 and genes in ALS family in *C. albicans* encode adhesins, whereas in *C. glabrata*, adherence is controlled by GPI-anchored cell wall proteins in the epithelial adhesions (EPA) gene family (Filler 2006). Notably, EPA1 is a lectin that binds to host glycoproteins. The *C. glabrata* genome contains several EPA-related genes, and deletion of EPA1 drastically reduces adherence in vitro, which implies that this is an important virulence factor (Salazar et al. 2018; Cormack et al. 1999).

The frequency of resistance to antifungal drugs within the pathogenic *Candida* species has increased and is responsible for increasing hospitalizations and deaths due to *Candida* infections (Arendrup and Patterson 2017). Azoles, echinocandins, polyenes, nucleoside analogs, and allylamines are all used to treat candidiasis, but the efficacy varies depending on the species resistance mechanism as well as its related pathology (Whaley et al. 2016). Azoles are the most used drugs against *Candida* infections and act by inhibiting the enzyme lanosterol 14a-demethylase (used in the biosynthesis of ergosterol) codified by the gene Erg11 (Flowers et al. 2015). Point mutations in Erg11 confer resistance to azoles for both *C. albicans* and *C. tropicalis* and are a common mechanism responsible for azole resistance. Moreover, the overexpression of Erg11 via mutation of the zinc-cluster transcriptional regulator Upc2p is another mechanism that confers azole resistance (Macpherson et al. 2005). This mechanism has been observed in other species such as *C. tropicalis* and *C. parapsilosis* species complex; however it is absent in *C. glabrata*. In addition, Erg3 is associated with antifungal resistance in *C. albicans*. The Erg3 enzyme is sterol Δ5,6 desaturase which catalyzes the final steps in the ergosterol biochemical pathway (Martel et al. 2010). The accumulation of 14α-methylergosta-8,24(28)-dien-3β,6α-diol is toxic for fungal cells, and the activation or deletion of the ERG3 gene, therefore, prevents such toxic compounds from being produced. Overexpression of efflux pumps, ERG1 and its regulator Upc1, and ERG6 has been shown to increase fluconazole resistance in *C. parapsilosis* (Silva et al. 2011).

Additionally, the overexpression of two other genes encoded for drug efflux pumps named Mdr1p and Cdr1p/Cdr2p is also an important mechanism for azole resistance in *C. albicans* (White 1997). Cdr1p/Cdr2 encodes for an ATP-binding cassette (ABC)-containing protein and is regulated by TAC1 (transcriptional activator of CDR genes) (Tsao et al. 2009). Genome-wide analyses have shown at least

nine overexpressed TAC1 alleles that were shown to be acquired by LOH mechanisms (Coste et al. 2007). Another important gene conversion due LOH is the MRR1 (multidrug resistance regulator 1) and was identified by comparing the transcriptomes of fluconazole-resistant isolates that overexpressed MDR1 (White 1997). MRR1 knockouts have decreased fluconazole MICs, while introduction of point mutations in this gene rescues the fluconazole-susceptible phenotype in the native strains of *C. albicans* (Silva et al. 2011). In *C. krusei*, population and comparative genomics revealed novel transporters that may play a role in drug resistance or adaptation to different environments in this pathogen (Whaley et al. 2016; Cuomo et al. 2017). This pathogen is heterozygous, and large genomic regions are affected by LOH and may be associated with antifungal resistance (Cuomo et al. 2017).

Finally, one pathogenic yeast in the *Candida* genus that must be mentioned due to its global and recent emergence as the cause of severe infections on three continents is *C. auris*. This species harbors multiple drug-resistant strains/populations and was first described in Japan in 2009. This invasive fungus with the ability to cause fungemia and infect tissues, such as inner ear canal, is a new threat to public health (Vallabhaneni et al. 2016). *C. auris* seems to be highly resistant to fluconazole, other azoles, and amphotericin B (Lockhart et al. 2017). This azole resistance may come from orthologs of the *C. albicans* genes ERG11 and SC5314 previously implicated in azole resistance in other *Candida* species (Whaley et al. 2016; Lockhart et al. 2017). Comparing these two orthologs, nine amino acid substitutions were identified that were also identified in azole-resistant *C. albicans*. All nine plus three more substitutions were identified in *C. auris* isolates that showed extreme azole resistance. These amino acid substitutions are thought to play a major role in antifungal resistance in this emerging pathogen.

Each species of *Candida* has varying degrees of pathogenicity and antifungal resistance, which is why identifying the organism causing invasive candidiasis down to the species level is important for treatment outcome (Lockhart et al. 2017). Most clinicians tend not to perform species conformation when treating candidiasis, which could have serious consequences. *Candida albicans* is very susceptible (>98% of isolates) to fluconazole as a therapy, but 10–12% of *Candida glabrata* isolates are resistant to fluconazole (Lockhart et al. 2017). With the emergence of *Candida auris*, which is often multidrug resistant, it is even more vital to identify the infectious species for the best treatment outcome. Most *C. auris* isolates are naturally resistant to fluconazole, 50% of isolates are resistant to amphotericin B, and three isolates to date have showed resistance to the three main classes of antifungal drugs (azoles, polyenes, echinocandins) (Lockhart et al. 2017). Genomic epidemiology can be a useful tool to identify antifungal-resistant genes and better treat these infections.

A clinical isolate of *C. auris* that showed resistance to all echinocandin drugs but was susceptible to azoles (except for fluconazole) and also resistant to 5-flucytosine was recently sequenced. In other fungal species, mutations in the FKS1 (encodes 1,3-β-D-glucan synthase which catalyzes the synthesis of crucial cell wall components) gene have led to the resistance to echinocandins (Jiménez-Ortigosa et al. 2017). This *C. auris* isolate had a non-synonymous SNP in the FKS1 genes that

Species complex	Population structure/Cryptic species	References
• *Candida albicans*	• Genetic Cluster 1-18 and A-E	• McManus & Coleman, 2014; Ropars et al., 2018
• *Candida dubliniensis*	• Genetic Cluster C1-C3	• McManus et al., 2018
• *Candida tropicalis*	• 71 clonal complexes	• Scordino et al., 2018
• *Candida parapsilosis sensu lato*	• *Candida parapsilosis sensu strictu* • *Candida orthopsilosis* • *Candida metapsilosis (genotypes I and II)*	• Tavanti et al., 2005;
• *Candida glabrata*	• Genetic Cluster I-VII	• Dodgson et al., 2003; Carreté et al., 2018
• *Candida kruseii*	• Genetic Cluster I-IV	• Jacobsen et al., 2007

Fig. 1 *Candida* sp. population structure

caused a serine to tyrosine substitution (Rhodes et al. 2017). Another mutation, which caused a phenylalanine to isoleucine substitution, was located in the FUR1 (encodes uracil phosphoribosyltransferase which synthesizes UMP from uracil) gene (commonly associated with 5-flucytosine resistance). These mutations were not described in any other *Candida* species (Rhodes et al. 2017). Non-synonymous SNPs seem to be the source of resistance to echinocandin and 5-flucytosine. Interestingly, no mutations specific for amphotericin B resistance were identified, which the authors suggest may be acquired through a non-mutation mechanism.

In summary, the genus *Candida* contains a diverse array of human pathogenic and commensal fungi. Genomics has helped to understand the basis of antifungal resistance and the evolution of pathogenesis. See Fig. 1 for explanation of genetic population structure and cryptic species of *Candida*.

3 Human Fungal Pathogens from the Genus *Aspergillus*

Aspergillus fumigatus is a primary and opportunistic pathogen, as well as considered a major allergen. The human respiratory tract is constantly being exposed to high volumes of the ubiquitous conidia, which are produced continuously by *A. fumigatus* in the environment (Kwon-Chung and Sugui 2013). In immunocompromised patients, *A. fumigatus* infections can reach up to 50% mortality rate, depending on patient group and type of infection (Lin et al. 2001). The genome of *A. fumigatus* was compared to its close nonpathogenic relative *Neosartorya fischeri*; 700 genes were found in *A. fumigatus*, which were either absent from or significantly diverged in *N. fischeri*. Many of these genes encode for pathogenic phenotypes and imply that *A. fumigatus* diverged to be a streamlined pathogen (Nierman et al. 2005).

Aspergillus fumigatus is an ubiquitous saprophyte in the environment. It is assumed that genes associated with a pathogenic lifestyle are upregulated during host infection (Chung et al. 2014). The genome of *A. fumigatus* was compared with a common food biotechnology fungus and relative *Aspergillus oryzae*, as well as a less

pathogenic relative *Aspergillus nidulans* (Galagan et al. 2005). Approximately 500 genes were found to be unique to *A. fumigatus*. These genes were then compared with other pathogenic fungi such as *Candida* and *Cryptococcus*, and none of these 500 genes were common virulence factors (Ronning et al. 2005). This suggests that the ability to infect and survive in a human body is not uniquely present within the genome but is a consequence of being a successful environmental saprophyte adapted to high temperature and stressful environments, such as compost piles (Kwon-Chung and Sugui 2013; Cramer 2016). Mechanisms to adapt to extreme environmental conditions are also critical for pathogenicity. *A. fumigatus* has the ability to grow well at 37°C, can grow in low oxygen, has a proteome that produces metabolites that helps the fungus compete with other organisms, has mechanisms that combat oxidative stressors, and activates efflux pumps regularly to transport toxins out of cells quickly (Ronning et al. 2005; Grahl et al. 2012). These metabolic factors are useful for survival and competition in the environment as well as the harsh environment of the human body. Tolerance of higher temperatures, oxidative stress (reactive oxygen species produced by neutrophils), and osmotic stress are obstacles that *A. fumigatus* has to overcome in order for the conidia to germinate into mycelia and establish infection in a living host (Takahashi et al. 2017; Cramer 2011). A total of 266 stress response genes were highly expressed when *A. fumigatus* was exposed to one of the environmental conditions above, and 77 have orthologs in *S. cerevisiae*, which suggests that these genes are needed for environmental stress adaptation and could increase pathogenicity (Takahashi et al. 2017). Studies have examined the stress response genes that are expressed under hypoxic conditions showing that 35% of genes were differentially expressed under 30–120 min of hypoxic conditions and that *A. fumigatus* has 32 hypoxia-specific genes that were not expressed under normal oxygen conditions (Losada et al. 2014). Hypoxia induces a genomic stress response to which *A. fumigatus* has adapted and gives *A. fumigatus* the ability to establish long-term chronic infections in hypoxic lung nodules. Thermotolerance is a critical trait needed to survive in a mammalian host. Certain genes in *A. fumigatus* were upregulated when exposed to higher temperatures (37–48°C). Many of these genes code for specific heat shock proteins that are different than the normal stress response genes of other fungi (Nierman et al. 2005). This suggests that *A. fumigatus* has unique mechanisms to thrive at higher temperatures and in many different environments. *Aspergillus fumigatus* also has more described allergens than in all other fungi combined (Ronning et al. 2005). *A. nidulans* and *A. oryzae* are also common allergen-producing fungi but are less common causes of allergies than *A. fumigatus*. Genes such as Asp f1, which codes for a ribonuclease toxin, and the metalloproteinase Asp f5 are two well-studied potent allergens (Ronning et al. 2005).

Aspergillus fumigatus has long been thought to reproduce solely asexually despite the tremendous amount of genetic diversity within the species and genetic evidence for sexual reproduction. Initially, the *MAT1-1* (α-box domain) and *MAT1-2* (HMG domain) mating-type genes of *A. fumigatus* were discovered, and the idiomorphic structure suggested heterothallic sexual development (Paoletti et al. 2005). Moreover, the authors found several pheromone-producing, pheromone-processing,

and pheromone-detecting genes associated with sexual reproduction in the genome. Recombination was previously detected within *A. fumigatus* species (Brenier-Pinchart et al. 1998), and coupled with this genomic data, this provided strong evidence that this fungus undergoes sexual reproduction. The existence of a sexual life cycle creates several implications for the virulence and pathogenicity of *A. fumigatus*. Heterothallic mating systems increase genetic variation within a population and can result in rapid evolution in response to a changing environment (Paoletti et al. 2005). The ability to adapt to novel environments may have led to the success of *A. fumigatus* as a pathogen. The presence of a sexual life cycle could also mean that within a population there could be a significant amount of gene flow. Four years after discovering the genomic potential for sexual reproduction in *A. fumigatus*, the sexual cycle in this species was discovered in laboratory setting (O'Gorman et al. 2009). Opposite mating-type strains were co-incubated in oatmeal agar medium for up to 6 months or until the observation of fruiting bodies. *A. fumigatus* produces a globose cleistothecia varying from yellow to white pigmentation and with diameter having no more than 150 μm (O'Gorman et al. 2009). The ascospores from each mating crossing were isolated, and random amplified polymorphic DNA (RAPD) markers and MAT idiomorph analysis of the progeny identified recombining individuals. Sexual reproduction could lead to inheritance of antifungal-resistant genes and genes that increase virulence expanding throughout a population (Paoletti et al. 2005).

Like *A. fumigatus*, *Aspergillus niger* is an opportunistic fungal pathogen that is also ubiquitous in the environment and is found worldwide. In the environment, it grows in a filamentous fungus on decaying plant matter, in litter, and in the soil and is able to grow at temperatures ranging from 6°C to 47°C (Schuster et al. 2002). It is also economically important as it yields a high amount of citric acid during fermentation that is used in a wide array of industrial applications (Schuster et al. 2002). Genome sequencing and analysis of *A. niger* revealed that this species harbors at least 8,695 genes and shares 6,755 orthologous genes with *A. nidulans* and *A. fumigatus* (Pel et al. 2007). Infections from *A. niger* are not as common as infections from *A. fumigatus*; however, immunocompromised patients are at risk for developing life-threatening mycoses from *A. niger*. A study in Italy showed that out of 194 patients that suffered from invasive aspergillosis, eight were due to *A. niger*, and all eight patients died due to complications from the fungal infection (Fianchi et al. 2004). Certain strains also produce a mycotoxin called ochratoxin A that can cause nephropathy (kidney disease) and other renal diseases (Schuster et al. 2002).

Aspergillus flavus is a common soil saprobe as well as a pathogen of insects, humans, and several different types of plants (Amaike and Keller 2011). Some strains of this species produce the highly toxic carcinogenic secondary metabolite aflatoxin (Geiser et al. 2000). Ingesting crops that have been contaminated with aflatoxin as well as being infected with an aflatoxin-producing strain of *A. flavus* causes serious health problems worldwide. It is also the second most common cause of aspergillus infections in humans and animals behind *A. fumigatus* (Hedayati et al. 2007). *A. flavus* expresses virulence genes that activate when exposed to

temperatures that correlate to mammalian and avian body temperatures (Yu et al. 2005). These genes encode for certain heat shock proteins. Several enzymes have been identified with pathogenicity of both plants and animals such as pectinase P2c, many proteases, and hydrolytic enzymes (amylases, glucanases). An important biochemical pathway produces aflatoxin. There are 29 gene clusters which have been identified in the production of this toxin. The polyketide synthase gene is the most significant because no other PKS gene is known to be involved in aflatoxin biosynthesis (Yu et al. 2005). There is evidence that several mitogen-activated protein kinases are involved in aflatoxin synthesis and increased virulence (Yu et al. 2005). There are many different genomic components that contribute to pathogenicity and mycotoxin production. The variability in virulence factors that are expressed may be due to the host that *A. flavus* infects (plant, insects, humans, etc.), which reflects the wide dynamic host range.

4 Human Fungal Pathogens from the *Cryptococcus* Genus

The genus *Cryptococcus* is composed of two (but most likely more) universally recognized species of pathogenic basidiomycete yeasts: *Cryptococcus neoformans* and *C. gattii* (Kwon-Chung et al. 2014; Hagen et al. 2015). See Fig. 2 for an itemization of the species complex and proposed species of *Cryptococcus*. These yeasts are ubiquitous in the environment but can become highly pathogenic in humans upon infection by aerosolized yeast or sexual spores. *Cryptococcus* spp. account for millions of opportunistic infections worldwide every year, and most of these infections take place in patients that have a compromised immune system such as those infected with HIV/AIDS or undergoing immunosuppressive therapies (May et al. 2016). Recently there has been an emergence of *Cryptococcus* infections in healthy individuals in northwestern North America, despite the previous assumptions that the fungus was native to tropical or subtropical regions of the world. Recent genomic analysis shows that this is a new, more virulent strain that evolved the ability to survive in an environment different from their tropical relatives.

Cryptococcus neoformans sensu lato is an opportunistic fungal pathogen that is emerging worldwide due to an increased number of immunocompromised individuals infected with HIV/AIDS and on immunosuppressive therapies. *C. neoformans* is the leading cause of secondary fungal infections and the leading cause of death among HIV-infected individuals (Liu et al. 2008). At least two species are recognized within *C. neoformans* species complex: *C. neoformans* sensu stricto genotypes (serotype A, AFLP1/VNI, AFLP1A/VNB/VNII, and AFLP1B/VNII) and *C. deneoformans* (serotype D, AFLP2/VNIV). There are many virulence factors that give the ability of *C. neoformans* sensu lato to be a successful opportunistic pathogen. The three most important are the ability to survive and proliferate at 37°C, a polysaccharide capsule that surrounds the fungal cell and increases infectivity, and the ability to melanize (Kent and Juneann 1998).

Old nomenclature	**Proposed nomenclature**
• ***Cryptococcus neoformans*** sensu lato	• ***Cryptococcus neoformans*** **sensu stricto** *Cryptococcus neoformans* var. *grubii* *Clade F, AFLP1, VNI* *Clade G, AFLP1A/VNBi, VNII* *Clade H, AFLP1B, VNII* • ***Cryptococcus deneoformans*** *Cryptococcus neoformans* var. *neoformans* *Clade I, AFLP2 VNIV*
• ***Cryptococcus gattii*** sensu lato	• ***Cryptococcus gattii*** **sensu strictu** Clade D, AFLP4, VGI • ***Cryptococcus bacillisporus*** Clade C, AFLP5, VGIII • ***Cryptococcus deuterogattii*** Clade A, AFLP6, VGII • ***Cryptococcus tetragattii*** Clade E, AFLP7, VGIV • ***Cryptococcus decagattii*** Clade B, AFLP10, VGIV/VGIIIc

Fig. 2 Proposed new species of *Cryptococcus neoformans* and *C. gattii* complexes based on population genomics

The reference *C. neoformans* genomes were sequenced, assembled, and annotated by two different groups (Loftus et al. 2005; Janbon et al. 2014). The genomes of *C. deneoformans* JEC21 and B-3501A were sequenced and annotated and assembled into a 19–18.5 Mb genome on 14 chromosomes (Loftus et al. 2005). The JEC21 genomes harbor 6,572 genes, and most of the of genes share >98% nucleotide identity; however chromosomal translocation and segmental duplication were observed comparing the assemblies of the strains JEC21 and B-3501A. Strain-specific genes were identified in both B-3501A (Ras guanosine triphosphatase-activating protein and 2 proteins of unknown function) and JEC21 (4 proteins of unknown function and 20 duplicate genes presented in the segmental duplication). The genome and transcriptome of *C. neoformans* var. *grubii* H99 strain provided an improved assembly of this species (Janbon et al. 2014). The genome of H99 also has 14 chromosomes (18.9 Mb) of varied sizes and encodes for 6,962 proteins, slightly more compared to *C. deneoformans*. The overall sequence divergence between orthologs shared between *C. neoformans* and *C. deneoformans* was high (on average 7%).

A distinctive feature of *Cryptococcus* spp. is the polysaccharide capsule that surrounds and makes up the outermost layer of the cell and is required for virulence. There are around 30 genes that are involved in biosynthesis of the polysaccharide capsule. Two gene families are unique to basidiomycetes and are not present in other nonpathogenic yeast relatives, the CAP64 and CAP10 gene families (Loftus et al. 2005). Recent gene knockout experiments reveal 5 of the 30 are novel genes that are crucial for capsule biosynthesis. GAT201 and SSN801 are important transcription factors. HOS2 and SET302 are chromatin genes that are important in gene expression, and CPL1 codes for proteins involved in capsule formation (Liu et al. 2008). Gene deletions showed decreased virulence in murine models, showing the importance of the capsule in infecting host cells. Melanization is another important virulence factor for *C. neoformans.* The ability to produce melanin is important to pathogens because it reduces the susceptibility of pathogens to host immune defenses. Thirty-three novel genes are identified in melanin production. Knockout experiments were done to show the role of these genes, and one gene in particular seemed to play a significant role in the infection process. The Rim101 gene, when knocked out, caused a major lack of melanin production and decreased pathogenesis (Liu et al. 2008). This is evidence that melanin plays an important role in the infection process of *C. neoformans.* There are a number of other factors that play a role in infectivity. These include genes that regulate urease production, oxidase production, phospholipase production, and increased iron uptake and genes that control resistance to some species of nitrogen and oxygen (Alspaugh 2015; Liu et al. 2008; Kronstad et al. 2012; Brown et al. 2007; Chun et al. 2007). When genes associated with these processes are deleted using mutagenesis, the fungus is more susceptible to certain chemicals and is unable to grow under the conditions needed for infection of the host.

Many of PAMPs associated with *C. neoformans* must be transported from intracellular sites to the cell surface (Rodrigues et al. 2008). These factors regularly lack signaling peptides; so *C. neoformans* developed alternative mechanisms to get these molecules to the cell surface. The polysaccharides that compose the capsule are examples of these virulence-associated molecules. Membrane-bound microvesicles have been observed inside and outside of the cell, evidence that these vesicles are one of the mechanisms responsible for transporting virulence-associated molecules throughout the fungal cell (Alspaugh 2015). These microvesicles have been shown to contain many of these virulence-associated molecules including capsule precursors, melanin, and secreted enzymes (Rodrigues et al. 2008). Microvesicles act directly on the extracellular surface as well as influence the interactions with the host cells (Alspaugh 2015). There is evidence that suggests that microvesicles can alter the host blood-brain barrier to allow the fungus to enter the central nervous system (Jong et al. 2008). *C. neoformans* appears to have evolved these microvesicles to increase virulence and evade the host immune response (Huang et al. 2012).

When a pathogen enters a host, it encounters a variety of host-derived stressors. Adaptation to these stressors is an essential task to becoming a successful pathogen. These adaptations lead to microevolution while infecting a host (Janbon et al. 2014).

C. neoformans has evolved many cellular mechanisms to survive these stressors. Transcriptome analysis of *C. neoformans* showed that most cryptococcal genes have introns (nucleotide sequence removed by RNA splicing) and alternatively spliced mRNA (Alspaugh 2015). These introns and mRNA variants are thought to be a means to rapid adaptation to host response. Another important adaptation that a pathogen must possess is the ability to rapidly repair DNA. The accumulation of damage to DNA will be detrimental to the organism, but it is thought that a low level of DNA damage is beneficial to *C. neoformans* as it allows for increased genetic diversity of the infection population which can lead to the adaptation to new stressors within that particular host. This microevolution due to host-derived stressors can allow for persistent infection by the population.

The genetic machinery that controls these microvesicles is poorly understood, although recent evidence suggests that the 14-3-3 proteins (which are highly abundant in *Cryptococcus* microvesicles) play a role in the regulation of these vesicles (Li et al. 2016). The 14-3-3 proteins are ubiquitous and conserved protein among eukaryotic organisms. 14-3-3 proteins are a family of dimeric proteins that attach to phosphorylated serine and threonine residues and play a critical role in maintaining the cell cycle checkpoints, DNA repair, prevention of apoptosis, coordination of cell adherence, and many other functions across all eukaryotes (Wilker and Yaffe 2004). When the 14-3-3 gene was knocked out in *C. neoformans*, total microvesicle protein was reduced, and laccase and acid phosphatase (both enzymes are associated with *Cryptococcus* microvesicles) activity was reduced (Li et al. 2016). This suggests that loss of 14-3-3 function results in a decline of microvesicle secretion, resulting in a smaller capsule and a drop in virulence.

The HIV/AIDS pandemic has created a large population of people that are immunocompromised susceptible to severe fungal infections (Chastain et al. 2017). *C. neoformans* is one of the main culprits of these debilitating infections, and meningitis cases may require long-term therapy (Charalambous et al. 2018). The long-term use of antifungal therapies during infection in immunocompromised patients can lead to microevolution of the fungus within the host (Rhodes et al. 2017). Patients that appear to be successfully treated and show a resolution in symptoms may later have a reoccurrence of disease due to a persistent infection of *C. neoformans*, in some cases exhibiting evidence of resistance to certain antifungals (azoles). A recent study investigating if relapsed infections are due to the original infection or a subsequent new infection showed evidence that relapse infection is a result of the original organism, but the original strain has changed consistently with microevolution (Rhodes et al. 2017). Aneuploidy (deviation of normal chromosome number leading to the loss or gain of chromosomes) events are known to occur in *Cryptococcus* (Lengeler et al. 2001). Rhodes et al. observed that the aneuploidy event (in this case triplication of the chromosome arm) was on different regions of chromosome 12, which has 327 genes. Of those genes, the SFB2 gene stood out as a potential virulence gene. This gene is involved in the conservation in sterol regulatory binding element pathway (an important transcription factor regulating sterol, a component of the fungal cell membrane) and producing an alcohol dehydrogenase, which is shown to be protective against host immune defense (Chang et al. 2009; De

Jesus-Berrios et al. 2003). It was also shown that there was an enrichment of genes on chromosome 12 arm that are involved in the metabolism of drugs (Rhodes et al. 2017). Evidence suggests that change in ploidy over the course of infection is an adaptive mechanism that allows microevolution of *C. neoformans* within the host, which causes persistent infections to return after antifungal therapy.

Cryptococcus gattii is mainly found in tropical and subtropical parts of the world and infects both immunocompetent or immunosuppressed people. *C. gattii* is phylogenetically distinct from the *C. neoformans*/*C. deneoformans* complex independent of the molecular marker used for analysis (Gillece et al. 2011; D'Souza et al. 2011; Fraser et al. 2005). *C. gattii* exhibits a unique teleomorph and yeast features and distinctive biochemical properties used for routine laboratory differential diagnostics (Kwon-Chung et al. 2014). Similar to *C. neoformans*/*deneoformans*, *C. gattii* has 14 chromosomes; centromere locations are preserved, but notable inversions and balanced translocations are observed (Janbon et al. 2014). The overall sequence divergence between *C. neoformans* and *C. gattii* orthologs is about 11% higher than between *C. neoformans* and *C. deneoformans*. Lastly, comparative genomics suggest that in contrast to *C. neoformans*, *C. gattii* VGII lineage has lost the RNAi genomic apparatus (Billmyre et al. 2013).

In the early 2000s, there was an outbreak of cryptococcosis in northwestern North America, and the causative organism was identified as a more virulent and clonal strain of *C. gattii* (Stephen et al. 2002; Kidd et al. 2004). This *C. gattii* genotype infects otherwise healthy individuals with no underlying conditions such as cancer or HIV/AIDs, as is more common with *C. neoformans*. The genomes of the worldwide strain of *C. gattii* (WM276) and the North American strain (R265) were compared to determine the cause of the increase in virulence and the geographic expansion. When the genomes of the two strains were compared, chromosomal rearrangement and regions with inversions, as well as a 7.6% divergence of nucleotide sequences, were observed (D'Souza et al. 2011). This suggests speciation within *C. gattii*, and the emergence of this new hypervirulent species may be the result of adaptation to a new ecological niche (Engelthaler et al. 2014). Analysis of the emergent Pacific Northwest strain of *C. gattii* showed multiple single nucleotide polymorphism (SNP) mutations that gave rise to unique alleles, which support niche adaptation and possibly explain changes in virulence.

The gene content of the two strains was compared revealing that 445 of WM276 (worldwide strain) genes did not map to the genome of strain R265 (North American strain). Perhaps the most important genes that were present in strain WM276 but absent in strain R265 were Ago1 and Ago2, which encode for Argonaute proteins (D'Souza et al. 2011). These Argonaute proteins are RNA silencing proteins that are responsible for the phenomena of gene silencing. This could provide evidence that strain R265 is more virulent because virulent genes are not silenced by Argonaute proteins. There are other examples of gene loss in the R265 strain of *C. gattii* that may be due to selective pressures, and selective loss of genes or gene functions caused this strain to become more virulent (D'Souza et al. 2011). There has been evidence that suggests that pathogens become adapted to selection pressures of the host by inactivating anti-virulence genes (Alves et al. 2014; D'Souza et al. 2011).

Another hypothesis for the emergence of *C. gattii* in North America has been proposed. There is evidence that the outbreak of *C. gattii* on Vancouver Island is due to same-sex mating by the fungus (Fraser et al. 2005). There are two genotypes of *C. gattii* on Vancouver Island, and the more virulent majority genotype seems to be produced through same mating-type sexual cycle. The researchers in this study propose the dramatic geographical shift is due to airborne sexual spores produced by same mating-type (α-mating type) parents. In laboratory studies it has been shown that the close relative *C. neoformans* can undergo same-sex mating between two α-mating-type individuals (Fraser et al. 2005). *C. gattii* isolates from Vancouver Island can reproduce sexually but all of the isolates are α-mating type. Fraser et al. (2005) analyzed the mating-type locus and showed that the majority of the outbreak isolates were produced by an α-α sexual cycle. The authors tested virulence in murine models, and the majority genotype (R265) is highly virulent, whereas the minority genotype (WM276) was avirulent. Same-sex mating could contribute to the global expansion of this fungus by eliminating the need for a sexual partner and producing a proliferation of aerosolized spores and would produce identical clones of the parents. If spores from a highly virulent parent strain of *C. gattii* landed on Vancouver Island, the ability to undergo same-sex mating would allow this strain to spread across the island infecting an abundance of naïve hosts. In *T. gondii* (causative agent of toxoplasmosis), same mating-type sexual life cycles can alter pathogenicity and lead to progeny with enhanced virulence (Grigg et al. 2001). Thus, the cryptic sexual life cycle of *C. gattii* could account for the geographic expansion and hypervirulence on Vancouver Island.

There are four distinct lineages of *C. gattii* (VGI-IV) that are so genetically variable that some consider them four separate species, but they have the ability to mate and exchange genetic material with one another (Farrer et al. 2015). These species complexes were renamed as follows: *C. gattii* sensu stricto (VGI), *Cryptococcus bacillisporus* (VGIII), *Cryptococcus deuterogattii* (VGII), *Cryptococcus tetragattii* (VGIV), and *Cryptococcus decagattii* – a hybrid VGIVj/VGIIIck (Hagen et al. 2015). *C. gattii* can form hybrids with *C. neoformans.* The population structure of *C. gattii* consists of VGI in Europe, VGII in North and South America, and VGIV in Africa (Farrer et al. 2015). All lineages cause infection, but VGI and VGII seem to be the most virulent. For example, the emerging hypervirulent strain of *C. gattii* in the American Pacific Northwest is in the VGII lineage. This lineage is associated with higher rates of respiratory symptoms during infection and has the ability to proliferate inside host macrophages, which differs from the other lineages (Farrer et al. 2015). When these authors compared the genomes of isolates from VGI and VGII lineages, there were several gene presence and absence of polymorphisms. Each lineage has a set of unique genes that may influence both virulence and outcome of disease. For example, VGI has a unique set of genes coding for ferric reductase enzymes that are involved in the production of melanin, a known virulence factor that contributes to azole drug resistance (Farrer et al. 2015). VGII has more secretory carrier membrane proteins than other lineages. These proteins are involved in distributing macromolecules throughout the cell. Additionally, the Prmt1 chromatin-associated protein domain and the heat

shock protein 70 (HSP70) domain that are found on chaperone proteins are all expanded in VGII (Farrer et al. 2015). The presence or absence in certain genes between *C. gattii* lineages reveals targets for different strategies to initiate and maintain infections. Changes in environment and hosts are likely drivers of gene polymorphism between lineages of *C. gattii*.

RNA interference (RNAi) is a process in which molecules of RNA inhibit transcription or translation by neutralizing mRNA molecules. This conserved eukaryotic process aids in genome stability and repression of transposable elements (Billmyre et al. 2013). Genomic analysis of *C. deuterogattii* strain R265 shows that both Argonaute genes (key components of the RNAi-induced silencing complex) are missing (Feretzaki et al. 2016). When this strain was compared to *Cryptococcus* genomes from other linages, 14 genes were found to be missing, four of which (RDP1, AGO1, DCR1, ZNF3) are key components of the RNAi pathway. The loss of RNAi could be associated with increased phenotypic and genotypic diversity that might lead to the increased virulence of *C. deuterogattii* (Feretzaki et al. 2016).

5 Human Fungal Pathogens from the *Onygenales* Order

The order *Onygenales* harbors many dimorphic fungal pathogens responsible for endemic and systemic mycosis worldwide (Sil and Andrianopoulos 2014). The order includes several species complexes nested within the genera *Paracoccidioides*, *Coccidioides*, *Histoplasma*, *Blastomyces*, *Emmonsia*, *Emergomyces*, and *Lacazia. Coccidioides immitis* and *C. posadasii* are members of the family *Onygenaceae* sensu stricto, but all other genera are placed within the *Ajellomycetaceae* family (Untereiner et al. 2004; Dukik et al. 2017). See Fig. 3 for population structure, proposed species, and cryptic species of dimorphic fungal pathogens. Beyond humans, these species can potentially naturally infect every mammalian species that comes in contact with the infectious propagules (primarily asexual conidia), and the primary affected organs are the lungs (Bagagli et al. 2006; Kohler et al. 2017). The teleomorph of these fungi was previously characterized as *Ajellomyces* for the genera *Emmonsia*, *Blastomyces*, and *Histoplasma* and comprises heterothallic species that, under specific conditions, form complex ascomata (gymnothecia) with coiled appendages, evanescent asci harboring oblate ascospores (McDonough and Lewis 1967; Kwon-Chung 1972). The infections caused by these fungi vary from asymptomatic to mild pneumonia that are resolved after a short period of time (Hage et al. 2012). However, the disease often may progress to a chronic pulmonary infection or disseminate to different body sites, including the meninges. Many of these fungal infections may be fatal, especially in immunocompromised patients (Brown et al. 2014), or debilitating if incorrectly diagnosed or treated (i.e., pulmonary fibrosis) (Hardie et al. 2009). Current standard diagnostic methods may take several weeks, and fast and accurate diagnostic tools are scarce.

Except for *Lacazia loboi*, which is considered an obligate pathogen, these species live in the soil as saprobes likely decomposing animal-derived matter in a filamentous form (Emmons and Ashburn 1942). These fungi produce large numbers of

Genus	Species proposed	Population structure/Cryptic species	References
• *Histoplasma*	• *Histoplasma capsulatum strictu senso* (Panama) • *Histoplasma mississippiense* (NAm1) • *Histoplasma ohiense* (NAm2) • *Histoplasma suramericanum* (LAm1)	• North American 1 (NAm1) • North American 2 (NAm2) • LAm A1 • LAm A2 • RJ • LAm B1 • LAm B2 • Panama • BAC1 • Australia, • Netherlands, • Africa	• Kasuga et al., 2003; Teixeira et al., 2016; Sepulveda et al., 2017
• *Coccidioides*	• *Coccidioides immitis* • *Coccidioides posadasii*	• San Joaquin Valley, • San Diego/Mexico • Washington • *Arizona* • *Texas/Mexico/South America* • *Guatemala*	• Fisher et al., 2002; Teixeira et al., 2016; Engelthaler et al., 2016
• *Paracoccidioides*	• *Paracoccidioides brasiliensis* • *Paracoccidioides lutzii* • *Paracoccidioides venezuelensis* • *Paracoccidioides americana* • *Paracoccidioides restrepiensis*	• *S1a* • *S1b* • *Pb01-like* • *PS4* • *PS2* • *PS3*	• *Matute et al., 2006; Teixeira et al., 2009; Turissini et al., 2017*
• *Blastomyces*	• *Blastomyces dermatitidis* • *Blastomyces gilchristii* • *Blastomyces percursus*	• *Population 1-4* • *Population 1-4* • *ND*	• *Brown et al., 2013a, b; McTaggart et al., 2016; Dukik et al., 2017*
• *Emmonsia*	• *Emmonsia parva* • *Emmonsia crescens* • *Emmonsia helica*	• *ND* • *Eurasian* • *North America* • *ND*	• *Sigler et al., 1996; Dukik et al., 2017*
• *Emergomyces*	• *Emergomyces pasteuriana* • *Emergomyces africanus* • *Emergomyces orientalis* • *Emergomyces canadensis*	• *ND* • *ND* • *ND* • *ND*	• *Dukik et al., 2017*

Fig. 3 Genomic population structure and cryptic species of dimorphic fungi

airborne-dispersed conidia that upon inhalation by a susceptible host are accumulated into alveoli in the lungs. In the lung, or under specific in vitro conditions (36–39°C and specific media), they convert into budding yeast-like cells, adiaspores, or endosporulating spherules (Sil and Andrianopoulos 2014). This dimorphic transition is thought to promote expression of several virulence factors that allow the fungi to colonize the lung tissues and spread to other organs by different mechanisms (Boyce and Andrianopoulos 2015). According to Gauthier (2015), membrane fluidity and lipid dynamics have a direct effect on dimorphism because lower temperatures reduce membrane plasticity by a decrease in the ratio of saturated to unsaturated fatty acids. The opposite occurs at higher temperatures.

Several studies have been conducted on dimorphic fungi focused specifically on molecular mechanisms that underlie this morphological transition. Temperature, oxidative stress, modifications in carbon dioxide tension, and hormones were found to be important (Tavares et al. 2015; Edwards et al. 2013; Whiston et al. 2012). Signaling pathways have been shown to induce dimorphism and promote yeast growth at 37°C. The two-component signaling system regulated via DRK1 (dimorphism-regulating kinase) was the first mechanism discovered to be important for thermo-dimorphism in *Onygenales*. Gene knockouts for *drk1* gene in *B. dermatitidis* and *H. capsulatum* resulted in avirulent phenotypes in murine models of infection and impaired the capacity to convert into pathogenic yeast form. Moreover, those mutants grow as mycelia at 37°C, suggesting that this gene is a key regulator for this process (Nemecek et al. 2006).

The DRK1 ortholog in *P. brasiliensis* is highly expressed in the virulent yeast phase, and is fundamental in the mycelial to yeast transition, suggesting a potential new drug target (Camacho and Nino-Vega 2017). In *H. capsulatum*, another group of transcription factor genes was found to be important for the dimorphism, and such genes are preserved in other pathogenic *Onygenales*. Those genes were named RYP1–4 (required for yeast phase), and gene knockout studies also produced a truncated hyphal phase at 37°C, as well as directly affecting the expression of other known virulence factors (Nguyen and Sil 2008; Webster and Sil 2008). Using comparative genomics and transcriptomics, none of the homologs of these regulators (Ryp1, Ryp2, Ryp3, and Drk1 genes) are expressed more during the spherule phase in *Coccidioides* sp., suggesting that this particular genus may have alternative ways to induce the parasitic phase (Whiston et al. 2012).

Genetic studies revealed that all species, in both filamentous and pathogenic forms, are haploid. However, a single cell may contain multiple nuclei. So far, genomes of several species from *Onygenales* have been completed by different sequencing methods. Recently published comparative genomic studies provide information about the evolutionary adaptation of these pathogens to animal hosts or environments enriched with animal-derived compounds, such as animal burrows or caves (Sharpton et al. 2009; Whiston and Taylor 2015; Munoz et al. 2018). By comparing the entire genomes of onygenalean pathogenic fungi and related environmental species, several authors identified a significant gain of genes involved in the degradation of proteins (i.e., keratinases, subtilases, metalloproteinases) and loss of several genes involved in the degradation of plant-derived material enriched in

different carbohydrate molecules (i.e., glycosyl hydrolases and pectin lyases) (Sharpton et al. 2009). Fungal-specific kinases (FunK) are also overrepresented in those fungal pathogens and may be important to pathogenesis and survival inside the host (Desjardins et al. 2011; Munoz et al. 2015).

Species nested within the genus *Paracoccidioides* are the causative agents of paracoccidioidomycosis (PCM), a mycosis endemic to Latin America, which ranges from southern Mexico to northern Argentina (Bocca et al. 2013). The disease has a high prevalence in Brazil, Colombia, Venezuela, and Argentina. The annual incidence rate in Brazil is 10–30 infections per million inhabitants, and the mean mortality rate is 1.4 per million inhabitants per year, the highest cause of mortality among deep mycoses (Martinez 2017). The incidence of PCM is elevated in rural and endemic areas, and the disease mainly affects those individuals that pursue agricultural or hunting activities (especially armadillos). The infection primarily affects the lungs after inhalation of conidia upon soil disturbance. A multi-budding yeast cell characterizes the pathogenic form, and both cell types vary in shape and size according to the isolate/species (Tavares et al. 2015).

Paracoccidioides brasiliensis was considered an orphan species for at least 100 years despite several genetic studies demonstrating a high level of genetic diversity between isolates obtained from different endemic areas of the disease. According to phenotypical, MSLT, phylogenomics and population genetics, five species were recently determined within the *Paracoccidioides* genus. *Paracoccidioides lutzii* is a single monophyletic species formed by isolates so-called Pb01-like due the genetic similarities with the isolate Pb01 (Teixeira et al. 2009, 2014a, b). This species was the second most diagnosed after *P. brasiliensis* was discovered by Adolpho Lutz in 1909 and reclassified in 1930 by Floriano P. Almeida (Lutz 1908; Almeida 1930). *P. lutzii* is found in central and northwestern Brazil, especially in the states of Goiás, Mato Grosso, and Rondônia and outside the Brazilian borders, although a single case was detected in Ecuador (Teixeira et al. 2014a, b). Conidial cells are usually more elongated compared to those produced by other *Paracoccidioides* species (Teixeira et al. 2014a, b). *P. brasiliensis* sensu lato (former S1) is the most dispersed species and has been found in almost all endemic areas in Brazil, Argentina, and Uruguay (Matute et al. 2006). This species is comprised of at least two populations called S1a and S1b (Turissini et al. 2017; Munoz et al. 2016). S1a harbors mostly isolates recovered from São Paulo and Rio de Janeiro states of Brazil as well as those from Argentina. Recent population genomic studies revealed that there is a strong geographic isolation between isolates from southeast Brazil and Argentina. S1b, on the other hand, is composed of isolates from Mato Grosso do Sul, Parana (Brazil), Paraguay, and Argentina; however, recent studies suggest that these two populations potentially overlap (Munoz et al. 2016; Theodoro et al. 2012). *P. americana* (former PS2) is a less frequently isolated species from both humans and armadillos, and a few cases have been reported in the São Paulo and Rio de Janeiro states of Brazil and a single site in Venezuela (Turissini et al. 2017; Matute et al. 2006). The most intriguing aspect is the fact that both *P. brasiliensis* and *P. americana* share the same niche, since both species were recovered in the same hospital in Rio de Janeiro

city or in nine-banded armadillos from a single municipality of São Paulo state (Theodoro et al. 2012; De Macedo et al. 2016). A single isolation from a dog that tested positive for this species represents the only canine isolate obtained for *Paracoccidioides* (Theodoro et al. 2012; De Farias et al. 2011). *P. venezuelensis* (former PS4) and *P. restrepiensis* (former PS3) are two other monophyletic species that are geographically restricted to Venezuela and Colombia, respectively (Turissini et al. 2017; Matute et al. 2006). Those species cause both acute and chronic forms of PCM, and no clinical or morphological differences were observed (Shikanai-Yasuda et al. 2017).

P. brasiliensis, *P. americana*, and *P. restrepiensis* have been recovered routinely from armadillos, although several attempts to isolate *P. lutzii*, the most divergent species, from these hosts have failed (Hrycyk et al. 2018). These results indicate that *P. lutzii* may have an alternative host or may have a lower virulence compared to other *Paracoccidioides* species. The diagnosis via serology between *P. lutzii* and the other related species of *Paracoccidioides* must be performed differently, due the high polymorphism of secreted antigens. The most used antigen glycoprotein 43 (GP43), as well as total antigenic extracts, results in differential serological test results between *P. lutzii* and other *Paracoccidioides* species (Teixeira et al. 2014a, b; Gegembauer et al. 2014). Previous studies suggest that infections caused by *P. lutzii* had lymphoabdominal forms as the most prevalent manifestation, while those carrying *P. brasiliensis* would have more classical pulmonary and mucosal involvement. More studies aiming to characterize the clinical relevance in the species context are needed (Shikanai-Yasuda et al. 2017; Teixeira et al. 2014a, b).

The genomes of *Paracoccidioides* range from 29.1 Mb to 32.9 Mb and code for 7,610 to 8,130 genes. The *Paracoccidioides* genomes are highly syntenic for the species investigated. *P. brasiliensis* and *P. americana* share a higher percentage of sequence similarity (around 96%) to each other compared to *P. lutzii* (around 90%). The *Paracoccidioides* genomes display unique specific fungal kinases, reduction of carbohydrate-degrading enzymes, and expansion of enzymes responsible for degrading animal biomass. The transposable element content between these species varies significantly: 8% in *P. americana*, 9% *P. brasiliensis*, and 16% in *P. lutzii* (Desjardins et al. 2011). The most abundant are class I elements (retrotransposons), LTR retrotransposons, and LINEs, but no SINE elements were identified. The *P. lutzii* genome accumulated a twofold greater content of LTR elements compared to *P. brasiliensis* and *P. americana*, while fewer LINE elements were found in *P. lutzii* compared to *P. brasiliensis* and *P. americana* (Desjardins et al. 2011). Additional analysis showed a rapid evolution of dimorphism-related genes as compared to *Histoplasma capsulatum*. Proteins with zinc- and DNA-binding motifs, especially transcription factors, represented the majority of genes under positive selection (Munoz et al. 2016).

Paracoccidioides was considered a strict asexual and clonal fungus, but molecular data using MLST showed evidence for recombination within both *P. brasiliensis* and *P. lutzii*. The presence of two mating-type idiomorphs and other mating- or meiosis-specific genes was also deciphered using comparative genomic tools. Most genes involved in chromosome cohesion and recombination

are conserved among sexual eukaryotes and are found in *Paracoccidioides* with the exception of HOP2. This gene is missing not only in *Paracoccidioides* but also in *H. capsulatum* and *A. fumigatus*. In vitro mating crosses showed evidence for sexual development, but so far no gymnothecia or ascospores have been observed in vitro (Teixeira et al. 2013; Desjardins et al. 2011).

Coccidioides immitis and *C. posadasii* are the etiological agents of coccidioidomycosis or valley fever, which is a deep systemic mycosis that affects animals in arid and semiarid regions of the Americas (Lewis et al. 2015). Coccidioidomycosis is a notifiable disease in both California and Arizona states which account for the majority of the cases of the disease in the United States (Brown et al. 2013b). The majority of the cases are asymptomatic (60%), and 250,000 new infections per year are predicted to occur in the Americas (Odio et al. 2017). Beyond the southwestern United States, the northern region of Mexico is highly affected by this disease (Gaona-Flores et al. 2016). In Central America, cases have been reported in Guatemala and Honduras, and in South America, the disease is present in the arid regions of Argentina, Brazil, Paraguay, and Venezuela (Campins 1970). Both species live as saprotrophs in the environment, and upon the inhalation of infectious arthroconidia into the lung, the fungus switches to the parasitic phase. In a mammalian lung, the arthroconidia swell into immature spherules and, after multiple rounds of mitosis, form mature and endosporulating spherules (Lewis et al. 2015). Any activity linked with soil perturbation in endemic areas or even natural events such as haboobs, earthquakes, and tornados may be associated with the incidence of coccidioidomycosis (Brown et al. 2013b).

C. immitis was considered a single species for almost a century, but with the advances of molecular tools, the identification of a second species, *C. posadasii*, was confirmed (Fisher et al. 2001, 2002). Recently, applying sophisticated phylogenetic and population genetic tools identified at least six populations classified within *C. immitis* and *C. posadasii*. The *C. immitis* isolates recovered from San Diego and Northern Mexico are genetically distinct compared to those recovered from the Central Valley of California (Teixeira and Barker 2016; Engelthaler et al. 2016). Moreover, recently a new endemic area of coccidioidomycosis was identified in the eastern part of the Washington state, USA, which harbors a cryptic *C. immitis* population (Litvintseva et al. 2015). *C. posadasii*, on the other hand, is found in every other state in the western part of the United States, Mexico, as well as in Central and South America. *C. posadasii* clinical isolates from Arizona are genetically different from those recovered from Texas, Mexico, and South America. Genomic analysis also identified a third cryptic population composed by isolates from Guatemala (Teixeira and Barker 2016; Engelthaler et al. 2016). So far no significant morphological differences have been observed between populations or species. The only known phenotypic difference is the fact that *C. posadasii* grows faster in high concentrations of salt compared to *C. immitis* (Fisher et al. 2002). The genomic survey on 18 *Coccidioides immitis* and *Coccidioides posadasii* isolates revealed that hybridization and genetic introgression recently took place between the two species and may be the main forces of genetic variation of this pathogen. This study showed that gene flow from *C. posadasii* into *C. immitis* is more common, and

at least 8% of the genes found within the *C. immitis* population were recently introgressed from *C. posadasii* genomes (Neafsey et al. 2010).

Coccidioides comparative genomics represents one of the most complete population genomic studies so far among onygenalean fungi based on deposited sequences and published manuscripts (Engelthaler et al. 2016; Whiston and Taylor 2014). Initial comparative genomics found that *C. posadasii* C735 strain had a 27 Mb genome and 7,229 genes, while the *C. immitis* RS genome size totalized 28.9 Mb and 10,355 genes (Sharpton et al. 2009). The authors suggested that the variation in the gene counts was due to different methods applied to both genomes. This study identified that the *Coccidioides* lineage experienced gene family contractions for genes related to plant material degradation and acquired additional copies of genes responsible for degradation of animal proteins (i.e., metalloproteases families M35 and M36), especially keratinases. Moreover, this study showed that *Coccidioides* gained genes involved in metabolism, membrane biology, and mycotoxin biosynthesis that potentially enable associations with living animal hosts. Genes that code for secreted proteins, metabolism, and secondary metabolism also were found to be under positive selection suggesting that *Coccidioides* may adapt to the host immune system defense (Sharpton et al. 2009).

Histoplasma capsulatum sensu lato is a dimorphic fungal pathogen causing histoplasmosis, a mild pulmonary to a disseminated disease that is often fatal for HIV patients (Kauffman 2007). Different than coccidioidomycosis or paracoccidioidomycosis, this disease is found in all continents with the exception of Antarctica (Bahr et al. 2015). The fungus lives in the soil, especially in humid and dark environments, and is frequently associated with the presence of bird and bat guano. The disease also deeply affects other mammal species such as dogs and cats and is periodically isolated from bats. This may contribute to the dispersion of this fungus (Vite-Garin et al. 2014). The disease is triggered by the inhalation of micro- or macroconidia by a susceptible host that, under body temperatures, start to switch into a single-budding yeast form; every single mammalian species is potentially naturally infected by this fungus (Kauffman 2007).

Since this fungus is found in almost all continents, a complex population structure is expected. Initial MLST evaluating 137 strains from 20 countries distributed in six continents revealed the existence of at least seven phylogenetic species named North American 1 (NAm 1), North American 2 (NAm 2), Latin American A (LAm A), Latin American B (LAm B), Australia, Netherlands, and Africa (Kasuga et al. 2003). Additionally, a cryptic clade composed of clinical isolates from England, China, Thailand, and India was found to be derived from LAm A species. The old nomenclature proposed to this fungus suggested the existence of three varieties, *Histoplasma capsulatum* var. *capsulatum*, *H. capsulatum* var. *duboisii*, and *H. capsulatum* var. *farciminosum*, and according to phylogenetic analysis is meaningless since those isolates were found in multiple phylogenetic species. By using molecular clock and DNA mutation rate estimates, these authors found that *H. capsulatum* sensu lato started its radiation from 3.2 to 13.0 MYA (Kasuga et al. 2003).

Recently, the phylogenetic distribution of *H. capsulatum* was reevaluated by increasing the number of taxa. MLST and population genetic analyses identified additional cryptic species: (1) the LAm A clade was split into LAm A1, LAm A2 (composed primarily of isolates from Mexico, Argentina, Colombia, and Honduras), and RJ (composed of isolates from Rio de Janeiro and São Paulo); (2) the LAm B species was also separated into two groups, LAm B1 and LAm B2, according to its geographic location; (3) a phylogenetic species composed only by bat-derived strains was identified as BAC1; and (4) a cryptic clade within NAm 1 was identified and harbors only cat-derived strains. A series of additional cryptic monophyletic clades were also identified in this study, and instead of eight clades proposed by Kasuga et al. 2003, the *H. capsulatum* complex harbors at least reciprocally monophyletic clades (Teixeira et al. 2016).

Recent investigation into the population genomics of the *Histoplasma* genus suggests that there at least four species that are genetically isolated and rarely interbreed with other species. There are at least five discrete genetic clusters within the genus; these clusters appear to cluster based on geography. There are two North American clusters (NAm 1 and NAm 2), Latin American cluster (LAm A), Panamanian cluster, and an African cluster (Sepulveda et al. 2017). These clusters seem to have little genetic overlap (allele frequencies) and little gene flow between clusters. Using whole genome analysis, it was found that all genetic clusters appear to be monophyletic and that the concordance factors are high enough to consider each clade a different species (Sepulveda et al. 2017). It was calculated that the divergence between the different cryptic species is approximately 1.7 million years ago. Gene flow and the exchange of genes between the cryptic species were measured and found that some gene flow and hybridization have occurred but the magnitude is too small to impede species limitations.

In order to better understand the taxonomy and species boundaries in *H. capsulatum* sensu lato, phylogenomic and population-based approaches were employed using whole genome sequencing. By using non-discordance tree profiles, four phylogenetic species diagnosed by Kasuga et al. (2003) were formally described: *H. capsulatum* sensu stricto (formerly known as Panama/H81 lineage), *Histoplasma mississippiense* (formerly known as NAm 1), *Histoplasma ohiense* (formerly known as NAm 2), and *Histoplasma suramericanum* (formerly known as LAm A). These authors showed that those species have limited gene flow, and introgression may also take place between species since *Histoplasma mississippiense* and *Histoplasma ohiense* are potentially sympatric (Sepulveda et al. 2017). Phenotypic difference between species from *Histoplasma* occurs (Sepulveda et al. 2014), but an assessment of relationships between the newly described species and clinical/mycological phenotypes is needed.

Histoplasma spp. are also haploid fungi, and the first three genome assemblies (strains 186R, 217B, and WU24) for this genus were published by Sharpton et al. (2009) along with two *Coccidioides* genomes. The genome of *Histoplasma ohiense* (NAm 1) assembled into 33 Mb and possesses 9,390 genes. When compared to other *Onygenales*, *Histoplasma* did not present the same protease expansions observed in *Coccidioides*. However, a reduction of plant-degrading enzymes was also observed

as for other *Onygenales* (Sharpton et al. 2009; Desjardins et al. 2011). Other genomes were also sequenced and annotated such as the *H. capsulatum* African clade (H143 – 33.17 Mb and 9,547 genes; and H88 – 37.68 Mb and 9,445 genes), *H. capsulatum* sensu stricto (G186AR – 30.28 Mb and 9,254 genes), and a second *H. ohiense* (41.28 Mb but no annotation data is available – see https://genome.jgi.doe.gov/Hisca1/Hisca1.home.html). No comparative genomic studies aiming to compare *H. capsulatum* with related species have been executed so far.

Blastomyces is a genus of fungi responsible for pulmonary infections in areas of the United States and Canada surrounded by the Ohio and Mississippi River valleys as well as the Great Lakes (Castillo et al. 2016). This fungus also lives as a saprobe and produces blastoconidia that upon inhalation by humans and other mammals (such as dogs) differentiates into the yeast pathogenic phase (Smith and Kauffman 2010). The fungus is found mostly in humid soils enriched in organic matter and usually associated with watercourses. Blastomycosis, as other fungal infections caused by dimorphic fungi, varies from a mild pneumonia to a severe disseminated mycosis that may be fatal if not diagnosed and treated correctly, since the disease is often confounded with bacterial and viral infections (Saccente and Woods 2010).

For several years *Blastomyces dermatitidis* was believed to be a single species, and outside the American territory, few cases were observed in Africa. The African blastomycosis is now known to be caused by two closely related fungi in the *Emergomyces* genus (Kenyon et al. 2013). Population genetics and MLST analysis suggested that *Blastomyces* was composed of at least two different species: *B. dermatitidis* and *B. gilchristii*. Phylogeographic data reveals that *B. gilchristii* is restricted to northwestern Ontario, Wisconsin, and Minnesota, and *B. dermatitidis* is found in central and southern Ontario. By using molecular clock analysis on nuclear markers, *B. dermatitidis* and *B. gilchristii* diverged about 1.9 MYA during the Pleistocene. Population analysis reveals that those species harbor cryptic populations and more species within both complexes may exist (Brown et al. 2013a; McTaggart et al. 2016). Recently two other *Blastomyces* species were described as causing disseminated human infections, but not related to any of the previous species. *B. helicus* was isolated from a patient suffering from chronic leukemia with disseminated infection in Canada and is genetically distant from *B. dermatitidis* and *B. gilchristii* (Schwartz et al. 2017). *B. percursus* is another recently discovered species of *Blastomyces*. It was isolated from a granulomatous lip lesion from a patient with severe disseminated infection (Dukik et al. 2017).

A deep comparative genomic analysis between *B. dermatitidis* and *B. gilchristii* was performed, and a better understanding of the overall genetics of these fungi was achieved (Munoz et al. 2015). The species from the *Blastomyces* genus are haploid, and the assembled genomes are double the size of other *Onygenales*. *B. dermatitidis* assemblies ranged from 66.6 Mb for the ER-3 strain to 75.4 Mb for the *B. gilchristii* strain SLH14081. However, the number of genes was predicted to be similar to other *Onygenales*, ranging from 9,180 in the ATCC26199 to 10,187 in the strain ATCC18188. By deeply investigating the chromosomal organization of both *Blastomyces* species, the author provided evidence that these genomes are composed of large isochore-like regions containing high and low GC content. The low GC

content regions are structured into large AT-repeat segments as well as transposable elements (especially *gypsy*) and contain very few genes (Munoz et al. 2015).

Comparative gene family evolution data reveals that fungi in the *Ajellomycetaceae* family have experienced significant loss of polyketide synthase (PKS) domains and consequently fewer PKS gene clusters compared to other *Onygenales*. Moreover, the *Ajellomycetaceae* have fewer classes of peptidases (M36, M43, S8) as well as its associated inhibitor (I9, inhibitor of S8 protease), variable copy number of LysM-domain proteins, and a higher number of fungal-specific protein kinases (FunK1). Finally, *Blastomyces* and *Coccidioides* have gained specific genes required for zinc uptake while such genes are depleted in *Paracoccidioides* and *Histoplasma*, suggesting that those fungi have evolved different mechanisms for zinc acquisition (Munoz et al. 2015, 2018).

Emmonsia is a genus nested within the *Ajellomycetaceae* family that comprises mammalian fungal pathogens rarely associated with human infections (Schwartz et al. 2015). *Emmonsia* spp. are filamentous and saprotrophic fungi and the causative agents of adiaspiromycosis (Anstead et al. 2012). *E. parva* was first isolated and characterized by Chester Emmons as *Chrysosporium parvum* from rodents in Arizona (Emmons and Ashburn 1942). The first human case was reported in France in 1964, and the disease is also found in Honduras, Brazil, the Czech Republic, Russia, the United States, and Guatemala (Schwartz et al. 2015). The fungus is believed to be found in rodent burrows, which are enriched in animal-derived material and the probable niche for this fungal species. By using phylogenetic analysis and mating assays, *Emmonsia parva* was split into two different species (Sigler 1996). Those two species were reciprocally monophyletic and sexually not compatible. Both species are thermo-dimorphic; however, *E. crescens* produces classical budding yeasts, while *E. parva* usually produces adiaspores (enlarged double-walled yeast-like cells) that may reach 200 um in diameter (Dukik et al. 2017; Sigler 1996).

Recently, *Emmonsia*-like organisms were isolated from invasive human infections, and by using phylogenetic analysis, a new species was proposed: *Emmonsia pasteuriana* (Kenyon et al. 2013). These studies are extremely important in addressing questions about disseminated yeast-like infections in HIV patients caused by dimorphic fungi that aren't classical *Blastomyces* and *Histoplasma* (Schwartz et al. 2015). *Histoplasma* and *Blastomyces* also produce small (2–5 um) yeast cells during infection and cannot be discriminated by classical mycological methods from other infections such as *E. pasteuriana* and *E. crescens*. With increased attention on these emerging pathogens, new phylogenetic and phylogenomic methods were applied to better understand the systematics and taxonomy of new *Emmonsia*-like fungus that was isolated from patients worldwide. Phylogenetic analysis of five loci (ITS, LSU, rPB2, TEF3, and TUB2) revealed that *E. crescens* was grouped in a single branch, while *E. parva* was closely related with other *Blastomyces* species such as *B. dermatitidis* and *B. gilchristii* (Dukik et al. 2017). Kenyon et al. (2013) reported another disease among HIV-positive patients in South Africa caused by an *Emmonsia*-like fungus different from *E. pasteuriana*. Using the same five loci and whole genome phylogenies, Dukik et al. (2017) proposed a new genus called

Emergomyces that contained *E. pasteurianus* and the new fungus reported by Kenyon et al. (2013), *Emergomyces africanus*. Given the attention to this new fungal taxon, novel species within *Emergomyces* were described: *Emergomyces orientalis* (Wang et al. 2017a) and *Emergomyces canadensis* (Schwartz et al. 2018). So far, 80 human infections with *E. africanus* in South Africa have been reported, mostly immunocompromised patients with cutaneous dissemination in 95% of patients and mortality reaching 50% (Maphanga et al. 2017). Based on phylogenetic analysis, *E. crescens* maintains a single branch of *Ajellomycetaceae*, *E. parva* was placed within the *Blastomyces*, and a new genus (*Emergomyces*) was proposed that includes *E. pasteurianus*, *E. africanus*, *E. orientalis*, and *E. canadensis.* This species needs special attention due to its rapid emergence and high rates of mortality among HIV-infected patients (Crombie et al. 2018).

The estimated genome sizes of *E. parva* and *E. crescens* were deduced based on genome assemblies totaling 30.35 and 30.36 Mb, respectively, which is a similar size range to other onygenalean species but about half the size of *B. dermatitidis* and *B. gilchristii* genomes (Munoz et al. 2015). The number of predicted genes for *E. parva* and *E. crescens* is similar to other *Onygenales*, 8,563 and 9,444, respectively, and both exhibit a lower repeat sequence content as compared to *Blastomyces* sp., 9.9% (3.0 Mb) and 5.4% (1.6 Mb), respectively. Because *Emmonsia* has a lower pathogenicity profile as compared to *Blastomyces*, a comparative genomic analysis between *Blastomyces* sp. and *Emmonsia* sp. was conducted. At least 552 orthologs were identified in all *Blastomyces* isolates and found absent in both *Emmonsia* isolates where 92% of unique hits for *Blastomyces* had no PFAM domain records and included the *Blastomyces* yeast phase-specific gene 1 (BYS1). One gene absent only in *E. crescens* was the siderophore iron transporter mirB, while most of the dimorphic fungal pathogens harbor two iron transporters (mirB and mirC) and are overexpressed under iron starvation (Munoz et al. 2015). More recently the genomes of two *Emergomyces* species were sequenced, and the assemblies indicate that the genome of *E. africanus* harbors 29.7 Mb and 32.4 Mb in *E. pasteurianus*. As expected for other *Onygenales*, 8,769 and 8,950 protein-coding genes were discovered for *E. africanus* and *E. pasteurianus*, respectively (Dukik et al. 2017). To date, no comparative genomics has been performed exclusively for the genus *Emergomyces*.

Dermatophytes are a diverse group of onygenalean fungi that include human and animal pathogens, as well as free-living environmental fungi (Persinoti et al. 2018). *Trichophyton rubrum* is the most common human pathogen that infects the skin of immunocompetent individuals (causative agent of athlete's foot). There are other species of dermatophytes that are localized to other areas of the body including *T. tonsurans* and *Microsporum canis*, which infect the skin on the head (Kohler et al. 2017). The economic impact of dermatophytes is vast but not well catalogued; one estimate shows that 500 million dollars was spent every year on the treatment of these infections worldwide in the 1990s (Kane et al. 1997).

The genomes of dermatophytes are enriched with four gene classes that may contribute to the ability to infect the skin of many organisms. These gene classes include genes that encode for secreted proteases (degradation of skin), kinases

(involved in signaling and the adaption to skin), secondary metabolites (involved in the interactions between fungus and host), and LysM proteins that mask cell wall components to avoid detection by the host immune system (Martinez et al. 2012). There are homologous of Argonaute and dicer genes in all dermatophyte species, indicating that RNA interference (constrains transposable elements within the genome) is an important adaptation in the infection process. When the genomes of environmental dermatophytes were compared to zoophiles and anthropophiles, there were orthologs unique to the pathogens and absent in the environmental organisms. These are kinase domains, secondary metabolism domains, major facilitator superfamily I (MFS-1) (which is related to secondary metabolite production), and the production of zinc finger proteins (Martinez et al. 2012). These orthologs suggest that signaling and regulation may determine the ecological niche and host specificity of these fungi.

The sexual life cycle has been observed in some dermatophyte species but not in human pathogens including *T. rubrum* (Persinoti et al. 2018). The presence of one of two of the idiomorphs at a single mating-type (MAT) locus, which is the case in other ascomycete species, determines mating type (Fraser and Heitman 2003). In environmental dermatophytes, such as *M. gypseum*, sexual reproduction occurs when isolates of opposite mating type (MAT1-1 and MAT1-2) mate and produce recombinant meiospores (Li et al. 2010). Recently, 79 out of 80 isolates of the *T. rubrum* species complex were shown to contain the α-domain gene at the MAT1-1 locus, and a small subset of a Turkish population contained HMG gene at the MAT1-2 (Persinoti et al. 2018). This suggests that *T. rubrum* has the ability to undergo sexual recombination, but due to a high frequency of the MAT1-1 mating type circulating within populations, it is likely that clonal growth is the main mode of reproduction.

Members of the dermatophyte family have additional copies of genes that may aid in drug resistance and evading the host immune system. ERG4 is a gene that encodes for an enzyme that catalyzes the final step in the biosynthesis of ergosterol, an important component of the fungal cell wall (Martinez et al. 2012). *T. interdigitale* has an extra copy of the ERG4 gene, which indicates that more enzyme is being synthesized and the reaction is not rate limited, suggesting that this could lead to resistance of antifungals that target the biosynthesis of ergosterol by outcompeting competitive inhibitors (Persinoti et al. 2018). LysM-domain proteins are involved in dulling the host recognition of chitin, which is present in the fungal cell wall (De Jonge et al. 2010). Dermatophyte species have increased copy numbers of these LysM genes than closely related onygenalean fungi (Martinez et al. 2012). Specifically, *M. canis* has 31 copies of LysM genes, and *T. rubrum* has 16–18 copies, one of which encodes for a polysaccharide deacetylase involved in the breakdown of chitin, suggesting that changing the cell wall may decrease recognition by host immune cells and is an important way that these pathogens evade the host immune system (Persinoti et al. 2018).

6 Human Fungal Pathogens from the *Pneumocystis* Genus

Pneumocystis species are a group of opportunistic fungal pathogens of mammals. These pathogens can infect humans and can cause severe pneumonia, most frequently in individuals who are immunocompromised (Hawksworth 2007). These obligate pathogens require a living host in order to proliferate. The life cycle consists of a metabolically active trophic stage and asci stage which consists of sexual spores (the sexual cycle takes place only within mammals) (Cushion et al. 2018). *Pneumocystis jirovecii* is the main species that infects humans. Infections by this fungus have high mortality rates in immunocompromised people (Wang et al. 2017b). These infections are clinically complicated to treat due to the lack of treatment options and the evolution of drug-resistant strains (Cisse et al. 2018).

Fungal parasitism generally materializes in two ways, necrotrophy (feeding on dead host cells) and biotrophy (feeding on host cells when host is still alive (Kemen and Jones 2012)). *P. jirovecii* lives almost exclusively in human lungs where it feeds on living cells without causing extensive cell death (Cisse et al. 2014). The genome of *P. jirovecii* reveals a lack of virulence machinery that is usually common among fungal pathogens such as a glyoxylate cycle, toxin-producing pathways, and secondary metabolites (Cushion et al. 2007). This suggests that gene loss contributed to the evolution of biotrophy in *P. jirovecii*.

Comparative genomics was used to investigate the hypothesis that gene loss may be associated with adaptation to biotrophy in *Pneumocystis.* It was shown that there are 2,324 genes that are present in the most recent common ancestor of the genus but are lost in *Pneumocystis* spp., and out of 183 enzymes encoded by these genes, 42% were involved in amino acid and purine metabolism (Cisse et al. 2014). This implies that *P. jirovecii* scavenges these compounds, which it is unable to synthesize, from the lungs of the host it is infecting. Another 19% of the identified enzymes that are lost in *Pneumocystis* are used in purine (synthesis and degradation of adenine and guanine) metabolism; however the precursors of purine and pyrimidine (inosine 5-phosphate and uridine 5-phosphate) synthesis were identified. The actual enzymes to complete purine synthesis were not identified in *P. jirovecii* (Cisse et al. 2014).

The metabolism of inorganic nitrogen and sulfur is essential for life, but *Pneumocystis* spp. lack the key enzymes (nitrite and sulfite reductases) needed for nitrogen and sulfur metabolism (Cisse et al. 2014). *Pneumocystis* has a lower number of proteases used for the breakdown of the extracellular matrix in the lungs, which was thought to be an important virulence factor for fungal pathogens. *P. jirovecii*, however, has a Clp protease that is involved in the hydrolysis of proteins (Cisse et al. 2014). The evolution of biotrophy in the genus *Pneumocystis* likely contributed to the loss of certain gene families (Cisse et al. 2018). The ability to acquire nutrients and molecules from the host, rather than synthesize de novo, is evidence that these fungi are evolved to be obligate pathogens, and the results from this analysis of genome reduction is consistent with this lifestyle.

7 Future Directions

In this chapter, we surveyed the genomic structure and composition of several human fungal pathogens. Each pathogen discussed has unique mechanisms to evade the immune system of humans. However, shared evolutionary patterns such as streamlined genomes and reduction of plant-based carbohydrate degrading enzymes and expansion of gene families related to adhesion, infection, and degradation of host cells are shared among several species of distantly related genera. We observed that these fungi could generate genetic diversity using either sexual or parasexual cycles. Thus, complex population genetic structure may result in a wide range of phenotypes, including variation in virulence and antifungal resistance. Specific point mutations and loss of heterozygosity play important roles in antifungal resistance and other medically relevant phenotypes. In addition to intraspecific variation, the occurrence of hybrid genotypes between species of human fungal pathogens can be revealed with advances in genomic science and may be a novel source of extreme genetic reassortment and phenotypic diversity.

Despite the great advances in comparative genomics and molecular systematics of fungal human pathogens, there are key questions that still need to be answered: (1) despite efforts on global molecular epidemiology, there are broad unsampled regions of disease. South and Central America, Africa, and Asia are still poorly sampled for many of these pathogens. (2) Many reference genomes are based on old sequencing technologies (e.g., Sanger sequencing) or obsolete next-generation sequencing pipelines (e.g., Roche 454, Illumina GAII), and assembly errors are frequent (Munoz et al. 2014). Long-read sequencing using PacBio or Nanopore instruments coupled with high-coverage short-read sequencing has become a powerful technology to close microbial genomes with high repetitive content. (3) Automatic ab initio gene predication and annotation also need attention from the fungal community. The use of mRNA and proteomic-based data for more accurate model prediction of intron-exon boundaries is advised. There are new platforms specific for fungal genome annotation such as FunGAP and funannotate. (4) More phenotypic data is needed to advance functional significance to the exponential growth of genome sequencing projects. Genome-wide association studies (GWAS) are useful to associate loci in a given population genome dataset with phenotypes (e.g., antifungal susceptibility, thermotolerance, dimorphic switch, virulence) using pipelines adapted from other eukaryotic systems. These modern molecular methods will be useful when determining evolutionary relationships of these pathogens, improving rapid diagnostics in a clinical setting, and identifying mechanisms of antifungal resistance and pathogenesis.

References

Almeida F. Estudos comparativos do granuloma coccidioidico nos Estados Unidos e no Brasil. Novo gênero para o parasito brasileiro. An Fac Med S Paulo. 1930;5:18.

Almeida RS, Brunke S, Albrecht A, Thewes S, Laue M, Edwards JE, Filler SG, Hube B. the hyphal-associated adhesin and invasin Als3 of Candida albicans mediates iron acquisition from host ferritin. PLoS Pathog. 2008;4:e1000217.

Alspaugh JA. Virulence mechanisms and Cryptococcus neoformans pathogenesis. Fungal Genet Biol. 2015;78:55–8.

Alves CT, Wei XQ, Silva S, Azeredo J, Henriques M, Williams DW. Candida albicans promotes invasion and colonisation of Candida glabrata in a reconstituted human vaginal epithelium. J Infect. 2014;69:396–407.

Amaike S, Keller NP. Aspergillus flavus. Annu Rev Phytopathol. 2011;49:107–33.

Anstead GM, Sutton DA, Graybill JR. Adiaspiromycosis causing respiratory failure and a review of human infections due to Emmonsia and Chrysosporium spp. J Clin Microbiol. 2012;50:1346–54.

Arendrup MC, Patterson TF. Multidrug-resistant Candida: epidemiology, molecular mechanisms, and treatment. J Infect Dis. 2017;216:S445–51.

Ariyachet C, Solis NV, Liu Y, Prasadarao NV, Filler SG, McBride AE. SR-like RNA-binding protein Slr1 affects Candida albicans filamentation and virulence. Infect Immun. 2013;81:1267–76.

Bagagli E, Bosco SM, Theodoro RC, Franco M. Phylogenetic and evolutionary aspects of Paracoccidioides brasiliensis reveal a long coexistence with animal hosts that explain several biological features of the pathogen. Infect Genet Evol. 2006;6:344–51.

Bahr NC, Antinori S, Wheat LJ, Sarosi GA. Histoplasmosis infections worldwide: thinking outside of the Ohio River valley. Curr Trop Med Rep. 2015;2:70–80.

Barberan A, Ladau J, Leff JW, Pollard KS, Menninger HL, Dunn RR, Fierer N. Continental-scale distributions of dust-associated bacteria and fungi. Proc Natl Acad Sci U S A. 2015;112:5756–61.

Basenko E, Pulman J, Shanmugasundram A, Harb O, Crouch K, Starns D, Warrenfeltz S, Aurrecoechea C, Stoeckert C, Kissinger J, Roos D, Hertz-Fowler C. FungiDB: an integrated bioinformatic resource for fungi and oomycetes. J Fungi. 2018;4:39.

Bennett RJ. Coming of age – sexual reproduction in Candida species. PLoS Pathog. 2010;6: e1001155.

Bennett RJ. The parasexual lifestyle of Candida albicans. Curr Opin Microbiol. 2015;28:10–7.

Billmyre RB, Calo S, Feretzaki M, Wang X, Heitman J. RNAi function, diversity, and loss in the fungal kingdom. Chromosom Res. 2013;21:561–72.

Bocca AL, Amaral AC, Teixeira MM, Sato PK, Shikanai-Yasuda MA, Soares Felipe MS. Paracoccidioidomycosis: eco-epidemiology, taxonomy and clinical and therapeutic issues. Future Microbiol. 2013;8:1177–91.

Boekhout T, Gueidan C, Hoog GS, Samson RA, Varga J, Walther G. Fungal taxonomy: new developments in medically important fungi. Curr Fungal Infect Rep. 2009;3:9.

Bongomin F, Gago S, Oladele RO, Denning DW. Global and multi-national prevalence of fungal diseases-estimate precision. J Fungi (Basel). 2017;3:57.

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.

Bowman BH, Taylor JW, Brownlee AG, Lee J, Lu SD, White TJ. Molecular evolution of the fungi: relationship of the Basidiomycetes, Ascomycetes, and Chytridiomycetes. Mol Biol Evol. 1992a;9:285–96.

Bowman BH, Taylor JW, White TJ. Molecular evolution of the fungi: human pathogens. Mol Biol Evol. 1992b;9:893–904.

Boyce KJ, Andrianopoulos A. Fungal dimorphism: the switch from hyphae to yeast is a specialized morphogenetic adaptation allowing colonization of a host. FEMS Microbiol Rev. 2015;39:797–811.

Brenier-Pinchart MP, Lebeau B, Devouassoux G, Mondon P, Pison C, Ambroise-Thomas P, Grillot R. Aspergillus and lung transplant recipients: a mycologic and molecular epidemiologic study. J Heart Lung Transplant. 1998;17:972–9.

Brown SM, Campbell LT, Lodge JK. Cryptococcus neoformans, a fungus under stress. Curr Opin Microbiol. 2007;10:320–5.

Brown GD, Denning DW, Gow NA, Levitz SM, Netea MG, White TC. Hidden killers: human fungal infections. Sci Transl Med. 2012;4:165rv13.

Brown EM, McTaggart LR, Zhang SX, Low DE, Stevens DA, Richardson SE. Phylogenetic analysis reveals a cryptic species Blastomyces gilchristii, sp. nov. within the human pathogenic fungus Blastomyces dermatitidis. PLoS One. 2013a;8:e59237.

Brown J, Benedict K, Park BJ, Thompson GR 3rd. Coccidioidomycosis: epidemiology. Clin Epidemiol. 2013b;5:185–97.

Brown GD, Meintjes G, Kolls JK, Gray C, Horsnell W, Working Group from The, E.-A. R. M. W, Achan B, Alber G, Aloisi M, Armstrong-James D, Beale M, Bicanic T, Black J, Bohjanen P, Botes A, Boulware DR, Brown G, Bunjun R, Carr W, Casadevall A, Chang C, Chivero E, Corcoran C, Cross A, Dawood H, Day J, De Bernardis F, de Jager V, De Repentigny L, Denning D, Eschke M, Finkelman M, Govender N, Gow N, Graham L, Gryschek R, Hammond-Aryee K, Harrison T, Heard N, Hill M, Hoving JC, Janoff E, Jarvis J, Kayuni S, King K, Kolls J, Kullberg BJ, Lalloo DG, Letang E, Levitz S, Limper A, Longley N, Machiridza TR, Mahabeer Y, Martinsons N, Meiring S, Meya D, Miller R, Molloy S, Morris L, Mukaremera L, Musubire AK, Muzoora C, Nair A, Nakiwala Kimbowa J, Netea M, Nielsen K, O'Hern J, Okurut S, Parker A, Patterson T, Pennap G, Perfect J, Prinsloo C, Rhein J, Rolfes MA, Samuel C, Schutz C, Scriven J, Sebolai OM, Sojane K, Sriruttan C, Stead D, Steyn A, Thawer NK, Thienemann F, von Hohenberg M, Vreulink JM, Wessels J, Wood K, Yang YL. AIDS-related mycoses: the way forward. Trends Microbiol. 2014;22:107–9.

Butler G. Fungal sex and pathogenesis. Clin Microbiol Rev. 2010;23:140–59.

Butler G, Rasmussen MD, Lin MF, Santos MA, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL, Agrafioti I, Arnaud MB, Bates S, Brown AJ, Brunke S, Costanzo MC, Fitzpatrick DA, de Groot PW, Harris D, Hoyer LL, HUBE B, Klis FM, Kodira C, Lennard N, Logue ME, Martin R, Neiman AM, Nikolaou E, Quail MA, Quinn J, Santos MC, Schmitzberger FF, Sherlock G, Shah P, Silverstein KA, Skrzypek MS, Soll D, Staggs R, Stansfield I, Stumpf MP, Sudbery PE, Srikantha T, Zeng Q, Berman J, Berriman M, Heitman J, Gow NA, Lorenz MC, Birren BW, Kellis M, Cuomo CA. Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature. 2009;459:657–62.

Cain CW, Lohse MB, Homann OR, Sil A, Johnson AD. A conserved transcriptional regulator governs fungal morphology in widely diverged species. Genetics. 2012;190:511–21.

Calcagno AM, Bignell E, Warn P, Jones MD, Denning DW, Muhlschlegel FA, Rogers TR, Haynes K. Candida glabrata STE12 is required for wild-type levels of virulence and nitrogen starvation induced filamentation. Mol Microbiol. 2003;50:1309–18.

Camacho E, Nino-Vega GA. Paracoccidioides spp.: virulence factors and immune-evasion strategies. Mediat Inflamm. 2017;2017:5313691. https://doi.org/10.1155/2017/5313691.

Campins H. Coccidioidomycosis in South America. A review of its epidemiology and geographic distribution. Mycopathol Mycol Appl. 1970;41:25–34.

Carreté L, Ksiezopolska E, Pegueroles C, Gómez-Molero E, Saus E, Iraola-Guzmán S, Loska D, Bader O, Fairhead C, Gabaldón T. Patterns of genomic variation in the opportunistic pathogen candida glabrata suggest the existence of mating and a secondary association with humans. Curr Biol. 2018;28:15–27.

Castillo CG, Kauffman CA, Miceli MH. Blastomycosis. Infect Dis Clin N Am. 2016;30:247–64.

Chang YC, Ingavale SS, Bien C, Espenshade P, Kwon-Chung KJ. Conservation of the sterol regulatory element-binding protein pathway and its pathobiological importance in Cryptococcus neoformans. Eukaryot Cell. 2009;8:1770–9.

Charalambous LT, Premji A, Tybout C, Hunt A, Cutshaw D, Elsamadicy AA, Yang S, Xie J, Giamberardino C, Pagadala P, Perfect JR, Lad SP. Prevalence, healthcare resource utilization and overall burden of fungal meningitis in the United States. J Med Microbiol. 2018;67:215–27.

Chastain DB, Henao-Martinez AF, Franco-Paredes C. Opportunistic invasive mycoses in AIDS: cryptococcosis, histoplasmosis, coccidiodomycosis, and talaromycosis. Curr Infect Dis Rep. 2017;19:36.

Chun CD, Liu OW, Madhani HD. A link between virulence and homeostatic responses to hypoxia during infection by the human fungal pathogen Cryptococcus neoformans. PLoS Pathog. 2007;3:e22.

Chung D, Barker BM, Carey CC, Merriman B, Werner ER, Lechner BE, Dhingra S, Cheng C, Xu W, Blosser SJ, Morohashi K, Mazurie A, Mitchell TK, Haas H, Mitchell AP, Cramer RA. ChIP-seq and in vivo transcriptome analyses of the Aspergillus fumigatus SREBP SrbA reveals a new regulator of the fungal hypoxia response and virulence. PLoS Pathog. 2014;10: e1004487.

Cisse OH, Pagni M, Hauser PM. Comparative genomics suggests that the human pathogenic fungus Pneumocystis jirovecii acquired obligate biotrophy through gene loss. Genome Biol Evol. 2014;6:1938–48.

Cisse OH, Ma L, Wei Huang D, Khil PP, Dekker JP, Kutty G, Bishop L, Liu Y, Deng X, Hauser PM, Pagni M, Hirsch V, Lempicki RA, Stajich JE, Cuomo CA, Kovacs JA. Comparative population genomics analysis of the mammalian fungal pathogen pneumocystis. MBio. 2018;9:e00381–18.

Cormack BP, Ghori N, Falkow S. An adhesin of the yeast pathogen Candida glabrata mediating adherence to human epithelial cells. Science. 1999;285:578–82.

Coste A, Selmecki A, Forche A, Diogo D, Bougnoux ME, D'Enfert C, Berman J, Sanglard D. Genotypic evolution of azole resistance mechanisms in sequential Candida albicans isolates. Eukaryot Cell. 2007;6:1889–904.

Cramer RA. Secretion stress and fungal pathogenesis: a new, exploitable chink in fungal armor? Virulence. 2011;2:1–3.

Cramer RA. In vivo veritas: Aspergillus fumigatus proliferation and pathogenesis – conditionally speaking. Virulence. 2016;7:7–10.

Crombie K, Spengane Z, Locketz M, Dlamini S, Lehloenya R, Wasserman S, Maphanga TG, Govender NP, Kenyon C, Schwartz IS. Paradoxical worsening of Emergomyces africanus infection in an HIV-infected male on itraconazole and antiretroviral therapy. PLoS Negl Trop Dis. 2018;12:e0006173.

Cuomo CA. Harnessing whole genome sequencing in medical mycology. Curr Fungal Infect Rep. 2017;11:52–9.

Cuomo CA, Shea T, Yang B, Rao R, Forche A. Whole genome sequence of the heterozygous clinical isolate Candida krusei 81-B-5. G3 (Bethesda). 2017;7:2883–9.

Cushion MT, Smulian AG, Slaven BE, Sesterhenn T, Arnold J, Staben C, Porollo A, Adamczak R, Meller J. Transcriptome of Pneumocystis carinii during fulminate infection: carbohydrate metabolism and the concept of a compatible parasite. PLoS One. 2007;2:e423.

Cushion MT, Ashbaugh A, Hendrix K, Linke MJ, Tisdale N, Sayson SG, Porollo A. Gene expression of Pneumocystis murina after treatment with anidulafungin results in strong signals for sexual reproduction, cell wall integrity, and cell cycle arrest, indicating a requirement for ascus formation for proliferation. Antimicrob Agents Chemother. 2018;62:e02513–7.

D'Souza CA, Kronstad JW, Taylor G, Warren R, Yuen M, Hu G, Jung WH, Sham A, Kidd SE, Tangen K, Lee N, Zeilmaker T, Sawkins J, McVicker G, Shah S, Gnerre S, Griggs A, Zeng Q, Bartlett K, Li W, Wang X, Heitman J, Stajich JE, Fraser JA, Meyer W, Carter D, Schein J, Krzywinski M, Kwon-Chung KJ, Varma A, Wang J, Brunham R, Fyfe M, Ouellette BF, Siddiqui A, Marra M, Jones S, Holt R, Birren BW, Galagan JE, Cuomo CA. Genome variation in Cryptococcus gattii, an emerging pathogen of immunocompetent hosts. MBio. 2011;2: e00342–10.

De Farias MR, Condas LA, Ribeiro MG, Bosco SDM, Muro MD, Werner J, Theodoro RC, Bagagli E, Marques SA, Franco M. Paracoccidioidomycosis in a dog: case report of generalized lymphadenomegaly. Mycopathologia. 2011;172:147–52.

De Jesus-Berrios M, Liu L, Nussbaum JC, Cox GM, Stamler JS, Heitman J. Enzymes that counteract nitrosative stress promote fungal virulence. Curr Biol. 2003;13:1963–8.

De Jonge R, Van Esse HP, Kombrink A, Shinya T, Desaki Y, Bours R, Van Der Krol S, Shibuya N, Joosten MH, Thomma BP. Conserved fungal LysM effector Ecp6 prevents chitin-triggered immunity in plants. Science. 2010;329:953–5.

De Macedo PM, Almeida-Paes R, De Medeiros Muniz M, Oliveira MM, Zancope-Oliveira RM, Costa RL, Do Valle AC. Paracoccidioides brasiliensis PS2: first autochthonous paracoccidioidomycosis case report in Rio de Janeiro, Brazil, and literature review. Mycopathologia. 2016;181:701–8.

Desjardins CA, Champion MD, Holder JW, Muszewska A, Goldberg J, Bailao AM, Brigido MM, Ferreira ME, Garcia AM, Grynberg M, Gujja S, Heiman DI, Henn MR, Kodira CD, Leon-Narvaez H, Longo LV, Ma LJ, Malavazi I, Matsuo AL, Morais FV, Pereira M, Rodriguez-Brito S, Sakthikumar S, Salem-Izacc SM, Sykes SM, Teixeira MM, Vallejo MC, Walter ME, Yandava C, Young S, Zeng Q, Zucker J, Felipe MS, Goldman GH, Haas BJ, McEwen JG, Nino-Vega G, Puccia R, San-Blas G, Soares CM, Birren BW, Cuomo CA. Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. PLoS Genet. 2011;7:e1002345.

Dodgson AR, Pujol C, Denning DW, Soll DR, Fox AJ. Multilocus sequence typing of Candida glabrata reveals geographically enriched clades. J Clin Microbiol. 2003;41:5709–17.

Dukik K, Munoz JF, Jiang Y, Feng P, Sigler L, Stielow JB, Freeke J, Jamalian A, Gerrits Van Den Ende B, McEwen JG, Clay OK, Schwartz IS, Govender NP, Maphanga TG, Cuomo CA, Moreno LF, Kenyon C, Borman AM, De Hoog S. Novel taxa of thermally dimorphic systemic pathogens in the Ajellomycetaceae (Onygenales). Mycoses. 2017;60:296–309.

Edwards JA, Chen C, Kemski MM, Hu J, Mitchell TK, Rappleye CA. Histoplasma yeast and mycelial transcriptomes reveal pathogenic-phase and lineage-specific gene expression profiles. BMC Genomics. 2013;14:695.

Eggimann P, Garbino J, Pittet D. Epidemiology of Candida species infections in critically ill non-immunosuppressed patients. Lancet Infect Dis. 2003;3:685–702.

Emmons CW, Ashburn LL. The isolation of Haplosporangium parvum n. sp and Coccidioides immitis from wild rodents. Their relationship to coccidioidomycosis. Public Health Rep. 1942;57:1715–27.

Engelthaler DM, Hicks ND, Gillece JD, Roe CC, Schupp JM, Driebe EM, Gilgado F, Carriconde F, Trilles L, Firacative C, Ngamskulrungroj P, Castaneda E, Lazera Mdos S, Melhem MS, Perez-Bercoff A, Huttley G, Sorrell TC, Voelz K, May RC, Fisher MC, Thompson GR 3rd, Lockhart SR, Keim P, Meyer W. Cryptococcus gattii in North American Pacific Northwest: whole-population genome analysis provides insights into species evolution and dispersal. MBio. 2014;5:e01464–14.

Engelthaler DM, Roe CC, Hepp CM, Teixeira M, Driebe EM, Schupp JM, Gade L, Waddell V, Komatsu K, Arathoon E, Logemann H, Thompson GR 3rd, Chiller T, Barker B, Keim P, Litvintseva AP. Local population structure and patterns of western hemisphere dispersal for Coccidioides spp., the fungal cause of valley fever. MBio. 2016;7:e00550–16.

Farrer RA, Desjardins CA, Sakthikumar S, Gujja S, Saif S, Zeng Q, Chen Y, Voelz K, Heitman J, May RC, Fisher MC, Cuomo CA. Genome evolution and innovation across the four major lineages of Cryptococcus gattii. MBio. 2015;6:e00868–15.

Fedorova ND, Khaldi N, Joardar VS, Maiti R, Amedeo P, Anderson MJ, Crabtree J, Silva JC, Badger JH, Albarraq A, Angiuoli S, Bussey H, Bowyer P, Cotty PJ, Dyer PS, Egan A, Galens K, Fraser-Liggett CM, Haas BJ, Inman JM, Kent R, Lemieux S, Malavazi I, Orvis J, Roemer T, Ronning CM, Sundaram JP, Sutton G, Turner G, Venter JC, White OR, Whitty BR, Youngman P, Wolfe KH, Goldman GH, Wortman JR, Jiang B, Denning DW, Nierman WC. Genomic islands in the pathogenic filamentous fungus Aspergillus fumigatus. PLoS Genet. 2008;4:e1000046.

Feretzaki M, Billmyre RB, Clancey SA, Wang X, Heitman J. Gene network polymorphism illuminates loss and retention of novel RNAi silencing components in the Cryptococcus pathogenic species complex. PLoS Genet. 2016;12:e1005868.

Fianchi L, Picardi M, Cudillo L, Corvatta L, Mele L, Trape G, Girmenia C, Pagano L. Aspergillus niger infection in patients with haematological diseases: a report of eight cases. Mycoses. 2004;47:163–7.

Filler SG. Candida-host cell receptor-ligand interactions. Curr Opin Microbiol. 2006;9:333–9.

Fisher MC, Koenig GL, White TJ, San-Blas G, Negroni R, Alvarez IG, Wanke B, Taylor JW. Biogeographic range expansion into South America by Coccidioides immitis mirrors New World patterns of human migration. Proc Natl Acad Sci U S A. 2001;98:4558–62.

Fisher MC, Koenig GL, White TJ, Taylor JW. Molecular and phenotypic description of Coccidioides posadasii sp. nov., previously recognized as the non-California population of Coccidioides immitis. Mycologia. 2002;94:73–84.

Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, McCraw SL, Gurr SJ. Emerging fungal threats to animal, plant and ecosystem health. Nature. 2012;484:186–94.

Flowers SA, Colon B, Whaley SG, Schuler MA, Rogers PD. Contribution of clinically derived mutations in ERG11 to azole resistance in Candida albicans. Antimicrob Agents Chemother. 2015;59:450–60.

Fraser JA, Heitman J. Fungal mating-type loci. Curr Biol. 2003;13:R792–5.

Fraser JA, Giles SS, Wenink EC, Geunes-Boyer SG, Wright JR, Diezmann S, Allen A, Stajich JE, Dietrich FS, Perfect JR, Heitman J. Same-sex mating and the origin of the Vancouver Island Cryptococcus gattii outbreak. Nature. 2005;437:1360–4.

Fu C, Sun S, Billmyre RB, Roach KC, Heitman J. Unisexual versus bisexual mating in Cryptococcus neoformans: consequences and biological impacts. Fungal Genet Biol. 2015;78:65–75.

Gabaldon T, Martin T, Marcet-Houben M, Durrens P, Bolotin-Fukuhara M, Lespinet O, Arnaise S, Boisnard S, Aguileta G, Atanasova R, Bouchier C, Couloux A, Creno S, Almeida Cruz J, Devillers H, Enache-Angoulvant A, Guitard J, Jaouen L, Ma L, Marck C, Neuveglise C, Pelletier E, Pinard A, Poulain J, Recoquillay J, Westhof E, Wincker P, Dujon B, Hennequin C, Fairhead C. Comparative genomics of emerging pathogens in the Candida glabrata clade. BMC Genomics. 2013;14:623.

Gabaldon T, Naranjo-Ortiz MA, Marcet-Houben M. Evolutionary genomics of yeast pathogens in the Saccharomycotina. FEMS Yeast Res. 2016;16 https://doi.org/10.1093/femsyr/fow064.

Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglou S, Lee SI, Basturkmen M, Spevak CC, Clutterbuck J, Kapitonov V, Jurka J, Scazzocchio C, Farman M, Butler J, Purcell S, Harris S, Braus GH, Draht O, Busch S, D'Enfert C, Bouchier C, Goldman GH, Bell-Pedersen D, Griffiths-Jones S, Doonan JH, Yu J, Vienken K, Pain A, Freitag M, Selker EU, Archer DB, Penalva MA, Oakley BR, Momany M, Tanaka T, Kumagai T, Asai K, Machida M, Nierman WC, Denning DW, Caddick M, Hynes M, Paoletti M, Fischer R, Miller B, Dyer P, Sachs MS, Osmani SA, Birren BW. Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. Nature. 2005;438:1105–15.

Gaona-Flores VA, Campos-Navarro LA, Cervantes-Tovar RM, Alcala-Martinez E. The epidemiology of fungemia in an infectious diseases hospital in Mexico city: a 10-year retrospective review. Med Mycol. 2016;54:600–4.

Gauthier GM. Dimorphism in fungal pathogens of mammals, plants, and insects. PLoS Pathog. 2015;11:e1004608.

Gegembauer G, Araujo LM, Pereira EF, Rodrigues AM, Paniago AM, Hahn RC, De Camargo ZP. Serology of paracoccidioidomycosis due to Paracoccidioides lutzii. PLoS Negl Trop Dis. 2014;8:e2986.

Geiser DM, Dorner JW, Horn BW, Taylor JW. The phylogenetics of mycotoxin and sclerotium production in Aspergillus flavus and Aspergillus oryzae. Fungal Genet Biol. 2000;31:169–79.

Gillece JD, Schupp JM, Balajee SA, Harris J, Pearson T, Yan Y, Keim P, Debess E, Marsden-Haug N, Wohrle R, Engelthaler DM, Lockhart SR. Whole genome sequence analysis

of Cryptococcus gattii from the Pacific Northwest reveals unexpected diversity. PLoS One. 2011;6:e28550.
Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. Life with 6,000 genes. Science. 1996;274(546):563–7.
Gomez-Raja J, Andaluz E, Magee B, Calderone R, Larriba G. A single SNP, G929T (Gly310Val), determines the presence of a functional and a non-functional allele of HIS4 in Candida albicans SC5314: detection of the non-functional allele in laboratory strains. Fungal Genet Biol. 2008;45:527–41.
Grahl N, Shepardson KM, Chung D, Cramer RA. Hypoxia and fungal pathogenesis: to air or not to air? Eukaryot Cell. 2012;11:560–70.
Grigg ME, Bonnefoy S, Hehl AB, Suzuki Y, Boothroyd JC. Success and virulence in Toxoplasma as the result of sexual recombination between two distinct ancestries. Science. 2001;294:161–5.
Hage CA, Knox KS, Wheat LJ. Endemic mycoses: overlooked causes of community acquired pneumonia. Respir Med. 2012;106:769–76.
Hagen F, Khayhan K, Theelen B, Kolecka A, Polacheck I, Sionov E, Falk R, Parnmen S, Lumbsch HT, Boekhout T. Recognition of seven species in the Cryptococcus gattii/Cryptococcus neoformans species complex. Fungal Genet Biol. 2015;78:16–48.
Hardie WD, Glasser SW, Hagood JS. Emerging concepts in the pathogenesis of lung fibrosis. Am J Pathol. 2009;175:3–16.
Hawksworth DL. Pandora's mycological box: molecular sequences vs. morphology in understanding fungal relationships and biodiversity. Rev Iberoam Micol. 2006;23:127–33.
Hawksworth DL. Responsibility in naming pathogens: the case of Pneumocystis jirovecii, the causal agent of pneumocystis pneumonia. Lancet Infect Dis. 2007;7:3–5; discussion 5.
Hawksworth DL, Lucking R. Fungal diversity revisited: 2.2 to 3.8 million species. Microbiol Spectr. 2017;5 https://doi.org/10.1128/microbiolspec.FUNK-0052-2016.
Hawksworth DL, Rossman AY. Where are all the undescribed fungi? Phytopathology. 1997;87:888–91.
Hedayati MT, Pasqualotto AC, Warn PA, Bowyer P, Denning DW. Aspergillus flavus: human pathogen, allergen and mycotoxin producer. Microbiology. 2007;153:1677–92.
Heitman J. Evolution of eukaryotic microbial pathogens via covert sexual reproduction. Cell Host Microbe. 2010;8:86–99.
Hirakawa MP, Chyou DE, Huang D, Slan AR, Bennett RJ. Parasex generates phenotypic diversity de novo and impacts drug resistance and virulence in Candida albicans. Genetics. 2017;207:1195–211.
Holland LM, Schroder MS, Turner SA, Taff H, Andes D, Grozer Z, Gacser A, Ames L, Haynes K, Higgins DG, Butler G. Comparative phenotypic analysis of the major fungal pathogens Candida parapsilosis and Candida albicans. PLoS Pathog. 2014;10:e1004365.
Hrycyk MF, Garcia Garces H, Bosco SMG, De Oliveira SL, Marques SA, Bagagli E. Ecology of Paracoccidioides brasiliensis, P. Lutzii and related species: infection in armadillos, soil occurrence and mycological aspects. Med Mycol. 2018; https://doi.org/10.1093/mmy/myx142.
Huang SH, Wu CH, Chang YC, Kwon-Chung KJ, Brown RJ, Jong A. Cryptococcus neoformans-derived microvesicles enhance the pathogenesis of fungal brain infection. PLoS One. 2012;7:e48570.
Jackson AP, Gamble JA, Yeomans T, Moran GP, Saunders D, Harris D, Aslett M, Barrell JF, Butler G, Citiulo F, Coleman DC, De Groot PW, Goodwin TJ, Quail MA, McQuillan J, Munro CA, Pain A, Poulter RT, Rajandream MA, Renauld H, Spiering MJ, Tivey A, Gow NA, Barrell B, Sullivan DJ, Berriman M. Comparative genomics of the fungal pathogens Candida dubliniensis and Candida albicans. Genome Res. 2009;19:2231–44.
Jacobsen MD, Gow NAR, Maiden MCJ, Shaw DJ, Odds FC. Strain typing and determination of population structure of Candida krusei by multilocus sequence typing. J Clin Microbiol. 2007;45:317–23.

Janbon G, Ormerod KL, Paulet D, Byrnes EJ 3rd, Yadav V, Chatterjee G, Mullapudi N, Hon CC, Billmyre RB, Brunel F, Bahn YS, Chen W, Chen Y, Chow EW, Coppee JY, Floyd-Averette A, Gaillardin C, Gerik KJ, Goldberg J, Gonzalez-Hilarion S, Gujja S, Hamlin JL, Hsueh YP, Ianiri G, Jones S, Kodira CD, Kozubowski L, Lam W, Marra M, Mesner LD, Mieczkowski PA, Moyrand F, Nielsen K, Proux C, Rossignol T, Schein JE, Sun S, Wollschlaeger C, Wood IA, Zeng Q, Neuveglise C, Newlon CS, Perfect JR, Lodge JK, Idnurm A, Stajich JE, Kronstad JW, Sanyal K, Heitman J, Fraser JA, Cuomo CA, Dietrich FS. Analysis of the genome and transcriptome of Cryptococcus neoformans var. grubii reveals complex RNA expression and microevolution leading to virulence attenuation. PLoS Genet. 2014;10:e1004261.

Jiménez-Ortigosa C, Perez WB, Angulo D, Borroto-Esoda K, Perlin DS. *De novo* acquisition of resistance to SCY-078 in *Candida glabrata* involves FKS mutations that both overlap and are distinct from those conferring echinocandin resistance. Antimicrob Agents Chemother. 2017;61:e00833–17. https://doi.org/10.1128/AAC.00833-17.

Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT, Davis RW, Scherer S. The diploid genome sequence of Candida albicans. Proc Natl Acad Sci U S A. 2004;101:7329–34.

Jong A, Wu CH, Prasadarao NV, Kwon-Chung KJ, Chang YC, Ouyang Y, Shackleford GM, Huang SH. Invasion of Cryptococcus neoformans into human brain microvascular endothelial cells requires protein kinase C-alpha activation. Cell Microbiol. 2008;10:1854–65.

Kane J, Summerbell R, Sigler L, Krajden S, Land G. Laboratory handbook of dermatophytes. A clinical guide and laboratory manual of dermatophytes and other filamentous fungi from skin, hair and nails. Belmont: Star; 1997.

Kasuga T, White TJ, Koenig G, McEwen J, Restrepo A, Castaneda E, Da Silva Lacaz C, Heins-Vaccari EM, De Freitas RS, Zancope-Oliveira RM, Qin Z, Negroni R, Carter DA, Mikami Y, Tamura M, Taylor ML, Miller GF, Poonwan N, Taylor JW. Phylogeography of the fungal pathogen Histoplasma capsulatum. Mol Ecol. 2003;12:3383–401.

Kauffman CA. Histoplasmosis: a clinical and laboratory update. Clin Microbiol Rev. 2007;20:115–32.

Kaur R, Domergue R, Zupancic ML, Cormack BP. A yeast by any other name: Candida glabrata and its interaction with the host. Curr Opin Microbiol. 2005;8:378–84.

Kemen E, Jones JD. Obligate biotroph parasitism: can we link genomes to lifestyles? Trends Plant Sci. 2012;17:448–57.

Kent LB, Juneann WM. What makes Cryptococcus neoformans a pathogen? Emerg Infect Dis J. 1998;4:71.

Kenyon C, Bonorchis K, Corcoran C, Meintjes G, Locketz M, Lehloenya R, Vismer HF, Naicker P, Prozesky H, Van Wyk M, Bamford C, Du Plooy M, Imrie G, Dlamini S, Borman AM, Colebunders R, Yansouni CP, Mendelson M, Govender NP. A dimorphic fungus causing disseminated infection in South Africa. N Engl J Med. 2013;369:1416–24.

Kidd SE, Hagen F, Tscharke RL, Huynh M, Bartlett KH, Fyfe M, MacDougall L, Boekhout T, Kwon-Chung KJ, Meyer W. A rare genotype of Cryptococcus gattii caused the cryptococcosis outbreak on Vancouver Island (British Columbia, Canada). Proc Natl Acad Sci U S A. 2004;101:17258–63.

Kohler JR, Hube B, Puccia R, Casadevall A, Perfect JR. Fungi that infect humans. Microbiol Spectr. 2017;5 https://doi.org/10.1128/microbiolspec.FUNK-0014-2016.

Kronstad J, Saikia S, Nielson ED, Kretschmer M, Jung W, Hu G, Geddes JM, Griffiths EJ, Choi J, Cadieux B, Caza M, Attarian R. Adaptation of Cryptococcus neoformans to mammalian hosts: integrated regulation of metabolism and virulence. Eukaryot Cell. 2012;11:109–18.

Kwon-Chung KJ. Sexual stage of Histoplasma capsulatum. Science. 1972;175:326.

Kwon-Chung KJ, Sugui JA. Aspergillus fumigatus – what makes the species a ubiquitous human fungal pathogen? PLoS Pathog. 2013;9:e1003743.

Kwon-Chung KJ, Fraser JA, Doering TL, Wang Z, Janbon G, Idnurm A, Bahn YS. Cryptococcus neoformans and Cryptococcus gattii, the etiologic agents of cryptococcosis. Cold Spring Harb Perspect Med. 2014;4:a019760.

Lengeler KB, Cox GM, Heitman J. Serotype Ad strains of Cryptococcus neoformans are diploid or aneuploid and are heterozygous at the mating-type locus. Infect Immun. 2001;69:115–22.

Lewis ER, Bowers JR, Barker BM. Dust devil: the life and times of the fungus that causes valley fever. PLoS Pathog. 2015;11:e1004762.

Li W, Metin B, White TC, Heitman J. Organization and evolutionary trajectory of the mating type (MAT) locus in dermatophyte and dimorphic fungal pathogens. Eukaryot Cell. 2010;9:46–58.

Li J, Chang YC, Wu CH, Liu J, Kwon-Chung KJ, Huang SH, Shimada H, Fante R, Fu X, Jong A. The 14-3-3 gene function of Cryptococcus neoformans is required for its growth and virulence. J Microbiol Biotechnol. 2016;26:918–27.

Lin SJ, Schranz J, Teutsch SM. Aspergillosis case-fatality rate: systematic review of the literature. Clin Infect Dis. 2001;32:358–66.

Litvintseva AP, Marsden-Haug N, Hurst S, Hill H, Gade L, Driebe EM, Ralston C, Roe C, Barker BM, Goldoft M, Keim P, Wohrle R, Thompson GR 3rd, Engelthaler DM, Brandt ME, Chiller T. Valley fever: finding new places for an old disease: Coccidioides immitis found in Washington state soil associated with recent human infection. Clin Infect Dis. 2015;60:e1–3.

Liu Y, Filler SG. Candida albicans Als3, a multifunctional adhesin and invasin. Eukaryot Cell. 2011;10:168–73.

Liu OW, Chun CD, Chow ED, Chen C, Madhani HD, Noble SM. Systematic genetic analysis of virulence in the human fungal pathogen Cryptococcus neoformans. Cell. 2008;135:174–88.

Lockhart SR, Etienne KA, Vallabhaneni S, Farooqi J, Chowdhary A, Govender NP, Colombo AL, Calvo B, Cuomo CA, Desjardins CA, Berkow EL, Castanheira M, Magobo RE, Jabeen K, Asghar RJ, Meis JF, Jackson B, Chiller T, Litvintseva AP. Simultaneous emergence of multidrug-resistant Candida auris on 3 continents confirmed by whole-genome sequencing and epidemiological analyses. Clin Infect Dis. 2017;64:134–40.

Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA, Allen JE, Bosdet IE, Brent MR, Chiu R, Doering TL, Donlin MJ, D'Souza CA, Fox DS, Grinberg V, Fu J, Fukushima M, Haas BJ, Huang JC, Janbon G, Jones SJ, Koo HL, Krzywinski MI, Kwon-Chung JK, Lengeler KB, Maiti R, Marra MA, Marra RE, Mathewson CA, Mitchell TG, Pertea M, Riggs FR, Salzberg SL, Schein JE, Shvartsbeyn A, Shin H, Shumway M, Specht CA, Suh BB, Tenney A, Utterback TR, Wickes BL, Wortman JR, Wye NH, Kronstad JW, Lodge JK, Heitman J, Davis RW, Fraser CM, Hyman RW. The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. Science. 2005;307:1321–4.

Losada L, Barker BM, Pakala S, Pakala S, Joardar V, Zafar N, Mounaud S, Fedorova N, Nierman WC, Cramer RA. Large-scale transcriptional response to hypoxia in *Aspergillus fumigatus* observed using RNAseq identifies a novel hypoxia regulated ncRNA. Mycopathologia. 2014;178:331–9.

Loza L, Fu Y, Ibrahim AS, Sheppard DC, Filler SG, Edwards JE Jr. Functional analysis of the Candida albicans ALS1 gene product. Yeast. 2004;21:473–82.

Lutz A. A pseudococcidic mycosis localized in the mouth and observed in Brazil. Bras Med. 1908;22:20.

Macpherson S, Akache B, Weber S, De Deken X, Raymond M, Turcotte B. Candida albicans zinc cluster protein Upc2p confers resistance to antifungal drugs and is an activator of ergosterol biosynthetic genes. Antimicrob Agents Chemother. 2005;49:1745–52.

Magee BB, Magee PT. Induction of mating in Candida albicans by construction of MTLa and MTLalpha strains. Science. 2000;289:310–3.

Maguire SL, Oheigeartaigh SS, Byrne KP, Schroder MS, O'Gaora P, Wolfe KH, Butler G. Comparative genome analysis and gene finding in Candida species using CGOB. Mol Biol Evol. 2013;30:1281–91.

Maphanga TG, Britz E, Zulu TG, Mpembe RS, Naicker SD, Schwartz IS, Govender NP. In vitro antifungal susceptibility of yeast and mold phases of isolates of dimorphic fungal pathogen emergomyces africanus (formerly Emmonsia sp.) from HIV-infected South African patients. J Clin Microbiol. 2017;55:1812–20.

Martel CM, Parker JE, Bader O, Weig M, Gross U, Warrilow AG, Rolley N, Kelly DE, Kelly SL. Identification and characterization of four azole-resistant erg3 mutants of Candida albicans. Antimicrob Agents Chemother. 2010;54:4527–33.

Martinez R. New trends in paracoccidioidomycosis epidemiology. J Fungi. 2017;3:1.

Martinez DA, Oliver BG, Graser Y, Goldberg JM, Li W, Martinez-Rossi NM, Monod M, Shelest E, Barton RC, Birch E, Brakhage AA, Chen Z, Gurr SJ, Heiman D, Heitman J, Kosti I, Rossi A, Saif S, Samalova M, Saunders CW, Shea T, Summerbell RC, Xu J, Young S, Zeng Q, Birren BW, Cuomo CA, White TC. Comparative genome analysis of Trichophyton rubrum and related dermatophytes reveals candidate genes involved in infection. MBio. 2012;3:e00259–12.

Matute DR, McEwen JG, Puccia R, Montes BA, San-Blas G, Bagagli E, Rauscher JT, Restrepo A, Morais F, Nino-Vega G, Taylor JW. Cryptic speciation and recombination in the fungus Paracoccidioides brasiliensis as revealed by gene genealogies. Mol Biol Evol. 2006;23:65–73.

May RC, Stone NR, Wiesner DL, Bicanic T, Nielsen K. Cryptococcus: from environmental saprophyte to global pathogen. Nat Rev Microbiol. 2016;14:106–17.

McDonough ES, Lewis AL. Blastomyces dermatitidis: production of the sexual stage. Science. 1967;156:528–9.

McManus BA, Coleman DC. Molecular epidemiology, phylogeny and evolution of Candida albicans. Infect Genet Evol. 2014;21:166–78.

McManus BA, Coleman DC, Moran G, Pinjon E, Diogo D, Bougnoux M-E, Borecká-Melkusova S, et al. Multilocus sequence typing reveals that the population structure of Candida dubliniensis is significantly less divergent than that of Candida albicans. J Clin Microbiol. 2018;46:652–64.

McTaggart LR, Brown EM, Richardson SE. Phylogeographic analysis of Blastomyces dermatitidis and Blastomyces gilchristii reveals an association with North American freshwater drainage basins. PLoS One. 2016;11:e0159396.

Miranda I, Silva-Dias A, Rocha R, Teixeira-Santos R, Coelho C, Goncalves T, Santos MA, Pina-Vaz C, Solis NV, Filler SG, Rodrigues AG. Candida albicans CUG mistranslation is a mechanism to create cell surface variation. MBio. 2013;4:e00285–13.

Munoz JF, Gallo JE, Misas E, Priest M, Imamovic A, Young S, Zeng Q, Clay OK, McEwen JG, Cuomo CA. Genome update of the dimorphic human pathogenic fungi causing paracoccidioidomycosis. PLoS Negl Trop Dis. 2014;8:e3348.

Munoz JF, Gauthier GM, Desjardins CA, Gallo JE, Holder J, Sullivan TD, Marty AJ, Carmen JC, Chen Z, Ding L, Gujja S, Magrini V, Misas E, Mitreva M, Priest M, Saif S, Whiston EA, Young S, Zeng Q, Goldman WE, Mardis ER, Taylor JW, McEwen JG, Clay OK, Klein BS, Cuomo CA. The dynamic genome and transcriptome of the human fungal pathogen Blastomyces and close relative Emmonsia. PLoS Genet. 2015;11:e1005493.

Munoz JF, Farrer RA, Desjardins CA, Gallo JE, Sykes S, Sakthikumar S, Misas E, Whiston EA, Bagagli E, Soares CM, Teixeira MM, Taylor JW, Clay OK, McEwen JG, Cuomo CA. Genome diversity, recombination, and virulence across the major lineages of Paracoccidioides. mSphere. 2016;1 https://doi.org/10.1128/mSphere.00213-16.

Munoz JF, McEwen JG, Clay OK, Cuomo CA. Genome analysis reveals evolutionary mechanisms of adaptation in systemic dimorphic fungi. Sci Rep. 2018;8:4473.

Naglik JR, Challacombe SJ, Hube B. Candida albicans secreted aspartyl proteinases in virulence and pathogenesis. Microbiol Mol Biol Rev. 2003a;67:400–28. table of contents.

Naglik JR, Rodgers CA, Shirlaw PJ, Dobbie JL, Fernandes-Naglik LL, Greenspan D, Agabian N, Challacombe SJ. Differential expression of Candida albicans secreted aspartyl proteinase and phospholipase B genes in humans correlates with active oral and vaginal infections. J Infect Dis. 2003b;188:469–79.

Neafsey DE, Barker BM, Sharpton TJ, Stajich JE, Park DJ, Whiston E, Hung CY, McMahan C, White J, Sykes S, Heiman D, Young S, Zeng Q, Abouelleil A, Aftuck L, Bessette D, Brown A, Fitzgerald M, Lui A, Macdonald JP, Priest M, Orbach MJ, Galgiani JN, Kirkland TN, Cole GT, Birren BW, Henn MR, Taylor JW, Rounsley SD. Population genomic sequencing of Coccidioides fungi reveals recent hybridization and transposon control. Genome Res. 2010;20:938–46.

Nemecek JC, Wuthrich M, Klein BS. Global control of dimorphism and virulence in fungi. Science. 2006;312:583–8.

Nguyen VQ, Sil A. Temperature-induced switch to the pathogenic yeast form of Histoplasma capsulatum requires Ryp1, a conserved transcriptional regulator. Proc Natl Acad Sci U S A. 2008;105:4880–5.

Ni M, Feretzaki M, Sun S, Wang X, Heitman J. Sex in fungi. Annu Rev Genet. 2011;45:405–30.

Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, Berriman M, Abe K, Archer DB, Bermejo C, Bennett J, Bowyer P, Chen D, Collins M, Coulsen R, Davies R, Dyer PS, Farman M, Fedorova N, Fedorova N, Feldblyum TV, Fischer R, Fosker N, Fraser A, Garcia JL, Garcia MJ, Goble A, Goldman GH, Gomi K, Griffith-Jones S, Gwilliam R, Haas B, Haas H, Harris D, Horiuchi H, Huang J, Humphray S, Jimenez J, Keller N, Khouri H, Kitamoto K, Kobayashi T, Konzack S, Kulkarni R, Kumagai T, Lafon A, Latge JP, Li W, Lord A, Lu C, Majoros WH, May GS, Miller BL, Mohamoud Y, Molina M, Monod M, Mouyna I, Mulligan S, Murphy L, O'Neil S, Paulsen I, Penalva MA, Pertea M, Price C, Pritchard BL, Quail MA, Rabbinowitsch E, Rawlins N, Rajandream MA, Reichard U, Renauld H, Robson GD, Rodriguez De Cordoba S, Rodriguez-Pena JM, Ronning CM, Rutter S, Salzberg SL, Sanchez M, Sanchez-Ferrero JC, Saunders D, Seeger K, Squares R, Squares S, Takeuchi M, Tekaia F, Turner G, Vazquez De Aldana CR, Weidman J, White O, Woodward J, Yu JH, Fraser C, Galagan JE, Asai K, Machida M, Hall N, Barrell B, Denning DW. Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. Nature. 2005;438:1151–6.

O'Gorman CM, Fuller HT, Dyer PS. Discovery of a sexual cycle in the opportunistic fungal pathogen Aspergillus fumigatus. Nature. 2009;457:471–4.

Odio CD, Marciano BE, Galgiani JN, Holland SM. Risk factors for disseminated Coccidioidomycosis, United States. Emerg Infect Dis. 2017;23:308.

Oladele RO, Ayanlowo OO, Richardson MD, Denning DW. Histoplasmosis in Africa: an emerging or a neglected disease? PLoS Negl Trop Dis. 2018;12:e0006046.

Paoletti M, Rydholm C, Schwier EU, Anderson MJ, Szakacs G, Lutzoni F, Debeaupuis JP, Latge JP, Denning DW, Dyer PS. Evidence for sexuality in the opportunistic fungal pathogen Aspergillus fumigatus. Curr Biol. 2005;15:1242–8.

Park BJ, Sigel K, Vaz V, Komatsu K, McRill C, Phelan M, Colman T, Comrie AC, Warnock DW, Galgiani JN, Hajjeh RA. An epidemic of coccidioidomycosis in Arizona associated with climatic changes, 1998-2001. J Infect Dis. 2005;191:1981–7.

Pel HJ, De Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, Turner G, De Vries RP, Albang R, Albermann K, Andersen MR, Bendtsen JD, Benen JA, Van Den Berg M, Breestraat S, Caddick MX, Contreras R, Cornell M, Coutinho PM, Danchin EG, Debets AJ, Dekker P, Van Dijck PW, Van Dijk A, Dijkhuizen L, Driessen AJ, D'Enfert C, Geysens S, Goosen C, Groot GS, De Groot PW, Guillemette T, Henrissat B, Herweijer M, Van Den Hombergh JP, Van Den Hondel CA, Van Der Heijden RT, Van Der Kaaij RM, Klis FM, Kools HJ, Kubicek CP, Van Kuyk PA, Lauber J, Lu X, Van Der Maarel MJ, Meulenberg R, Menke H, Mortimer MA, Nielsen J, Oliver SG, Olsthoorn M, Pal K, Van Peij NN, Ram AF, Rinas U, Roubos JA, Sagt CM, Schmoll M, Sun J, Ussery D, Varga J, Vervecken W, Van De Vondervoort PJ, Wedler H, Wosten HA, Zeng AP, Van Ooyen AJ, Visser J, Stam H. Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88. Nat Biotechnol. 2007;25:221–31.

Persinoti GF, Martinez DA, Li W, Dogen A, Billmyre RB, Averette A, Goldberg JM, Shea T, Young S, Zeng Q, Oliver BG, Barton R, Metin B, Hilmioglu-Polat S, Ilkit M, Graser Y, Martinez-Rossi NM, White TC, Heitman J, Cuomo CA. Whole-genome analysis illustrates global clonal population structure of the ubiquitous dermatophyte pathogen trichophyton rubrum. Genetics. 2018;208:1657–69.

Porman AM, Alby K, Hirakawa MP, Bennett RJ. Discovery of a phenotypic switch regulating sexual mating in the opportunistic fungal pathogen Candida tropicalis. Proc Natl Acad Sci U S A. 2011;108:21158–63.

Queiroz-Telles F, Fahal AH, Falci DR, Caceres DH, Chiller T, Pasqualotto AC. Neglected endemic mycoses. Lancet Infect Dis. 2017;17:e367–77.

Rhodes J, Beale MA, Vanhove M, Jarvis JN, Kannambath S, Simpson JA, Ryan A, Meintjes G, Harrison TS, Fisher MC, Bicanic T. A population genomics approach to assessing the genetic basis of within-host microevolution underlying recurrent cryptococcal meningitis infection. G3 (Bethesda). 2017;7:1165–76.

Rodrigues ML, Alvarez M, Fonseca FL, Casadevall A. Binding of the wheat germ lectin to Cryptococcus neoformans suggests an association of chitinlike structures with yeast budding and capsular glucuronoxylomannan. Eukaryot Cell. 2008;7:602–9.

Rodrigues CF, Silva S, Henriques M. Candida glabrata: a review of its features and resistance. Eur J Clin Microbiol Infect Dis. 2014;33:673–88.

Ronning CM, Fedorova ND, Bowyer P, Coulson R, Goldman G, Kim HS, Turner G, Wortman JR, Yu J, Anderson MJ, Denning DW, Nierman WC. Genomics of Aspergillus fumigatus. Rev Iberoam Micol. 2005;22:223–8.

Ropars J, Maufrais C, Diogo D, Marcet-Houben M, Perin A, Sertour N, Mosca K, et al. Gene flow contributes to diversification of the major fungal pathogen Candida albicans. Nat Commun. 2018;9:2253.

Saccente M, Woods GL. Clinical and laboratory update on blastomycosis. Clin Microbiol Rev. 2010;23:367–81.

Salazar SB, Wang C, Munsterkotter M, Okamoto M, Takahashi-Nakaguchi A, Chibana H, Lopes MM, Guldener U, Butler G, Mira NP. Comparative genomic and transcriptomic analyses unveil novel features of azole resistance and adaptation to the human host in Candida glabrata. FEMS Yeast Res. 2018;18:fox079.

Santos MA, Cheesman C, Costa V, Moradas-Ferreira P, Tuite MF. Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in Candida spp. Mol Microbiol. 1999;31:937–47.

Schuster E, Dunn-Coleman N, Frisvad JC, Van Dijck PW. On the safety of Aspergillus niger – a review. Appl Microbiol Biotechnol. 2002;59:426–35.

Schwartz IS, Kenyon C, Feng P, Govender NP, Dukik K, Sigler L, Jiang Y, Stielow JB, Munoz JF, Cuomo CA, Botha A, Stchigel AM, De Hoog GS. 50 years of Emmonsia disease in humans: the dramatic emergence of a cluster of novel fungal pathogens. PLoS Pathog. 2015;11:e1005198.

Schwartz IS, Wiederhold NP, Patterson TF, Sigler L. Blastomyces helicus, an emerging dimorphic fungal pathogen causing fatal pulmonary and disseminated disease in humans and animals in western Canada and United States. Open Forum Infect Dis. 2017;4:2.

Schwartz IS, Sanche S, Wiederhold NP, Patterson T, Sigler L. Emergomyces canadensis, a dimorphic fungus causing fatal systemic human disease in North America. Emerg Infect Dis. 2018;24:758–61.

Scordino F, Giuffrè L, Barberi G, Merlo FM, Orlando MG, Giosa D, Romeo O. Multilocus sequence typing reveals a new cluster of closely related Candida tropicalis genotypes in Italian patients with neurological disorders. Front Microbiol. 2018;9:679.

Sepulveda VE, Williams CL, Goldman WE. Comparison of phylogenetically distinct Histoplasma strains reveals evolutionarily divergent virulence strategies. MBio. 2014;5:e01376–14.

Sepulveda VE, Marquez R, Turissini DA, Goldman WE, Matute DR. Genome sequences reveal cryptic speciation in the human pathogen Histoplasma capsulatum. MBio. 2017;8:e01339–17.

Sharma KK. Fungal genome sequencing: basic biology to biotechnology. Crit Rev Biotechnol. 2016;36:743–59.

Sharpton TJ, Stajich JE, Rounsley SD, Gardner MJ, Wortman JR, Jordar VS, Maiti R, Kodira CD, Neafsey DE, Zeng Q, Hung CY, McMahan C, Muszewska A, Grynberg M, Mandel MA, Kellner EM, Barker BM, Galgiani JN, Orbach MJ, Kirkland TN, Cole GT, Henn MR, Birren BW, Taylor JW. Comparative genomic analyses of the human fungal pathogens Coccidioides and their relatives. Genome Res. 2009;19:1722–31.

Sheppard DC, Yeaman MR, Welch WH, Phan QT, Fu Y, Ibrahim AS, Filler SG, Zhang M, Waring AJ, Edwards JE Jr. Functional and structural diversity in the Als protein family of Candida albicans. J Biol Chem. 2004;279:30480–9.

Shikanai-Yasuda MA, Mendes RP, Colombo AL, Queiroz-Telles F, Kono ASG, Paniago AM, Nathan A, Valle A, Bagagli E, Benard G, Ferreira MS, Teixeira MM, Silva-Vergara ML, Pereira RM, Cavalcante RS, Hahn R, Durlacher RR, Khoury Z, Camargo ZP, Moretti ML, Martinez R. Brazilian guidelines for the clinical management of paracoccidioidomycosis. Rev Soc Bras Med Trop. 2017;50:715–40.

Sigler L. Ajellomyces crescens sp. nov., taxonomy of Emmonsia spp., and relatedness with Blastomyces dermatitidis (teleomorph Ajellomyces dermatitidis). J Med Vet Mycol. 1996;34:303–14.

Sil A, Andrianopoulos A. Thermally dimorphic human fungal pathogens – polyphyletic pathogens with a convergent pathogenicity trait. Cold Spring Harb Perspect Med. 2014;5:a019794.

Silva AP, Miranda IM, Guida A, Synnott J, Rocha R, Silva R, Amorim A, Pina-Vaz C, Butler G, Rodrigues AG. Transcriptional profiling of azole-resistant Candida parapsilosis strains. Antimicrob Agents Chemother. 2011;55:3546–56.

Smith JA, Kauffman CA. Blastomycosis. Proc Am Thorac Soc. 2010;7:173–80.

Stephen C, Lester S, Black W, Fyfe M, Raverty S. Multispecies outbreak of cryptococcosis on southern Vancouver Island, British Columbia. Can Vet J. 2002;43:792–4.

Takahashi H, Kusuya Y, Hagiwara D, Takahashi-Nakaguchi A, Sakai K, Gonoi T. Global gene expression reveals stress-responsive genes in Aspergillus fumigatus mycelia. BMC Genomics. 2017;18:942.

Tautz D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res. 1989;17:6463–71.

Tavanti A, Davidson AD, Gow NAR, Maiden MCJ, Odds FC. Candida orthopsilosis and Candida metapsilosis spp. nov. to replace Candida parapsilosis groups II and III. J Clin Microbiol. 2005;43:284–92.

Tavares AH, Fernandes L, Bocca AL, Silva-Pereira I, Felipe MS. Transcriptomic reprogramming of genus Paracoccidioides in dimorphism and host niches. Fungal Genet Biol. 2015;81:98–109.

Taylor JW. Evolutionary perspectives on human fungal pathogens. Cold Spring Harb Perspect Med. 2014;5:a019588.

Taylor JW, Ellison CE. Mushrooms: morphological complexity in the fungi. Proc Natl Acad Sci U S A. 2010;107:11655–6.

Taylor JW, Geiser DM, Burt A, Koufopanou V. The evolutionary biology and population genetics underlying fungal strain typing. Clin Microbiol Rev. 1999;12:126–46.

Taylor JW, Jacobson DJ, Kroken S, Kasuga T, Geiser DM, Hibbett DS, Fisher MC. Phylogenetic species recognition and species concepts in fungi. Fungal Genet Biol. 2000;31:21–32.

Taylor JW, Turner E, Townsend JP, Dettman JR, Jacobson D. Eukaryotic microbes, species recognition and the geographic limits of species: examples from the kingdom Fungi. Philos Trans R Soc Lond Ser B Biol Sci. 2006;361:1947–63.

Taylor JW, Hann-Soden C, Branco S, Sylvain I, Ellison CE. Clonal reproduction in fungi. Proc Natl Acad Sci U S A. 2015;112:8901–8.

Teixeira MM, Barker BM. Use of population genetics to assess the ecology, evolution, and population structure of Coccidioides. Emerg Infect Dis. 2016;22:1022–30.

Teixeira MM, Theodoro RC, De Carvalho MJ, Fernandes L, Paes HC, Hahn RC, Mendoza L, Bagagli E, San-Blas G, Felipe MS. Phylogenetic analysis reveals a high level of speciation in the Paracoccidioides genus. Mol Phylogenet Evol. 2009;52:273–83.

Teixeira MDM, Theodoro RC, Derengowski Lda S, Nicola AM, Bagagli E, Felipe MS. Molecular and morphological data support the existence of a sexual cycle in species of the genus Paracoccidioides. Eukaryot Cell. 2013;12:380–9.

Teixeira MDM, Theodoro RC, Oliveira FF, Machado GC, Hahn RC, Bagagli E, San-Blas G, Soares Felipe MS. Paracoccidioides lutzii sp. nov.: biological and clinical implications. Med Mycol. 2014a;52:19–28.

Teixeira MM, Theodoro RC, Nino-Vega G, Bagagli E, Felipe MS. Paracoccidioides species complex: ecology, phylogeny, sexual reproduction, and virulence. PLoS Pathog. 2014b;10: e1004397.

Teixeira MDM, Patane JS, Taylor ML, Gomez BL, Theodoro RC, De Hoog S, Engelthaler DM, Zancope-Oliveira RM, Felipe MS, Barker BM. Worldwide phylogenetic distributions and population dynamics of the genus Histoplasma. PLoS Negl Trop Dis. 2016;10:e0004732.

Theodoro RC, Teixeira Mde M, Felipe MS, Paduan Kdos S, Ribolla PM, San-Blas G, Bagagli E. Genus paracoccidioides: species recognition and biogeographic aspects. PLoS One. 2012;7:e37694.

Tsao S, Rahkhoodaee F, Raymond M. Relative contributions of the Candida albicans ABC transporters Cdr1p and Cdr2p to clinical azole resistance. Antimicrob Agents Chemother. 2009;53:1344–52.

Turissini DA, Gomez OM, Teixeira MM, McEwen JG, Matute DR. Species boundaries in the human pathogen Paracoccidioides. Fungal Genet Biol. 2017;106:9–25.

Turner SA, Butler G. The Candida pathogenic species complex. Cold Spring Harb Perspect Med. 2014;4:a019778.

Untereiner WA, Scott JA, Naveau FA, Sigler L, Bachewich J, Angus A. The Ajellomycetaceae, a new family of vertebrate-associated Onygenales. Mycologia. 2004;96:812–21.

Vallabhaneni S, Kallen A, Tsay S, Chow N, Welsh R, Kerins J, Kemble SK, Pacilli M, Black SR, Landon E, Ridgway J, Palmore TN, Zelzany A, Adams EH, Quinn M, Chaturvedi S, Greenko J, Fernandez R, Southwick K, Furuya EY, Calfee DP, Hamula C, Patel G, Barrett P, MSD, Lafaro P, Berkow EL, Moulton-Meissner H, Noble-Wang J, Fagan RP, Jackson BR, Lockhart SR, Litvintseva AP, Chiller TM. Investigation of the first seven reported cases of Candida auris, a globally emerging invasive, multidrug-resistant fungus – United States, May 2013-August 2016. MMWR Morb Mortal Wkly Rep. 2016;65:1234–7.

Vite-Garin T, Estrada-Barcenas DA, Cifuentes J, Taylor ML. The importance of molecular analyses for understanding the genetic diversity of Histoplasma capsulatum: an overview. Rev Iberoam Micol. 2014;31:11–5.

Vos P, Hogers R, Bleeker M, Reijans M, Van De Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, et al. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 1995;23:4407–14.

Wang P, Kenyon C, De Hoog S, Guo L, Fan H, Liu H, Li Z, Sheng R, Yang Y, Jiang Y, Zhang L, Xu Y. A novel dimorphic pathogen, Emergomyces orientalis (Onygenales), agent of disseminated infection. Mycoses. 2017a;60:310–9.

Wang RJ, Miller RF, Huang L. Approach to fungal infections in human immunodeficiency virus-infected individuals: Pneumocystis and beyond. Clin Chest Med. 2017b;38:465–77.

Webster RH, Sil A. Conserved factors Ryp2 and Ryp3 control cell morphology and infectious spore formation in the fungal pathogen Histoplasma capsulatum. Proc Natl Acad Sci U S A. 2008;105:14573–8.

Whaley SG, Berkow EL, Rybak JM, Nishimoto AT, Barker KS, Rogers PD. Azole antifungal resistance in Candida albicans and emerging non-albicans Candida species. Front Microbiol. 2016;7:2173.

Whiston E, Taylor JW. Genomics in Coccidioides: insights into evolution, ecology, and pathogenesis. Med Mycol. 2014;52:149–55.

Whiston E, Taylor JW. Comparative phylogenomics of pathogenic and nonpathogenic species. G3 (Bethesda). 2015;6:235–44.

Whiston E, Zhang Wise H, Sharpton TJ, Jui G, Cole GT, Taylor JW. Comparative transcriptomics of the saprobic and parasitic growth phases in Coccidioides spp. PLoS One. 2012;7:e41034.

White TC. Increased mRNA levels of ERG16, CDR, and MDR1 correlate with increases in azole resistance in Candida albicans isolates from a patient infected with human immunodeficiency virus. Antimicrob Agents Chemother. 1997;41:1482–7.

White TJ, Bruns TD, Lee SB, Taylor JW. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: PCR protocols: a guide to methods and applications. Cambridge: Academic Press; 1990.

Wiley EO. The evolutionary species concept reconsidered. Syst Biol. 1978;27:17–26.

Wilker E, Yaffe MB. 14-3-3 proteins – a focus on cancer and human disease. J Mol Cell Cardiol. 2004;37:633–42.

Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. 1990;18:6531–5.

Yan Z, Hull CM, Sun S, Heitman J, Xu J. The mating type-specific homeodomain genes SXI1 alpha and SXI2a coordinately control uniparental mitochondrial inheritance in Cryptococcus neoformans. Curr Genet. 2007;51:187–95.

Yu J, Cleveland TE, Nierman WC, Bennett JW. Aspergillus flavus genomics: gateway to human and animal health, food safety, and crop resistance to diseases. Rev Iberoam Micol. 2005;22:194–202.

Zakikhany K, Naglik JR, Schmidt-Westhausen A, Holland G, Schaller M, Hube B. In vivo transcript profiling of Candida albicans identifies a gene essential for interepithelial dissemination. Cell Microbiol. 2007;9:2938–54.

Zhu W, Filler SG. Interactions of Candida albicans with epithelial cells. Cell Microbiol. 2010;12:273–82.

Yeast Population Genomics Goes Wild: The Case of *Saccharomyces paradoxus*

Mathieu Hénault, Chris Eberlein, Guillaume Charron, Éléonore Durand, Lou Nielly-Thibault, Hélène Martin, and Christian R. Landry

Abstract Speciation and adaptation are important processes that are difficult to study in the invisible microbial world because of the lack of easily identifiable characters that can be correlated with species boundaries and adaptive traits. Genomic tools can be used to assess and measure the genetic and genomic bases of species and population differentiation. This allows for the identification of the genes that are potential targets of natural selection and thus that underlie adaptation to specific environments. Here, we illustrate how useful this approach is by describing recent progress on microbial genomics empowered by studying *Saccharomyces paradoxus* in the wild. These studies have revealed the spatial and temporal scales at which fungal populations diverge, a quantification of the life history parameters of this yeast and its mechanisms of speciation, which include allopatric speciation driven by geographical barriers and hybrid speciation driven by chromosomal reorganization. Altogether, these studies establish *S. paradoxus* as an extremely powerful model in microbial population genomics.

Keywords Adaptation · Hybridization · Introgression · Population genomics · *Saccharomyces paradoxus* · Speciation · Yeast

1 Introduction

Fungi are among the most diverse and ubiquitous eukaryotes. Recent surveys of fungi across various environments showed that we are only beginning to appreciate the extent of this diversity, which could include more than one million species, and

Mathieu Hénault and Chris Eberlein contributed equally to this work.

M. Hénault, C. Eberlein, G. Charron, É. Durand, L. Nielly-Thibault, H. Martin, and C.R. Landry (✉)
Département de Biologie and Département de Biochimie, Microbiologie et Bio-informatique, Institut de Biologie Intégrative et des Systèmes, PROTEO, Université Laval, Quebec City, QC, Canada
e-mail: christian.landry@bio.ulaval.ca

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2017_4,

even more (Newsham et al. 2016; Mora et al. 2011). One field of fungal biology that has progressed extremely rapidly over the past decade is the genomics of fungi associated with human activity, including wine and beer yeasts (*Saccharomyces* spp.) (Marsit et al. 2017); filamentous fungi used in the food industry, for instance, cheese production (Cheeseman et al. 2014); and plant and human pathogens such as *Magnaporthe* sp., *Candida* sp., and *Cryptococcus* sp. (Desjardins et al. 2017; Ford et al. 2015; Gabaldon et al. 2013; Chiapello et al. 2015). These studies offered unprecedented insight into genome evolution by revealing that fungal genomes are extremely plastic in terms of structure, gene content, and regulation and as a consequence, they can adapt rapidly to challenging conditions. However, we are still lagging behind in terms of understanding what are the ecological, demographic, and historical factors that drive fungal diversity in nature.

While human-associated fungal species offer great study systems in genomics and the genetics of adaptation, the underlying genetic changes that have accumulated over time may not reflect processes that normally take place in nature. Accordingly, they may only provide limited information on the evolutionary forces that have shaped the diversity of fungi over the past millions of years. To explore the interplay between historical events that have shaped fungal phylogeography, ecology, and genomic variation, it is imperative that we develop tractable model systems that have been minimally affected by human activities. A few free-living models have emerged in the field of ecological genomics for this purpose over the past decade. One of them is *Neurospora crassa*, a classical model in genetics with a nearly worldwide distribution and a large diversity of habitats (Turner et al. 2001). The analysis of *Neurospora* genomes, for instance, identified populations locally adapted to ambient temperatures and this, despite its broad-scale distribution (Ellison et al. 2011). Another model that has emerged recently is the wild yeast *Saccharomyces paradoxus*, sister species of the budding yeast *S. cerevisiae* (Replansky et al. 2008).

Yeast population genomics benefited from many aspects of the recent progress on *S. cerevisiae*'s cell biology, systems biology, and genetics, including the ability to be handled in the lab, to be subjected to high-throughput phenotyping, and to have its genome manipulated in several ways, including by whole chromosome synthesis (Richardson et al. 2017). While *S. cerevisiae* is the prime fungal model in experimental and evolutionary biology in the laboratory (Marsit et al. 2017), its wild sister species has become a very promising model for population genomics by providing opportunities to explore yeast ecology and the genomics of adaptation and speciation. Here, we review recent research performed in ecological genomics of *S. paradoxus* to illustrate how population genomics illuminates the ecological and historical factors that shape fungal genome diversity.

2 Distribution and Ecology of *S. paradoxus*

S. paradoxus is associated with deciduous forests in the northern hemisphere across the globe (Fig. 1). Some strains have been isolated in Australia and New Zealand but these appear to be due to recent migration from Europe (Zhang et al. 2010).

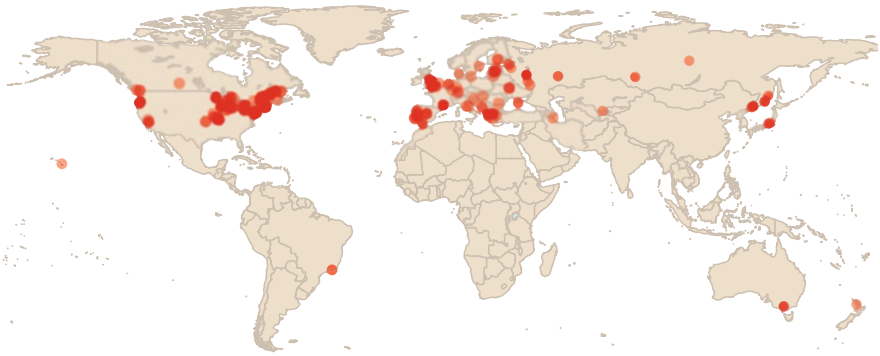

Fig. 1 Worldwide distribution of *Saccharomyces paradoxus*. *S. paradoxus* has been sampled in many countries around the world (>500 strains) on various substrates including tree bark, flowers, insects, and soil. Most strains were sampled in Europe and North America, and only a few strains have been isolated from the southern hemisphere. Sampling regions were retrieved from Zhang et al. (2010), Naumov et al. (1998), Gonçalves et al. (2011), Charron et al. (2014a), Robinson et al. (2016), Johnson et al. (2004), Koufopanou et al. (2006), Sniegowski et al. (2002), Liti et al. (2009), Samani et al. (2015), Leducq et al. (2016), Xia et al. (2017), Almeida et al. (2017), Hyma and Fay (2013), Redzepovic et al. (2002), and Sampaio and Gonçalves (2008). The map was drawn with R using the package maps (version 2.3–9). Red intensities reflect the number of strains isolated by sampling site

Although the extent of the specificity of *S. paradoxus* to its host still requires full investigation, the fact that it is commonly found on the bark and in the soil associated with trees such as oaks (*Quercus* spp.) and maples (*Acer* spp.) suggests that these are likely its natural habitats (Sniegowski et al. 2002; Naumov et al. 1998; Kowallik and Greig 2016; Charron et al. 2014a). In addition, a recent broad study of yeast diversity in Wisconsin (USA) showed that *S. paradoxus* appears indeed to be significantly more often associated with oak trees than with any other tree species (Sylvester et al. 2015). Another survey performed recently in Europe confirms that *S. paradoxus* prefers oak trees but that leaf litter could be its usual substrate, not tree bark (Kowallik and Greig 2016). *S. paradoxus* could therefore occupy similar niches across its range. This habitat appears to be shared with other species of the genus, which often have overlapping geographical distributions. For instance, *S. paradoxus* and *S. cerevisiae* are frequently found on the same tree species in North America where the two yeasts' distributions overlap (Sniegowski et al. 2002). The ecological niches of *S. paradoxus* and its closest relatives within the genus are therefore not entirely distinct, which suggests that the radiation of the *Saccharomyces* genus may have taken place in this type of habitat and conditions. As discussed below, this gives rise to opportunities for interspecies hybridization, as recently observed in several contexts using population genomics (Barbosa et al. 2016; Leducq et al. 2016; Peris et al. 2016).

As is the case for most microbes, the ecological significance of *S. paradoxus* in its ecosystem is largely unknown but one can assume that it is a commensal

saprophyte, feeding on tree exudates, leaf litters, and soils associated with trees. Conversely, our understanding of the role of ecological factors in driving *S. paradoxus'* distribution and success on specific hosts has progressed in recent years. Summer temperatures appear to be major determinants of its geographical distribution. Laboratory experiments showed that the different species of the genus have distinct maximum growth temperatures, from 34°C for *S. uvarum*, 37–38°C for *S. paradoxus*, and up to 42°C for *S. cerevisiae* (Gonçalves et al. 2011). Surveys of *S. paradoxus* and *S. cerevisiae* in the wild generally confirmed these laboratory observations. A study by Charron et al. (2014a) examined the presence of *S. cerevisiae* and *S. paradoxus* in eastern Canada and showed that *S. paradoxus* tends to be found alone in northern deciduous forests, while *S. cerevisiae* is absent at these latitudes, in accordance with their respective maximum growth temperatures. A worldwide analysis of sampling sites also confirmed that summer temperatures and optimal growth temperatures are predictors of the *S. paradoxus* geographical distribution (Robinson et al. 2016). In addition to these observations in the field, measurements of growth rates and cell survival in the laboratory showed that, as predicted from the local adaptation hypothesis, southern populations of *S. paradoxus* grow faster at higher temperatures than northern populations and that survival to freezing correlates with the number of oscillations above and below 0°C at sampling locations (Leducq et al. 2014). As for many other fungi (Sylvester et al. 2015; Newsham et al. 2016), temperature therefore appears to be a major factor driving the geographical distribution of *S. paradoxus*. This makes *S. paradoxus* a powerful model system to examine how temperature determines the geographical distributions of fungi.

3 Population Genetics and Life Cycle

Several aspects of microbial ecology in the wild remain to be investigated to fully understand how microbes adapt to local conditions across their geographical ranges. One of these aspects is how much time individuals spend in each phase of their life cycle, including the frequency of sexual reproduction. These parameters are key determinants for evolution because they define how selection and genetic drift shape genome architecture and limit the level of genetic polymorphism by affecting linkage among beneficial and deleterious mutations (McDonald et al. 2016; Lang et al. 2013).

It can be extremely difficult to directly quantify yeast life cycles in natural populations because it cannot be directly observed. Population genetics approaches can however provide invaluable insights into this question by allowing the quantification of the extent of heterozygosity and recombination in natural samples. Thanks to several years of research on the model *S. cerevisiae*, the life cycle of species of the *Saccharomyces* genus has been described in great details. As in *S. cerevisiae*, wild strains of *S. paradoxus* are usually diploid and homothallic (Johnson et al. 2004): haploid spores can switch mating types after a cell division

to mate with daughter cells. Consequently, the opportunity for inbreeding is extremely high and may limit the rate of adaptation by favoring the accumulation of deleterious mutations in small, local populations.

One of the first surveys of genetic diversity in *S. paradoxus* was performed in a 10 km^2 forest area in England and looked at 7 kb of DNA sequences in 28 isolates (Johnson et al. 2004). The first observation that was made on a subset of strains revealed that none of the isolates were heterozygous, suggesting that outbreeding occurred indeed extremely rarely but was happening, as inferred through traces of recombination among loci. In addition, identical genotypes were identified in different samples, indicating that mitotic growth and thus clonal reproduction are important. The same genotypes could also be found on the same tree, in a proximity of 5 cm, implying that mating among clones is physically possible, which further enhances the opportunity for inbreeding (Koufopanou et al. 2006). The extension of this survey to the sequencing of the entire third chromosome in 20 isolates allowed Tsai and colleagues to quantify the different steps of its life cycle (Tsai et al. 2008). Using population genetics analyses, Tsai and colleagues confirmed that outcrossing is rare and estimated that sexual cycles occur only about once every 1,000 cell divisions. Their estimates also suggest that 94% of mating events occur between spores of the same tetrad (haploid meiotic products), 5% of matings occur within a clone after mating type switching, and only 1% of matings occur between spores of different tetrads (potential outcrossing events). Sexual reproduction is therefore infrequent, and inbreeding may be a dominant factor in driving local population differentiation. Finally, a recent study showed that genotypes at a local site could be resident over time, for instance, from year to year, such that limited migration could allow adaptation to local conditions (Xia et al. 2017). How the factors promoting inbreeding interfere with the opportunity for local adaptation created by the existence of stable population structure over time remains to be investigated.

A survey of genetic diversity of a handful of loci across geographical scales confirmed that populations of *S. paradoxus* are differentiated. Genetic differentiation was shown to increase with geographical distance. Sampling sites within a tree on a centimeter scale are more similar than samples from different continents, for instance, between Europe, North America, and Far East Asia (Koufopanou et al. 2006). These results, along with experimental crosses demonstrating that American and European strains show partial reproductive isolation (Kuehne et al. 2007), confirmed that *S. paradoxus* is not a panmictic species but rather displays local populations that diverged over a long period across continents, leading to partial reproductive isolation and potentially to allopatric species formation. This global pattern of genomic differentiation by geographic origins was later confirmed by whole genome sequencing of a large number of strains sampled worldwide (Liti et al. 2009). This survey revealed that on a global scale, there are four distinct lineages of *S. paradoxus* corresponding to European, Far Eastern, American, and Hawaiian populations that show divergence on the order of a few percent at the nucleotide level. A more recent study using long-read sequencing showed that these lineages also diverge in terms of protein-coding gene content due to chromosomal rearrangements in the chromosomal cores, including the deletion of some stress

response genes and the duplication of sulfite transporters in specific lineages (Yue et al. 2017), which could contribute to their phenotypic differentiation (Fig. 2). The initial genome-wide survey (Liti et al. 2009) also included a large set of *S. cerevisiae* strains, which failed to show perfectly consistent geographic population structures, confirming that *S. paradoxus* may be a more suitable model for elucidating the role of ecological and evolutionary forces on natural populations.

To be relevant for local adaptation and speciation, these global genetic differences uncovered by whole genome sequencing have to translate into phenotypic variation. A global survey of growth rates across conditions, including temperature gradients, ethanol concentration, pH, and various drugs, suggested that indeed phenotypic variation is extensive across worldwide samples of *S. paradoxus* (Liti et al. 2009), although phenotypic similarity among the strains did not necessarily reflect their phylogenetic relationships. While this result does not support the role of local adaptation in the phenotypic divergence of the strains, it does not reject it either because the conditions surveyed, with few exceptions, do not represent conditions that natural yeast populations would encounter in their natural habitats. However, phenotypic divergence between the North American and European populations was recently addressed for relevant metabolic traits (Samani et al. 2015).

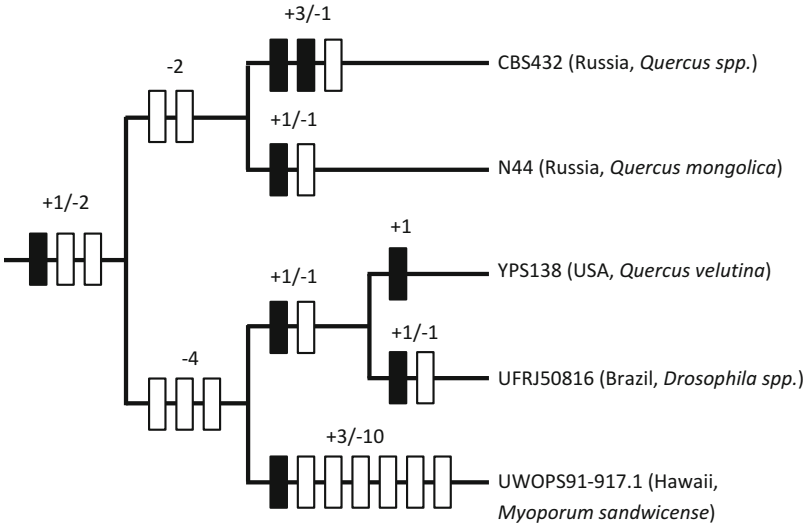

Fig. 2 Gene gain and loss by unbalanced structural rearrangements among the major lineages of *S. paradoxus*. Yue et al. (2017) used long-read sequencing to produce de novo assemblies of five *S. paradoxus* genomes from various origins and sources. The resulting high-quality genomes allowed for the detection of large structural unbalanced rearrangements responsible for gene number variation among strains. Events that occurred in nuclear chromosomal cores – i.e., considering nuclear genomes without subtelomeres and chromosome ends – are shown with a filled (insertion or duplication) or empty (deletion) rectangle. The total number of gene gains and losses is indicated above each branch of the tree; one event could affect several genes at the same time

The ability of 45 wild isolates to use various carbon sources, which are expected to vary in relative abundance in nature temporally and/or spatially, was measured in the laboratory. This study revealed that both populations could readily metabolize sugars such as glucose, fructose, galactose, mannose, sucrose, turanose, and isomaltulose, indicating that these could be among their main carbon sources if they are available on natural substrates. However, a strong difference in yield was observed on substrates of the pentose phosphate pathway, with North American strains performing better than European strains. These analyses demonstrate that there is a population differentiation at the level of carbon source metabolism between these two main lineages of *S. paradoxus*, indicating that differentiation between continents also accumulates at the phenotypic level. Whether these phenotypic differences result from the neutral accumulation of mutations by genetic drift or from adaptation to local conditions remains to be investigated.

One of the major challenges in microbial ecological genomics is to identify what are the fitness determinants in natural conditions, for instance, what are the actual sources of limiting nutrients exploited by free-living cells in nature. As *Saccharomyces* yeasts are associated with deciduous trees including oak and maple trees, sap exudates are suspected to provide nutrients used for growth. Because it is used in the production of maple syrup in northeastern America, maple sap is readily available to do experiments and examine these questions in the laboratory. Filteau et al. (2016) tested the growth of *S. paradoxus* wild strains in maple sap and observed variation in growth rates among strains from different locations along a northeast to southwest gradient. Using a functional genomics approach based on the yeast deletion collection and barcode sequencing, Filteau and colleagues identified the allantoin degradation pathway to be required for optimal growth rate on maple sap. Using knockout strains of *S. paradoxus* for the genes involved in the utilization of allantoate, a metabolite found in the same pathway, the authors demonstrated that allantoate is indeed one of the main limiting nitrogen sources available in maple sap and that growth depends on the ability of strains to use this nitrogen source. The ability to do so could therefore be a major determinant of fitness in wild population feeding on maple trees. Variation in growth rate among strains from diverse geographical origins therefore reflects potentially adaptive standing genetic variation in nitrogen metabolism.

4 Genomics of Speciation in North American *S. paradoxus* Populations

Studies of *S. paradoxus* on a global scale showed that geographic barriers play a predominant role in shaping its population structure, with continents representing the major lineages (North American, Far Eastern, American, Hawaiian). Geographical barriers alone could therefore contribute to microbial speciation. However, analyses on these larger scales make it difficult to estimate the contribution of local

ecological factors to this differentiation because long divergence time can be confounded with other factors such as climatic conditions. Ideally, one would study the early step of divergence among populations. A study by Kuehne and colleagues (2007) uncovered distinct populations of *S. paradoxus* within a close geographic range within North America. The study of several kilobases of DNA sequence in a set of 62 isolates from the east coast of the United States revealed three lineages with potentially overlapping distributions, with two highly abundant lineages, *SpA* and *SpB*. Sequence analyses showed that *SpB* appears to be specific to North America, while the *SpA* group was nearly genetically identical to European strains, suggesting that it recently migrated to North America. In addition, a single isolate from a newly discovered lineage, called *SpC*, was identified, suggesting that this rare lineage, could have diverged from *SpB* within North America and this, without any obvious geographical barriers.

The population structure and evolution of these North American populations were addressed by whole genome sequencing of more than 150 strains from a large region in the northeast of North America. The analysis confirmed the presence of three genetically distinct lineages in North America, *SpA*, *SpB*, and *SpC*, and revealed in addition that the *SpC* lineage is not a rare lineage found only in Pennsylvania but is broadly occupying the northeast of the *S. paradoxus* distribution (Leducq et al. 2016). Another lineage, closely related to *SpC*, labeled *SpC**, was also identified in this region (see below for details on its origin). Across this geographic range, strains display variation at several levels, including the utilization of carbon and nitrogen sources, confirming what was observed between continents, i.e., that populations of *S. paradoxus* diverge in their ability to use various substrates for growth in the laboratory. *SpB* outperforms *SpC* on almost all of the substrates except a few cases, notably when asparagine, proline, or lysine is the sole nitrogen source (Fig. 3). The ecological significance and the molecular basis of these differences are unknown, but since proline plays a key role in plant biology, including stress response, its availability on plant substrates and leaf litter may vary geographically with local conditions (Hayat et al. 2012) and thus be a key element to determining fitness locally.

The northeast of North America was covered by an ice sheet up until about 10,000 years ago, which implies that the divergence between *SpB* and *SpC* could have been initiated during or prior to the last ice age (~110,000–10,000 years ago). This event likely shaped the evolution of a large number of microbes, plants, and animals. For instance, this region comprises many species pairs or semi-species pairs of animals that show similar patterns of divergence (April et al. 2013; Wong Miller et al. 2017). The two lineages *SpB* and *SpC* thus would have been isolated during the last glaciation in separate glacial refugia and are now partly sympatric following the glacial retreat.

The current distribution, which shows almost no *SpC* strains in the south and almost no *SpB* in the north, suggests that ecological factors are limiting migration. One reason why the data suggests that effective migration is limited by ecological factors is that the dispersal capabilities of budding yeast is likely high and 10,000 years would have been enough to homogenize the two distributions if

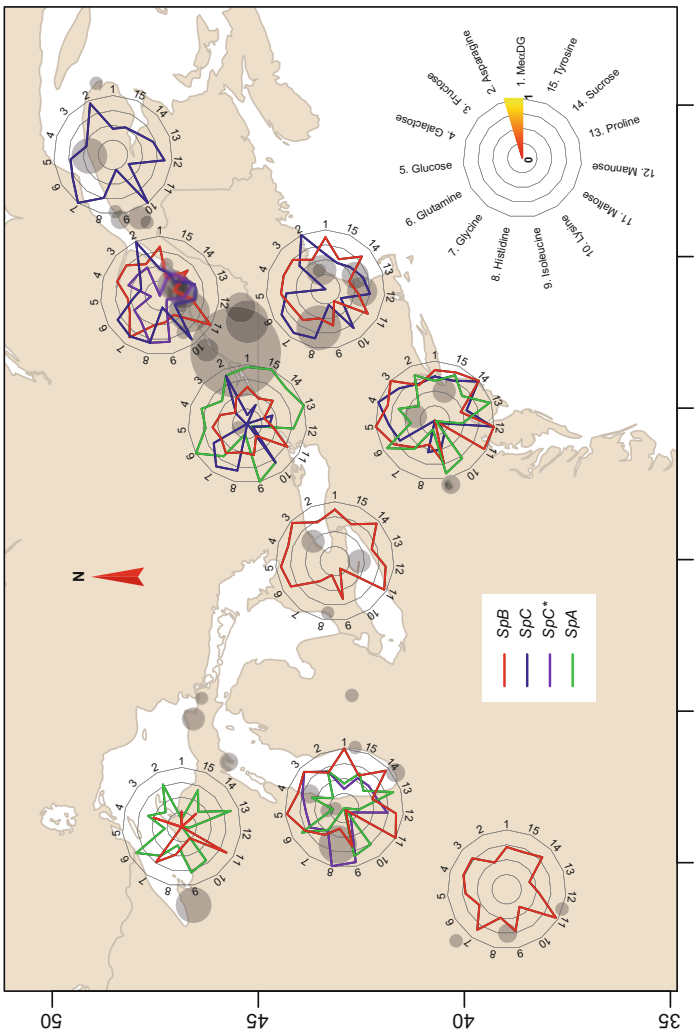

Fig. 3 North American *S. paradoxus* populations exhibit extensive phenotypic variation across their distribution area. Radar charts show the median phenotypes of strains of a given lineage (*SpB*: red, *SpC*: blue, *SpC**: purple, *SpA*: green) sampled in each subregion. The axes of environmental conditions are scaled independently. The outer points represent the highest median phenotypic values across all subregions and strains. Shaded gray circles represent the individual sampling locations, with the area of each circle being proportional to the number of strains sampled. "MeαDG" stands for methyl-alpha-D-glucopyranoside. Growth rates were measured by monitoring colony sizes on solid culture media (Leducq et al. 2016)

selection was not acting. For instance, *SpA* is thought to have been introduced in North America in the last 600 years (Kuehne et al. 2007), and it is now found over a large geographic region overlapping Pennsylvania, Québec, and Ontario, and this, in the absence of genetic diversity, implying that it was introduced as a single or a few related clones. The mode of dispersion of budding yeasts is largely unknown, but their association with insects (Stefanini et al. 2012) allows to believe that they may disperse along by these vectors over relatively long distances. The recent demonstration that local populations could be stable at least over a 2-year period (Xia et al. 2017) however argues against an extremely high dispersal rate, although this study examined two consecutive years only. Whether any of the phenotypic traits that differentiate the two lineages (growth at high temperature and on various carbon and nitrogen sources) contribute to this ecological barrier remains to be determined. The analyses of the *SpB* and *SpC* coding genomes recently provided insight into this question.

5 Genomic Divergence Between Lineages, Ecological Specialization and Reproductive Isolation

Genomic divergence between the North American *S. paradoxus* lineages was driven by the isolation period during the last ice age and evolutionary forces such as genetic drift and natural selection caused by variation in local environmental conditions. Drift likely has played an important role because these natural populations, especially *SpC*, show limited polymorphism, which may reflect small effective population sizes (Leducq et al. 2016), at least partly due to the frequent opportunities for inbreeding. The main lineages in Northern America are distinguished by genomic divergence of about 2–3%, providing the opportunity to identify regions of the genome that are rapidly evolving. A study by Eberlein et al. (2017) examined the degree of divergence of the coding genomes between *SpB* and *SpC* by investigating 17 genomes from representatives of both lineages. The authors assessed the divergence of these lineages in about 4,400 genes since their separation from the European sister clade *SpA* (Fig. 4). The analysis revealed that both lineages tend to be under purifying selection across the proteome, and only a few genes showed signs of positive selection. Moreover, they identified genes that accumulated significantly different numbers of protein-altering mutations between *SpB* and *SpC*.

The resulting 76 candidate genes included *GRS2*, which has been acquired during the whole genome duplication in yeasts ~100 Mya (Kellis et al. 2004). *GRS2*, a glycyl-tRNA synthetase, was long thought to be pseudogene-like (Turner et al. 2000), but Chen et al. (2012) showed that the transcription of this paralog is actually induced under various stress conditions such as 37°C in *S. cerevisiae*. For *S. paradoxus*, Eberlein and colleagues (2017) were able to link the rapid evolution of *GRS2* in the *SpC* lineage to relaxed selection. Their hypothesis was that survival

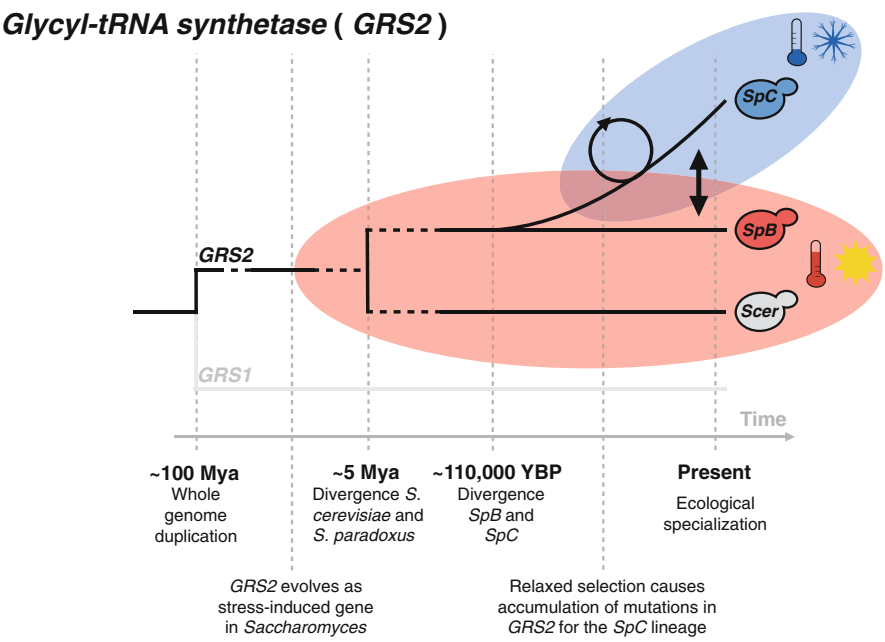

Fig. 4 Protein-coding divergence between *SpB* and *SpC* and ecological specialization. A study of the coding genomes of 17 representative strains of the lineages *SpB* and *SpC* using the European lineage *SpA* as an outgroup identified candidate genes that potentially evolved asymmetrically or were under positive selection. One of the candidate genes is *GRS2*, a paralog that originated during the yeast whole genome duplication and potentially neofunctionalized for stress response in the *Saccharomyces* clade prior to the divergence between *S. paradoxus* and *S. cerevisiae*. Eberlein et al. (2017) proposed a model in which this gene evolved under relaxed selection in the *SpC* lineage, which correlates with the inability of *SpC* strains to grow at high temperature. Fitness assays performed in the laboratory confirmed that the GRS2 *SpB* allele performs better than the *SpC* allele at high temperature. This result indicates that relaxed selection could be an important factor in ecological specialization

in *SpC*'s northern habitat did not require high temperature tolerance, which resulted in the relaxation of selection on some genes in this lineage, including *GRS2*. These could therefore now contribute to the poor growth of *SpC* at high temperature and thus to its inability to migrate further south. This study highlights the importance of relaxed selection rather than adaptive changes in ecological specialization and the importance of paralogous genes as a driving force for ecological divergence between closely related species (Sanchez-Perez et al. 2008).

The distinct geographical distributions of *SpB* and *SpC* and their monophyly based on whole genome sequencing also suggest that they are reproductively isolated. Prezygotic reproductive isolation is thought to be limited in budding yeast because the mating systems are simple (Hittinger 2013). Crosses between distantly related species of the genus can be performed with success, showing that intrinsic prezygotic barriers are inexistent or weak, although evidence for

prezygotic isolation was recently reported between populations (Murphy and Zeyl 2015). The major mechanism of reproductive isolation appears to be postzygotic and can be detected by measuring spore survival in interspecies crosses. This was examined in crosses between *SpB* and *SpC* strains. Charron et al. (Charron et al. 2014b) showed that *SpB* and *SpC* are indeed partially reproductively isolated such that their recent divergence was enough for the accumulation of reproductive incompatibilities, most likely due to genomic rearrangements. Variation in chromosomal structure also seemed to correlate with the extent of partial sterility, even within *SpC*, which displayed particularly strong variation in spore survival in within-lineage crosses (see *SpC** below).

Reproductive incompatibilities between *SpB* and *SpC* could decrease gene flow between these two incipient species. However, the distributions of these lineages do overlap significantly, and postzygotic reproductive isolation cannot alone prevent the formation of F1 hybrids in this region. In spite of this, no F1 hybrids have been sampled so far, suggesting that if hybridization occurs, it is rare. Another mechanism that could contribute to diminish the presence of *SpB-SpC* hybrids is selection against them, for instance, through poor growth performance. This hypothesis was recently tested by using high-throughput screening of yeast growth rates over multiple environmental conditions (Fig. 5) (Charron and Landry 2017). The growth of *SpB-SpC* hybrids was compared to their parents in order to assess the mode of inheritance of the parental phenotypes by calculating the degree of dominance. The results showed that, in the majority of the cases, hybrids between the two diverging lineages display overdominant or partially dominant growth phenotypes of the fittest parent. *SpB-SpC* hybrids could therefore outperform the least fit or even both of their parents in several growth conditions.

Charron and Landry (2017) concluded that postzygotic extrinsic isolation (selection against hybrids) likely does not act as a barrier to gene flow because the overdominance observed in hybrids across a wide range of environmental conditions could have the effect of promoting hybridization. The authors suggest hypotheses for the rarity of hybrids in the wild such as the colonization of a novel and unsampled habitat or the specialization of each lineage for specific host trees, which would effectively keep the lineages in allopatry, even if their geographical distributions overlap. Other studies suggest that this is a general phenomenon and have reported similar hybrid superiority for other inter- or intraspecific crosses within *Saccharomyces*, although with limited biological or ecological contexts (Shapira et al. 2014; Bernardes et al. 2016). In the case of *SpB-SpC* crosses, the extent of dominance of the growth phenotypes is not correlated with the fertility of crosses, suggesting that they are caused by different mechanisms (Fig. 5).

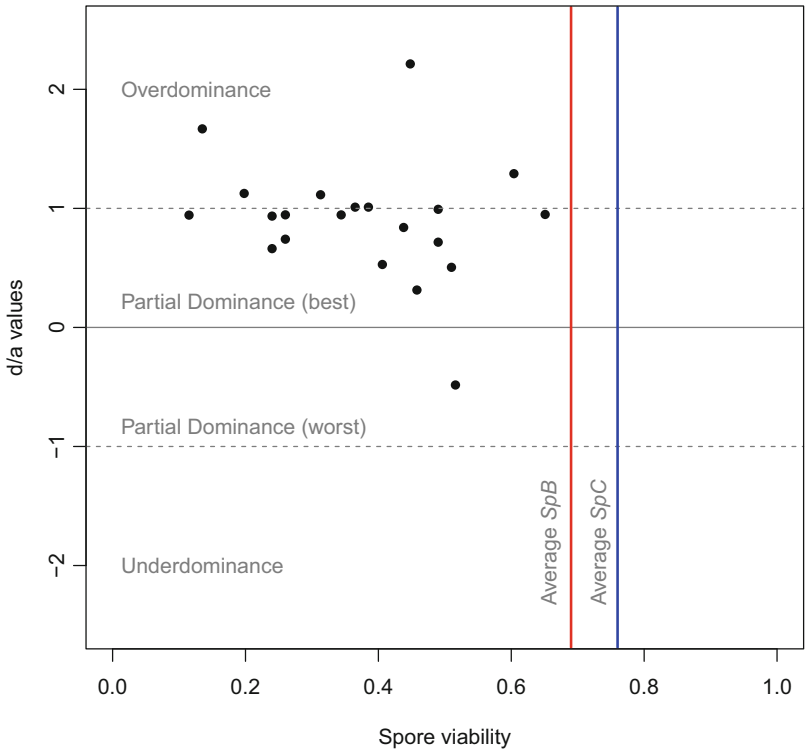

Fig. 5 Postzygotic intrinsic and extrinsic reproductive isolation between incipient species of *S. paradoxus SpB* and *SpC*. Degree of dominance, shown as *d*/*a* values here, is derived from the measurement of growth rates of hybrids and their parental strains on solid medium. The median values of 32 growth conditions are shown (Charron and Landry 2017) for 21 crosses. Each zone delimited by dotted lines corresponds to the different modes of inheritance based on d/a values. Spore viability for the same 21 *SpB*-*SpC* hybrids is shown on the x-axis. The colored vertical lines correspond to the average spore survival for within group crosses

6 Hybrid Speciation

The study of speciation in microorganisms has benefited the most from the development of genomics tools because it allows to detect barriers to gene flow without the need for distinguishing species a priori based on phenotypic traits, which is particularly challenging in microbes. Population genomics analyses of the North American *S. paradoxus* revealed the existence of a cryptic lineage within *SpC*, called *SpC** (Fig. 6) (Leducq et al. 2016). In a genome-wide and windows-based analysis of divergence between *SpB* and *SpC*, some small regions showed abnormally small F_{ST} values, indicating balancing selection or gene flow between *SpB* and *SpC*. These low divergence regions were found to be caused by genomic segments in the *SpC** lineage, accounting for 2–6% of the genome, that were highly similar to *SpB*'s, suggesting that this lineage could have arisen through the recent

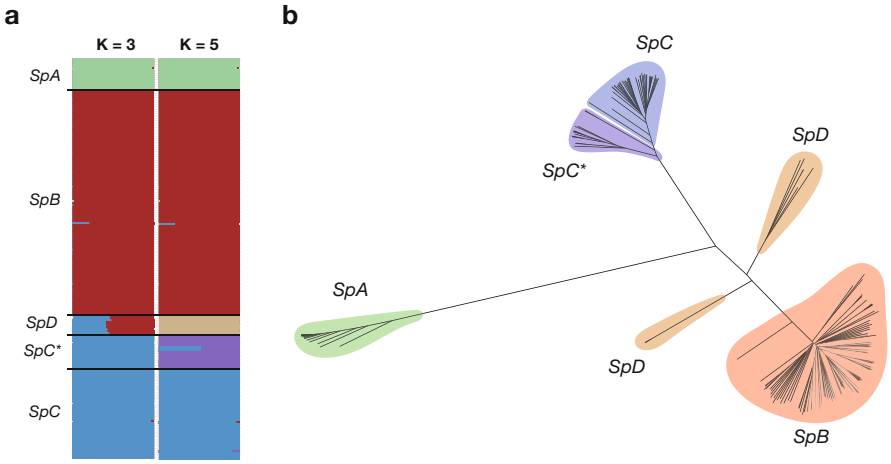

Fig. 6 *S. paradoxus* population structure in northeastern America. More than 170 strains of *S. paradoxus* have been sequenced in the northeast of North America. (**a**) Ancestry proportions inferred by fastSTRUCTURE (Raj et al. 2014) assuming $K = 3$ and 5 clusters. Each individual is represented by a line partitioned into *K*-colored segments displaying the individual's estimated membership fractions in K clusters. (**b**) Unrooted neighbor-joining tree showing the global relationships among strains. Both analyses were performed on 21,890 randomly selected SNPs from 178 strains. *SpA* corresponds to strains of European ancestries. Other strains are from North America

hybridization of *SpB* and *SpC* strains. Further support for the hybridization hypothesis came from genomic data that revealed particular chromosomal rearrangements that were found in both *SpB* and *SpC** strains but were absent from *SpC*. Geographical data also supported this scenario as *SpC** was mostly isolated in the zone of sympatry between *SpB* and *SpC*. The analysis of the divergence between *SpC** and *SpC* allowed to estimate that the *SpC** lineage initially diverged about 10,000 years ago, while the glaciers were retreating in this region. In the first study on the reproductive isolation between *SpB* and *SpC* (Charron et al. 2014b), these *SpC** (considered as *SpC* at the time) contributed disproportionally to the variance in spore survival in crosses performed within *SpC*, revealing that they are partially reproductively isolated. A larger number of crosses within and among *SpC**, *SpC*, and *SpB* revealed that *SpC** is partially reproductively isolated from both its putative parental lineages (Leducq et al. 2016). The emergence and persistence of *SpC** therefore represent an incipient hybrid speciation event that occurred after the allopatric speciation event that gave rise to *SpB* and *SpC*.

The regions introgressed from *SpB* to *SpC** and the partial reproductive isolation with *SpC* showed that *SpC** is a hybrid species. Remarkably, speciation by hybridization was shown to occur in experimental crosses between closely related species where a small fraction of the surviving spores are reproductively isolated from the two parental species and yet interfertile among themselves (Greig et al. 2002). The definition of hybrid species requires that the mechanisms of reproductive isolation

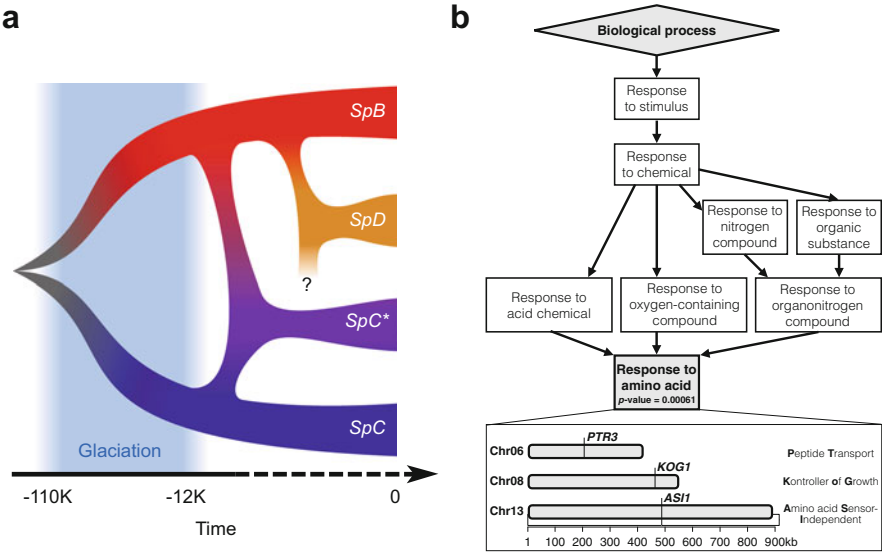

Fig. 7 Hybrid speciation and gene ontology enrichment in introgressed regions of *SpC**. (**a**) Speciation by hybridization between *SpB* and *SpC* gave rise to the *SpC** hybrid species. Another admixed population recently identified in Ontario, Canada (Xia et al. 2017), *SpD*, may result from the hybridization between *SpB* and *SpC* or *SpC** or both. (**b**) The fixed introgressed regions are defined as (Leducq et al. 2016) genomic blocks distributed over six chromosomes that were inherited from *SpB* and that are present in all *SpC** strains. These blocks comprise 105 genes and are enriched for response to amino acid with the genes *PTR3*, *KOG1*, and *ASI1*. GO enrichment was performed with Gorilla (Eden et al. 2009)

are caused by hybridization itself (Schumer et al. 2014). In the case of *SpC**, support from this association comes from a correlation between spore survival in *SpB-SpC* crosses and the segregation of introgressed regions, particularly chromosomal fusions (Leducq et al. 2016). In a recent survey of genomic variation in Ontario (Canada), Xia et al. (2017) discovered a new group of strains, the lineage *SpD*, which may be another admixed population (Figs. 6 and 7a). This observation further strengthens the previous evidence supporting the role of hybridization in shaping genome diversity in *S. paradoxus*.

7 Consequence of Hybridization on Genome Organization

Introgression and admixture are initiated by the formation of F1 hybrids between partially isolated populations and species. The first-generation yeast hybrid can undergo several fates. The F1 hybrid could reproduce like the parental species by meiosis, sporulation, and mating, including backcrosses with the parental species (see life cycle above). In the case of highly diverged parental lineages, F1 hybrids

would be sterile, limiting their potential to contribute to further generations. However, another possibility is that the F1 hybrid could divide mitotically and lose heterozygosity, either through mitotic recombination or the initiation of meiosis and return to growth (Laureau et al. 2016). Such loss of heterozygosity would eventually create a largely homozygous mosaic genome, which could restore fertility. The hybrid species *SpC** gradually lost elements of the *SpB* genome by the first or second scenario. Only a few percent of the *SpC** genome originated from the *SpB* parental species (2–6%) (Leducq et al. 2016). Although small, these regions appear to contribute disproportionally to the traits of *SpC** because in many conditions, this lineage is phenotypically more similar to *SpB* than to *SpC* or intermediate between them. It is not clear why these regions in particular were maintained. Neutral mechanisms such as a low recombination rate in some regions could have contributed. Another possibility is that these regions contain *SpB* alleles that confer an advantage to *SpC**. A gene ontology enrichment analysis of genes occurring in the 1.6% *SpB*-like regions revealed a significant enrichment for genes involved in the response to amino acids (Fig. 7b), a function that repeatedly show signs of divergence among populations (Fig. 3).

8 Effect of Hybridization on the Mitochondrial Genome

The study of nuclear markers and whole nuclear genome sequencing has revealed both the existence of genetically distinct populations and evidence of extensive genetic exchanges across populations in the worldwide distribution of *S. paradoxus*. The mitochondrial genome makes no exception to this pattern, as recent population genomic studies revealed that mitochondrial DNA (mtDNA) sequence, structure, and content vary greatly among populations of *S. paradoxus* (Leducq et al. 2017). The phylogenetic relationships among North American *S. paradoxus* mtDNAs mostly fit those of the nuclear genomes, revealing a clearly defined population structure. However, even at the intrapopulation level, mtDNA content varies in terms of elements like introns and genes coding for homing endonucleases and maturases. Furthermore, the comparison of mtDNAs showed important rearrangements that occurred since the divergence between the North American and Asian populations of *S. paradoxus* (Fig. 8a). A recent study performed long-read whole genome sequencing and de novo assembly of five *S. paradoxus* genomes, including the mitochondrial genomes (Yue et al. 2017). They found a very similar pattern of rearrangements between mtDNAs of two Russian strains and three American strains from the continental United States, Hawaii, and Brazil. They also confirmed the presence/absence polymorphism of introns in the *COX1* and *COB* mitochondrial genes.

Unlike what is known in most animals and plants, yeast cell biology allows biparental inheritance of mtDNAs, i.e., the mitochondria of both parental gametes can be found at least initially in the zygote (Breton and Stewart 2015). Recombination between mtDNA molecules inherited from the two parental cells in yeast

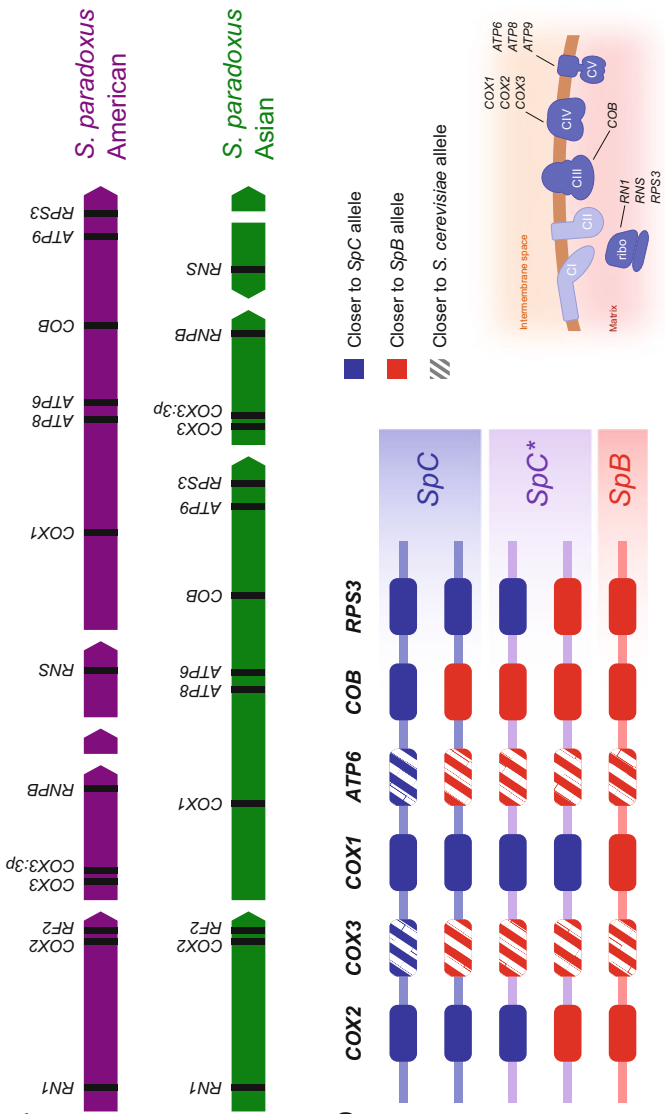

Fig. 8 Mitochondrial genome introgression and rearrangements among *S. paradoxus* populations. (**a**) Translocations and inversions differentiate the American and Asian *S. paradoxus* mitochondrial genomes. The mitochondrial genome sequence is schematized as a linear map showing the approximate position of each gene. (**b**) Examples of introgression for six protein-coding mitochondrial genes into *SpC* and *SpC** lineages. The color of the boxes indicates to which lineage/species the gene sequence is mostly similar. MtDNAs of many *SpC** and one *SpC* strains contain introgressions from *SpB*, with different patterns found within *SpC**. *COX3* and *ATP6* seem to have been inherited from *S. cerevisiae* in all North American *S. paradoxus* lineages. Many mitochondrial gene products are components of the same molecular machineries, like complexes of the oxidative phosphorylation pathway and the mitochondrial ribosome

controlled crosses is well known (Shannon et al. 1972) and suggests that it is likely to occur in the context of natural hybridization. In line with this hypothesis, mtDNA sequences of natural *S. paradoxus* populations exhibit evidence of genetic exchange with other populations and even with other *Saccharomyces* species. For instance, the *COX3* and *ATP6* mitochondrial genes of the North American lineages exhibit close sequence similarity with the alleles of the same genes in *S. cerevisiae*, suggesting ancient hybridization and mtDNA recombination between those clades (Fig. 8b) (Leducq et al. 2017; Peris et al. 2017a). In addition, mtDNAs of *SpC** strains exhibit various patterns of introgression between *SpB* and *SpC* mtDNAs, as revealed by gene-by-gene phylogenetic analysis (Fig. 8b) (Leducq et al. 2017). In many cases, genes encoding components of the same mitochondrial complexes (notably, the complexes of oxidative phosphorylation and the mitochondrial ribosome) were inherited from both *SpB* and *SpC*. The alleles of these genes, which evolved independently since the divergence of *SpB* and *SpC*, were thus combined in the same mtDNA haplotype in *SpC**, leading to a situation in which *SpB*-*SpC* chimeric complexes have to assemble in *SpC**. Given the fact that any heterogeneous mtDNA population within a cell lineage ultimately fixes a single haplotype (i.e., homoplasmy is reached) (Birky 2001), fixation of such mosaic mtDNAs could promote the emergence of incompatibilities among mitochondrial loci or with interacting nuclear loci. A recent study supports this hypothesis by showing an association between recombinant mtDNAs in *S. paradoxus* hybrids and increased phenotypic variation in a condition requiring mitochondrial metabolism (Leducq et al. 2017).

9 Perspective

Population genomics in wild yeast populations has revealed their population structure and the existence of cryptic species. These studies have laid the groundwork for the study of microbial speciation and adaptation in the wild. They have also opened the door to further studies that could be empowered by the tools that were recently adapted in the genetics and genomics model *S. cerevisiae*, including genome editing. These tools could allow to examine the molecular mechanisms responsible for the fitness effects of mutations among populations and incipient species. For instance, CRISPR-Cas9 genome editing was recently used to engineer the budding yeast genome to express candidate adaptive alleles from *S. paradoxus* for adaptation to elevated temperatures (Eberlein et al. 2017). In another recent study, the authors used the Cas9 enzyme to manipulate gene expression levels and elucidate the adaptive bases of gene expression differences between *S. paradoxus* and *S. cerevisiae* (Naranjo et al. 2015). The ability to relocate single alleles or entire molecular pathways from one genetic background to another and to measure the phenotypic and fitness consequences is extremely promising.

Technical advances in the laboratory will however not be sufficient to understand what are the fitness determinants of yeast in nature. Additional tools in

microbial ecology are needed to be able to measure fitness in natural conditions, for instance, through reciprocal transplant experiments. Such approaches have recently been developed and allowed for the detection of fitness differences among yeast strains on leaf litter (Boynton et al. 2017). These tools, combined with the ability to introduce tractable DNA barcodes in the yeast genomes of any strain of interest (Maclean et al. 2017), could allow to profile the fitness of entire populations of known genotypes in natural conditions. Because the distributions of yeast species and populations appear to be largely determined by the ambient temperature, *S. paradoxus* could be a powerful tool to study the migration and evolution of genotypes through time, for instance, in the context of climate change or through annual environmental fluctuations.

Hybridization appears to be a driving force in yeast genome evolution. However, its contribution to yeast genomic diversity in nature is still largely unknown. Most yeast hybrids identified up until recently were associated with human activities (Hittinger 2013; Marsit et al. 2017), including those developed for biotechnological purposes (Peris et al. 2017b). However, population genomics of natural populations has revealed that this is happening relatively frequently without the need for direct human intervention (Leducq et al. 2016; Peris et al. 2016; Barbosa et al. 2016). The growth advantage observed in yeast hybrids (heterosis) suggests that hybridization could be favored because it has an immediate effect on fitness. However, outcrossing is rare, and yeast hybrids generally suffer from strong postzygotic fitness reduction. For instance, *S. cerevisiae*-*S. paradoxus* hybrids show heterosis but as little as 1% spore viability (Greig et al. 2002). Traces of hybridization between the two species was recently reported in South American populations (Barbosa et al. 2016), showing that this strong reproductive barrier can indeed be overcome and allow for introgression to proceed. By which mechanisms these barriers are overcome to allow for heterosis and other genetic interactions to favor the maintenance of genomic introgression appears to be a particularly promising avenue of research. Once again, the ability to borrow tools from *S. cerevisiae* to study hybrid genome instability (Herbst et al. 2017), protein-protein interactions (Piatkowska et al. 2013; Leducq et al. 2012), and transcriptional networks (Tirosh et al. 2009; Swain Lenz et al. 2014) will accelerate discoveries in this field.

Although chromosomal rearrangements were pointed out as an important molecular mechanism contributing to the emergence of reproductive isolation among incipient species (Charron et al. 2014b; Leducq et al. 2016), the accumulation of negative epistatic interactions in hybrids (first modeled by Bateson, Dobzhansky, and Mueller) stands as an important driver of speciation and is supported by solid empirical evidence (Presgraves 2010). Notably, many genetic interactions found to be responsible for the fitness decrease of hybrids between yeast species involve genes in both the nuclear and mitochondrial genomes (Chou and Leu 2010; Jhuang et al. 2017), suggesting an important role for cytonuclear genetic interactions in the generation of hybrid incompatibilities http://www.annualreviews.org/doi/abs/10.1146/annurev-ecolsys-110512-135758. The availability of population-scale genomic data, the vast body of knowledge on yeast systems biology, and the ability to perform high-throughput phenotypic measurements hold promising avenues for investigating how previously unmatched genetic variation interacts within hybrids

and translates into phenotypic changes at various levels. For instance, the use of the latest yeast genome-editing tools will make it possible to introduce heterozygosity at very specific loci in diploid yeast strains. This enables to dissect the gene-by-gene phenotypic consequences of hybridization in an ecologically realistic scenario, for instance, by using genetic variation segregating in natural populations.

Acknowledgments The authors thank Anna Fijarczyk and Souhir Marsit for comments on the manuscript. This work was supported by a NSERC Discovery Grant to CRL. CRL holds the Canada Research Chair in Evolutionary Cell and Systems Biology.

References

Almeida P, Barbosa R, Bensasson D, Goncalves P, Sampaio JP. Adaptive divergence in wine yeasts and their wild relatives suggests a prominent role for introgressions and rapid evolution at noncoding sites. Mol Ecol. 2017;26(7):2167–82. https://doi.org/10.1111/mec.14071.

April J, Hanner RH, Dion-Cote AM, Bernatchez L. Glacial cycles as an allopatric speciation pump in north-eastern American freshwater fishes. Mol Ecol. 2013;22(2):409–22. https://doi.org/10.1111/mec.12116.

Barbosa R, Almeida P, Safar SV, Santos RO, Morais PB, Nielly-Thibault L, Leducq JB, Landry CR, Goncalves P, Rosa CA, Sampaio JP. Evidence of natural hybridization in Brazilian wild lineages of Saccharomyces cerevisiae. Genome Biol Evol. 2016;8(2):317–29. https://doi.org/10.1093/gbe/evv263.

Bernardes J, Stelkens RB, Greig D. Heterosis in hybrids within and between yeast species. J Evol Biol. 2016;30(3):538–48. https://doi.org/10.1111/jeb.13023.

Birky CW Jr. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Annu Rev Genet. 2001;35:125–48. https://doi.org/10.1146/annurev.genet.35.102401.090231.

Boynton PJ, Stelkens R, Kowallik V, Greig D. Measuring microbial fitness in a field reciprocal transplant experiment. Mol Ecol Resour. 2017;17(3):370–80. https://doi.org/10.1111/1755-0998.12562.

Breton S, Stewart DT. Atypical mitochondrial inheritance patterns in eukaryotes. Genome. 2015;58(10):423–31. https://doi.org/10.1139/gen-2015-0090.

Charron G, Landry CR. No evidence for extrinsic post-zygotic isolation in a wild Saccharomyces yeast system. Biol Lett. 2017;13(6):20170197. https://doi.org/10.1098/rsbl.2017.0197.

Charron G, Leducq JB, Bertin C, Dube AK, Landry CR. Exploring the northern limit of the distribution of *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* in North America. FEMS Yeast Res. 2014a;14(2):281–8. https://doi.org/10.1111/1567-1364.12100.

Charron G, Leducq JB, Landry CR. Chromosomal variation segregates within incipient species and correlates with reproductive isolation. Mol Ecol. 2014b;23(17):4362–72. https://doi.org/10.1111/mec.12864.

Cheeseman K, Ropars J, Renault P, Dupont J, Gouzy J, Branca A, Abraham AL, Ceppi M, Conseiller E, Debuchy R, Malagnac F, Goarin A, Silar P, Lacoste S, Sallet E, Bensimon A, Giraud T, Brygoo Y. Multiple recent horizontal transfers of a large genomic region in cheese making fungi. Nat Commun. 2014;5:2876. https://doi.org/10.1038/ncomms3876.

Chen SJ, YH W, Huang HY, Wang CC. *Saccharomyces cerevisiae* possesses a stress-inducible glycyl-tRNA synthetase gene. PLoS One. 2012;7(3):e33363. https://doi.org/10.1371/journal.pone.0033363.

Chiapello H, Mallet L, Guerin C, Aguileta G, Amselem J, Kroj T, Ortega-Abboud E, Lebrun MH, Henrissat B, Gendrault A, Rodolphe F, Tharreau D, Fournier E. Deciphering genome content and

evolutionary relationships of isolates from the fungus Magnaporthe oryzae attacking different host plants. Genome Biol Evol. 2015;7(10):2896–912. https://doi.org/10.1093/gbe/evv187.

Chou JY, Leu JY. Speciation through cytonuclear incompatibility: insights from yeast and implications for higher eukaryotes. BioEssays. 2010;32(5):401–11. https://doi.org/10.1002/bies.200900162.

Desjardins CA, Giamberardino C, Sykes SM, Yu CH, Tenor JL, Chen Y, Yang T, Jones AM, Sun S, Haverkamp MR, Heitman J, Litvintseva AP, Perfect JR, Cuomo CA. Population genomics and the evolution of virulence in the fungal pathogen Cryptococcus neoformans. Genome Res. 2017;27(7):1207–19. https://doi.org/10.1101/gr.218727.116.

Eberlein C, Nielly-Thibault L, Maaroufi H, Dube AK, Leducq JB, Charron G, Landry CR. The rapid evolution of an ohnolog contributes to the ecological specialization of incipient yeast species. Mol Biol Evol. 2017;34(9):2173–86. https://doi.org/10.1093/molbev/msx153.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinf. 2009;10:48. https://doi.org/10.1186/1471-2105-10-48.

Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, Taylor JW. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proc Natl Acad Sci U S A. 2011;108(7):2831–6. https://doi.org/10.1073/pnas.1014971108.

Filteau M, Charron G, Landry CR. Identification of the fitness determinants of budding yeast on a natural substrate. ISME J. 2016;11(4):959–71. https://doi.org/10.1038/ismej.2016.170.

Ford CB, Funt JM, Abbey D, Issi L, Guiducci C, Martinez DA, Delorey T, Li BY, White TC, Cuomo C, Rao RP, Berman J, Thompson DA, Regev A. The evolution of drug resistance in clinical isolates of Candida albicans. elife. 2015;4:e00662. https://doi.org/10.7554/eLife.00662.

Gabaldon T, Martin T, Marcet-Houben M, Durrens P, Bolotin-Fukuhara M, Lespinet O, Arnaise S, Boisnard S, Aguileta G, Atanasova R, Bouchier C, Couloux A, Creno S, Almeida Cruz J, Devillers H, Enache-Angoulvant A, Guitard J, Jaouen L, Ma L, Marck C, Neuveglise C, Pelletier E, Pinard A, Poulain J, Recoquillay J, Westhof E, Wincker P, Dujon B, Hennequin C, Fairhead C. Comparative genomics of emerging pathogens in the Candida glabrata clade. BMC Genomics. 2013;14:623. https://doi.org/10.1186/1471-2164-14-623.

Goncalves P, Valerio E, Correia C, de Almeida JM, Sampaio JP. Evidence for divergent evolution of growth temperature preference in sympatric Saccharomyces species. PLoS One. 2011;6(6):e20739. https://doi.org/10.1371/journal.pone.0020739.

Greig D, Louis EJ, Borts RH, Travisano M. Hybrid speciation in experimental populations of yeast. Science. 2002;298(5599):1773–5. https://doi.org/10.1126/science.1076374.

Hayat S, Hayat Q, Alyemeni MN, Wani AS, Pichtel J, Ahmad A. Role of proline under changing environments: a review. Plant Signal Behav. 2012;7(11):1456–66. https://doi.org/10.4161/psb.21949.

Herbst RH, Bar-Zvi D, Reikhav S, Soifer I, Breker M, Jona G, Shimoni E, Schuldiner M, Levy AA, Barkai N. Heterosis as a consequence of regulatory incompatibility. BMC Biol. 2017;15(1):38. https://doi.org/10.1186/s12915-017-0373-7.

Hittinger CT. Saccharomyces diversity and evolution: a budding model genus. Trends Genet. 2013;29(5):309–17. https://doi.org/10.1016/j.tig.2013.01.002.

Hyma KE, Fay JC. Mixing of vineyard and oak-tree ecotypes of *Saccharomyces cerevisiae* in North American vineyards. Mol Ecol. 2013;22(11):2917–30. https://doi.org/10.1111/mec.12155.

Jhuang HY, Lee HY, Leu JY. Mitochondrial-nuclear co-evolution leads to hybrid incompatibility through pentatricopeptide repeat proteins. EMBO Rep. 2017;18(1):87–101. 10.15252/embr.201643311.

Johnson LJ, Koufopanou V, Goddard MR, Hetherington R, Schafer SM, Burt A. Population genetics of the wild yeast *Saccharomyces paradoxus*. Genetics. 2004;166(1):43–52.

Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature. 2004;428(6983):617–24. https://doi.org/10.1038/nature02424.

Koufopanou V, Hughes J, Bell G, Burt A. The spatial scale of genetic differentiation in a model organism: the wild yeast *Saccharomyces paradoxus*. Philos Trans R Soc Lond Ser B Biol Sci. 2006;361(1475):1941–6. https://doi.org/10.1098/rstb.2006.1922.

Kowallik V, Greig D. A systematic forest survey showing an association of *Saccharomyces paradoxus* with oak leaf litter. Environ Microbiol Rep. 2016;8(5):833–41. https://doi.org/10.1111/1758-2229.12446.

Kuehne HA, Murphy HA, Francis CA, Sniegowski PD. Allopatric divergence, secondary contact, and genetic isolation in wild yeast populations. Curr Biol. 2007;17(5):407–11. https://doi.org/10.1016/j.cub.2006.12.047.

Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, Desai MM. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature. 2013;500(7464):571–4. https://doi.org/10.1038/nature12344.

Laureau R, Loeillet S, Salinas F, Bergstrom A, Legoix-Ne P, Liti G, Nicolas A. Extensive recombination of a yeast diploid hybrid through meiotic reversion. PLoS Genet. 2016;12(2):e1005781. https://doi.org/10.1371/journal.pgen.1005781.

Leducq JB, Charron G, Diss G, Gagnon-Arsenault I, Dube AK, Landry CR. Evidence for the robustness of protein complexes to inter-species hybridization. PLoS Genet. 2012;8(12):e1003161. https://doi.org/10.1371/journal.pgen.1003161.

Leducq JB, Charron G, Samani P, Dube AK, Sylvester K, James B, Almeida P, Sampaio JP, Hittinger CT, Bell G, Landry CR. Local climatic adaptation in a widespread microorganism. Proc Biol Sci. 2014;281(1777):20132472. https://doi.org/10.1098/rspb.2013.2472.

Leducq JB, Nielly-Thibault L, Charron G, Eberlein C, Verta JP, Samani P, Sylvester K, Hittinger CT, Bell G, Landry CR. Speciation driven by hybridization and chromosomal plasticity in a wild yeast. Nat Microbiol. 2016;1:15003. https://doi.org/10.1038/nmicrobiol.2015.3.

Leducq JB, Henault M, Charron G, Nielly-Thibault L, Terrat Y, Fiumera HL, Jesse Shapiro B, Landry CR. Mitochondrial recombination and introgression during speciation by hybridization. Mol Biol Evol. 2017;34(8):1947–59. https://doi.org/10.1093/molbev/msx139.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, Tsai IJ, Bergman CM, Bensasson D, O'Kelly MJ, van Oudenaarden A, Barton DB, Bailes E, Nguyen AN, Jones M, Quail MA, Goodhead I, Sims S, Smith F, Blomberg A, Durbin R, Louis EJ. Population genomics of domestic and wild yeasts. Nature. 2009;458(7236):337–41. https://doi.org/10.1038/nature07743.

Maclean CJ, Metzger BPH, Yang JR, Ho WC, Moyers B, Zhang J. Deciphering the genic basis of yeast fitness variation by simultaneous forward and reverse genetics. Mol Biol Evol. 2017;34(10):2486–502. https://doi.org/10.1093/molbev/msx151.

Marsit S, Leducq JB, Durand E, Marchant A, Filteau M, Landry CR. Evolutionary biology through the lens of budding yeast comparative genomics. Nat Rev Genet. 2017;18(10):581–98.

McDonald MJ, Rice DP, Desai MM. Sex speeds adaptation by altering the dynamics of molecular evolution. Nature. 2016;531(7593):233–6. https://doi.org/10.1038/nature17143.

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. https://doi.org/10.1371/journal.pbio.1001127.

Murphy HA, Zeyl CW. A potential case of reinforcement in a facultatively sexual unicellular eukaryote. Am Nat. 2015;186(2):312–9. https://doi.org/10.1086/682071.

Naranjo S, Smith JD, Artieri CG, Zhang M, Zhou Y, Palmer ME, Fraser HB. Dissecting the genetic basis of a complex cis-regulatory adaptation. PLoS Genet. 2015;11(12):e1005751. https://doi.org/10.1371/journal.pgen.1005751.

Naumov GI, Naumova ES, Sniegowski PD. *Saccharomyces paradoxus* and *Saccharomyces cerevisiae* are associated with exudates of North American oaks. Can J Microbiol. 1998;44(11):1045–50.

Newsham K, Hopkins D, Carvalhais L, Fretwell P, Rushton S, O'Donnell A, Dennis P. Relationship between soil fungal diversity and temperature in the maritime Antarctic. Nat Clim Chang. 2016;6:182–6.

Peris D, Langdon QK, Moriarty RV, Sylvester K, Bontrager M, Charron G, Leducq JB, Landry CR, Libkind D, Hittinger CT. Complex ancestries of lager-brewing hybrids were shaped by standing variation in the wild yeast Saccharomyces eubayanus. PLoS Genet. 2016;12(7): e1006155. https://doi.org/10.1371/journal.pgen.1006155.

Peris D, Arias A, Orlic S, Belloch C, Perez-Traves L, Querol A, Barrio E. Mitochondrial introgression suggests extensive ancestral hybridization events among Saccharomyces species. Mol Phylogenet Evol. 2017a;108:49–60. https://doi.org/10.1016/j.ympev.2017.02.008.

Peris D, Moriarty RV, Alexander WG, Baker E, Sylvester K, Sardi M, Langdon QK, Libkind D, Wang QM, Bai FY, Leducq JB, Charron G, Landry CR, Sampaio JP, Goncalves P, Hyma KE, Fay JC, Sato TK, Hittinger CT. Hybridization and adaptive evolution of diverse Saccharomyces species for cellulosic biofuel production. Biotechnol Biofuels. 2017b;10:78. https://doi.org/10.1186/s13068-017-0763-7.

Piatkowska EM, Naseeb S, Knight D, Delneri D. Chimeric protein complexes in hybrid species generate novel phenotypes. PLoS Genet. 2013;9(10):e1003836. https://doi.org/10.1371/journal.pgen.1003836.

Presgraves DC. The molecular evolutionary basis of species formation. Nat Rev Genet. 2010;11 (3):175–80. https://doi.org/10.1038/nrg2718.

Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 2014;197(2):573–89. https://doi.org/10.1534/genetics.114.164350.

Redzepovic S, Orlic S, Sikora S, Majdak A, Pretorius IS. Identification and characterization of *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* strains isolated from Croatian vineyards. Lett Appl Microbiol. 2002;35(4):305–10.

Replansky T, Koufopanou V, Greig D, Bell G. Saccharomyces sensu stricto as a model system for evolution and ecology. Trends Ecol Evol. 2008;23(9):494–501. https://doi.org/10.1016/j.tree.2008.05.005.

Richardson SM, Mitchell LA, Stracquadanio G, Yang K, Dymond JS, Di Carlo JE, Lee D, Huang CL, Chandrasegaran S, Cai Y, Boeke JD, Bader JS. Design of a synthetic yeast genome. Science. 2017;355(6329):1040–4. https://doi.org/10.1126/science.aaf4557.

Robinson HA, Pinharanda A, Bensasson D. Summer temperature can predict the distribution of wild yeast populations. Ecol Evol. 2016;6(4):1236–50. https://doi.org/10.1002/ece3.1919.

Samani P, Low-Decarie E, McKelvey K, Bell T, Burt A, Koufopanou V, Landry CR, Bell G. Metabolic variation in natural populations of wild yeast. Ecol Evol. 2015;5(3):722–32. https://doi.org/10.1002/ece3.1376.

Sampaio JP, Goncalves P. Natural populations of Saccharomyces kudriavzevii in Portugal are associated with oak bark and are sympatric with *S. cerevisiae* and *S. paradoxus*. Appl Environ Microbiol. 2008;74(7):2144–52. https://doi.org/10.1128/AEM.02396-07.

Sanchez-Perez G, Mira A, Nyiro G, Pasic L, Rodriguez-Valera F. Adapting to environmental changes using specialized paralogs. Trends Genet. 2008;24(4):154–8. https://doi.org/10.1016/j.tig.2008.01.002.

Schumer M, Rosenthal GG, Andolfatto P. How common is homoploid hybrid speciation? Evolution. 2014;68(6):1553–60. https://doi.org/10.1111/evo.12399.

Shannon C, Rao A, Douglass S, Criddle RS. Recombination in yeast mitochondrial DNA. J Supramol Struct. 1972;1(2):145–52. https://doi.org/10.1002/jss.400010207.

Shapira R, Levy T, Shaked S, Fridman E, David L. Extensive heterosis in growth of yeast hybrids is explained by a combination of genetic models. Heredity (Edinb). 2014;113(4):316–26. https://doi.org/10.1038/hdy.2014.33.

Sniegowski PD, Dombrowski PG, Fingerman E. *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. FEMS Yeast Res. 2002;1(4):299–306.

Stefanini I, Dapporto L, Legras JL, Calabretta A, Di Paola M, De Filippo C, Viola R, Capretti P, Polsinelli M, Turillazzi S, Cavalieri D. Role of social wasps in *Saccharomyces cerevisiae* ecology and evolution. Proc Natl Acad Sci U S A. 2012;109(33):13398–403. https://doi.org/10.1073/pnas.1208362109.

Swain Lenz D, Riles L, Fay JC. Heterochronic meiotic misexpression in an interspecific yeast hybrid. Mol Biol Evol. 2014;31(6):1333–42. https://doi.org/10.1093/molbev/msu098.

Sylvester K, Wang QM, James B, Mendez R, Hulfachor AB, Hittinger CT. Temperature and host preferences drive the diversification of Saccharomyces and other yeasts: a survey and the discovery of eight new yeast species. FEMS Yeast Res. 2015;15(3):fov002. https://doi.org/10.1093/femsyr/fov002.

Tirosh I, Reikhav S, Levy AA, Barkai N. A yeast hybrid provides insight into the evolution of gene expression regulation. Science. 2009;324(5927):659–62. https://doi.org/10.1126/science.1169766.

Tsai IJ, Bensasson D, Burt A, Koufopanou V. Population genomics of the wild yeast *Saccharomyces paradoxus*: quantifying the life cycle. Proc Natl Acad Sci U S A. 2008;105(12):4957–62. https://doi.org/10.1073/pnas.0707314105.

Turner RJ, Lovato M, Schimmel P. One of two genes encoding glycyl-tRNA synthetase in *Saccharomyces cerevisiae* provides mitochondrial and cytoplasmic functions. J Biol Chem. 2000;275(36):27681–8. https://doi.org/10.1074/jbc.M003416200.

Turner BC, Perkins DD, Fairfield A. Neurospora from natural populations: a global study. Fungal Genet Biol. 2001;32(2):67–92. https://doi.org/10.1006/fgbi.2001.1247.

Wong Miller KM, Bracewell RR, Eisen MB, Bachtrog D. Patterns of genome-wide diversity and population structure in the Drosophila athabasca species complex. Mol Biol Evol. 2017;34(8):1912–23. https://doi.org/10.1093/molbev/msx134.

Xia W, Nielly-Thibault L, Charron G, Landry CR, Kasimer D, Anderson JB, Kohn LM. Population genomics reveals structure at the individual, host-tree scale and persistence of genotypic variants of the undomesticated yeast *Saccharomyces paradoxus* in a natural woodland. Mol Ecol. 2017;26(4):995–1007. https://doi.org/10.1111/mec.13954.

Yue JX, Li J, Aigrain L, Hallin J, Persson K, Oliver K, Bergstrom A, Coupland P, Warringer J, Lagomarsino MC, Fischer G, Durbin R, Liti G. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nat Genet. 2017;49(6):913–24. https://doi.org/10.1038/ng.3847.

Zhang H, Skelton A, Gardner RC, Goddard MR. *Saccharomyces paradoxus* and *Saccharomyces cerevisiae* reside on oak trees in New Zealand: evidence for migration from Europe and interspecies hybrids. FEMS Yeast Res. 2010;10(7):941–7. https://doi.org/10.1111/j.1567-1364.2010.00681.x.

Part IV
Population Genomics of Viruses

Population Genomics of Plant Viruses

Israel Pagán and Fernando García-Arenal

Abstract For more than one century, studies of plant viruses have broken paths in many fields of biology. More recently, studies of plant viruses have also been pioneer in population genomics. In the past few decades, there has been a significant advance in the number, sophistication, and quality of molecular techniques and bioinformatics tools for the genetic characterization of virus populations. This has broadened current knowledge on the mechanisms that generate genetic diversity and on the evolutionary forces and ecological factors that shape the genetic structure and dynamics of plant virus populations. This chapter aims at summarizing this knowledge, and it is structured around three major levels at which plant virus populations have been studied:

1. The within-host level, that is, the analysis of the genetic diversity of virus populations during plant colonization and of how phenomena such as co-/superinfection exclusion and population bottlenecks determine population structure
2. The between-host level, which includes studies on genetic diversity of virus populations in the host plant population and on the ecological factors shaping the genetic structure of the virus populations
3. The community level, which adddresses current studies on the genetic diversity of virus communities in multiple infected hosts and of multi-host-multivirus interactions

In sum, we provide an overview of current understanding on the population genomics of plant viruses at every level of population organization.

I. Pagán · F. García-Arenal (✉)
Centro de Biotecnología y Genómica de Plantas UPM-INIA, Universidad Politécnica de Madrid, Pozuelo de Alarcón, Madrid, Spain

E.T.S. Ingeniería Agronómica, Alimentaria y de Biosistemas, Madrid, Spain
e-mail: fernando.garciaarenal@upm.es

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_15,

Keywords Ecosystem biodiversity · Genomics of plant viruses · Multiplicity of infection · Plant virome assembly · Plant-virus coevolution · Time scale of plant virus evolution · Virus coinfection · Virus-virus interactions

1 Introduction

Most plant viruses have genome sizes ranging between 5 and 15 kb, and one of the defining traits of viruses with small genomes is their high capacity to generate genetic diversity (Holmes 2009). This capacity has been proposed to be one of the main reasons for their biological success. Indeed, plant viruses have been found to be ubiquitous in ecosystems inhabited by plants (Roossinck 2017), which has been linked to the continuous appearance of new virus genotypes or species that colonize new areas or previously non-infected host populations (García-Arenal and McDonald 2003; Holmes 2009; Elena et al. 2014).

1.1 Mechanisms of Generation of Genetic Diversity in Plant Virus Populations

The high capacity to generate genetic diversity derives from a combination of factors. The first major factor is the high mutation rates of plant viruses. Most plant viruses have RNA genomes that encode RNA-dependent RNA polymerases, which lack proofreading activity resulting in high rates of nucleotide misincorporation (Drake and Holland 1999). Mutation rates in RNA plant viruses, first estimated for *Tobacco mosaic virus* (TMV) (Malpica et al. 2002), are in the range of 10^{-3} to 10^{-6} nucleotide substitutions per site per round of replication, similar to those reported for RNA viruses infecting bacteria or animals (Malpica et al. 2002; Sanjuán et al. 2009, 2010; Tromas and Elena 2010). Thus, viral mutation rates are several orders of magnitude higher than those of their host plants, estimated to be around 10^{-9} (Kay et al. 2006). High mutation rates allow viruses to explore large portions of the mutational space. The higher the mutation rate the larger the probability to generate virus genotypes fitter in new environments. However, it has been well established that most mutations in RNA plant viruses are highly deleterious, with an important fraction being lethal, whereas neutral mutations are considerably less frequent (Carrasco et al. 2007; Hillung et al. 2015). In this scenario, high mutation rates may lead to a high burden of deleterious mutations (i.e., mutational load) in the virus populations and ultimately to extinction (Chao 1990). Here the second major factor associated with the high genetic diversity of virus populations becomes important: plant viruses generally have large population sizes (García-Arenal et al. 2001). For instance, the number of infectious units of TMV has been estimated as about 10^7 per infected mesophyll cell of experimentally infected tobacco plants (Harrison 1956; Malpica et al. 2002), and the number of *Tobacco*

mild green mosaic virus (TMGMV) particles in field-infected *Nicotiana glauca* leaves has been estimated as 10^{11} (Moya et al. 1993). Also, the effective population size of TMGMV in *N. glauca* was estimated to be of about 10^5 (Moya et al. 1993), and that of 12 potyviruses have been estimated to be in the order of 10^4 (Hughes 2009). As a consequence, deleterious mutations have small chances to persist in the virus population, as these will be quickly purged by negative selection (Elena and Sanjuán 2005). However, it should be noted that factors, such as variation in replication potential among genotypes, differences in generation time among infected cells, and severe reductions in population size at various steps during the virus life cycle, might lead to effective population sizes (roughly, the number of individuals in the population that pass their genes to the next generation) much smaller than the census size of the population. Finally, the third major factor is the much shorter generation time of RNA viruses (minutes to hours) (Wu et al. 1994) than that of plants. Thus, virus evolution occurs in different time scales from that of their host plants.

Besides mutation, genetic diversity in virus populations can also be generated by recombination, that is, the exchange of genomic fragments between genotypes. Recombination can occur between genotypes of the same or of different virus species. Recombination rates have been estimated for plant viruses to be between 10^{-5} and 10^{-8} (Froissart et al. 2005; Tromas et al. 2014b), so that the contribution of recombination to the generation of genetic diversity would be in the same order as that of mutation. Recombination may represent an evolutionary advantage for viruses because (1) it can create fitter genotypes more rapidly than mutation and (2) it might purge deleterious mutations from virus populations, thereby preventing the decrease in overall fitness in clonal populations due to the accumulation of deleterious mutations (Muller's ratchet) (Pressing and Reanney 1984; Chao 1990; García-Arenal et al. 2001; Hull 2014; Moya et al. 2004). Genetic exchange may also result from reassortment of genomic segments in viruses with segmented or multipartite genomes, with similar genetic and evolutionary consequences as recombination sensu stricto. Indeed, segment reassortment is also called pseudorecombination by plant virologists.

1.2 Processes that Shape the Genetic Diversity of Plant Virus Populations

Mutation and recombination are a consequence of the mechanisms of virus replication, and the resulting new genotypes are therefore in principle randomly generated (but see Bujarski 2013 for exceptions). Central to understanding the population genomics of plant viruses is which mutations become fixed in the population and how fast they do so (Duffy et al. 2008) or, more generally, what determines the frequency in the populations of the genotypes generated through mutation and recombination. In the absence of migration, the number and frequency of these

genotypes in the population (i.e., the genetic structure of the population) is the result of two different evolutionary processes: genetic drift and selection (García-Arenal et al. 2001; Hartl and Clark 2007; Acosta-Leal et al. 2011). Genetic drift occurs when populations of organisms are not large enough to ensure that each genotype will have progeny in the next generation. As a consequence, the genotypes passed into the next generation are randomly sampled from the mother population, regardless of their relative fitness. Genetic drift may be particularly relevant in plant virus populations during the severe reductions in their population size (population bottlenecks) that may occur along the virus life cycle, for instance, at the infection of a new host population, a new host plant, or new organs within a host plant (Sacristán et al. 2003; Gutiérrez et al. 2012a; Fabre et al. 2014). Genetic drift reduces the genetic diversity of populations and increases the diversity among populations. Also, because the genotypes that start a new population are not selected according to their fitness, genetic drift counters the effects of selection (García-Arenal et al. 2001; Acosta-Leal et al. 2011). On the other hand, selection is a directional process by which genotypes that are fittest in a given environment will increase in frequency in the population (positive selection), whereas less fit genotypes will decrease in frequency (negative or purifying selection). As is the case for genetic drift, selection results in a decrease of the population diversity and may also cause an increased diversity between populations, if under different selection pressures, so that the effects of selection and genetic drift are often difficult to distinguish. When selection has been differentiated from genetic drift, selection has been associated with every life history trait of plant viruses, such as survival in the environment due to higher structural stability of the virus particles (Fraile et al. 2014), adaptation to the host plant resulting in more effective within-host multiplication (Hillung et al. 2015), and adaptation to the transmission mode resulting in more efficient between-host transmission (Hajimorad et al. 2011; Pagán et al. 2014).

This chapter aims at summarizing current knowledge on the interplay between the mechanisms that generate genetic diversity and the evolutionary forces that shape the genetic structure and dynamics of plant virus populations. We have structured the chapter around three major levels at which plant virus populations have been studied: the within-host level, the between-host level, and the community level.

2 Genomics of Within-Host Plant Virus Populations

2.1 Within-Host Virus Genetic Diversity

Early evidence for genetic heterogeneity of within-host plant virus populations was provided several decades ago (McKinney 1935; Rochow 1972). However, the analysis of the genetic structure of virus populations within the plant level became a topic of research considerably more recently. More recent research has shown that nucleic acid extracts obtained from plants systemically infected with both DNA and RNA viruses contain genetically heterogeneous virus populations, even if the

infections were generated from biologically active cDNA clones (e.g., García-Arenal et al. 2001; García-Arenal and Fraile 2008). Since then, accumulating evidence based on population genomic analyses has shown that within-plant virus populations may be genetically diverse and plastic.

The within-host genetic diversity of plant viruses has been analyzed in several plant-virus combinations, most of them involving crops. Thus, within-host populations of *Zucchini yellow mosaic virus* (ZYMV) in squash; of *Cucumber mosaic virus* (CMV) in tomato, pepper, and squash; and of *Plum pox virus* (PPV) and *Prunus necrotic ringspot virus* (PNRSV) in *Prunus* spp. trees have been found to consist of a cloud of genotypes (Jridi et al. 2006; Alí and Roossinck 2010; Simmons et al. 2012; Dunham et al. 2014; Kinoti et al. 2017). Analyses in wild plants reported the same trends. Populations of *Endive necrotic mosaic virus* (ENMV) infecting *Tragopogon pratensis* and of *Asclepias asymptomatic virus* in *Euphorbia marginata* showed high levels of genetic diversity within a single plant (Hackett et al. 2009; Piry et al. 2017). Interestingly, these analyses showed that such genetic diversity is not homogeneous across the plant, so that different parts of the plant host virus populations that differ between them. For instance, Jridi et al. (2006) found a nonrandom association between physical distance (distance between tree leaves from which samples were collected) and PPV genetic distance. Nucleotide sequence analyses of *Asclepias asymptomatic virus* by Hackett et al. (2009) showed differences in the genetic structure of the virus in the different *Euphorbia marginata* organs. In the same sense, several analyses using high-coverage deep sequencing data reported the presence of different viral genotypes depending on the plant organ sampled (Dunham et al. 2014; Kinoti et al. 2017).

Besides describing the within-host genetic structure of plant virus populations, some studies have also attempted to understand the evolutionary mechanisms involved. Dunham et al. (2014) reported that 80% of the ZYMV genotypes found in squash plants were sampled only once and consisted of synonymous mutations. This could be explained if most genotypes generated during plant colonization were deleterious. However, the authors suggested that, because most mutations were synonymous, their observations would be compatible with neutral evolution and genetic drift likely being the major drivers of ZYMV genetic diversity. Sacristán et al. (2003) also reported stochastic processes influencing TMV genotype composition in different systemically infected leaves of pepper. Alí and Roossinck (2010) obtained similar results when they analyzed CMV genotype composition during pepper, *Nicotiana benthamiana,* and squash colonization. However, genetic drift is not always the dominating evolutionary force. The relative importance of selection and genetic drift, and/or the sense and magnitude of selection pressures, might depend on the host-virus combination. For instance, CMV within-host populations in tomato and tobacco have been shown to be under negative selection (Li and Roossinck 2004; Alí and Roossinck 2010). Also, Kinoti et al. (2017) found that selection pressures in the PNRSV populations within *Prunus* spp. trees varied depending on the genomic segment analyzed, with genetic segment 1 being under neutral evolution and segments 2 and 3 mostly accumulating non-synonymous mutations. Analyses of the fitness of the genotypes generated during plant

colonization have been seldom reported. Perhaps the most detailed analysis is that of *Tobacco etch virus* (TEV) mutants generated during adaptation to a single *Arabidopsis thaliana* genotype (Hillung et al. 2015). Most of the mutations were deleterious or neutral and did not result in increased fitness, suggesting a major role of genetic drift in shaping TEV genetic diversity. In addition, the authors found that mutational effects were mostly multiplicative, with few cases of significant epistasis. Thus, the virus population was composed of genotypes that, regardless they had one or several mutations, had similar fitness levels. Complementary work by the same group demonstrated that the larger the plant colonization time, the greater the chances of fitter genotypes being selected (Zwart et al. 2014).

Together, reported analyses indicate that the within-host genetic diversity of plant viruses is spatially structured and reveal that both selection and genetic drift shape that structure, with the relative importance of these two evolutionary forces depending on the specific host-virus combination. Several factors associated with virus population dynamics have been proposed to determine the relative importance of selection and genetic drift in the within-host genetic structure of virus populations. Thus, severe population bottlenecks at different stages of plant colonization result in genetic drift, whereas competition for resources as a result of coinfection by more than one genotype of the same cell or organ could exert a selection pressure on the virus population (Frank 2001; Holmes 2009; García-Arenal and Fraile 2013). The next sections focus on these factors.

2.2 *Virus Coinfection and Superinfection Exclusion in Host Cells*

Evidence of a spatially heterogeneous distribution of virus genotypes within the plant was provided long before this genetic diversity could be characterized. McKinney (1935) reported the spatial separation of TMV genotypes causing common mosaic and yellow mosaic from tobacco leaves showing both symptoms. Similarly, Hull and Plaski (1970) reported the spatial separation of two strains of *Alfalfa mosaic virus* (AMV) that induced specific aggregation bodies in the cytoplasm of infected cells, on the basis of electron microscopy examination of samples from different parts of the same leaf. Later, Hall et al. (2001) reported nonuniform distribution of two strains of *Wheat streak mosaic virus* (WSMV) that multiplied to similar levels in coinfected wheat leaves. Molecular detection of each strain in disks from coinfected leaves allowed these authors to report the first quantitative description of genotype distribution of a virus in different leaf areas, as there were disks in which only one strain was detected and disks in which both strains were detected either in similar or in different amounts. The use of viruses labeled with different fluorescent proteins, coupled with detection of single-cell infection by confocal laser scanning microscopy, facilitated the location of leaf areas infected by the different virus genotypes. Thus, spatial separation of variously GFP- and RFP-labeled

genotypes has been reported for CMV, *Potato virus X* (PVX), *Plum pox virus* (PPV), TVMV, *Bean common mosaic virus* (BCMV), and *Apple latent spherical virus* (ALSV): red and green fluorescence occurred in discrete cell clusters, and only a small number of cells at the contact area of these clusters showed both red and green fluorescence, i.e., coinfection by both genotypes (Divéki et al. 2002; Dietrich and Maiss 2003; Takeshita et al. 2004; Takahashi et al. 2007).

The abovementioned studies suggest that infection by one virus of a given genotype results in some type of exclusion mechanism limiting super- and/or coinfection by a virus of a second genotype. To better describe virus exclusion, González-Jara et al. (2009) co-inoculated *N. benthamiana* plants with GFP- and RFP-labeled TMV and determined the fraction of protoplasts isolated from leaves at different times post-inoculation that fluoresced green or red. Only a small fraction of infected cells (2–5%) were coinfected by both TMV genotypes. The fraction of coinfected cells at later times after infection (2–3%) was smaller than expected had they been infected by random. In both single- and mixed-infected cells, the kinetics of the number of infected cells over time followed a logistic distribution. However, a plateau was reached much earlier in mixed- than in single-infected cells. As it could be expected that the fraction of coinfected cells would increase as foci of single-infected cells coalesce, the observation of an early plateau for the fraction of coinfected cells suggests the existence of mechanisms preventing superinfection of already infected cells. Using a similar approach, Miyashita and Kishino (2010) analyzed the colonization of barley leaves by *Soil-borne wheat mosaic virus* (SBWMV) using two YFP- and CFP-labeled genotypes. Although soon after inoculation both genotypes coinfected cells, their frequency decreased as the foci expanded, and as soon as 3 days post-inoculation, most cells at the periphery of the foci were single-infected. Moreover, 83% of the cells adjacent to the initially coinfected ones were coinfected, whereas only 61% of the cells in the subsequent rows were so. These results suggest again the operation of exclusion mechanisms. Cross-protection and RNA silencing (Roossinck 2005; Bergua et al. 2014; Donaire et al. 2016), and/or competitive displacement of one genotype in coinfected cells (González-Jara et al. 2009), have been proposed as the basis for these mechanisms. Competition among viruses of different genotypes in a coinfected cell may result in a smaller fitness of each genotype as compared with their fitness in single infection (Levontin 1970; Frank 2001). Thus, limiting coinfection may be a major driver of within-host population genetic diversity and a selective advantage for the virus.

2.3 Multiplicity of Infection During Plant Colonization

Despite the consistent evidence for spatial exclusion of virus genotypes within the infected cell, the work discussed above shows that cell coinfections do occur and not infrequently. These observations have boosted the analyses of the multiplicity of infection (MOI), i.e., the number of virus particles or genomes that simultaneously infect a cell during plant colonization. In this sense, MOI is different than the number

of virus particles required to start an infection, which was shown from early date to be one in viruses with monopartite genomes (reviewed in García-Arenal and Fraile 2011). MOI is a relevant parameter in virus epidemiology and evolution, as it determines processes such as genetic exchange through recombination or reassortment of genomic segments, complementation of deleterious mutants and hence selection intensity on viral genes, hyperparasitism by molecular parasites such as RNA satellites, or the evolution of segmented genomes (Nee and Maynard-Smith 1990; Chao 1991; Szathmáry 1992; Simon and Bujarski 1994; Roossinck 1997; Worobey and Holmes 1999; García-Arenal et al. 2001; Tepfer 2002; Froissart et al. 2004). However, MOI values have been rarely estimated for plant viruses or for any viruses (Table 1).

González-Jara et al. (2009) estimated MOI values for TMV during colonization of *N. benthamiana* leaves, assuming that the probability of infection of a cell by GFP- and RFP-labeled viruses was independent and followed a binomial distribution. Average MOI values in inoculated leaves were of around 1 along the monitored infection period (2–17 dpi) both in inoculated and in systemically infected leaves (González-Jara et al. 2013). Miyashita and Kishino (2010) also estimated the value of MOI for SBWMV from the frequency of double- and single-infected cells in infection foci initiated by two virus genotypes, assuming a binomial distribution of the probability of infection and a Poisson distribution of the value of MOI. Their reported MOI values are in the same range as those for TMV: five to six founder genomes infect a new cell from an adjacent one. MOI estimates for cells two cells apart from the initially infected ones (~5) were slightly lower than for cells adjacent to the initially infected ones (~6), suggesting a decrease of MOI as infection progressed. However, Miyashita and Kishino (2010) made the cautionary comment that the MOI decrease with leaf colonization could be an artifact of the model, because as infection foci expand each infected cell makes contact with a decreasing number of uninfected cells. The model used by González-Jara et al. (2009, 2013) was not sensitive to this factor, as it assumed an equal probability of infection of all abutting cells. Analyses for other plant viruses have yielded MOI values in the same range as for TMV and SBWMV. For instance, a MOI of <1.5 has been estimated for

Table 1 Multiplicity of infection (MOI) estimates for plant viruses

Virus	Host	Method[a]	MOI	Reference
CaMV	*Brassica rapa*	Genetic marker	2–13	Gutiérrez et al. (2010)
CTV	*Citrus macrophylla*	FP-labeling	1.06–1.07	Bergua et al. (2014)
SBWMV	*Triticum aestivum*	FP-labeling	5.02–5.97	Miyashita and Kishino (2010)
TBSV	*Nicotiana benthamiana*	Genetic marker	1.76–3.98	Donaire et al. (2016)
TEV	*Nicotiana tabacum*	FP-labeling	1.00–1.43	Tromas et al. (2014a, b)
TMV	*Nicotiana benthamiana*	FP-labeling	1.17–7.00	González-Jara et al. (2009)
TMV	*Nicotiana benthamiana*	FP-labeling	1.01–1.18	González-Jara et al. (2013)
TuMV	*Brassica rapa*	FP-labeling	21.7–41.5	Gutiérrez et al. (2015)

[a]Genetic marker, virus genotype-specific sequence tag; FP-labeling, virus genotypes labeled with fluorescence proteins

TEV in *Nicotiana tabacum* (Tromas et al. 2014a), values between 1.6 and 3.9 have been calculated for *Tomato bushy stunt virus* (TBSV) in *N. benthamiana* (Donaire et al. 2016), and *Citrus tristeza virus* (CTV) MOI in *Citrus macrophylla* was estimated to be around 1 (Bergua et al. 2014). On the other hand, higher MOI values have been reported for *Cauliflower mosaic virus* (CaMV) (13) (Gutiérrez et al. 2010) and for TuMV (21.7–41.5) (Gutiérrez et al. 2015). If these MOI values are considered as Poisson distributed over the infected cells, most of them would be coinfected at the end of the colonization period. This would provide ample opportunity for recombination or for the complementation of lethal or deleterious mutations, which may have large impact on the genetic composition of the virus population.

2.4 Population Bottlenecks During Systemic Infection

Another factor involved in the within-host spatial genetic structure of plant virus populations is the existence of population bottlenecks associated with the colonization of new organs during systemic movement. This hypothesis was first tested by Sacristán et al. (2003), who used the segregation in the two first systemically infected leaves of two TMV genotypes co-inoculated at the same infectivity dosage for estimating the effective population number under a binomial model for the probability of infection. Estimates indicated effective founder sizes of 1–15, i.e., several orders of magnitude less than the 10^7–10^9 census of the TMV population in an infected leaf. Using a similar approach, French and Stenger (2003) estimated the effective WSMV founder population size of new wheat tillers from the data of Hall et al. (2001) at about 4, a value in the same range as that reported for TMV. Li and Roossinck (2004) utilized an artificial population of 12 CMV mutants to show that the virus population diversity decreased significantly when the population moved from the inoculated leaves to primary systemically infected leaves (six to eight mutants recovered) and decreased further as the systemic infection progressed (three to five mutants recovered). The elimination of a majority of the mutants was a stochastic process attributed to population bottlenecks. Comparable founder sizes (6–8) have been recently estimated by Thébaud and Michalakis (2016) for TEV infection in tobacco using data extracted from Tromas et al. (2014a).

In contrast with these two reports, Monsion et al. (2008) did not find any significant difference in the genetic composition of *Cauliflower mosaic virus* (CaMV) populations in different systemically infected leaves of turnip plants infected with a mixture of six mutants. Based on analyses of population differentiation, they came to the conclusion that founder populations for systemically infected leaves were large, in the order of several hundreds, with confidence intervals for the estimates extending into the thousands. They interpreted these results as a trait of CaMV, a reverse-transcribing DNA virus, while all previous evidence or estimates of severe bottlenecks during systemic colonization derived from the analysis of RNA viruses. However, another factor that might influence the results of Monsion et al. (2008) is the way they handled the infected plants. At odds with the other founder

size estimates, all inoculated and systemically infected turnip leaves were eliminated at 13 dpi, and analyses were done 32 days later in the 10–15 leaves that formed afterward. Analyses at such late times after infection do not allow the identification of the sources for the virus infection of the systemically infected leaves, and pruning the plants could alter the photoassimilate source-sink dynamics in the plant and consequently the virus transport to the new growing leaves. Indeed, more recent estimates of CaMV population bottlenecks during turnip colonization, using conditions similar to those utilized by Hall et al. (2001), Sacristán et al. (2003), and Li and Roossinck (2004), yielded much similar founder sizes at the 5th and 21th leaves (8.8 and 10.8, respectively). Larger founder sizes were observed at the 16th leaf (127), but this estimate had a large confidence interval (19.1–908.9) (Gutiérrez et al. 2012b).

In summary, most evidence indicates that population bottlenecks are severe during systemic colonization of new plant organs and that these bottlenecks, regardless of other phenomena such as exclusion mechanisms discussed at the cell level, may explain the observations on the spatial genetic structure of virus populations within the organs of infected plants.

3 Genomics of Between-Host Plant Virus Populations

The increasing number of tools for comparative genomic analyses of plant virus populations, the constant increase in the sophistication of sequence methods, and the exponential reduction of sequencing costs have allowed enormous amounts of sequence information to be obtained with a reasonable investment of resources and time. This has boosted the analysis of the population genomics of plant viruses not only within an individual plant (see previous section) but also at the population level. In this section, we will focus on three major aspects of plant virus evolution that benefit from access to such a large amount of genomic information: the time scale of virus evolution, the effects of ecological factors on virus evolution, and the analyses of plant-virus coevolution.

3.1 Time Scale of Plant Virus Evolution

As mentioned above, many if not most plant virus populations have high mutation rates, which allow direct estimation of evolutionary rates. The evolutionary rates of plant virus populations can be estimated most directly by comparing samples collected at different times (Table 2). For instance, the combined analysis of mutations accumulated by *Wheat streak mosaic virus* (WSMV) during serial passaging in wheat, and of differences within its population in North America, gave an estimate that the virus population diverged at about 1.1×10^{-4} nucleotide substitutions per site per year (Stenger et al. 2002). Alternatively, and taking advantage

Table 2 Estimates of nucleotide substitution rates in plant viruses

Method[a]	Virus	Substitution rates[b]	Reference
Serial sampling	BBTV	3.9×10^{-4}	Almeida et al. (2009)
	MSV-1	$7.4–7.9 \times 10^{-4}$	van der Walt et al. (2008)
	MSV-2	$2.0–3.0 \times 10^{-4}$	Harkins et al. (2009)
Heterochronous sampling	BBTV	1.4×10^{-4}	Almeida et al. (2009)
	BYDV	$3.2–6.3 \times 10^{-4}$	Malmstrom et al. (2007), Wu et al. (2011)
	EACMV	$1.6–1.3 \times 10^{-4}$	Duffy and Holmes (2009)
	Geminivirus	$1.8 \times 10^{-3}–3.9 \times 10^{-4}$	Lefeuvre et al. (2011)
	Luteoviridae	$3.5 \times 10^{-2}–1.4 \times 10^{-4}$	Pagán and Holmes (2010)
	PepGMV	$2.4–3.7 \times 10^{-3}$	Rodelo-Urrego et al. (2013)
	PepMV	5.6×10^{-3}	Gómez et al. (2012)
	PHYVV	$1.7–3.9 \times 10^{-3}$	Rodelo-Urrego et al. (2013)
	PVY	$5.97–9.99 \times 10^{-5}$	Gibbs et al. (2017)
	RYMV	4.8×10^{-3}	Fargette et al. (2008a)
	Tobamovirus	$1.3 \times 10^{-3}–1 \times 10^{-5}$	Pagán et al. (2010a)
	TYLCV	2.9×10^{-4}	Duffy and Holmes (2008)
	ZYMV	5.0×10^{-4}	Simmons et al. (2008)
Node dating	Geminivirus	3.1×10^{-8}	Lefeuvre et al. (2011)
	Potyviruses	1.2×10^{-4}	Gibbs et al. (2008b)
	WSMV	1.1×10^{-4}	Stenger et al. (2002)
Co-divergence	Cereal mastrevirus	1.0×10^{-8}	Wu et al. (2008)
	TYMV	1.3×10^{-7}	Blok et al. (1987), Gibbs et al. (1986), Guy and Gibbs (1981)
	Tobamovirus	2.2×10^{-8}	Gibbs (1980), Gibbs et al. (2008a)
	Begomovirus	$<0.6 \times 10^{-6}$	Gibbs et al. (2010)

[a]Method for phylogenetic tree dating
[b]Nucleotide substitutions per site per year

of the combination of Bayesian analyses and coalescence theory, evolutionary rates are estimated most often from phylogenetic analyses of gene sequences obtained from natural populations. A key aspect of this phylogenetic methodology is dating the tree. For this, different approaches have been used. Serial sampling was utilized to estimate the rates of evolution of *Banana bunchy top virus* (BBTV) and *Maize streak virus* (MSV) populations, in which trees were dated from epidemiological data (Almeida et al. 2009; Harkins et al. 2009). Also, node dating through known events or co-divergence studies has been used to estimate substitution rates of potyviruses and *Turnip yellow mosaic virus* (TYMV) (Gibbs et al. 1986, 2008b). But perhaps the most used method is heterochronous sampling, i.e., tree calibration using the collection dates of sequences obtained during a sufficiently long period of time for evolutionary changes to have occurred (Drummond et al. 2003). Using this approach, estimates of evolutionary rates for RNA plant viruses such as species of the family *Luteoviridae*, *Rice yellow mottle virus* (RYMV), or ZYMV, among

others, have been obtained (Pagán and Holmes 2010; Fargette et al. 2008a; Simmons et al. 2008; Rodelo-Urrego et al. 2013). These studies reported substitution rates between 10^{-3} and 10^{-6} nucleotide substitutions per site per year, which are values in the same range than those obtained for RNA animal viruses (Jenkins et al. 2002). Similar rates have been estimated for plant viruses with small DNA genomes such as *East African cassava mosaic virus* (EACMV) and *Tomato yellow leaf curl virus* (TYLCV), both belonging to the genus *Begomovirus* (Duffy and Holmes 2008, 2009; Lefeuvre et al. 2011), and CaMV (Yasaka et al. 2014). However, there are notable exceptions to this general pattern. Estimates for tobamoviruses, *Begomovirus*-related sequences integrated in the host genome, and cereal mastreviruses revealed evolutionary rates of 10^{-8}–10^{-9} (Gibbs et al. 2008a; Wu et al. 2008; Lefeuvre et al. 2011), much closer to those of their hosts (but see Pagán et al. 2010a).

The introduction of time as a variable in phylogenetic analyses has allowed us to estimate evolutionary rates, as well as explore when the current virus diversity originated. Analyses of evolutionary scales of several virus species/families traced back the radiation of these viruses hundreds, if not thousands, of years ago. Pagán and Holmes (2010) dated the origin of the family *Luteoviridae* up to 10,000 years ago and the two major genera of the family (*Luteovirus* and *Polerovirus*) to about 1000–2000 years ago, with most of the species radiation in these two genera occurring in the last 500 years. Similarly, Gibbs et al. (2017) dated the origin of *Potato virus Y* (PVY) up to 7000 years ago, Gibbs et al. (2008b) traced the origin of the genus *Potyvirus* about 6600 years ago, and Fargette et al. (2008b) dated the divergence of the genus *Sobemovirus* up to 5000 years ago. All these estimates of origin or radiation times for virus taxa that include important crop pathogens find time points near to the origin or the expansion of agriculture, which has led to formulate the hypothesis that agriculture has provided the ecological driver for the radiation of pathogenic plant viruses. Despite this general agreement on plant virus evolutionary time scales, some reports have revealed that certain groupings of RNA plant viruses could be much more ancient. This is the case of the species in the genus *Tobamovirus*, the origin of which has been traced to around 100,000 years ago (Gibbs et al. 2008a), and the proposal that known tobamoviruses first diverged at the same time as their solanaceous hosts, namely, 1000 times earlier (Gibbs et al. 2015). Similarly, diversification of *Begomovirus* species through analyses of virus-related sequences integrated in the host genome has been traced back up to 80 million years ago (Lefeuvre et al. 2011). Besides these estimates of long-term evolutionary rates, coalescent Bayesian phylogenies have been also used to analyze the short-term evolutionary time scales. For instance, Fraile et al. (2011) traced the origin of tobamovirus epidemics in pepper crops of Spain to about 100–120 years before present and showed that in epidemic outbreaks tobamoviruses had much faster evolutionary rates (1×10^{-4} subs/site/year) than during their long-term evolution. Duffy and Holmes (2009) estimated similar evolutionary rates for EACMV epidemics in Central Africa and dated the origin of the epidemic 34–175 years ago based on sequences of the coat protein (CP) gene. Using also the CP gene, Rodelo-Urrego et al. (2013) dated the origin of the begomovirus epidemics in wild pepper or

chiltepin (*Capsicum annuum* var. *glabriusculum*) plants in Mexico to about 30–50 years ago. Importantly, these estimates were in agreement with existing epidemiological evidence, indicating that phylogenetics can also provide relevant information both on disease evolution and epidemiology. In this regard, further advances in the methodology for tree inference have allowed adding the spatial scale to phylogenetic reconstructions and contributed to understand the factors determining the epidemiology of plant viruses. Using spatial diffusion models, De Bruyn et al. (2012) showed that dispersion of EACMV from Africa to the Indian Ocean islands mostly resulted from human activity. Similarly, Rodelo-Urrego et al. (2013) reconstructed the migration patterns of begomoviruses infecting wild pepper populations across Mexico and observed that short-distance movements were mostly driven by viral vectors, whereas long-distance dispersal of the virus was likely to be due to human intervention (i.e., trade of infected plant material). Using a similar approach, Trovão et al. (2015) showed that the major factor affecting RYMV dispersal in Africa was the distance between rice fields, with faster dispersal rates in West Africa where agriculture is more intensive and extensive.

Together, these works have shown that the ecological context in which plant-virus interactions take place influences the epidemiology and evolution of plant virus populations.

3.2 *Effect of Ecosystem Biodiversity on Plant Virus Populations*

Changes in host ecology are among the most frequently identified causes of disease emergence (i.e., the increase of disease incidence following its appearance in a new, or previously existing, host population) (Morse and Schluederberg 1990; Jones 2009; Pagán et al. 2016). It is thought that changes in plant populations or communities caused by humans are a major driver of these ecological changes. A classical hypothesis in plant pathology states that ecological changes associated with agriculture favor epidemic infection dynamics of highly virulent pathogens and lead to disease emergence. Three main factors are considered to be involved in this process as a consequence of ecosystem simplification: the reduced species diversity, the reduced genetic diversity within species—both components of biodiversity—and the greater host density of agricultural ecosystems as compared with wild ones (Burdon and Chilvers 1982; Thresh 1982; Stukenbrock and McDonald 2008). Current decreases in biodiversity and the increasing number of emergent pathogens have resulted in a new interest on the relationship between ecosystem simplification and disease risk (Keesing et al. 2010). As a result, two hypotheses representing extreme situations have been proposed that relate ecosystem biodiversity to disease risk: diversity may be either positively ("amplification effect" hypothesis) or negatively ("dilution effect" hypothesis) correlated with disease risk, as greater biodiversity may result in either increased abundance of reservoirs for a focal host or in a

decreased abundance of the focal host hindering pathogen transmission (Keesing et al. 2006; Ostfeld and Keesing 2012). Hence, the effects of diversity on disease risk would be related to the host range of the pathogen: an amplification effect would require a generalist pathogen, while the more restricted the host range of the pathogen, or the higher the differences between shared hosts in their ability to amplify or transmit the pathogen, the higher the dilution effect. The general application of these concepts has been questioned (Randolph and Dobson 2012). For instance, a recent analysis of a multi-host-multivirus system at different spatial scales showed that the relationship between biodiversity and disease risk was both scale dependent and habitat dependent (McLeish et al. 2017).

The relationship between biodiversity and disease risk has been analyzed only in a few plant-virus interactions. Early analyses of *Cereal* and *Barley yellow dwarf viruses* (B/CYDV) in wild grassland ecosystems in the west of the USA mostly agree with the amplification effect hypothesis (Power and Mitchell 2004; Malmstrom et al. 2005a, b; Borer et al. 2009, 2010; Hall et al. 2010; Power et al. 2011). However, the effect of biodiversity changes on the genetic composition of B/CYDV populations was not addressed. More recently, a study has analyzed begomovirus incidence in Mexican populations of chiltepin growing in habitats along a gradient of human management, from wild to cultivated populations (Pagán et al. 2012). Increased human management was associated with an increase of virus infection risk, and the main predictor of disease risk was the biodiversity of the habitat, agreeing with the dilution effect hypothesis. Interestingly, cultivation resulted in the loss of the spatial genetic structure of the virus populations (Rodelo-Urrego et al. 2013). Moreover, the smaller biodiversity and greater host density characteristic of cultivated chiltepin populations favored the genetic diversification of the begomovirus populations (Rodelo-Urrego et al. 2015). Another observation of these authors was that the ecological factors associated with higher virus genetic diversity decreased the frequency of recombinant genotypes, suggesting that in cultivated chiltepin populations genetic diversity was mostly generated by mutation, whereas recombination had higher relative importance in wild populations (Rodelo-Urrego et al. 2015). Hence, biodiversity and plant density could also affect the relative importance of mutation and recombination in virus evolution. These two studies involved viruses that infect more than a single host, but as stated above the relationship between biodiversity and disease risk could depend on the virus host range as well as on community composition and spatial scales. Rodríguez-Nevado et al. (2017) utilized data on infection risk and population genetic diversity of the specialist virus Mediterranean ruda virus (MeRV) in its host *Ruta montana* L., collected in the native wild ecosystem where this host-virus interaction occurs, to test their association with biodiversity and host density. These authors showed that plant density, but not ecosystem biodiversity, was the major determinant of infection risk: the larger the host density, the greater the MeRV incidence. Both infection risk and host density were positively associated with MeRV population genetic diversity. Although these results would be compatible with the dilution effect hypothesis, they also show that the effects and relative importance of biodiversity and host density could depend on the identity of the host and virus involved. This would be in

accordance with further elaborations about the amplification/dilution effect hypotheses (Randolph and Dobson 2012; Johnson et al. 2015; Ostfeld and Keesing 2017).

It is worth noting that in wild populations of both chiltepin and *R. montana*, the infecting viruses increased host mortality, which may determine the demography of the host plant (Fraile et al. 2017; Rodríguez-Nevado et al. 2017). This being so, hosts might develop mechanisms to avoid/limit virus infection or to reduce its detrimental effects (Agnew et al. 2000; Little et al. 2010). Plant defense against viruses can manifest as resistance, i.e., the host ability to limit virus multiplication (Clarke 1986), and tolerance, i.e., the host ability to reduce the effect of infection on its fitness (Little et al. 2010). Because pathogen virulence and host resistance/tolerance have reciprocal effects on each other's fitness, it is generally assumed that host-virus interactions result in coevolutionary processes (Woolhouse et al. 2002).

3.3 Coevolution Between Plants and Viruses

Host-pathogen coevolution requires four conditions: (1) genetic variation in the relevant host and pathogen traits (e.g., resistance, tolerance, infectivity, virulence), (2) reciprocal effects of the relevant traits of the interaction on the fitness of host and pathogen, (3) dependence of the outcome of the host-pathogen interaction on the specific host and pathogen genotypes involved, and (4) changes in genotype frequencies in both the host and the pathogen populations (Woolhouse et al. 2002). Current evidence of plant-virus interactions that meet these conditions mostly derives from highly virulent viruses in crops and from the changes in the genetic structure of virus populations in agricultural ecosystems as a response to human manipulation of the genetics of the host by breeding resistance factors (Fraile and García-Arenal 2010). Demonstration of plant-virus coevolution in wild plant populations, where the genetic composition of the host may change in response to virus infection, is lacking, and evidence in support of plant-virus coevolution is very scant (but see Pagán et al. 2010b).

The first condition for plant-virus coevolution to occur is the genetic variation in host resistance/tolerance and in the pathogen's ability to infect and multiply in the host (i.e., infectivity) and to cause disease (i.e., pathogenicity). The available evidence that this is indeed the case for plant viruses derives mostly from analysis of the patterns of variability of plant resistance genes and of the viral pathogenicity genes in interactions resulting in qualitative resistance. This is the case for the different alleles of the *L* gene in *Capsicum*, which confer resistance to different genotypes/species of tobamoviruses and differ in a few non-synonymous mutations (Tomita et al. 2008). Similarly, field isolates of the tobamovirus species *Pepper mild mottle virus* (PMMoV) that overcome *L*-mediated resistance in pepper differ from the avirulent genotype by a few amino acid substitutions in the coat protein (Berzal-Herranz et al. 1995; Tsuda et al. 1998; Hamada et al. 2002, 2007). Another good example is provided by the rice-RYMV interaction. In this system, different mutations have been observed in the RYMV VPg determining infection on rice genotypes

carrying different resistance alleles at the *rymv* locus, encoding eIF(iso)4G (Pinel-Galzi et al. 2007; Poulicard et al. 2014). The interaction pepper-*Potato virus Y* (PVY), determined by the *pvr2* locus of *Capsicum* spp., encoding eIF4E, and the virus VPg has been also well characterized (Quenoiulle et al. 2013; Moury et al. 2014). High variability and positive selection has been observed at both the VPg and eIF4E encoding genes, and almost all amino acid changes are linked to gains of function (resistance of the plant or infectivity of the virus), which determines the genetic composition of both plant and virus populations. Interestingly, no equivalent analyses have been done in plant-virus interactions in wild ecosystems.

The second condition for plant-virus coevolution is that there must be reciprocal effects of the relevant traits of the interaction on the fitness of host and pathogen. For this to happen, infectivity and resistance must negatively affect the fitness of plants and viruses, respectively. Evidence that virus infection can have detrimental effects in the host mainly comes from crop-virus systems (Fraile and García-Arenal 2010; Froissart et al. 2010). A few experiments have shown such effects in wild plants under controlled conditions (e.g., Kelly 1994; Friess and Maillet 1996; Pagán et al. 2007), but evidence that plant viruses have a negative effect on plant fitness in natural ecosystems is still limited (but see Maskell et al. 1999; Power and Mitchell 2004; Prendeville et al. 2014; Fraile et al. 2017). Also, evidence of negative effects of plant resistance on virus fitness derives from crop-virus interactions (Fraile and García-Arenal 2010). Most of these studies quantify virus fitness as within-host multiplication and assume that virus multiplication and transmission are positively correlated. Although this seems to be the case for horizontally transmitted plant viruses (Sacristán and García-Arenal 2008; Froissart et al. 2010), it has been shown that this positive correlation does not hold for some vertically transmitted plant viruses (Pagán et al. 2014). A potential consequence of the reciprocal effects of plants and viruses on each other's fitness and evolution is the congruence of their phylogenies (Nieberding and Olivieri 2007). In various plant-virus systems, such congruence has been interpreted as evidence of coevolution, i.e., reciprocal evolutionary change in the host and the virus driven by natural selection (Thompson 2005), and as mentioned earlier even of co-divergence, co-divergence meaning that the host plant and the virus have shared their whole evolutionary history (Lartey et al. 1996; Wu et al. 2008; Gibbs et al. 2015; Stobbe et al. 2012). However, other phenomena such as population geographic isolation may result in host-virus phylogenetic congruence. Also, different evolutionary scales have been estimated for some viruses and their hosts (see Sect. 1.1), which do not support plant-virus co-divergence over extended periods of time (Pagán et al. 2010a; Rodelo-Urrego et al. 2013).

The third condition for plant-virus coevolution is dependence of the outcome of the interaction on the combination of host and virus genotypes involved. As discussed above, genetic diversity in genes determining resistance/susceptibility in plants and infectivity/pathogenicity in viruses has been described in a variety of plant-virus interactions (García-Arenal and Fraile 2013). In such interactions, the specific combination of plant and virus alleles determines whether the plant-virus interaction would be a compatible one, i.e., whether the virus is able to establish a

successful infection (Tomita et al. 2008; Quenoiulle et al. 2013; Ishibashi et al. 2014; Moury et al. 2014). Thus, there is evidence that the outcome of the plant-virus interaction depends on the specific genotypes of the interacting partners. Resistance to viruses in plants may also be quantitative, in which within-host multiplication of the virus is reduced. Examples of quantitative resistance depending on the plant-virus genotype × genotype interaction have also been reported. This is the case of the interaction between CMV, *Turnip crinkle virus* (TCV), CaMV and TuMV, and its natural host *A. thaliana* (Pagán et al. 2007, 2010b; Shuckla et al. 2018). The other major defense strategy of plants against pathogens' tolerance is determined by genotype × genotype interactions. Tolerance of *A. thaliana* to CMV, which depends on modification of life-history traits, varies according to the plant and virus genotype (Pagán et al. 2008, 2009). In summary, there is evidence that the outcome of plant-virus interactions depends on genotype × genotype interactions regardless of the type of plant defense or interaction model.

Finally, the fourth condition for plant-virus coevolution is the existence of changes in genotype frequencies in both the host and the pathogen populations. Even in agricultural systems, long-term analyses of these dynamics are rather rare, but they have demonstrated that the frequency of virus genotypes changes in response to the deployment of genetic resistance of crop varieties (Fraile et al. 2011).

Altogether, the studies discussed above highlight the contribution of genomic studies to understand the evolution of plant-virus interactions. Despite evidence for plant-virus coevolution is limited, these studies also suggest that this process influences the genetic composition of both interactive partners.

4 Genomics of Plant and Virus Communities

Most of the work discussed above focuses on single plant-virus combinations. However, in nature the interaction between a virus population and its host(s) is embedded in the context of the plant and virus communities within multi-host-multivirus systems (Malmstrom and Alexander 2016). As a consequence, the outcome of virus infection and the evolution of virus populations may be influenced by the presence of other viruses that share common hosts or vectors with the focal one and by the presence of plant species other than the focal host, both alternative hosts and nonhost species. These virus-virus, plant-virus, and plant-plant interactions are central to understand the genomics of plant virus ecology and evolution. Although scientists are just starting to explore interactions at the community level, virus population genomics has been pivotal in the pioneer works that have addressed these questions (Roossinck 2017). Because the previous section extensively dealt with plant-virus interactions, this section focuses in virus-virus and plant-plant interactions.

4.1 Virus-Virus Interactions

Although a large fraction of known plant viruses are multi-host pathogens (García-Arenal and Fraile 2013; Moury et al. 2017), the study of the distribution of multi-host plant viruses in an ecosystem has not been undertaken until recently. Malpica et al. (2006) first addressed this question by analyzing the prevalence and distribution of five vector-transmitted multi-host plant viruses in wild plant communities of Central Spain. This work raised two major conclusions: (1) the distribution of plant viruses was not random; rather, viruses preferred some host species to others; and (2) coinfection of a single plant by more than one virus was frequent. Since this seminal paper, accumulating evidence indicates that infection of the same plant by multiple viruses is common in nature (Moreno et al. 2004; Roossinck et al. 2010; Tugume et al. 2016). This being so, perhaps the most obvious questions to address when exploring the dynamics of plant virus communities are (1) how many viruses are in a plant assemblage (Roossinck 2011) and (2) how these viruses interact with each other (Syller 2012).

The advent of deep sequencing techniques has allowed the virome of a number of different ecosystems (e.g., Delwart 2007; Hurwitz and Sullivan 2013), including plant communities, to be explored. The majority of the analyses of plant community viromes have focused on wild ecosystems (reviewed by Roossinck 2012; Stobbe and Roossinck 2014), uncovering large (previously) unknown virus diversity (Roossinck 2011), in particular when it comes to double-stranded RNA viruses (Massart et al. 2014). These analyses have revealed that virus infections are common in wild plants, and virus diversity and identity in wild plants appear to be different from that in crops (Roossinck 2012; Wylie et al. 2013; Stobbe and Roossinck 2014). The few metagenomic analyses of crop plants have confirmed this observation, although these works tend to focus in already known plant viruses, which bias the comparison between wild ecosystems and agroecosystems (Coetzee et al. 2010; Giampetruzzi et al. 2012; Ng et al. 2011). Despite that metagenomics analyses have significantly broadened our understanding on the composition of plant virus communities, two main questions remain opened. The first question refers to how much of the true viral diversity is represented in the metagenomics analyses. The methodology used is a significant factor limiting how much of the viral diversity present in a plant community is captured. Since there are no universal genes for viruses, deep sequencing methods often use random priming for reverse transcription (RT) or PCR to obtain virus sequences. Therefore, the procedure to obtain template material for sequencing determines the information that can be extracted from metagenomics analyses (Roossinck et al. 2015; Blawid et al. 2017; Massart et al. 2017). One approach (Fig. 1) is to enrich the sample in viruslike particles and use the nucleic acids in these particles for deep sequencing. Although this method has yielded validated results, it may exclude genomes of viruses that do not form particles or those with labile particles. Some works have used methods that enrich nucleic acid extracts in double-stranded RNAs (dsRNA), as many plant viruses have RNA genomes and generate dsRNA intermediates during replication (Roossinck et al. 2010). However, this

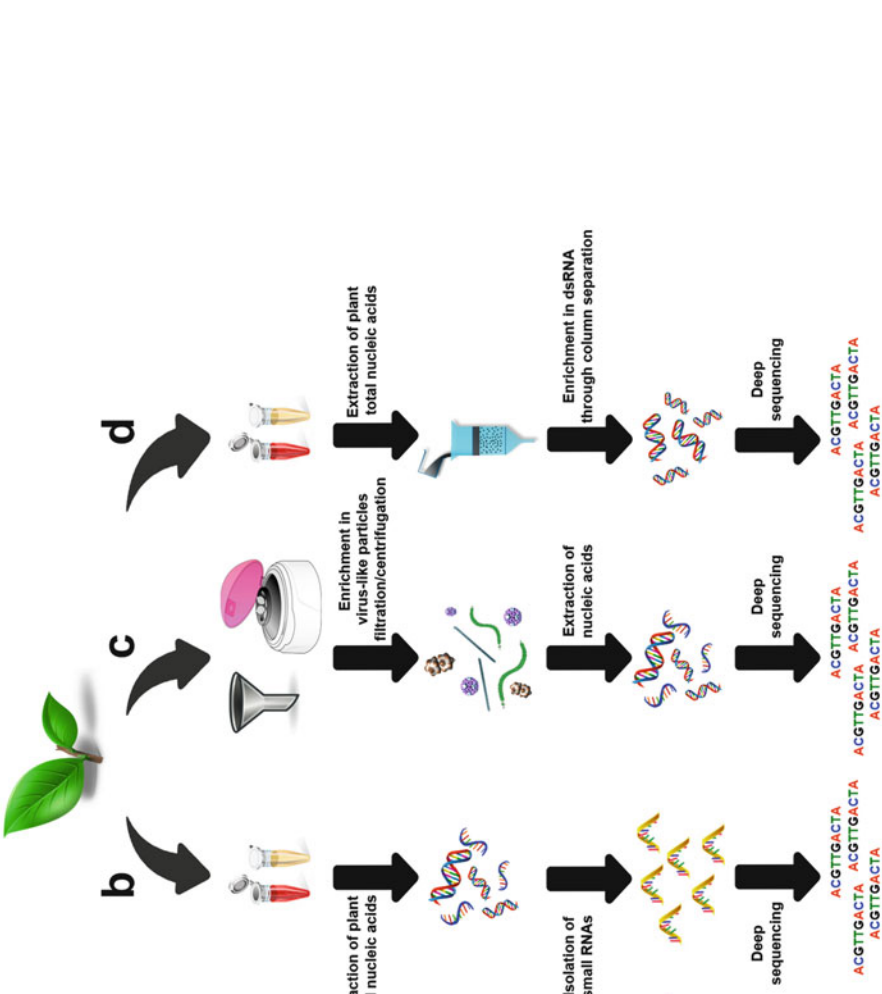

Fig. 1 Methods for deep sequencing-based characterization of plant virus genetic diversity. (**a**) Direct deep sequencing from total plant nucleic acids. (**b**) Deep sequencing of small interfering RNAs. (**c**) Deep sequencing of nucleic acids from preparations enriched in viruslike particles. (**d**) Deep sequencing of total plant nucleic acids enriched in dsRNAs

method could preferentially target persistent viruses, which often have dsRNA genomes (Roossinck 2012). Also, projects of plant virus discovery have taken advantage of the plant defense mechanisms based on gene silencing, which involves the generation of small interfering RNAs (siRNAs) derived from viral genomes that target viral sequences for degradation (Ghoshal and Sanfaçon 2015; Wu et al. 2015). These siRNAs have been used for discovery of RNA and DNA viruses in plants (Donaire et al. 2009; Kreuze et al. 2009). However, genome assembly through siRNAs is based on short sequences (21–24 nt), which makes repetitive regions difficult to resolve (Blawid et al. 2017; Massart et al. 2017). The second aspect brought by the increasing metagenomics information on plant viromes is how much of the described virus diversity is biologically meaningful. Although much progress has been done in bioinformatics methods for sequence assembly, particularly in the case of deep sequencing data (Hadidi et al. 2016; Blawid et al. 2017), plant virus diversity estimates are often based only on partial sequences. The available partial sequences often do not allow differentiating if they derive from infectious genomes or are just "environmental noise." Recently Koch's postulates have been revised to accommodate new pathogen discovery techniques, based on criteria for establishing causation organized into three levels (Lipkin 2013): Level 1, evidence of the disease agent by sequence analysis; Level 2, finding the agent in host cells, similarity of the agent to known pathogens, and finding the agent in numerous individuals with similar symptoms; and Level 3, prevention of the disease by an agent-specific drug, antibody, or vaccine. Not all of these criteria can be applied to the discovery of new plant viruses. Level 3 is probably not possible, although siRNAs could be used as corroborative evidence that the plant has mounted a defense response (Chen et al. 2015). On the other hand, levels 1 and 2 could be easily adapted to the discovery of new plant viruses even if they are asymptomatic in their host(s), as in many wild plants (Prendeville et al. 2012; Roossinck 2012). However, although metagenomic analyses of plant virus communities comply with Level 1, very few of the new viruses discovered have been biologically characterized and tested for infectivity (Massart et al. 2017; Rodríguez-Nevado et al. 2017).

Analyses of plant viromes therefore indicate that plant virus communities are populated by a large number of species, and it would be reasonable to expect that these species interact with each other. In this sense, another interesting observation is that multiple infections are frequent in wild and cultivated plants (Malpica et al. 2006; Roossinck et al. 2010; Tugume et al. 2016). The importance of mixed infections in the epidemiology and genetic diversity of plant virus populations was realized almost a century ago (Fawcett 1931). Early works indicated that mixed infections could lead to the appearance of new virus strains and to modify the fitness of some virus genotypes (Thompson 1961; Rochow 1972). Although initially considered as relatively infrequent, the increasing evidence that mixed infections are widespread has prompted the analyses of the effects of such virus-virus interactions for the population dynamics of plant viruses. These studies have indicated that mixed infections may have far-reaching consequences for virus populations. For instance, mixed infections may determine the severity of infection symptoms (Rodelo-Urrego et al. 2013; Tugume et al. 2016), alter plant viral load (Taiwo

et al. 2007), allow the survival of less fit genotypes and affect virus host range (Gómez et al. 2009), promote recombination events between coexisting viruses (García-Andrés et al. 2007), and alter virus transmission (de Assis Filho and Sherwood 2000; Salvaudon et al. 2013). Mixed infections may also modify the fitness of the interacting viruses, and both synergistic (Taiwo et al. 2007; Rentería-Canett et al. 2011; Nsa and Kareem 2015) and antagonistic (Martín and Elena 2009; Syller 2012) effects have been described. These effects may have deep impact on the prevalence of plant virus species and genotypes in the within-host population, therefore determining the genetic diversity/haplotype composition of plant virus populations (Gómez et al. 2009; Rodelo-Urrego et al. 2015). Altogether, these works underscore the relevance of considering virus-virus, and not only host-virus, interactions in determining the genetic structure of the population.

4.2 Plant-Plant Interactions

Plant virus communities exist in the context of the plant communities they infect. Therefore, plant virus communities should not be independent of the structure of the plant community where infections occur. Competition is one major factor determining community structure. In host-parasite systems, there are different competitive interactions: intra-class competition among uninfected hosts or among infected hosts and interclass competition between uninfected and infected hosts. Each interaction may have different effects on host life-history traits, resulting in a direct cost of infection, due to parasitism itself, and in an indirect cost, due to modification of the competitive ability of the infected host, both being modulated by host population density (Bedhomme et al. 2005). Theory predicts that fitness reduction will be higher under the combined effects of host population density and parasitism than under each factor separately (Hochberg 1991; Lively 2006). Although these indirect costs of virus infection may obviously occur at the plant population level (see Friess and Maillet 1996; Pagán et al. 2009 for examples), such costs may also apply for plant and virus communities. An obvious effect of virus infection on plant competition derives from the different susceptibility that the same host species may have to different viruses. It has been shown that B/CYDV infection has similar direct costs in invasive exotic grasses and native bunchgrasses in California. However, indirect costs of B/CYDV infection were much higher in native than in exotic grasses, suggesting that B/CYDV infection can have high impact on host population dynamics through indirect costs of infection (Malmstrom et al. 2005a). Interestingly, infected exotic grasses did not only compete better with native bunchgrasses, they were also more attractive for virus vectors, such that they amplified infection and increased virus incidence in the native host (Malmstrom et al. 2005b). Such flow of virus genotypes between hosts has been shown to play a role in both selectively favoring novel mutations as well as contributing variation to the pathogen population on the receptor host (Burdon and Thrall 2008). Moreover, changes in virus incidence may affect population size, affecting evolutionary rates (Scholle et al.

2013; Lanfear et al. 2014). Hence, plant-plant competition may have an effect on the genetics of virus population. Indirect costs of virus infection may also involve third-party parasites. For instance, indirect costs of ZYMV infection in populations consisting on mixtures of resistant and nonresistant squash genotypes were derived from the fact that susceptible plants, upon infection, were less competitive than resistant ones. However, this indirect cost was softened because ZYMV-infected plants became less attractive to vectors of pathogenic bacteria, so that plants resistant to viruses had greater burden of bacterial diseases than ZYMV susceptible plants (Sasu et al. 2009). Thus, the effect of virus infection on plant-plant interactions is not necessarily negative. In this sense, it has been shown that plant virus infection may reduce plant attractiveness to herbivores (Gibbs 1980; van Molken et al. 2012), which represents a competitive advantage against uninfected individuals of the same host species as well as against other coexisting species. Other viral infections have been shown to confer resistance to drought (Xu et al. 2008), which could be also interpreted as a benefit for infected individuals in plant communities. To our knowledge, information on how the properties conferred to the host affect the genetic structure of the virus are not available.

In summary, these observations indicate that plant-plant interactions may have a great impact on host population dynamics. Evidence of equivalent effects on the evolution of virus populations are only indirect, and whether this translates into changes of the genetic composition of the virus population is yet to be addressed.

5 Future Perspective

Plant virologists have benefited from remarkable advances in the number, sophistication, and quality of the methods for isolation of virus nucleic acids and for determining their nucleotide sequences and of bioinformatics tools for the genetic characterization of virus populations. The information obtained with these approaches, reviewed in this chapter, has significantly increased knowledge on the mechanisms and determinants of plant virus evolution at every level of population organization. Still, various aspects of the molecular and deterministic basis of changes in plant virus population structure, and the consequences for disease dynamics, remain poorly understood. We think that future studies would pay special attention to aspects that include, but are not restricted to:

– Studies on the genetic diversity of plant virus populations within a plant. Present studies have shown that it is structured according to the tissue and organ. However, current methodologies would allow questions to be studied such as: is the genetic diversity of virus populations structured at the cell level within a given tissue? Single-cell sequencing has been suggested as a possible way to address this question in other host-virus interactions (Rato et al. 2016), and it could be equally applicable to plant viruses. These analyses may be particularly interesting in the case of multipartite plant viruses. It has been recently shown

that, during within-host infection, some genomic components accumulate at higher frequency than others (Sicard et al. 2013), but is this a general phenomenon among plant viruses? If so, infection of susceptible cells by at least one copy of each segment would require the entry of a number of viral particles well above the estimated MOI (see Sect. 2.3). Thus, is it necessary that every genomic segment is present in every infected cell? (Sicard et al. 2016). If the answer is no, then the minimum unit at which virus genetic diversity within a plant is generated could be different for mono- and multipartite viruses.

– Analyses of the factors that shape the genetic diversity of plant virus populations. These are currently derived mostly from crop-virus interactions; analyses in hosts growing in wild ecosystems are scant. Analyses in wild ecosystems should consider exploring: (1) under which conditions do plants and viruses coevolve, and what are the consequences of plant-virus coevolution for the genetic composition of both plant and virus populations; (2) if epidemiological changes associated with biodiversity loss that determine the relationship between diversity and disease risk affect the genetic diversity of virus populations and how; (3) if estimates of the evolutionary rates for crop viruses can be extrapolated to wild ecosystems, as factors associated with virus evolutionary rates such as biodiversity, host population size, or virus prevalence (Burdon and Thrall 2008; Scholle et al. 2013) may broadly differ between wild ecosystems and agroecosystems (Malmstrom and Alexander 2016); and (4) if, as suggested by some studies (Rodelo-Urrego et al. 2015; Lima et al. 2017), the relative importance of mutation and recombination in generating genetic diversity of virus populations differs between viruses infecting wild and cultivated hosts. Answering these questions would contribute to understand the genetic plasticity of virus populations in the wide variety of environments in which they are present.
– Finally, technological improvements will allow better investigating the true diversity of plant virus communities. In this sense, recently developed procedures for fast and accurate sequencing of whole genomes (Wanunu 2012; Wang et al. 2017) would contribute to reduce, at least in part, the "environmental noise" associated with the virome assembly based on partial sequences. This would also help to better characterize the number of virus-virus interactions and identify the most relevant ones, in both wild and anthropic habitats. Obviously, the full description of plant viromes would need to combine the application of these technological advances with the biological characterization of the virus species identified in these ecosystems. Also, understanding plant-virus interactions at the community level, where viruses will be able to infect a variety of hosts, and hosts will be exposed/infected by a variety of viruses, would require utilizing theoretical models and bioinformatic tools designed to analyze mutualistic infection networks. Network models open the possibility of understanding how plant-virus interactions evolve at scales other than the single individual/population (Nuismer et al. 2013) and may allow exploration of the factors that shape virus and plant communities (Malmstrom and Alexander 2016).

We believe that the improvement of sequencing and in silico analysis of genomic data will continue in the near future, and this will further allow these questions to be addressed. This will be a most stimulating challenge for the next few years and will much contribute to understand the evolution and genomics of plant virus populations.

Acknowledgment IP was supported by grant (BIO2016-79165-R) and FGA was supported by grant (BFU2015-60418-R), both funded by Plan Nacional I + D + I, MINECO, Spain.

References

Acosta-Leal R, Duffy S, Xiong Z, Hammond RW, Elena SF. Advances in plant virus evolution: translating evolutionary insights into better disease management. Phytopathology. 2011;101:1136–48.

Agnew P, Koella JC, Michalakis Y. Host life-history responses to parasitism. Microbes Infect. 2000;2:891–6.

Alí A, Roossinck MJ. Genetic bottlenecks during systemic movement of *Cucumber mosaic virus* vary in different host plants. Virology. 2010;404:279–83.

Almeida RPP, Bennett GM, Anhalt MD, Tsai C-W, O'Grady P. Spread of an introduced vector-borne banana virus in Hawaii. Mol Ecol. 2009;18:136–46.

de Assis Filho F, Sherwood J. Evaluation of seed transmission of turnip yellow mosaic virus and tobacco mosaic virus in *Arabidopsis thaliana*. Phytopathology. 2000;90:1233–8.

Bedhomme S, Agnew P, Vital Y, Sidobre C, Michalakis Y. Prevalence-dependent costs of parasite virulence. PLoS Biol. 2005;2:e262.

Bergua M, Zwart MP, El-Mohtar C, Shilts T, Elena SF, Folimonova SY. A viral protein mediates superinfection exclusion at the whole-organism level but is not required for exclusion at the cellular level. J Virol. 2014;88:11327–38.

Berzal-Herranz A, de la Cruz A, Tenllado F, Díaz-Ruíz JR, López L, Sanz AI, Vaquero C, Serra MT, García-Luque I. The *Capsicum* L^3 gene-mediated resistance against the tobamoviruses is elicited by the coat protein. Virology. 1995;209:498–505.

Blawid R, Silva JMF, Nagata T. Discovering and sequencing new plant viral genomes by next-generation sequencing: description of a practical pipeline. Ann Appl Biol. 2017;170:301–14.

Blok J, Mackenzie A, Guy P, Gibbs AJ. Nucleotide sequence comparisons of turnip yellow mosaic isolates from Australia and Europe. Arch Virol. 1987;97:283–95.

Borer ET, Adams VT, Engler GA, Adams AL, Schumann CB, Seabloom EW. Aphid fecundity and grassland invasion: invader life history is the key. Ecol Appl. 2009;19:1187–96.

Borer ET, Seabloom EW, Mitchell CE, Power AG. Local context drives infection of grasses by vector-borne generalist viruses. Ecol Lett. 2010;13:810–8.

Bujarski J. Genetic recombination in plant-infecting messenger-sense RNA viruses: overview and research perspectives. Front Plant Sci. 2013;4:68.

Burdon JJ, Chilvers GA. Host density as a factor in plant-disease ecology. Annu Rev Phytopathol. 1982;20:143–66.

Burdon JJ, Thrall PH. Pathogen evolution across the agro-ecological interface: implications for disease management. Evol Appl. 2008;1:57–65.

Carrasco P, de la Iglesia F, Elena SF. Distribution of fitness and virulence effects caused by single-nucleotide substitutions in tobacco etch virus. J Virol. 2007;81:12979–84.

Chao L. Fitness of RNA virus decreased by Muller's ratchet. Nature. 1990;348:454–5.

Chao L. Levels of selection, evolution of sex in RNA viruses, and the origin of life. J Theor Biol. 1991;153:229–46.

Chen S, Huang Q, Wu L, Qian Y. Identification and characterization of a maize-associated mastrevirus in China by deep sequencing small RNA populations. Virol J. 2015;12:156.
Clarke DD. Tolerance of parasites and disease in plants and its significance in host-parasite interactions. Adv Plant Pathol. 1986;5:161–98.
Coetzee B, Freeborough M-J, Maree HJ, Celton J-M, Rees DJG, Burger JT. Deep sequencing analysis of viruses infecting grapevines: virome of a vineyard. Virology. 2010;400:157–63.
De Bruyn A, Villemot J, Lefeuvre P, Villar E, Hoareau M, Harimalala M, Abdoul-Karime AL, Abdou-Chakour C, Reynaud B, Harkins GW, Varsani A, Martin DP, Lett JM. East African cassava mosaic-like viruses from Africa to Indian Ocean islands: molecular diversity, evolutionary history and geographical dissemination of a bipartite begomovirus. BMC Evol Biol. 2012;12:228.
Delwart EL. Viral metagenomics. Rev Med Virol. 2007;17:115–31.
Dietrich C, Maiss E. Fluorescent labelling reveals spatial separation of potyvirus populations in mixed infected *Nicotiana benthamiana* plants. J Gen Virol. 2003;84:2871–6.
Divéki Z, Salánki K, Balázs E. Limited utility of blue fluorescent protein in monitoring plant virus movement. Biochimie. 2002;84:997–1002.
Donaire L, Wang Y, Gonzalez-Ibeas D, Mayer KF, Aranda MA, Llave C. Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. Virology. 2009;392:203–14.
Donaire L, Burgyán J, García-Arenal F. RNA silencing may play a role in but is not the only determinant of the multiplicity of infection. J Virol. 2016;90:553–61.
Drake JW, Holland JJ. Mutation rates among RNA viruses. Proc Natl Acad Sci U S A. 1999;96:13910–3.
Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. Trends Ecol Evol. 2003;18:481–8.
Duffy S, Holmes EC. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J Virol. 2008;82:957–65.
Duffy S, Holmes EC. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. J Gen Virol. 2009;90:1539–47.
Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9:267–76.
Dunham JP, Simmons HE, Holmes EC, Stephenson AG. Analysis of viral (zucchini yellow mosaic virus) genetic diversity during systemic movement through a *Cucurbita pepo* vine. Virus Res. 2014;191:172–9.
Elena SF, Sanjuán R. On the adaptive value of high mutation rates in RNA viruses: separating causes from consequences. J Virol. 2005;79:11555–8.
Elena SF, Fraile A, García-Arenal F. Evolution and emergence of plant viruses. Adv Virus Res. 2014;88:161–91.
Fabre F, Moury B, Johansen EI, Simon V, Jacquemond M, Senoussi R. Narrow bottlenecks affect *Pea seedborne mosaic virus* populations during vertical seed transmission but not during leaf colonization. PLoS Pathog. 2014;10:e1003833.
Fargette D, Pinel A, Rakotomalala M, Sangu E, Traoré O, Sérémé D, Sorho F, Issaka S, Hébrard E, Séré Y, Kanyeka Z, Konaté G. *Rice yellow mottle virus*, an RNA plant virus, evolves as rapidly as most RNA animal viruses. J Virol. 2008a;82:3584–9.
Fargette D, Pinel-Galzi A, Sérémé D, Lacombe S, Hébrard E, Traoré O, Konaté G. Diversification of *Rice yellow mottle virus* and related viruses spans the history of agriculture from the Neolithic to the present. PLoS Pathog. 2008b;4:e1000125.
Fawcett HS. The importance of investigations on the effects of known mixtures of microorganisms. Phytopathology. 1931;2:545–50.
Fraile A, García-Arenal F. The coevolution of plants and viruses: resistance and pathogenicity. Adv Virus Res. 2010;76:1–32.
Fraile A, Pagán I, Anastasio G, Sáez E, García-Arenal F. Rapid genetic diversification and high fitness penalties associated with pathogenicity evolution in a plant virus. Mol Biol Evol. 2011;28:1425–37.

Fraile A, Hily J-M, Pagán I, Pacios LF, García-Arenal F. Host resistance selects for traits unrelated to resistance-breaking that affect fitness in a plant virus. Mol Biol Evol. 2014;31:928–39.
Fraile A, McLeish MJ, Pagán I, González-Jara P, Piñero P, García-Arenal F. Environmental heterogeneity and the evolution of plant-virus interactions: viruses in wild pepper populations. Virus Res. 2017;241:68–76.
Frank SA. Multiplicity of infection and the evolution of hybrid incompatibility in segmented viruses. Heredity. 2001;87:522–9.
French R, Stenger DC. Evolution of wheat streak mosaic virus: dynamics of population growth within plants may explain limited variation. Annu Rev Phytopathol. 2003;41:199–214.
Friess N, Maillet J. Influence of cucumber mosaic virus infection on the intraspecific competitive ability and fitness of purslane (*Portulaca oleracea*). New Phytol. 1996;132:103–11.
Froissart R, Wilke CO, Montville R, Remold SK, Chao L, Turner PE. Co-infection weakens selection against epistatic mutations in RNA viruses. Genetics. 2004;168:9–19.
Froissart R, Roze D, Uzest M, Galibert L, Blanc S, Michalakis Y. Recombination every day: abundant recombination in a virus during a single multi-cellular host infection. PLoS Biol. 2005;3:e89.
Froissart R, Doumayrou J, Vuillaume F, Alizon S, Michalakis Y. The virulence-transmission trade-off in vector-borne plant viruses: a review of (non-)existing studies. Philos Trans R Soc B. 2010;365:1907–18.
García-Andrés S, Tomás DM, Sánchez-Campos S, Navas-Castillo J, Moriones E. Frequent occurrence of recombinants in mixed infections of tomato yellow leaf curl disease associated begomoviruses. Virology. 2007;365:210–9.
García-Arenal F, Fraile A. Questions and concepts in plant virus evolution: a historical perspective. In: Roossinck MJ, editor. Plant virus evolution. Berlin: Springer; 2008. p. 1–14.
García-Arenal F, Fraile A. Population dynamics and genetics of plant infection by viruses. In: Caranta C, Aranda MA, Tepfer M, Lopez-Moya JJ, editors. Recent advances in plant virology. Norfolk: Caister Academic Press; 2011. p. 263–81.
García-Arenal F, Fraile A. Trade-offs in host range evolution of plant viruses. Plant Pathol. 2013;62:S2–9.
García-Arenal F, McDonald BA. An analysis of the durability of resistance to plant viruses. Phytopathology. 2003;93:941–52.
García-Arenal F, Fraile A, Malpica JM. Variability and genetic structure of plant virus populations. Annu Rev Phytopathol. 2001;39:157–86.
Ghoshal B, Sanfaçon H. Symptom recovery in virus-infected plants: revisiting the role of RNA silencing mechanisms. Virology. 2015;479–480:167–79.
Giampetruzzi A, Roumi V, Roberto R, Malossini U, Yoshikawa N, La Notte P, Terlizzi F, Credi R, Saldarelli P. A new grapevine virus discovered by deep sequencing of virus- and viroid-derived small RNAs in Cv Pinot gris. Virus Res. 2012;163:262–8.
Gibbs AJ. A plant virus that partially protects its wild legume host against herbivores. Intervirology. 1980;13:42–7.
Gibbs AJ, Blok J, Coates DJ, Guy PL, Mackenzie A, Pigram N. Turnip yellow mosaic virus in an endemic Australian alpine *Cardamine*. In: Barlow BA, editor. Flora and Fauna of Alpine Australasia; ages and origins. Collingwood: CSIRO; 1986. p. 289–300.
Gibbs AJ, Gibbs MJ, Ohshima K, García-Arenal F. More plant virus evolution; past present and future. In: Domingo E, Parrish CR, Holland JJ, editors. Origin and evolution of viruses. 2nd ed. London: Academic Press; 2008a.
Gibbs AJ, Ohshima K, Phillips MJ, Gibbs MJ. The prehistory of potyviruses: their initial radiation was during the dawn of agriculture. PLoS One. 2008b;3:e2523.
Gibbs AJ, Fargette D, García-Arenal F, Gibbs MJ. Time—the emerging dimension of plant virus studies. J Gen Virol. 2010;91:13–22.
Gibbs AJ, Wood J, Garcia-Arenal F, Ohshima K, Armstrong JS. Tobamoviruses have probably co-diverged with their eudicotyledonous hosts for at least 110 million years. Virus Evol. 2015;1:vev019.

Gibbs AJ, Ohshima K, Yasaka R, Mohammadi M, Gibbs MJ, Jones RAC. The phylogenetics of the global population of potato virus Y and its necrogenic recombinants. Virus Evol. 2017;3: vex002.

Gómez P, Sempere RN, Elena SF, Aranda MA. Mixed infections of *Pepino mosaic virus* strains modulate the evolutionary dynamics of this emergent virus. J Virol. 2009;83:12378–87.

Gómez P, Sempere RN, Aranda MA, Elena SF. Phylodynamics of *Pepino mosaic virus*. Eur J Plant Pathol. 2012;134:445–9.

González-Jara P, Fraile A, Canto T, García-Arenal F. The multiplicity of infection of a plant virus varies during colonization of its eukaryotic host. J Virol. 2009;83:7487–94.

González-Jara P, Fraile A, Canto T, García-Arenal F. The multiplicity of infection of a plant virus varies during colonization of its eukaryotic host. Author's correction. J Virol. 2013;87:2374.

Gutiérrez S, Yvon M, Thébaud G, Monsion B, Michalakis Y, Blanc S. Dynamics of the multiplicity of cellular infection in a plant virus. PLoS Pathog. 2010;6:e1001113.

Gutiérrez S, Michalakis Y, Blanc S. Virus population bottlenecks during within-host progression and host-to-host transmission. Curr Op Virol. 2012a;2:546–55.

Gutiérrez S, Yvon M, Pirolles E, Garzo E, Fereres A, Michalakis Y, Blanc S. Circulating virus load determines the size of bottlenecks in viral populations progressing within a host. PLoS Pathog. 2012b;8:e1003009.

Gutiérrez S, Pirolles E, Yvon M, Baecker V, Michalakis Y, Blanc S. The multiplicity of cellular infection changes depending on the route of cell infection in a plant virus. J Virol. 2015;89:9665–75.

Guy P, Gibbs AJ. A tymovirus of *Cardamine* sp. from alpine Australia. Australas Plant Pathol. 1981;10:12–3.

Hackett J, Muthukumar V, Wiley GB, Palmer MW, Roe BA, Melcher U. Viruses in Oklahoma *Euphorbia marginata*. Proc Oklahoma Acad Sci. 2009;89:49–54.

Hadidi A, Flores R, Candresse T, Barba M. Next-generation sequencing and genome editing in plant virology. Front Microbiol. 2016;7:1325.

Hajimorad MR, Wen R-H, Eggenberger AL, Hill JH, Saghai Maroof MA. Experimental adaptation of an RNA virus mimics natural evolution. J Virol. 2011;85:2557–64.

Hall JS, French R, Hein GL, Morris TJ, Stenger DC. Three distinct mechanisms facilitate genetic isolation of sympatric wheat streak mosaic virus lineages. Virology. 2001;282:230–6.

Hall GS, Peters JS, Little DP, Power AG. Plant community diversity influences vector behavior and *Barley yellow dwarf virus* population structure. Plant Pathol. 2010;59:152–1158.

Hamada H, Takeuchi S, Kiba A, Tsuda S, Hikichi Y, Okuno T. Amino acid changes in *Pepper mild mottle virus* coat protein that affect L^3 gene-mediated resistance in pepper. J Gen Plant Pathol. 2002;68:155–62.

Hamada H, Tomita R, Iwadate Y, Kobayashi K, Minemura I, Takeuchi S, Hikichi Y, Suzuki K. Cooperative effect of two amino acid mutations in the coat protein of Pepper mild mottle virus overcomes L^3-mediated resistance in *Capsicum* plants. Virus Genes. 2007;34:205–14.

Harkins GW, Delport W, Duffy S, Wood N, Monjane AL, Owor BE, Donaldson L, Saumtally S, Triton G, Briddon RW, Shepherd DN, Rybicki EP, Martin DP, Varsani A. Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. Virol J. 2009;6:104.

Harrison BD. The infectivity of extracts made from leaves at intervals after inoculation with viruses. J Gen Microbiol. 1956;15:210–20.

Hartl DL, Clark AG. Principles of population genetics. 4th ed. Sunderland: Sinauer; 2007.

Hillung J, Cuevas JM, Elena SF. Evaluating the within-host fitness effects of mutations fixed during virus adaptation to different ecotypes of a new host. Philos Trans R Soc B. 2015;370:20140292.

Hochberg ME. Population dynamic consequences of the interplay between parasitism and intra-specific competition for host-parasite systems. Oikos. 1991;61:297–306.

Holmes EC. The evolution and emergence of RNA viruses. Oxford: Oxford University Press; 2009.

Hughes AL. Small effective population sizes and rare nonsynonymous variants in potyviruses. Virology. 2009;393:127–34.

Hull R. Plant virology. 5th ed. San Diego: Academic Press; 2014.
Hull R, Plaski A. Electron microscopy on the behaviour of two strains of *Alfalfa mosaic virus* in mixed infections. Virology. 1970;42:773–6.
Hurwitz BL, Sullivan MB. The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. PLoS One. 2013;8:e57355.
Ishibashi K, Kezuka Y, Kobayashi C, Kato M, Inoue T, Nonaka T, Ishikawa M, Matsumura H, Katoh E. Structural basis for the recognition-evasion arms race between *Tomato mosaic virus* and the resistance gene Tm-1. Proc Natl Acad Sci U S A. 2014;111:3486–95.
Jenkins GM, Rambaut A, Pybus OG, Holmes EC. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J Mol Evol. 2002;54:156–65.
Johnson PT, Ostfeld RS, Keesing F. Frontiers in research on biodiversity and disease. Ecol Lett. 2015;18:1119–33.
Jones RAC. Plant virus emergence and evolution: origins, new encounter scenarios, factors driving emergence, effects of changing world conditions, and prospects for control. Virus Res. 2009;141:113–30.
Jridi C, Martin JF, Marie-Jeanne V, Labonne G, Blanc S. Distinct viral populations differentiate and evolve independently in a single perennial host plant. J Virol. 2006;80:2349–57.
Kay KM, Whittall JB, Hodges SA. A survey of nuclear ribosomal internal transcribed spacer substitution rates across angiosperms: an approximate molecular clock with life history effects. BMC Evol Biol. 2006;6:36.
Keesing F, Holt RD, Ostfeld RS. Effects of species diversity on disease risk. Ecol Lett. 2006;9:485–98.
Keesing F, Belden LK, Daszk P, Dobson A, Harwell CD, Holt RD, Hudson P, Jolles A, Jones KE, Mitchell CE, Myers SS, Bogich T, Ostfeld RS. Impacts of biodiversity on the emergence and transmission of infectious diseases. Nature. 2010;468:647–52.
Kelly SE. Viral pathogens and the advantage of sex in the perennial grass *Anthoxanthum odoratum*: a review. Phil Trans R Soc Lond B. 1994;346:295–302.
Kinoti WM, Constable FE, Nancarrow N, Plummer KM, Rodoni B. Analysis of intra-host genetic diversity of *Prunus necrotic ringspot virus* (PNRSV) using amplicon next generation sequencing. PLoS One. 2017;12:e0179284.
Kreuze JF, Pérez A, Untiveros M, Quispe D, Fuentes S, Barker I, Simon R. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology. 2009;388:1–7.
Lanfear R, Kokko H, Eyre-Walker A. Population size and the rate of evolution. Trends Ecol Evol. 2014;29:33–41.
Lartey RT, Voss TC, Melcher U. Tobamovirus evolution: gene overlaps, recombination, and taxonomic implications. Mol Biol Evol. 1996;13:1327–38.
Lefeuvre P, Harkins GW, Lett J-M, Briddon RW, Chase MW, Moury B, Martin DP. Evolutionary time-scale of the begomoviruses: evidence from integrated sequences in the *Nicotiana* genome. PLoS One. 2011;6:e19193.
Levontin RC. The units of infection. Annu Rev Ecol Syst. 1970;1:1–18.
Li H, Roossinck MJ. Genetic bottlenecks reduce population variation in an experimental RNA virus population. J Virol. 2004;78:10582–7.
Lima ATM, Silva JCF, Silva FN, Castillo-Urquiza GP, Silva FF, Seah YM, Mizubuti ESG, Duffy S, Murilo Zerbini F. The diversification of begomovirus populations is predominantly driven by mutational dynamics. Virus Evol. 2017;3:vex005.
Lipkin WI. The changing face of pathogen discovery and surveillance. Nat Rev Microbiol. 2013;11:133–41.
Little TJ, Shuker DM, Colegrave N, Day N, Graham AL. The coevolution of virulence: tolerance in perspective. PLoS Pathog. 2010;6:e1001006.
Lively CM. The ecology of virulence. Ecol Lett. 2006;9:1089–95.
Malmstrom CM, Alexander HM. Effects of crop viruses on wild plants. Curr Op Virol. 2016;19:30–6.

Malmstrom CM, Hughes CC, Newton LA, Stoner CJ. Virus infection in remnant native bunchgrasses from invaded California grasslands. New Phytol. 2005a;168:217–30.
Malmstrom CM, McCullough AJ, Johnson HA, Newton LA, Borer ET. Invasive annual grasses indirectly increase virus incidence in California native perennial bunchgrasses. Oecologia. 2005b;145:153–64.
Malmstrom CM, Shu R, Linton EW, Newton LA, Cook MA. Barley yellow dwarf viruses (BYDVs) preserved in herbarium specimens illuminate historical disease ecology of invasive and native grasses. J Ecol. 2007;95:1153–66.
Malpica JM, Fraile A, Moreno I, Obies CI, Drake JW, García-Arenal F. The rate and character of spontaneous mutation in an RNA virus. Genetics. 2002;162:1505–11.
Malpica JM, Sacristán S, Fraile A, García-Arenal F. Association and host selectivity in multi-host pathogens. PLoS One. 2006;1:e41.
Martín S, Elena SF. Application of game theory to the interaction between plant viruses during mixed infections. J Gen Virol. 2009;90:2815–20.
Maskell LC, Raybould AF, Cooper JI, Edwards ML, Gray AJ. Effects of turnip mosaic virus and turnip yellow mosaic virus on the survival, growth and reproduction of wild cabbage (*Brassica oleracea*). Ann Appl Biol. 1999;135:401–7.
Massart S, Olmos A, Jijakli H, Candresse T. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res. 2014;188:90–6.
Massart S, Candresse T, Gil J, Lacomme C, Predajna L, Ravnikar M, Reynard JS, Rumbou A, Saldarelli P, Škorić D, Vainio EJ, Valkonen JP, Vanderschuren H, Varveri C, Wetzel T. A framework for the evaluation of biosecurity, commercial, regulatory, and scientific impacts of plant viruses and viroids identified by NGS technologies. Front Microbiol. 2017;8:45.
McKinney HH. Evidence of virus mutation in the common mosaic of tobacco. J Agric Res. 1935;51:951–81.
McLeish MJ, Sacristán S, Fraille A, García-Arenal F. Scale dependencies and generalism in host use shape virus prevalence. Proc. R. Soc. B, 2017;284: 20172066.
Miyashita S, Kishino H. Estimation of the size of genetic bottlenecks in cell-to-cell movement of soil-borne wheat mosaic virus and the possible role of the bottlenecks in speeding up selection of variations in trans-acting genes or elements. J Virol. 2010;84:1828–37.
van Molken T, de Caluwe H, Hordijk CA, Leon-Reyes A, Snoeren TA, van Dam NM, Stuefer JF. Virus infection decreases the attractiveness of white clover plants for a non-vectoring herbivore. Oecologia. 2012;170:433–44.
Monsion B, Froissart R, Michalakis Y, Blanc S. Large bottleneck size in *Cauliflower mosaic virus* populations during host plant colonization. PLoS Pathog. 2008;4:e1000174.
Moreno A, De Blas C, Biurrun R, Nebreda M, Palacios I, Duque M, Fereres A. The incidence and distribution of viruses infecting lettuce, cultivated Brassica and associated natural vegetation in Spain. Ann Appl Biol. 2004;144:339–46.
Morse SS, Schluederberg A. Emerging viruses: the evolution of viruses and viral diseases. J Infect Dis. 1990;162:1–7.
Moury B, Janzac B, Ruellan Y, Simon V, Ben Khalifa M, Fakhfakh H, Fabre F, Palloix A. Interaction patterns between *Potato virus Y* and eIF4E-mediated recessive resistance in the *Solanaceae*. J Virol. 2014;88:9799–807.
Moury B, Fabre F, Hébrard E, Froissart R. Determinants of host species range in plant viruses. J Gen Virol. 2017;98:862–73.
Moya A, Rodríguez-Cerezo E, García-Arenal F. Genetic structure of natural populations of the plant RNA virus tobacco mild green mosaic virus. Mol Biol Evol. 1993;10:449–56.
Moya A, Holmes EC, González-Candelas F. The population genetics and evolutionary epidemiology of RNA viruses. Annu Rev Microbiol. 2004;2:279–88.
Nee S, Maynard-Smith J. The evolutionary biology of molecular parasites. Parasitology. 1990;100: S5–S18.
Ng TFF, Duffy S, Polston JE, Bixby E, Vallad GE, Breitbart M. Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. PLoS One. 2011;6:e19050.

Nieberding CM, Olivieri I. Parasites: proxies for host genealogy and ecology? Trends Ecol Evol. 2007;22:156–65.
Nsa IY, Kareem KT. Additive interactions of unrelated viruses in mixed infections of cowpea (*Vigna unguiculata* L. Walp). Front Plant Sci. 2015;6:812.
Nuismer SL, Jordano P, Bascompte J. Coevolution and the architecture of mutualistic networks. Evolution. 2013;67:338–54.
Ostfeld RS, Keesing F. Effects of host diversity on infectious disease. Annu Rev Ecol Evol Syst. 2012;43:157–82.
Ostfeld RS, Keesing F. Is biodiversity bad for your health? Ecosphere. 2017;8:e01676.
Pagán I, Holmes EC. Long-term evolution of the *Luteoviridae*: time scale and mode of virus speciation. J Virol. 2010;84:6177–87.
Pagán I, Alonso-Blanco C, García-Arenal F. The relationship of within-host multiplication and virulence in a plant-virus system. PLoS One. 2007;2:e786.
Pagán I, Alonso-Blanco C, García-Arenal F. Host responses in life-history traits and tolerance to virus infection in *Arabidopsis thaliana*. PLoS Pathog. 2008;4:e1000124.
Pagán I, Alonso-Blanco C, García-Arenal F. Differential tolerance to direct and indirect density-dependent costs of viral infection in *Arabidopsis thaliana*. PLoS Pathog. 2009;5:e1000531.
Pagán I, Firth C, Holmes EC. Phylogenetic analysis reveals rapid evolutionary dynamics in the plant RNA virus genus *Tobamovirus*. J Mol Evol. 2010a;71:298–307.
Pagán I, Fraile A, Fernández-Fuello E, Montes N, Alonso-Blanco C, García-Arenal F. *Arabidopsis thaliana* as a model for plant-virus co-evolution. Philos Trans R Soc B. 2010b;365:1983–95.
Pagán I, González-Jara P, Moreno-Letelier A, Rodelo-Urrego M, Fraile A, Piñero D, García-Arenal F. Effect of biodiversity changes in disease risk: exploring disease emergence in a plant-virus system. PLoS Pathog. 2012;8:e1002796.
Pagán I, Montes N, Milgroom MG, García-Arenal F. Vertical transmission selects for reduced virulence in a plant virus and for increased resistance in the host. PLoS Pathog. 2014;10:e1004293.
Pagán I, Fraile A, García-Arenal F. Evolution of the interactions of viruses with their plant hosts. In: Weaver SC, Denison M, Roossink MJ, Vignuzzi M, editors. Virus evolution: current research and future directions. Norfolk: Caister Academic Press; 2016. p. 127–54.
Pinel-Galzi AS, Rakotomalala M, Sangu E, Sorho F, Kanyeka Z, Traoré O, Sérémé D, Poulicard N, Rabenantoandro Y, Sere Y, Konaté G, Ghesquiere A, Hébrard E, Fargette D. Theme and variations in the evolutionary pathways to virulence of an RNA plant virus species. PLoS Pathog. 2007;3:e180.
Piry S, Wipf-Scheibel C, Martin J-F, Galan M, Berthier K. High throughput amplicon sequencing to assess within- and between-host genetic diversity in plant viruses. BioRXiv. 2017. https://doi.org/10.1101/168773.
Poulicard N, Pinel-Galzi A, Fargette D, Hébrard E. Alternative mutational pathways, outside the VPg, of rice yellow mottle virus to overcome eIF(iso)4G-mediated rice resistance under strong genetic constraints. J Gen Virol. 2014;95:219–24.
Power AG, Mitchell CE. Pathogen spillover in disease epidemics. Am Nat. 2004;164:S79–89.
Power AG, Borer ET, Hosseini P, Mitchell CE, Seabloom EW. The community ecology of barley/cereal yellow dwarf viruses in Western US grasslands. Virus Res. 2011;159:95–100.
Prendeville HR, Ye XH, Morris TJ, Pilson D. Virus infections in wild plant populations are both frequent and often unapparent. Am J Bot. 2012;99:1033–42.
Prendeville HR, Tenhumberg B, Pilson D. Effects of virus on plant fecundity and population dynamics. New Phytol. 2014;202:1346–56.
Pressing J, Reanney DC. Divided genomes and intrinsic noise. J Mol Evol. 1984;20:135–46.
Quenoiulle J, Vassilakos N, Moury B. *Potato virus Y*: a major crop pathogen that has provided major insights into the evolution of viral pathogenicity. Mol Plant Pathol. 2013;14:439–52.
Randolph SE, Dobson DM. Pangloss revisited: a critique of the dilution effect and the biodiversity-buffers-disease paradigm. Parasitology. 2012;139:847–63.
Rato S, Golumbeanu M, Telenti A, Ciuffi A. Exploring viral infection using single-cell sequencing. Virus Res. 2016;239:55–68.

Rentería-Canett I, Xoconostle-Cázares B, Ruiz-Medrano R, Rivera-Bustamante RF. Geminivirus mixed infection on pepper plants: synergistic interaction between PHYVV and PepGMV. Virol J. 2011;8:104.
Rochow WF. The role of mixed infections in the transmission of plant viruses by aphids. Annu Rev Phytopatol. 1972;10:101–24.
Rodelo-Urrego M, Pagán I, González-Jara P, Betancourt M, Moreno-Letelier A, Ayllón MA, Fraile A, Piñero D, García-Arenal F. Landscape heterogeneity shapes host-parasite interactions and results in apparent plant-virus codivergence. Mol Ecol. 2013;22:2325–40.
Rodelo-Urrego M, García-Arenal F, Pagán I. The effect of ecosystem biodiversity on virus genetic diversity depends on virus species: a study of chiltepin-infecting begomoviruses in Mexico. Virus Evol. 2015;1:vev004.
Rodríguez-Nevado C, Montes N, Pagán I. Ecological factors affecting the infection risk and population genetic diversity of a novel potyvirus in its native wild ecosystem. Front Plant Sci. 2017;8:1958.
Roossinck MJ. Mechanisms of plant virus evolution. Annu Rev Phytopathol. 1997;35:191–209.
Roossinck MJ. Symbiosis versus competition in plant virus evolution. Nat Rev Microbiol. 2005;3:917–24.
Roossinck MJ. The big unknown: plant virus biodiversity. Curr Op Virol. 2011;1:63–7.
Roossinck MJ. Plant virus metagenomics: biodiversity and ecology. Annu Rev Genet. 2012;46:357–67.
Roossinck MJ. Deep sequencing for discovery and evolutionary analysis of plant viruses. Virus Res. 2017;239:82–6.
Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, Lai H, Chavarría F, Shen G, Roe BA. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol. 2010;19:81–8.
Roossinck MJ, Martin DP, Roumagnac P. Plant virus metagenomics: advances in virus discovery. Phytopatology. 2015;105:716–27.
Sacristán S, García-Arenal F. The evolution of virulence and pathogenicity in plant pathogen populations. Mol Plant Pathol. 2008;9:369–84.
Sacristán S, Malpica JM, Fraile A, García-Arenal F. Estimation of population bottlenecks during systemic movement of tobacco mosaic virus in tobacco plants. J Virol. 2003;77:9906–11.
Salvaudon L, De Moraes CM, Mescher MC. Outcomes of co-infection by two potyviruses: implications for the evolution of manipulative strategies. Proc R Soc Lon B. 2013;280:20122959.
Sanjuán R, Agudelo-Romero P, Elena SF. Upper-limit mutation rate estimation for a plant RNA virus. Biol Lett. 2009;5:394–6.
Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. J Virol. 2010;84:9733–48.
Sasu MA, Ferrari MJ, Du D, Winsor JA, Stephenson AG. Indirect costs of a nontarget pathogen mitigate the direct benefits of a virus-resistant transgene in wild *Cucurbita*. 2009;45:19067–71.
Scholle SO, Ypma RJ, Lloyd AL, Koelle K. Viral substitution rate variation can Arise from the interplay between within-host and epidemiological dynamics. Am Nat. 2013;182:494–513.
Shuckla A, Pagán I, García-Arenal F. Effective tolerance based on resource reallocation is a virus-specific defence in Arabidopsis thaliana. Mol Plant Pathol. Published on line 30 Jan 2018.
Sicard A, Yvon M, Timchenko T, Gronenborn B, Michalakis Y, Gutierrez S, Blanc S. Gene copy number is differentially regulated in a multipartite virus. Nat Commun. 2013;4:2248.
Sicard A, Michalakis Y, Gutierrez S, Blanc S. The strange lifestyle of multipartite viruses. PLoS Pathog. 2016;12:e1005819.
Simmons HE, Holmes EC, Stephenson AG. Rapid evolutionary dynamics of zucchini yellow mosaic virus. J Gen Virol. 2008;89:1081–5.
Simmons HE, Dunham JP, Stack JC, Dickins BJA, Pagán I, Holmes EC, Stephenson AG. Deep sequencing reveals persistence of intra- and inter-host genetic diversity in natural and greenhouse populations of zucchini yellow mosaic virus. J Gen Virol. 2012;93:1831–40.

Simon AE, Bujarski JJ. RNA-RNA recombination and evolution in virus infected plants. Annu Rev Phytopathol. 1994;32:337–62.
Stenger DC, Seifers DL, French R. Patterns of polymorphism in wheat streak mosaic virus: sequence space explored by a clade of closely related viral genotypes rivals that between the most divergent strains. Virology. 2002;302:58–70.
Stobbe AH, Roossinck MJ. Plant virus metagenomics: what we know and why we need to know more. Front Plant Sci. 2014;5:150.
Stobbe AH, Melcherl U, Palmer MW, Roossinck MJ, Shen G. Co-divergence and host-switching in the evolution of tobamoviruses. J Gen Virol. 2012;93:408–18.
Stukenbrock EH, McDonald BA. The origin of plant pathogens in agro- ecosystems. Annu Rev Phytopathol. 2008;46:75–100.
Syller J. Facilitative and antagonistic interactions between plant viruses in mixed infections. Mol Plant Pathol. 2012;13:204–16.
Szathmáry E. Viral sex, levels of selection, and the origin of life. J Theor Biol. 1992;159:99–109.
Taiwo MA, Kareem KT, Nsa IY, Hughes JD'A. Cowpea viruses: effect of single and mixed infections on symptomatology and virus concentration. Virol J. 2007;4:95.
Takahashi T, Sugawara T, Yamatsuta T, Isogai M, Natsuaki T, Yoshikawa N. Analysis of the spatial distribution or identical and two distinct virus populations differently labelled with cyan and yellow fluorescent proteins in coinfected plants. Phytopathology. 2007;97:1200–6.
Takeshita M, Shigemune N, Kikuhara K, Takanami Y. Spatial analysis for exclusive interactions between subgroups I and II of cucumber mosaic virus in cowpea. Virology. 2004;328:45–51.
Tepfer M. Risk assessment of virus-resistant transgenic plants. Annu Rev Phytopathol. 2002;40:467–91.
Thébaud G, Michalakis Y. Comment on "Large bottleneck size in *Cauliflower mosaic virus* populations during host plant colonization" by Monsion et al. (2008). PLoS Pathog. 2016;12: e1005512.
Thompson AD. Interactions between plant viruses. I Appearance of new strains after mixed infection with *Potato virus X* strains. Virology. 1961;13:507–14.
Thompson JN. The geographic mosaic of coevolution. Chicago: University of Chicago Press; 2005.
Thresh JM. Cropping practices and virus spread. Annu Rev Phytopathol. 1982;20:193–218.
Tomita R, Murai J, Miura Y, Ishikara H, Liu S, Kubotera Y, Honda A, Hatta R, Kuroda T, Hamada H, Sakamoto M, Munemura I, Nunomura O, Ishikawa K, Genda Y, Kawasaki S, Suzuki K, Meksem K, Kobayashi K. Fine mapping and DNA fiber FISH analysis locates the tobamovirus resistance gene L^3 of *Capsicum chinense* in a 400-kb region of R-like genes cluster embedded in highly repetitive sequences. Theor Appl Genet. 2008;117:1107–18.
Tromas N, Elena SF. The rate and spectrum of spontaneous mutations in a plant RNA virus. Genetics. 2010;185:983–9.
Tromas N, Zwart MP, Lafforgue G, Elena SF. Within-host spatiotemporal dynamics of plant virus infection at the cellular level. PLoS Genet. 2014a;10:e1004186.
Tromas N, Zwart MP, Poulain M, Elena SF. Estimation of the in vivo recombination rate for a plant RNA virus. J Gen Virol. 2014b;95:724–32.
Trovão NS, Baele G, Vrancken B, Bielejec F, Suchard MA, Fargette D, Lemey P. Host ecology determines the dispersal patterns of a plant virus. Virus Evol. 2015;1:vev016.
Tsuda S, Kirita M, Watanabe Y. Characterization of a pepper mild mottle tobamovirus strain capable of overcoming the L^3 gene-mediated resistance, distinct from the resistance-breaking Italian isolate. Mol Plant Microbe Interact. 1998;11:327–31.
Tugume AK, Mukasa SB, Valkonen JPT. Mixed infections of four viruses, the incidence and phylogenetic relationships of *Sweet potato chlorotic fleck virus* (*Betaflexiviridae*) isolates in wild species and sweetpotatoes in Uganda and evidence of distinct isolates in East Africa. PLoS One. 2016;11:e0167769.
van der Walt E, Martin DP, Varsani A, Polston JE, Rybicki EP. Experimental observations of rapid maize streak virus evolution reveal a strand-specific nucleotide substitution bias. Virol J. 2008;5:104.

Wang J, Moore NE, Deng Y-M, Eccles DA, Hall RJ. MinION nanopore sequencing of an influenza genome. Front Microbiol. 2017;6:766.

Wanunu M. Nanopores: a journey towards DNA sequencing. Phys Life Rev. 2012;9:125–58.

Woolhouse MEJ, Webster JP, Domingo E, Charlesworth B, Levin BR. Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nat Genet. 2002;32:569–77.

Worobey M, Holmes EC. Evolutionary aspects of recombination in RNA viruses. J Gen Virol. 1999;80:2535–43.

Wu X, Xu Z, Shaw JG. Uncoating of tobacco mosaic virus RNA in protoplasts. Virology. 1994;200:256–62.

Wu B, Melcher U, Guo X, Wang X, Fan L, Zhou G. Assessment of codivergence of mastreviruses with their plant hosts. BMC Evol Biol. 2008;8:335.

Wu B, Blanchard-Letort A, Liu Y, Zhou G, Wang X, Elena SF. Dynamics of molecular evolution and phylogeography of *Barley yellow dwarf virus*-PAV. PLoS One. 2011;6:e16896.

Wu Q, Ding SW, Zhang Y, Zhu S. Identification of viruses and viroids by next-generation sequencing and homology-dependent and homology-independent algorithms. Annu Rev Phytopathol. 2015;53:425–44.

Wylie SJ, Li H, Dixon KW, Richards H, Jones MGK. Exotic and indigenous viruses infect wild populations and captive collections of temperate terrestrial orchids (*Diuris* species) in Australia. Virus Res. 2013;171:22–32.

Xu P, Chen F, Mannas JP, Feldman T, Sumner LW, Roossinck MJ. Virus infection improves drought tolerance. New Phytol. 2008;180:911–21.

Yasaka R, Nguyen HD, Ho SYW, Duchêne S, Korkmaz S, Nikolaos K, Takahashi H, Gibbs AJ, Ohshima K. The temporal evolution and global spread of *Cauliflower mosaic virus*, a plant pararetrovirus. PLoS One. 2014;9:e85641.

Zwart MP, Willemsen A, Darós JA, Elena SF. Experimental evolution of pseudogenization and gene loss in a plant RNA virus. Mol Biol Evol. 2014;31:121–34.

Population Genomics of Human Viruses

Fernando González-Candelas, Juan Ángel Patiño-Galindo, and Carlos Valiente-Mullor

Abstract Viruses, and a few RNA viruses in particular, represent one of the greatest threats for human health. High mutation rates, large population sizes, and short generation times contribute to their typically fast evolutionary rates. However, many additional processes operate on their genomes, often in opposite directions, driving their evolution and allowing them to adapt to diverse host populations and antiviral drugs. Until recently, the high levels of genetic variation of most viruses have been explored only at a few genes or genome regions. The recent advent and increasing affordability of next-generation sequencing techniques have allowed obtaining complete genome sequences of large numbers of viruses, mainly HIV, HCV, influenza A, and others associated with emerging infections, such as Zika, chikungunya, or dengue virus. This opens the possibility to explore the effects of the different processes affecting viral diversity and evolution at the genome level. Consequently, population genomics provides the conceptual and empirical tools necessary to interpret genetic variation in viruses and its dynamics and drivers and to transform these results into information that may complement the epidemiological surveillance of the virus and its disease. This chapter provides an overview of human viruses from a population genomics perspective, with a special emphasis on RNA viruses, and the potential benefits of "genomic surveillance" to establish public health policies that improve the control and monitoring of the diseases caused by these viruses.

Keywords Complete genome · Epidemiology · Genetic variation · Mutation · Next-generation sequencing · Phylogeography · Reassortment · Recombination · Secondary structure

F. González-Candelas (✉), J. Á. Patiño-Galindo · C. Valiente-Mullor
Joint Research Unit "Infection and Public Health" FISABIO-Universitat de València, Institute for Integrative Systems Biology, I2SysBio (CSIC-UV), Valencia, Spain

CIBER in Epidemiology and Public Health, Madrid, Spain
e-mail: fernando.gonzalez@uv.es

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_31,

1 Introduction

The development of fast and efficient sequencing methodologies has brought the opportunity for obtaining complete sequences of hundreds, even thousands, of viral genomes at affordable costs. This has led to a new interest in the analysis of viral populations, which, until recently, was usually linked to outbreaks and other health emergencies. Most previous studies paid attention only to those fragments of the viral genome that were of interest from a clinical perspective, for diagnostics, surveillance, or similar applications. Furthermore, most insights into the population genetics of viral populations were drawn from markers likely under the influence of selective forces, thus leading to distorted or biased views of viral population genetics. This situation is rapidly changing, and the availability of complete genome sequences is shifting the perspective from "population genetics" to "population genomics," that is, the analysis of the processes and mechanisms that govern the population dynamics of genetic variation at the complete genome level and not only on a portion of it.

Although information on complete genomes is rapidly accumulating, there is still a huge gap between the number and diversity of viral population samples that have been analyzed in only one or a few genes and those with complete genome information. However, there is a shift of interest in using population genomics inference to better understand the intra-host and inter-host dynamics of epidemiologically and evolutionary relevant processes and to incorporate this information into surveillance systems. This shift has also benefited from recent methodological and technical advances, which have allowed the combination of different sources of information (temporal, geographical, genetic, and epidemiological) into a comprehensive framework, known as "genomic surveillance." Here, we review the current state of the art in the population genomics of human viruses and its relevance for the surveillance, monitoring, and control of the diseases they cause. Because of their rapid rate of evolution and the serious diseases they produce – AIDS, hepatitis C, Ebola, influenza, among many others – RNA viruses have received most attention until now and abundant information on their population genomics is accumulating. We will center this review on these viruses.

2 Evolutionary Processes in Viral Populations

Mutation is the ultimate source of variation in all living organisms, viruses included. However, the genetic diversity and the evolutionary rate of RNA viruses are influenced and shaped by other processes and factors apart from mutation. The action of natural selection and genetic drift, the mode of transmission, particular mechanisms for genetic exchange (such as recombination and reassortment), genome size, procedures for compressing genetic information, generation time, and population size are the most relevant such factors. In addition, we must also consider environmental

factors, resulting from differences among hosts and, occasionally, from antiviral treatments (Cuypers et al. 2016; Rambaut et al. 2008; Renzette et al. 2014; Simon-Loriere et al. 2013; Snoeck et al. 2011; Wilson et al. 2016).

Deciphering the mechanisms responsible for the production of spontaneous mutations in viruses has important applications for public health and for basic science (Geller et al. 2016) due to their critical role in virus evolution and genetic diversity (Cuevas et al. 2015). One defining feature of RNA viruses is their high mutation rates, in the range from 10^{-3} to 10^{-6} mutations/nucleotide/replication round, which result from low-fidelity replication (Simon-Loriere et al. 2013; Cuevas et al. 2015; Duffy et al. 2008; Sanjuan et al. 2010). This also leads to a very high evolutionary rate of 10^{-2} to 10^{-5} substitutions/site/year. These high mutation rates can be decomposed into several factors or mechanisms with complex interactions such as the fidelity of the RNA polymerase, the capacity for error correction, the propensity of ribonucleic acid to damage, or the edition by hosts' enzymes (Geller et al. 2016; Cuevas et al. 2015). These high mutation rates might explain the small genome size of RNA viruses (ranging from 3 to 29 kb) because in larger genomes, deleterious mutations would appear at such a high frequency that they would compromise virus survival (Duffy et al. 2008; Bradwell et al. 2013).

In general, RNA viruses have short generation times and large population sizes. These features favor fast evolutionary rates, which lead to genetically very diverse populations, with a high capacity for adaptation even under very strong selective pressures (Wilson et al. 2016). However, we must consider that virus evolutionary rates are limited by the frequency of deleterious mutations since virus mutation rates are very close to the error threshold beyond which deleterious mutations are so frequent that they lead to population extinction (Holmes 2003). In addition, some mechanisms for genetic exchange, which are present in some RNA viruses, favor the generation and maintenance of diversity. Two such mechanisms are genetic reassortment and recombination (homologous and nonhomologous).

Genetic reassortment occurs in segmented viruses, whose genome is distributed in individual segments, each carrying a different portion of the genetic information. Reassortment plays a major role at the epidemiological level in the evolution of influenza A virus (Rambaut et al. 2008; Wilson et al. 2016; Steel and Lowen 2014) and in other segmented viruses (McDonald et al. 2016; Nomikou et al. 2015). Recombination is also a frequent process in many viruses. In retroviruses, such as HIV, a nonhomologous type of recombination known as copy-choice recombination is common, and it can occur when two different viral strains simultaneously infect the same cell. In this form of recombination, the RNA polymerase "jumps" between two copies of single-stranded RNA, which makes up the genome of retroviruses, while it is still attached to the newly synthesized chain. This mechanism occurs only during RNA synthesis and the parental (donor) strand is not physically transferred to the recombined strand. It is likely that secondary structures of the RNA genomes are involved in controlling the "jump" between strands (Lai 1992; Negroni and Buc 2000; Simon-Loriere and Holmes 2011).

Natural selection may deplete genetic diversity from viral populations (negative or purifying selection) or increase its levels (some forms of positive selection) and,

consequently, may increase or decrease the rate of evolution. Therefore, there is a trade-off between the conservation of those regions that are essential for completing the viral cycle of replication and the genetic change and innovation that are involved in evading the immune system and responding to antiviral treatments. The former group includes genes encoding for slowly evolving enzymes and structural proteins as well as genome regions involved in the formation of secondary structures. This compromise is partially achieved through differential mutation rates along the viral genome (Geller et al. 2015, 2016).

Because of their high mutation rates, RNA viruses are under selection for small genome size. This is due to the deleterious effect on fitness of most mutations, which lead to an excessive genetic load in large genomes, which, in turn, leads to population extinction (Muller 1932). A small genome size represents a limitation for the generation of genetic diversity because (1) sequence lengths are limited and (2) using gene overlapping to compress genetic information implies an increase in the sensitivity to deleterious mutations in certain parts of the genome and, consequently, a larger role for purifying selection (Simon-Loriere et al. 2013). Although gene overlapping is present in all cellular organisms, mammals included (Veeramachaneni et al. 2004), it is very frequent only in viruses (Rogozin et al. 2002; Brandes and Linial 2016).

Another factor limiting the rate of evolution is the transmission between hosts. Each of these events represents a bottleneck that dramatically reduces the size of the viral population and, as a result, its genetic diversity (Gray et al. 2011; Grenfell et al. 2004; Joseph et al. 2015). Besides, founder viruses will generally be poorly adapted to the new environment because, in general, the specific adaptations to the immune system of the donor/source individual do not imply a higher fitness in another individual of the same species and they can even be penalized by natural selection (Kubinak et al. 2012).

Therefore, viruses mutate and may evolve very fast. It is crucial to understand the mechanisms by which they generate and maintain genetic diversity for the application of research on these organisms to health-related questions, such as the evasion of immune response, the development and spread of drug resistance mutations, virulence, species jumps, and the failure or success of vaccination campaigns (Wilson et al. 2016; Geller et al. 2016; Smyth et al. 2012). In this context, the rate of mutation should be considered not only as a mechanism generating diversity but also as a virulence factor (Cuevas et al. 2015).

3 Selective Pressures

In viruses, as in all living organisms, natural selection operates as a force that, on the one hand, may reduce genetic variability and, on the other hand, may increase genetic diversity (Snoeck et al. 2011). Hence, we start by describing the different types of natural selection that operate in viral populations.

Positive selection promotes an increase of the relative frequency of an allele or genetic variant in a population. Positively selected mutations confer higher fitness to

their carriers, resulting in an increased frequency of the corresponding allele. Two paradigmatic examples of positive selection are immune escape mutations and drug resistance mutations.

Positive selection may act as an evolutionary force that restricts the genetic diversity of a population (directional selection) or as a force that promotes an increase of genetic diversity (diversifying selection). Directional selection is commonly associated with selective sweeps. During selective sweeps, neutral or nearly neutral mutations increase their relative frequencies, even become fixed, in the population due to genetic linkage with positively selected variants (Maynard-Smith and Haigh 1974). The strength and scope of selective sweeps (which may act at the genome-wide level) (Rambaut et al. 2008) will also depend on the rate of recombination. For example, as detailed above, high recombination rates limit the scope of selective sweeps in HIV (Ramirez et al. 2008; Vuilleumier and Bonhoeffer 2015; Zanini et al. 2015). Additionally, the rate of emergence of adaptive mutations influences the intensity of selective sweeps.

When different adaptive mutations, which have either newly arisen or were previously present at low frequencies in a population, are selected simultaneously or nearly simultaneously, then soft sweeps will be produced as several mutations – located in different regions of the viral genome – propagate jointly. Such soft sweeps may result in clonal interference, which consists of competition between distinct lineages in the viral population carrying different adaptive mutations. Consequently, even mutations favored by natural selection may not be fixed in the population or, alternatively, they may get fixed at lower rates. Thus, clonal interference may slow down adaptation in a viral population (Miralles et al. 1999).

Soft selective sweeps have a minor impact on the loss of population genetic diversity. However, if adaptive mutations arise rarely in a population, then a single variant will increase its frequency along with its genetically linked neutral alleles. Consequently, a hard selective sweep will be produced, which implies a huge decrease in genomic diversity (Feder et al. 2016; Hermisson and Pennings 2005; Messer and Petrov 2013; Pennings et al. 2014).

The rate at which adaptive mutations emerge and become positively selected depends on the mutation rate, the population size, with small sizes resulting in strong genetic drift and reduced efficiency of natural selection, and the strength of the selective pressure. This complex process can be studied in individuals under antiviral drug therapy. Highly efficient drug treatments, consisting of a combination of antiviral drugs, reduce viral population size and the frequencies of drug-resistant alleles. Moreover, the number of permissive mutations needed for acquiring drug resistance (genetic barrier) may increase (Feder et al. 2016).

Diversifying selection occurs when selection favors different adaptive mutations over time and/or space, and it results in an increase in genetic diversity. Generally, we can observe this type of selection in viral responses to the hosts' immune systems. As a result, the genome regions coding for proteins targeted by the immune response (antigens) present much higher variability than the remainder of the genome. Antigenic drift – antigenic evolution in influenza A virus – exemplifies this phenomenon and should not be confused with genetic drift. Due to the interaction between influenza A

virus and the human immune system, mutations accumulate in antigenic regions encoding surface proteins such as hemagglutinin and neuraminidase. Influenza A viruses show episodic selection, that is, positive selection over long periods interspersed with purifying selection over short time periods. This process might explain the new seasonal antigenic variants of influenza A virus (Rambaut et al. 2008; Cobey and Koelle 2008; McHardy and Adams 2009). However, it must be noted that the two sides of positive selection are linked: antigenic drift is inevitably related to periodic selective sweeps (McHardy and Adams 2009).

Negative (or purifying) selection operates by removing deleterious alleles (i.e., mutations that decrease viral fitness). Negative effects of deleterious mutations can involve a reduction in the replication rate or increased susceptibility to the host immune response or to antiviral drugs. Most mutations arising in living organisms are deleterious. For instance, nearly 60% of the spontaneous mutations in vesicular stomatitis virus are deleterious (Duffy et al. 2008). Thus, purifying selection constitutes a force acting to preserve nucleotide or amino acid sequence. Therefore, negative selection constrains genetic diversity.

Purifying selection can be prominent in the viral genome, even in viruses, such as HIV, in which positive selection and neutral evolution have an important role at the intra-host level (Snoeck et al. 2011; Zanini et al. 2015; Pybus and Rambaut 2009; Ross and Rodrigo 2002). HCV is a clear example of predominance of negative selection. Despite the high levels of genetic variability in this virus, negative selection represents the main force acting on the HCV genome: more than 80% of the nucleotide sites in the viral genome are under negative selective pressure (Cuypers et al. 2016; Geller et al. 2016; Patiño Galindo and González-Candelas 2017).

Natural selection can be studied by comparing the synonymous substitution rate per synonymous site (dS) and the non-synonymous substitution rate per non-synonymous site (dN). The ratio of both rates ($\omega = \mathrm{d}N/\mathrm{d}S$) allows different types of selection to be distinguished throughout the viral genome. Under neutral evolution, all mutations are expected to have the same effect (i.e., none or negligible) on fitness, and, thus, ω will be around 1. Negative or purifying selection reduces dN – because non-synonymous substitutions lead to changes in the amino acid sequence and thus protein structure or function will likely be affected – whereas dS should not be affected. Therefore, the ratio ω will be lower than 1. In contrast, positive selection favors non-synonymous substitutions over synonymous substitutions, and, therefore, ω will be larger than 1 (Cobey and Koelle 2008; Jackowiak et al. 2014).

When interpreting the results of analyses based on this popular method for analyzing selection at the genome level, several caveats have to be considered. Firstly, the method was originally proposed to analyze selection acting over evolutionary large time scales, because it makes use of the rates of substitution, which implies the replacement and fixation of mutations in populations/species. This is not usually the case in viral populations, where we are dealing with constantly arising polymorphisms that, even when they are deleterious, will segregate in the population before selection removes them. This effect can be controlled for by considering only those mutations that can be mapped onto the internal branches of the phylogeny

whereas those at the external branches are excluded from the analyses. Secondly, these tests can be misleading if recombination occurs frequently (Anisimova et al. 2003), because it may alter the estimates of d*N* and d*S*, thus leading to incorrect estimates of ω.

4 How Selective Pressures Operate on Viral Genomes

Viruses are subjected to different types of selective pressures that drive their evolution and shape their genome diversity. Distinct selective pressures can increase or constrain genome variability. These selective pressures and evolutionary trade-offs drive virus evolution. They include interactions with the host's immune system as well as the need for immune escape, the pressures exerted by antiviral drug therapies, the trade-off between high viral mutation rates and genome size, the maintenance of protein structure and function, the maintenance of RNA secondary structures, and the presence of epistatic interactions between different parts of the genome. It is necessary to take into account these, sometimes opposite, forces for understanding viral genome evolution (Snoeck et al. 2011).

4.1 Interaction with the Host Immune System

Some of the most prevalent infectious diseases are caused by RNA viruses due to their high capacity for escaping their hosts' immune system by rapid antigenic evolution (Cobey and Koelle 2008). Viruses, as well as parasites, are involved in a constant "arms race" with their hosts. The former evolve to evade the immune system of the latter, while hosts' immune systems evolve to detect, control, and efficiently eliminate pathogens [the "Red Queen hypothesis" illustrates this situation (Van Valen 1973)]. The strong selective pressures exerted by the hosts' immune systems on viruses, along with their high genomic variability, result in rapid adaptation and constant evolution in coding genome regions involved in interaction with the hosts (Snoeck et al. 2011; Duffy et al. 2008; Kubinak et al. 2012; Jackowiak et al. 2014; Alizon and Fraser 2013). Thus, mutations that allow evading the immune system usually propagate rapidly through viral populations (Zanini et al. 2015).

Genome regions or segments involved in immune escape show high evolutionary rates due to positive selective pressures exerted by the hosts (Rambaut et al. 2008). Generally, these regions encode surface or viral envelope proteins. Therefore, these proteins act as targets for viral recognition by the host's immune system. Examples include the *env* region in HIV (Cobey and Koelle 2008; Alizon and Fraser 2013), the E1 and E2 genes in HCV (Thurner et al. 2004; Campo et al. 2008), and the hemagglutinin and neuraminidase segments in influenza A virus (Rambaut et al. 2008; Cobey and Koelle 2008; Pybus and Rambaut 2009; Neverov et al. 2015).

There are three types of canonical viral targets and, consequently, "hot spots" for positive selection. These are targets of neutralizing antibodies, CD4 T-cell and CD8 T-cell epitopes (i.e., regions of viral antigen recognized by molecules of the host immune system) (Zanini et al. 2015; Jackowiak et al. 2014). However, their relevance has been questioned. For example, CD4 T-cell epitopes seem to be conserved (i.e., under negative selection) in HCV, whereas CD8 T-cell epitopes are under positive selection and, consequently, drive immune evasion in this virus (Cuypers et al. 2016; Patiño Galindo and González-Candelas 2017). Another example is represented by the mapping of positively selected sites in the HIV genome and by considering different likely targets of selection, such as epitopes recognized by immune system cells, secondary structure of protein and nucleic acids, and particular dinucleotides targeted by antiviral proteins such as APOBEC3G/F (Snoeck et al. 2011). Antibody and CD4 T-cell epitopes were found to be under positive selection. However, no positive selection was detected on CD8 T-cell epitopes. Although this observation may suggest an absence of host selective pressures acting on CD8 T-cell epitopes, the authors suggest other explanations. On the one hand, positively selected escape variants without deleterious effects will fix rapidly in the viral population, thus becoming relatively conserved. On the other hand, T-cell epitopes could be under opposite selective pressures over chronic infection.

HCV is a good example of changing host selective pressures through chronic infection at the intra-host level. HCV populations progress through different stages. Firstly, right after infection, the viral population establishes under relaxed selective pressures, before triggering the immune response. As the viral population size increases, the immune response activates. Consequently, population diversity also increases, whereas escape variants appear and become fixed under positive selection. In the last stage, purifying selection predominates. This suggests that the virus has adapted steadily to its host (Jackowiak et al. 2014).

4.2 Antiviral Drug Therapies

The evolution of pathogenic microorganisms – including viruses – and the emergence of drug resistances are major concerns for public health. Drug resistance is usually related to treatment failure and results in increasing deaths, hospitalizations, and treatment duration as well as huge economic costs (Wilson et al. 2016; McGowan 2001; WHO Scientific Working Group 1983).

Some features of RNA viruses, as with immune escape, allow them to adapt rapidly in response to the strong selective pressures exerted by antiviral treatments. These features (including high mutation rates, large population sizes, and recombination or reassortment) facilitate the emergence of de novo resistance mutations. In the absence of drug-selective pressures, resistance mutations may be deleterious or, occasionally, neutral, which implies that their evolution will be governed mainly by genetic drift. For this reason, in the absence of treatment, drug-resistant variants are usually found as minority variants that increase their relative frequency in

the population only in the presence of antiviral drugs. Hence, the possibility of transmission of resistance mutations between hosts must be taken into account in order to predict the effectiveness of a particular antiviral therapy. Next-generation sequencing is necessary to detect resistance variants at low frequencies prior to the start of treatment. The development of drug resistance may depend on the presence of various permissive mutations in the same haplotype in order to decrease the genetic barrier (Wilson et al. 2016; Pybus and Rambaut 2009; Chabria et al. 2014).

It is expected that strong and directional positive selection, which is restricted to periods of time when a patient is undergoing antiviral treatment, will increase the relative frequency of resistance alleles, whereas the genetic variability of those regions close to selected loci will decrease due to selective sweeps (Renzette et al. 2014; Murrell et al. 2012). The evolution of HIV since the introduction of early antiretroviral therapies is a good example of this process. Modern treatments – highly active antiretroviral therapy (HAART) – are more effective than single drug-based early therapies. HAART consists in a customized combination of drugs. Therefore, several resistance mutations are necessary to develop simultaneous resistance against every drug included in the treatment. In contrast, early, single drug-based therapies were prone to the rapid emergence of drug resistance (Smyth et al. 2012; Martin et al. 2008). Due to the high efficiency of treatments consisting of different drugs, resistance mutations are uncommon and emerge rarely. Thus, positive selection results in strong selective sweeps that reduce genetic diversity and slow down virus evolution (Feder et al. 2016). The opposite situation was found in influenza A virus resistance to oseltamivir. One of several resistance mutations to oseltamivir (H274Y) underwent rapid and global spread during the influenza seasons between 2007 and 2009. However, the rapid increase in H274Y frequency did not substantially alter the viral genomic diversity. It is perhaps a consequence of emergence of different mutations conferring resistance to oseltamivir (Renzette et al. 2014).

4.3 Secondary RNA Structures: Protein Structure and Function

The presence of structural elements at the nucleotide and amino acid levels is of major significance for viral genome evolution because they contribute to increasing genome stability, controlling viral replication, and avoiding genome recognition by RNAses and innate antiviral defenses (Baird et al. 2006; Watts et al. 2009). Structural elements are often highly conserved. Mutations that disrupt RNA secondary structures or protein domains may have strong deleterious effects (Thurner et al. 2004; Simmonds et al. 2004).

Coding regions are under strong purifying selection, and, therefore, they are highly conserved at the amino acid level, particularly those involved in the maintenance of protein secondary structure and function (Snoeck et al. 2011). This is true

for genes or segments that code for RNA polymerase in different viruses (Rambaut et al. 2008; Zanini et al. 2015; Rothenberger et al. 2016).

RNA secondary structures are frequent in viral genomes, particularly in those of single-stranded RNA viruses. RNA secondary structures may be relevant for replication and transmission of the virus as well as for drug resistance and host interaction (Cuypers et al. 2016; Simon-Loriere et al. 2013; Thurner et al. 2004; Simmonds et al. 2004; Sanjuán and Bordería 2011). In this case, purifying selection operates at the nucleotide sequence level. As nucleotide changes driven by positive selection might disrupt RNA secondary structures, this will result in conflict between purifying selection and positive selection acting on coding regions (Snoeck et al. 2011; Sanjuán and Bordería 2011). In other words, the maintenance of RNA secondary structures may restrict protein evolution, and, in turn, selection at the protein level may restrict the pairing of nucleotides that maintain RNA secondary structures. The disruption of RNA secondary structures produced by amino acid changes could explain the fitness decrease in drug-resistant viruses in the absence of selective pressure by antiviral therapies (Sanjuán and Bordería 2011).

The case of HIV illustrates this situation. Although HIV evolution is largely driven by positive selection, more than 60% of its amino acid sites are strongly conserved. RNA secondary structures and α-helix domains mainly determine conservation in the HIV genome (Snoeck et al. 2011).

4.4 Genome Size and Gene Overlapping

Due to their high mutation rates, RNA viruses are under selective pressures favoring small genome sizes. Because most spontaneous mutations are deleterious, high mutation rates in large genomes result in excessive mutational load that may lead to population extinction. More deleterious and even lethal mutations emerge in large genomes per replication cycle than in small genomes although mutation rates can be similar. Moreover, a trend toward small genome size may also be influenced by the rate of replication, because selection favoring rapid replication will, in turn, favor viruses with minimal genome sizes (Simon-Loriere et al. 2013; Duffy et al. 2008; Bradwell et al. 2013).

Small genome size implies two problems for viral evolution: firstly, the need for storing all the genetic information in a limited space and, secondly, the need for generating genetic novelty while maintaining a small genome size. Consequently, RNA viruses often use gene overlapping in order to compress genetic information and avoid the aforementioned problems without increasing their genome size. However, gene overlapping leads to hypersensitivity to deleterious mutations (i.e., an increase in the deleterious effects of mutations in overlapping genome regions) as they affect more than one gene. Therefore, strong purifying selection operates in these regions, resulting in a reduced evolutionary rate and adaptation in RNA viruses. Despite this, the negative effects of gene overlapping on evolutionary rate depend on the type of overlapping where internal overlapping (i.e., a single gene that

contains another gene within its nucleotide sequence) is associated with stronger negative selection (Simon-Loriere et al. 2013).

In conclusion, small genome sizes limit the generation of genetic diversity as the nucleotide sequence space is limited and the use of gene overlapping as a mechanism of genome compression leads to hypersensitivity to deleterious mutations in certain regions of the genome, thus resulting in stronger purifying selection.

4.5 Epistasis

Epistasis has been described as an evolutionary phenomenon in which the fitness of a mutation depends on its genetic background (Phillips 2008). In other words, different loci along the viral genome interact with each other and determine fitness. Consequently, the phenotypic effects of a mutation may change in the presence or absence of certain genetic elements. Therefore, epistasis can significantly influence how certain mutations navigate the adaptive landscape (Wilson et al. 2016; Cobey and Koelle 2008; Assis 2014). Epistasis can be relevant in the fitness effects of RNA secondary structures, drug resistance mutations, and recombination or reassortment events. Thus, epistasis must be taken into account in order to predict the success of mutations in a viral population.

A simple form of epistasis occurs in the secondary structures of RNA viruses. The maintenance of these structures depends on base pairing between sites located on a single-stranded RNA genome. Nucleotide pairings usually follow the classical Watson-Crick model (guanine-cytosine [G-C] and adenine-uracil [A-U]). As expected, any mutation disrupting Watson-Crick pairs will alter highly conserved RNA secondary structures. Thus, they are often deleterious, and we expect that strong purifying selection operates on Watson-Crick sites, resulting in a reduced rate of evolution. This pattern has been observed in HIV, HCV, and influenza A virus. However, G-U pairs are also stable and they can maintain RNA structures. Although G-U pairs usually show fewer effects on fitness than Watson-Crick pairs, the fitness difference is relatively small. G-U pairs can operate as intermediates between adaptive peaks (i.e., G-C and A-U pairs), thus relaxing negative selective pressures on Watson-Crick sites. Moreover, G-U can remain in the population because, after all, G-U pairs show higher fitness than unpaired nucleotides (Assis 2014).

Epistasis is also relevant for the emergence of drug resistance. The fate of new drug resistance mutations depends on their efficiency in avoiding antiviral drugs effects and on their deleterious effects, mainly on viral replication. However, a permissive mutation can interact epistatically with drug resistance mutations in order to increase their fitness and, therefore, their relative frequencies in the viral population (Wilson et al. 2016; Chabria et al. 2014). The emergence of oseltamivir resistance in influenza A virus during the influenza seasons of 2007–2009 illustrates this phenomenon (see Sect. 4.2). Highly deleterious effects were predicted for the H274Y drug resistance mutation. However, H274Y spread rapidly and globally, thanks to two permissive mutations that made the mutant fitness equal to that of the

non-mutated genotype in the absence of oseltamivir (Neverov et al. 2015; Duan et al. 2014; Kryazhimskiy et al. 2011).

Influenza A virus can also be used as an example to highlight the relevance of genetic background for genetic exchange between different strains. Most segment combinations resulting from genetic reassortment are probably deleterious due to epistatic interactions (Rambaut et al. 2008; Renzette et al. 2014; Sobel Leonard et al. 2017).

In conclusion, epistatic interactions must be taken into account in order to predict virus evolution and, specifically, the epidemiological consequences of drug resistance mutations. Complete genome sequencing can be used in this context to detect epistatic interactions between distant genome regions (Rambaut et al. 2008; Wilson et al. 2016).

5 Mutation Rate and Natural Selection

Mutation is a key factor in the generation of genetic variability. In addition, the rate of mutation is a viral character evolving under natural selection. Natural selection favors high mutation rates in viruses as they increase their adaptive capacity, particularly regarding infection, host adaptation, and immune escape. In this light, viral mutation rates might be considered a virulence factor. The presence of local RNA secondary structures in the viral genome may operate as a mechanism of modulation for genome variability. RNA secondary structures flank hypervariable regions, which are prone to low-fidelity replication because they are usually located in single-stranded segments, thus focusing higher mutation rates in genomic regions involved in immune escape (Geller et al. 2016; Cuevas et al. 2015; Duffy et al. 2008; Sanjuán and Bordería 2011).

However, variability in viral genomes has an upper limit. Mutation rates are often close to the error threshold. Beyond the error threshold, deleterious mutations emerge too frequently, resulting in population extinction (error catastrophe). Therefore, purifying selection purges variants exceeding certain mutation rates. In this context, it must be noted that viral genome hypermutation exerted by host deaminases constitutes a potential mechanism against viral infection. This is apparently the case in HIV infection (Snoeck et al. 2011; Cuevas et al. 2015; Duffy et al. 2008; Holmes 2003; Neogi et al. 2013; Noguera-Julian et al. 2016).

6 Within and Among Patient Diversification

The evolutionary dynamics of genetic diversity in RNA viruses can differ markedly between levels of biological organization, within individuals (intra-host), and at the epidemiological level (inter-hosts). This prominent feature has been analyzed in depth in some viruses that produce chronic or persistent infections, such as HIV or

hepatitis C virus (HCV). However, it is also possible to analyze the genetic changes at the intra- and inter-hosts levels in viruses that produce acute infections, such as influenza A virus. Viral evolution during chronic infection occurs simultaneously in different parts of the genome and, depending on the virus, independently in the segments. Hence, it is important to analyze genetic diversity in complete genomes, because different genome regions can be under distinct, even opposed, selective pressures (Pybus and Rambaut 2009; Holmes 2004; Luciani and Alizon 2009; Lythgoe and Fraser 2012; Sobel Leonard et al. 2016).

Viral infections usually start by a founder virus or a population of a few viral units with very similar genomes (Joseph et al. 2015; Jackowiak et al. 2014; Sobel Leonard et al. 2016). It is unlikely that there is only one genome sequence in the founder population shared by all the viruses. However, among the many variants present in the source individual, the fittest phenotypes for transmission will be more represented in the infecting population. Shortly after the infection, the process known as clonal expansion starts. This process results from the rapid replication of the virus that leads to an increasingly diverse population in which new mutations accumulate from the initial sequence. This genetically diverse population is usually known as a viral quasispecies (Eigen 1996), a set of highly diverse, evolutionarily close, nonidentical haplotypes (because they derive from the same virus or a reduced population) undergoing diversification, competition, and selection (Chabria et al. 2014; Domingo et al. 2012; Khiabanian et al. 2014). In later stages of infection, the initially homogeneous viral population will be more diverse. This indicates that, during transmission, there are several bottlenecks that reduce diversity at the inter-host level (Gray et al. 2011; Joseph et al. 2015).

Many pathogens produce chronic infections that evolve so rapidly that late variants in the infection are very different from the genetic variants in the founders (Luciani and Alizon 2009; Vrancken et al. 2015). During the early stages of chronic infection by RNA viruses, such as HIV, mutations that contribute to evade the host's immune system may appear and increase in frequency (Goonetilleke et al. 2009; Kearney et al. 2009; Liu et al. 2011). Hence, chronically infecting viral populations become adapted to their hosts and this may compromise their capacity for transmission (Wright et al. 2010; Brockman et al. 2010).

During infection, viral populations explore the adaptive landscape – the set of variants close to a given genotype that might increase the fitness of the population – around the founder virus. This is supported by the fact that the same reversions are observed in unrelated individuals. In HIV, some nucleotide substitutions produced during intra-host evolution are reversions to that global consensus sequence (Zanini et al. 2015; Li et al. 2007). This trend suggests that in chronic infections, directional natural selection is the main evolutionary force determining the diversity of the viral population. But most mutations are neutral or reduce rather than increase fitness. Nevertheless, in populations with high recombination rates, such as in HIV, adaptation to the host may be concurrent with a sustained exploration of the adaptive landscape. This trend is more evident for globally conserved genome positions, and it can also be observed in viruses producing acute infections (Zanini et al. 2015; Sobel Leonard et al. 2016; Wang et al. 2014; Gire et al. 2014). However, the number

of positions under directional positive selection in the HIV genome is limited. Most of the genome is under purifying selection or accumulates neutral mutations. The action of diversifying selection, which acts in an opposite sense to directional selection, and the emergence of neutral mutations may disguise the convergence toward a global consensus sequence in positively selected positions (Snoeck et al. 2011; Ross and Rodrigo 2002).

Selective pressures acting on a viral population can differ intra- or inter-host and can often have opposing effects, leading to a trade-off. At the intra-host level, natural selection favors fast replicating variants, those that can evade the immune response, and, if the patient is being treated, those with resistance mutations against the corresponding drugs. At the inter-host level, natural selection will favor variants that can propagate rapidly in the host population, that is, those that are more easily transmitted from one host to another (Alizon and Fraser 2013).

One of the most remarkable differences between intra- and inter-host dynamics is the faster evolutionary rate associated with intra-host differentiation compared to the inter-host rate of evolution (Alizon and Fraser 2013; Lythgoe and Fraser 2012; Khiabanian et al. 2014). Intra-host evolutionary rates can be from two to six times higher than those among hosts (Lythgoe and Fraser 2012). Viral evolutionary rates show a trend to slow down in the long term. This trend is reinforced by the bottlenecks and selective pressures operating at transmission events (Zanini et al. 2015). Due to their dependence on infecting other hosts, inter-host evolutionary rates are also dependent on the transmission rate (Gray et al. 2011).

The difference in intra- and inter-host evolutionary rates means that, in chronic infections, viral populations are not homogeneous in their capacity for transmission to another host. If this were the case, we would not observe such different values between the corresponding rates (Alizon and Fraser 2013; Lythgoe and Fraser 2012). To explain this difference, we should also take into account that the viral population needs to adapt to the immune system of a specific host after each transmission. Therefore, intra-host evolution is governed by strong, continuous selective pressures leading to fast evolutionary dynamics with high evolutionary rates. Furthermore, the heterogeneity of the viral population and the different lineages that can coinfect an individual may affect the action of the immune system and, in consequence, the viral evolutionary dynamics (Grenfell et al. 2004).

However, although the intra-host rate of evolution is generally higher throughout the genome of these viruses, the pattern of evolution and the intra- and inter-host differences vary among genomic regions (Alizon and Fraser 2013). In some viruses, different genome regions can evolve independently due to recombination, such as in HIV (Zanini et al. 2015), thus minimizing the effect of selective sweeps (see below). For instance, some genes encoding for viral proteins targeted by the immune response show a faster intra-host evolution, with high levels of positive selection as a consequence of the selective pressures by the host's immune system (Gray et al. 2011; Sobel Leonard et al. 2016).

The reasons for the differences between intra- and inter-host evolutionary rates are not fully understood. Among potential alternatives, we can mention the following: (a) preferential transmission of slow-evolving lineages, (b) reduced intra-host rate of evolution over time, (c) reversion to genotypes similar to the founder virus that are likely better adapted to infecting other hosts, and (d) changes in selective pressures over the course of infection (Gray et al. 2011; Pybus and Rambaut 2009; Lythgoe and Fraser 2012). In HCV, it has been shown that the large differences between intra- and inter-host evolutionary rates in genome regions related to evasion from the immune system can be explained by reversions of host-specific adaptations to genotypes similar to those of the founder virus. The hypothesis of a preferential transmission of slow-evolving lineages seems to be quite unlikely, at least for HCV (Gray et al. 2011). In other viruses, such as HIV, the contribution of reversions to evolution has not been studied in detail (Zanini et al. 2015). Another contributing factor is that inter-host evolution is shaped by many bottlenecks produced in every transmission event (Gray et al. 2011; Joseph et al. 2015), which act reducing the evolutionary rate. As a consequence, phylogenies including isolates serially sampled within patients usually present long external but short internal branches, the latter corresponding to evolutionary changes occurring among patients.

7 Conflict Between Selective Pressures Within and Among Hosts

Intra- and inter-host selective pressures can be in conflict because mutations favoring adaptations to exploit the host, that is, those that are favored at the intra-host level (including immune system evasion and resistance mutations), are unlikely to also increase transmissibility to other hosts. Consequently, such mutations will be neutral, or selection at the inter-host level may act against them. The viral population evolves at the intra-host level during infection, becoming adapted to each new host. However, genotypes carrying host-specific adaptations do not seem to be the most efficient in being transmitted to new hosts (Alizon and Fraser 2013). The study of this conflict, known as "short-sighted evolution," was initiated in the last decade of the past century and applied to different pathogens (Levin and Bull 1994). Although until recently this conflict had been studied at the genomic scale only in HIV, its presence in other viruses such as HCV or Marburg virus has led to question whether this is a common feature of RNA viruses (Gray et al. 2011). Would it be possible then that less fit variants, presumably purged by natural selection or belonging to minority classes, persist and be transmitted in a population?

Several mechanisms have been proposed to explain the transmission of those less fit variants (intra-host) to new hosts. For instance, HIV populations "archive" resistance variants in latent T-cells, which act as reservoirs of variants that can be transmitted later. Alternatively, mutations reverting to the founder virus, the one initially infecting the host and presumably fitter for transmission (Joseph et al. 2015;

Zanini et al. 2015; Jackowiak et al. 2014; Alizon and Fraser 2013; Chabria et al. 2014), might be transmitted preferentially to variants better adapted to the current host. These mechanisms might help to explain the persistence and transmission of resistance mutations to drugs in untreated hosts because, in an analogous way, resistance mutations usually reduce viral fitness in the host in the absence of selective pressure by drugs (Chabria et al. 2014).

When studying the virus rate of replication, we find a trade-off that represents a nice example of the conflict between selection pressures at the intra- and inter-host levels. The rate of replication of the founder virus is an important factor for the epidemiological success of the disease as well as for the natural history of the viral population in the infected individual. The rate of replication influences the interaction between the viral population and the immune system of the host, which is a key factor determining the outcome of the infection (acute or chronic). The rate of replication is a quantitative trait that also evolves throughout an infection. There are observations of groups of variants in subpopulations, both within and among hosts, with different RNA polymerase activity. Hence, diverse variants with different ranges in their rates of replication can coexist in the same individual (Luciani and Alizon 2009). High rates of growth lead to a stronger immune response against the virus. Consequently, at the inter-host level, the prevalence of slow-replicating variants is favored by natural selection, because it allows a longer time of infection in the host and, as a result, maximizes the reproductive number (R_0) of the infection. In epidemiology, R_0 is defined as the number of new infections caused by an infected individual in a susceptible population and is very closely related to the intrinsic rate of growth of a population in ecological models. However, at the intra-host level, variants with a high rate of replication are favored, because they allow a faster exploitation of the host's resources.

Therefore, it seems likely that this trade-off results in variants with intermediate rates of replication, which maximize the number of infected individuals from a single host and the exploitation of resources, being favored by natural selection (Luciani and Alizon 2009; Alizon et al. 2009).

Studying the intra- and inter-host dynamics and variation provides relevant information about the transmission and epidemiology of infectious diseases. This is highly relevant in the case of outbreaks, because groups of patients that share similar and even identical viral genotypes usually also show patterns of transmission coincident in time and suggest links that can help to determine the origin or the routes of transmission of the outbreak (Gire et al. 2014). In addition, understanding the evolution and diversity of viruses and their intra- and inter-host dynamics is relevant at the clinical level. The viral diversity and its dynamics are crucial for the design of vaccines (Cuypers et al. 2016; Gaschen et al. 2002) and for determining whether an infection leads to a chronic or acute disease or the chances of success of the antiviral therapy (Gray et al. 2011; Chabria et al. 2014).

The recent advances in sequencing technologies, more specifically in high-throughput sequencing (HTS), have led to significant improvements for the analysis of viral diversity and how it affects intra- and inter-host dynamics. The development of ultra-deep sequencing has been very important for research on chronic viral

infections, which can show high levels of intra-host diversity such as HIV and HCV. Its higher sensitivity compared to traditional Sanger sequencing allows a deeper analysis of viral diversity, identifying minority variants and rare polymorphisms that, on the one hand, are invisible for classical techniques, which usually involve reconstructing consensus sequences, and, on the other hand, can be very relevant for basic and applied research (Chabria et al. 2014; Khiabanian et al. 2014). Furthermore, the capability of HTS to sequence a large number of molecules in parallel allows obtaining large datasets, which also help in reducing the economic costs of sequencing (Hall 2007; Churko et al. 2013).

The efforts to investigate evolutionary dynamics at the genome level have focused mainly on RNA viruses causing chronic infections, for which the study of changes in genomic diversity at the intra-host level is more relevant. Among these, HIV and HCV have received most attention due to their evident clinical and epidemiological relevance for humans. In addition, there have also been studies at the genome level aimed at relating intra- and inter-host dynamics in acute disease-causing viruses such as influenza A (Sobel Leonard et al. 2016). Hence, lack of representative data for some viruses is still a major obstacle for studying their population evolution and dynamics.

8 Spatial Distribution of Viruses

The spatial distribution of rapidly evolving viruses depends on ecological and evolutionary processes that interact with each other. In RNA viruses, ecological processes, such as spatial spread, and epidemiological processes occur in a similar time scale to that of evolutionary processes, as a result of their high mutation and evolutionary rates (Holmes 2008). This makes them very appropriate model organisms to study the dynamics of microevolutionary changes, because these can be observed "in real time." In addition, there is a bias toward studying RNA viruses rather than those with a DNA genome that derives not only from their fast evolution (Duffy et al. 2008) but because, in general, they are more relevant in epidemics and emerging diseases (Holmes 2004; WHO 2017).

Avise (2000) defined phylogeography as the field of study concerned with the principles and processes governing the distribution of geographical lineages at the intraspecific level as well as the interspecific level for related species. In other words, from a more applied perspective, phylogeography includes studies using phylogenetic trees to combine genetic data with spatial information and analyze the spatial patterns suggested in these trees (Holmes 2004; Pybus et al. 2015). Holmes (2004) used a wider definition in which phylogeography incorporates spatial and temporal patterns as well as their interactions. The rapid evolution of viruses can generate enough genetic variation, even at the intra-host level, in just a few days to perform phylogenetic analyses at the infected individual level. This allows applying phylogenetic methods to emerging diseases and to build highly resolved phylogenetic trees (Holmes 2004; Avise 2000; Pybus et al. 2015). The most basic way to

integrate spatial and genetic information consists of localizing cases of infections and associating them to different variants (subtypes, genotypes, etc.) of the disease-causing virus (Pybus et al. 2015).

Phylogeographic methods are a powerful tool to infer migration and transmission routes and to reconstruct the evolutionary history of a lineage from genetic data. When applied to viruses, these methods are useful to track the origin of outbreaks and the source of emerging diseases and to reconstruct transmission histories not only between individual hosts but also among social groups of the hosts, among host species, and even their dispersion within body compartments within an individual (De Maio et al. 2015; Alcala et al. 2016).

Due to the coincidence of time scales between molecular evolution and ecological processes that shape their diversity, virus phylogenies provide not only spatial information (i.e., lineages that cluster in geographically defined clades) but also temporal information (i.e., lineages ordered according to sampling times). The molecular clock is a statistical model that establishes a relationship between time and genetic distances in nucleotide sequences. If samples are identified with known dates, then the branching events and the common ancestor in a phylogeny can be placed in a temporal scale. This information can be integrated with spatial information to reconstruct the dispersal history of a virus, linking each branch of the phylogeny with its geographic location. Therefore, with models based on the molecular clock, it is possible to analyze the spread of an epidemic (in months or years) complementing the phylogeny of the isolates with a time scale (Pybus and Rambaut 2009; Pybus et al. 2015). The simplest models for the molecular clock, also known as "strict clock" models, assume a single, constant evolutionary rate for all the lineages. However, more complex, "relaxed clock" models have incorporated variation in the evolutionary rate among lineages or through time (Drummond et al. 2006).

However, the application of phylogeographic tools is valuable only if the spatial epidemiology leaves a signal in the viral genome. This depends both on the rate of molecular evolution and on the rate of transmission in space. If the genome accrues diversity too quickly compared to the rate of spatial spread, then the information provided by phylogeographic analyses is lost as a result of mutation saturation at informative positions (Emmett et al. 2015; Pybus et al. 2015).

Using specific genes or regions to build phylogenetic trees is still a current and complementary approach to analyzing complete genomes (Shen et al. 2016), especially when these genome regions are important sources of predictive information because they encode antigenic proteins (McHardy and Adams 2009). However, the analysis at the genome level is very important to obtain a more complete and unbiased information. Mechanisms such as recombination and reassortment may generate genomes in which different portions thereof have different evolutionary histories (Rambaut et al. 2008; McHardy and Adams 2009; Holmes 2004; Pybus et al. 2015), and this has to be considered when analyzing complete genomes. Next-generation sequencing methods have advanced to the "subnucleotide" level in the analysis of viral sequences. This implies considering the infected individuals as viral populations rather than repetitive collections of the same consensus genome

and, additionally, detecting variability within individuals, even very low-frequency variants (subclonal variants). Studying the intra-host and subclonal variability can improve the resolution of phylogenetic analyses and, when combined with epidemiological information, provide a very valuable information to track transmission chains during an outbreak, especially when the transmission rate is very fast, even higher than the viral evolutionary rate (Emmett et al. 2015).

9 Transmission Dynamics

In order to study and understand the dynamics of viral epidemics, we need an approach combining the methods and theories of evolutionary biology, epidemiology, and human geography.

For obligate parasites, such as viruses, which are usually unable to survive for a long time outside their hosts, the mobility and movement patterns of the host are crucial for understanding their transmission dynamics (Alcala et al. 2016; Hufnagel et al. 2004; Pybus et al. 2015). This is closely related to the density and communication between susceptible populations because for virus transmission, a certain proximity between hosts or hosts and vectors is necessary. The smaller the population size of the host, the less likely transmission will be and, consequently, the more difficult to be sustained long enough to cause acute infections. However, large, dense host populations can easily sustain a virus that causes short, virulent infections. In this context, the analysis of the basic reproductive number (R_0) is highly relevant. This number depends on several factors, such as the number of contacts with susceptible individuals, the probability of transmission, and the length of the infectious period (Dietz 1993). This value is very useful to estimate the speed of propagation of an infection in a susceptible population (Ridenhour et al. 2014). The interest in estimating this parameter and its application to the analysis of outbreaks and epidemics and the design of public health strategies gained momentum during the influenza A pandemics of 2009 (Fraser et al. 2009; Ridenhour et al. 2014).

Therefore, the spatial distribution of human viruses will reflect, at least partially, the spatial distribution of human populations, which will also influence the virulence of the disease. However, we must also consider whether the virus can infect other animal species or whether they represent a reservoir for human infections (zoonoses). This is the case for some viruses, such as Ebola virus, with reservoirs in animal species but also capable of being transmitted from person to person. Furthermore, even in RNA viruses well-adapted to humans, there is the possibility of relatively frequent zoonotic contacts, such as in influenza A and MERS-CoV, which are usually associated with the emergence of epidemics and pandemics as a result of genetic exchanges between strains from different species. For vector-borne viruses, we must consider not only human geography but also the geographic distribution of the corresponding vectors, such as different mosquitos of the genera *Aedes* and *Culex*, which are vectors for Zika, dengue, or chikungunya viruses. Spatial distribution analyses should also include ecological features, life history, or migration

potential of the vectors (Holmes 2004; Faria et al. 2017; Shen et al. 2016; Bullivant and Martinou 2017; Cunha and Opal 2014).

In the study of the mobility and geographic distribution of humans for understanding the distribution and spread of human viruses, it is necessary to take into account social factors such as international trade and air traffic. The global communications and interrelationships of human populations are growing continuously and represent new opportunities for the transmission, propagation, and colonization of new regions by viruses and their vectors. These can move viruses across geographic barriers and bring into contact with previously isolated populations. This process has contributed to the emergence and reemergence of viral epidemics such as Zika, dengue, and chikungunya. However, we are just starting to understand the effects of global mobility of people and goods on the genetic diversity and evolution of viruses (Alcala et al. 2016; Pybus et al. 2015). To better control epidemics and to understand the evolution and ecology of viruses, it will be necessary to integrate spatial and genomic information along with information about human mobility in a single mathematical framework (Pybus et al. 2015). One example in this direction is BEAST, a framework for Bayesian statistical analysis that allows inference of phylogeographic relationships including spatial and temporal dynamics of migration (Lemey et al. 2009; Drummond and Rambaut 2007).

Clear examples of the relevance of this approach are the analyses of emerging viral epidemics such as SARS or Zika virus. The international spread of Zika virus is likely due to a global increase in air traffic. Specifically, using phylogeographic methods, the origin of the epidemics has been traced to Brazil, where it was detected in 2015, dating its origin in this country between 2013 and 2014 (Worobey 2017). These dates were coincident with several events that brought an important flow of international air traffic to Brazil, such as the 2014 FIFA World Cup (June–July 2014) and the 2013 FIFA Confederations Cup (June 2013) of football (Faria et al. 2016). This highlights the importance of integrating genomic and epidemiologic information about the global movement of persons when surveillance systems are implemented. The large-scale patterns of people's movements can suggest useful hypotheses to study the introduction of viruses and the emergence of epidemics (Faria et al. 2016, 2017; Shen et al. 2016).

From a population genomics perspective, how does this increase in international trade and movements impact the spread of infectious diseases, the population dynamics of viruses, and their genetic diversity?

Isolation and subsequent secondary contact of viral populations are common in natural host populations and can occur at short time scales. These events, facilitated by a higher mobility and contact among human populations, are usually associated to epidemics and pandemics. This has been observed in viruses such as influenza A virus, HIV, and human cytomegalovirus. Furthermore, these processes are important for understanding the evolutionary trajectory of zoonotic viruses, such as Ebola virus.

While they are isolated, viral populations from the same species diverge and adapt to the specific features of their host populations. Hence, during this period, natural selection and demographic changes, such as expansions and bottlenecks, acting on

either the viral or the host population will affect the evolution of the virus. After the viral populations are connected again, gene flow, recombination, or reassortment will influence the evolution of the virus, leading to a "mixture" at the genome level. Although selection and demographic changes still act during the reconnection, the other processes act more intensely and rapidly. This mixture impacts on diversity at the genome level: isolated populations have evolved independently, diverging and adapting to the specific conditions of their host populations. After reconnection, the diversity that has accumulated separately increases, which also leads to higher adaptive potential since recombination and reassortment allow the combination of polymorphisms selected in different environments into the same genome. If these polymorphisms are compatible in that particular genomic context, this opens the opportunity for the development of new features which might have not developed (or do so only after very long periods) by just mutation and selection. The second consequence of this shared genetic diversity is a progressive trend toward the homogenization of the populations. Due to the increase in human mobility, these events are expected to be more frequent in the future (Alcala et al. 2016).

10 Epidemiological Surveillance and Genomic Surveillance

Phylogenetic and phylogeographic analyses complement each other, and both are used in epidemiological surveillance systems to control infectious diseases. Phylogeographic information can be used to confirm the source(s) of epidemic outbreaks, and it can also provide valuable information when surveillance is not well implemented or the data it generates are uncertain, unavailable, or insufficient to reconstruct or predict the propagation of the virus (Faria et al. 2017; Pybus et al. 2015). It is even possible to talk about "genomic surveillance" (Emmett et al. 2015) in which the sequencing and analysis of complete genomes contribute to tracking evolution at the genome level as the disease spreads. On the other hand, phylogenetic analysis combined with epidemiological information is useful to study the routes of infection in human populations or the number of introductions that have caused an epidemic (Blackley et al. 2016; Gire et al. 2014; Shen et al. 2016; Drummond et al. 2006; Emmett et al. 2015; Faria et al. 2016).

Another goal of virus phylogeography is to ascertain the future propagation of the organisms and the potential for epidemics by asking which variants are more likely to become predominant and which places are more likely to be colonized and through which ways. This implies building a predictive framework integrating social and environmental factors associated to virus movement and transmission along with genomic and epidemiological information (McHardy and Adams 2009; Pybus et al. 2015).

Influenza A is a good example of how a well-established, global epidemiological surveillance system provides useful information for disease control and vaccine design. It also facilitates the collection of genome sequences at temporal and spatial scales that can be used in evolutionary and phylogeographic analyses. Conversely, at

the beginning of the Zika virus epidemics in Brazil in 2015, the country lacked a surveillance system for this virus, and, 1 year later, this task still rested on the passive diagnostics of the disease. This problem, along with the added difficulties for the diagnosis of Zika due to its coexistence with dengue and chikungunya virus, has been a major hurdle in the epidemiological study of the disease and the gathering of abundant genomic information. The example of Zika reinforces the relevance of epidemiological surveillance for the phylogeographic analysis of the virus (Faria et al. 2016; Worobey 2017; Metsky et al. 2017).

The phylogenetic analysis of a virus can help in evaluating the efficiency of surveillance systems. Estimating the most recent common ancestor of a group of sequences can inform about the delay in the detection and notification of the pathogen with respect to the moment of its introduction in the population (Pybus et al. 2015).

One of the limitations in the phylogeographic analysis of viruses is the choice of the correct model. A wrong model selection can lead to erroneous inferences about the transmission history of the pathogen. As epidemiological investigations rely increasingly on genome sequencing to study the origin and spread of infections, the use of accurate phylogeographic methods will be crucial to stop their propagation and design public health preventive measures. De Maio et al. (2015) review different models used to infer transmission rates and spread patterns for viruses, and they illustrate a trade-off between computational costs and speed, on the one hand, and the reliability of the conclusions, on the other hand. The more reliable approaches (continuous models) are, in general, the slowest and most costly with regard to computational resources.

Recently, and partly to fulfill the need for a fast response in cases of outbreaks and emerging epidemics, the so-called discrete character models have gained popularity (Gire et al. 2014). These models treat locations as if they were discrete traits evolving as alleles in a locus. This approach allows a much faster analysis; however, its results are not reliable. They are very sensitive to sampling bias and not robust to scarce genetic data. Different models can yield very different results for the same dataset and, in general, lead to very different and wrong biological interpretations when applied to the study of virus transmission (rather than to the evolution of discrete traits, their original target). De Maio et al. (2015) suggested a model for phylogeographic analysis that combines the advantages of both approaches, discrete and continuous (reliability and precision along with speed and computational efficiency). This model has been used recently in the study of emerging epidemics, such as Zika in Brazil (Faria et al. 2017).

Another limitation for phylogeographic analyses is the public availability of sequences. This depends, in part, on the relevance of the disease caused by the virus, the implementation of an efficient epidemiological surveillance, and the stage of the epidemics. For instance, in 2016 the number of genome sequences for Zika virus available in GenBank was very limited (Shen et al. 2016) as a result of being a recent epidemic and inefficient surveillance. On the contrary, the availability in the public domain of influenza A virus sequences is much higher (Pybus et al. 2015). The rapid publication of genome sequences during emerging epidemics is important

to improve genomic and epidemiological surveillance and to monitor the spread of the disease and the adaptive processes in the virus (Gire et al. 2014).

11 Conclusions

Viruses, especially those with RNA genomes, have high mutation rates, short generation times, and large population sizes and are under strong selective pressures. These factors make these viruses organisms with fast evolutionary rates, high genetic variability, and great adaptive capacity.

Understanding the mechanisms that allow human viruses to generate and maintain genetic diversity and to adapt to the host's selective pressure is fundamental for human health. A better knowledge of the evolution of human viruses at the genome level can shed light on questions such as the evasion of immune response, the development and transmission of resistance mutations, vaccine design, the evolution and virulence of the disease or the control of outbreaks, epidemics, and emerging diseases.

In general, viruses, as any other pathogen, are under strong selective pressure by the immune system of their hosts. In addition, human viruses are usually under the additional pressure of antiviral drugs and treatments. These pressures result in high mutation rates in those genome regions involved in the interaction with the host and in those that encode the targets of antiviral drugs. This leads to the development of drug resistance and of mechanisms to evade the immune system.

The typically high mutation rates of RNA viruses are, most likely, another consequence of these selective pressures because in a stable environment (very different from a host infected by the virus), natural selection will favor a low mutation rate (Kamp et al. 2002). This common feature of RNA viruses is a key factor to explain their adaptation, and, simultaneously, it keeps viral populations at the extinction threshold by accumulating an excessive number of deleterious mutations. Recently, it has been observed that the human immune system might take advantage of this feature to fight viral infections by forcing hypermutation in the viral genome.

The genomic diversity is also limited by different constraints: the need to keep a small genome size, RNA secondary structures at the genome level, structural domains of proteins to sustain their function, and gene overlapping. In some viruses, such as HCV, these negative selection pressures might be the main factor driving evolution. In others, such as HIV, positive selection has a more relevant role.

In this context, it is important to consider how the interactions between genome positions can affect the "displacement" of different mutations through the adaptive landscape. Mutations that could be considered as deleterious, such as some resistance mutations or those that disrupt secondary structures, can be retained in a population and even spread rapidly depending on the genome context where they appear.

The population dynamics of RNA viruses are different depending on the level of biological organization at which they are analyzed. Selective pressures acting at the

intra-host and inter-hosts levels can differ and often act in opposite directions. Frequently, these selective pressures conflict between the need to adapt to the host and the ability for transmission to other hosts. Those variants that are favored by selection within hosts – mutations for evading the immune system and drug resistance – may diminish the capacity for transmission of the virus and, in consequence, will be selected against at the inter-host level. In addition, every transmission event represents a bottleneck that reduces drastically the population size of the virus and, consequently, also its genetic diversity. This leads to slower evolutionary rates at the inter-host level. For instance, in HIV there seems to be an inverse relationship between transmission and evolutionary rate (Berry et al. 2007). We must also consider how and by which means is the virus transmitted. Transmission rates are higher in air-transmitted virus, such as influenza A, than in those that use the sexual route. Similarly, those viruses that use arthropod species as vectors have lower rates of evolution, a cost associated to their need for replication in different hosts (Holmes 2004; Woelk and Holmes 2002).

Another consequence of the high rates of evolution is that ecological and evolutionary processes acting on viral populations occur at similar time scales. Their interaction affects their spatial distribution. The combination of complete viral genomes and phylogeographic methods is very useful for tracking the origin of epidemic outbreaks, locating reservoirs that may act as sources of infection for humans or of new potentially virulent strains (such as influenza A), to reconstruct transmission histories and to monitor the spread of an epidemics. These applications are very relevant nowadays, in an increasingly connected planet in which trade and air traffic bring geographically distant populations close and erase natural barriers for the transmission of diseases. Furthermore, human impacts on previously intact ecosystems are helping the emergence and global spread of new infectious, as illustrated by the recent epidemics of Zika and Ebola viruses.

12 Future Perspectives

The development of population genomics is closely linked to advances in sequencing technologies. Standard techniques, based on deriving consensus sequences, miss the presence of minor or subclonal variants (low-frequency polymorphisms) which might be important to understand the dynamics of viral populations as well as the evolution and spread of the disease. Next-generation sequencing techniques allow the detection of rare polymorphisms and minor variants and lead to consideration of infected hosts as viral populations rather than “collections” of the same consensus genome. Consequently, these methods provide a better view of viral diversity, which enables an improvement in the study of the epidemiology and evolution of human viruses. A more widespread use of these technologies to characterize genome variation will provide increased information about the intra-host dynamics and the relationship between viral diversity and infection outcome (Liu et al. 2012; Farci et al. 2000), the inter-host transmission and dynamics (reservoirs for better-

transmitted variants), the development of resistance and the failure of antiviral treatments, and the building of highly resolved phylogenies and transmission histories during epidemic outbreaks. In addition, advances in sequencing technologies have also allowed the fast and in-depth analysis of complete genomes. The evolution and accumulation of genetic variation occur differently and simultaneously throughout the genome. Separate regions of the same genome can interact with each other (epistasis) and, even, evolve independently and show different phylogenetic histories. Hence, the possibility of analyzing complete genomes – as opposed to the analysis of individual loci or isolated genome regions – provides a more complete, resolved, and less biased view of genomic variation, the phylogeny and population dynamics of the virus.

Finally, an important limitation in the population genomic study of virus populations is the availability of genomic information for many viruses. This is intimately related to the clinical and epidemiological relevance of the disease caused by most viruses. Human diseases with high prevalence and important consequences such as HIV, hepatitis C, or influenza receive much attention in the public health realm and have a more efficient surveillance. This translates in higher availability of viral genomes and epidemiological information, which are necessary for the evolutionary analysis of virus populations.

The evolutionary analysis of viral genomes and epidemiological surveillance are, in consequence, necessarily complementary. Implementing a "genomic surveillance" can contribute to control and monitor the spread of infectious diseases and to design better public health strategies to achieve these goals.

Acknowledgments This work was supported by projects BFU2014-58656R and BFU2017-89594R from MINECO (Spanish Government) and PROMETEO2016-0122 from Generalitat Valenciana.

References

Alcala N, Jensen JD, Telenti A, Vuilleumier S. The genomic signature of population reconnection following isolation: from theory to HIV. G3 (Bethesda). 2016;6(1):107–20.

Alizon S, Fraser C. Within-host and between-host evolutionary rates across the HIV-1 genome. Retrovirology. 2013;10(1):49.

Alizon S, Hurford A, Mideo N, van Baalen M. Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future. J Evol Biol. 2009;22(2):245–59.

Anisimova M, Nielsen R, Yang Z. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics. 2003;164(3):1229–36.

Assis R. Strong epistatic selection on the RNA secondary structure of HIV. PLoS Pathog. 2014;10 (9):e1004363.

Avise JC. Phylogeography. The history and formation of species. 1st ed. Cambridge: Harvard University Press; 2000.

Baird HA, Galetto R, Gao Y, Simon-Loriere E, Abreha M, Archer J, et al. Sequence determinants of breakpoint location during HIV-1 intersubtype recombination. Nucleic Acids Res. 2006;34 (18):5203–16.

Berry IM, Ribeiro R, Kothari M, Athreya G, Daniels M, Lee HY, et al. Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases. J Virol. 2007;81(19):10625–35.

Blackley DJ, Wiley MR, Ladner JT, Fallah M, Lo T, Gilbert ML, et al. Reduced evolutionary rate in reemerged Ebola virus transmission chains. Sci Adv. 2016;2(4):e1600378.

Bradwell K, Combe M, Domingo-Calap P, Sanjuán R. Correlation between mutation rate and genome size in riboviruses: mutation rate of bacteriophage Qb. Genetics. 2013;195(1):243–51.

Brandes N, Linial M. Gene overlapping and size constraints in the viral world. Biol Direct. 2016;11 (1):1–15.

Brockman MA, Brumme ZL, Brumme CJ, Miura T, Sela J, Rosato PC, et al. Early selection in Gag by protective HLA alleles contributes to reduced HIV-1 replication capacity that may be largely compensated for in chronic infection. J Virol. 2010;84(22):11937–49.

Bullivant G, Martinou AF. Ascension Island: a survey to assess the presence of Zika virus vectors. J R Army Med Corps. 2017;163(5):347–54.

Campo DS, Dimitrova Z, Mitchell RJ, Lara J, Khudyakov Y. Coordinated evolution of the hepatitis C virus. Proc Natl Acad Sci U S A. 2008;105(28):9685–90.

Chabria SB, Gupta S, Kozal MJ. Deep sequencing of HIV: clinical and research applications. Annu Rev Genomics Hum Genet. 2014;15(1):295–325.

Churko JM, Mantalas GL, Snyder MP, Wu JC. Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. Circ Res. 2013;112(12):1613–23.

Cobey S, Koelle K. Capturing escape in infectious disease dynamics. Trends Ecol Evol. 2008;23 (10):572–7.

Cuevas JM, Geller R, Garijo R, Lòpez-Aldeguer J, Sanjuán R. Extremely high mutation rate of HIV-1 *in vivo*. PLoS Biol. 2015;13(9):e1002251.

Cunha CB, Opal SM. Middle East respiratory syndrome (MERS): a new zoonotic viral pneumonia. Virulence. 2014;5(6):650–4.

Cuypers L, Li G, Neumann-Haefelin C, Piampongsant S, Libin P, Van Laethem K, et al. Mapping the genomic diversity of HCV subtypes 1a and 1b: implications of structural and immunological constraints for vaccine and drug development. Virus Evol. 2016;2(2):vew024.

De Maio N, Wu CH, O'Reilly KM, Wilson D. New routes to phylogeography: a Bayesian structured coalescent approximation. PLoS Genet. 2015;11(8):e1005421.

Dietz K. The estimation of the basic reproduction number for infectious diseases. Stat Methods Med Res. 1993;2(1):23–41.

Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. Microbiol Mol Biol Rev. 2012;76 (2):159–216.

Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007;7:214.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006;4(5):e88.

Duan S, Govorkova EA, Bahl J, Zaraket H, Baranovich T, Seiler P, et al. Epistatic interactions between neuraminidase mutations facilitated the emergence of the oseltamivir-resistant H1N1 influenza viruses. Nat Commun. 2014;5:5029.

Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9(4):267–76.

Eigen M. On the nature of virus quasispecies. Trends Microbiol. 1996;4(6):216–8.

Emmett KJ, Lee A, Khiabanian H, Rabadan R. High-resolution genomic surveillance of 2014 Ebolavirus using shared subclonal variants. PLoS Curr. Outbreaks 2015;7.

Farci P, Shimoda A, Coiana A, Diaz G, Peddis G, Melpolder JC, et al. The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. Science. 2000;288:339–44.

Faria NR, Sabino EC, Nunes MRT, Alcantara LCJ, Loman NJ, Pybus OG. Mobile real-time surveillance of Zika virus in Brazil. Genome Med. 2016;8(1):97.

Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017;546:406–10.
Feder AF, Rhee SY, Holmes SP, Shafer RW, Petrov DA, Pennings PS. More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. Elife. 2016;5:e10670.
Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, et al. Pandemic potential of a strain of influenza A (H1N1): early findings. Science. 2009;324 (5934):1557–61.
Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, et al. Diversity considerations in HIV-1 vaccine selection. Science. 2002;296(5577):2354–60.
Geller R, Domingo-Calap P, Cuevas JM, Rossolillo P, Negroni M, Sanjuan R. The external domains of the HIV-1 envelope are a mutational cold spot. Nat Commun. 2015;6:8571.
Geller R, Estada Ú, Peris JB, Andreu I, Bou JV, Garijo R, et al. Highly heterogeneous mutation rates in the hepatitis C virus genome. Nat Microbiol. 2016;1(7):16045.
Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345 (6202):1369–72.
Goonetilleke N, Liu MKP, Salazar-Gonzalez JF, Ferrari G, Giorgi E, Ganusov VV, et al. The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. J Exp Med. 2009;206(6):1253–72.
Gray R, Parker J, Lemey P, Salemi M, Katzourakis A, Pybus O. The mode and tempo of hepatitis C virus evolution within and among hosts. BMC Evol Biol. 2011;11(1):131.
Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. Science. 2004;303(5656):327–32.
Hall N. Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol. 2007;210(9):1518–25.
Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 2005;169(4):2335–52.
Hickerson MJ, Carstens BC, Cavender-Bares J, Crandall KA, Graham CH, Johnson JB, et al. Phylogeography's past, present, and future: 10 years after. Mol Phylogenet Evol. 2010;54 (1):291–301.
Holmes EC. Error thresholds and the constraints to RNA virus evolution. Trends Microbiol. 2003;11(12):543–6.
Holmes EC. The phylogeography of human viruses. Mol Ecol. 2004;13(4):745–56.
Holmes EC. Evolutionary history and phylogeography of human viruses. Annu Rev Microbiol. 2008;62(1):307–28.
Hufnagel L, Brockmann D, Geisel T. Forecast and control of epidemics in a globalized world. Proc Natl Acad Sci U S A. 2004;101(42):15124–9.
Jackowiak P, Kuls K, Budzko L, Mania A, Figlerowicz M, Figlerowicz M. Phylogeny and molecular evolution of the hepatitis C virus. Infect Genet Evol. 2014;21(1):67–82.
Joseph SB, Swanstrom R, Kashuba AD, Cohen MS. Bottlenecks in HIV-1 transmission: insights from the study of founder viruses. Nat Rev Microbiol. 2015;13(7):414–25.
Kamp C, Wilke CO, Adami C, Bornholdt S. Viral evolution under the pressure of an adaptive immune system: optimal mutation rates for viral escape. Complexity. 2002;8(2):28–33.
Kearney M, Maldarelli F, Shao W, Margolick JB, Daar ES, Mellors JW, et al. Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. J Virol. 2009;83(6):2715–27.
Khiabanian H, Carpenter Z, Kugelman J, Chan J, Trifonov V, Nagle E, et al. Viral diversity and clonal evolution from unphased genomic data. BMC Genomics. 2014;15(6):S17.
Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB. Prevalence of epistasis in the evolution of influenza A surface proteins. PLoS Genet. 2011;7(2):e1001301.
Kubinak JL, Ruff JS, Hyzer CW, Slev PR, Potts WK. Experimental viral evolution to specific host MHC genotypes reveals fitness and virulence trade-offs in alternative MHC types. Proc Natl Acad Sci U S A. 2012;109(9):3422–7.

Lai MM. RNA recombination in animal and plant viruses. Microbiol Rev. 1992;51(1):61–79.
Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009;5(9):e1000520.
Levin BR, Bull JJ. Short-sighted evolution and the virulence of pathogenic microorganisms. Trends Microbiol. 1994;2(3):76–81.
Li B, Gladden AD, Altfeld M, Kaldor JM, Cooper DA, Kelleher AD, et al. Rapid reversion of sequence polymorphisms dominates early human immunodeficiency virus type 1 evolution. J Virol. 2007;81(1):193–201.
Liu Y, McNevin JP, Holte S, McElrath MJ, Mullins JI. Dynamics of viral evolution and CTL responses in HIV-1 infection. PLoS One. 2011;6(1):e15639.
Liu L, Fisher BE, Thomas DL, Cox AL, Ray SC. Spontaneous clearance of primary acute hepatitis C virus infection correlated with high initial viral RNA level and rapid HVR1 evolution. Hepatology. 2012;55(6):1684–91.
Luciani F, Alizon S. The evolutionary dynamics of a rapidly mutating virus within and between hosts: the case of hepatitis C virus. PLoS Comput Biol. 2009;5(11):e1000565.
Lythgoe KA, Fraser C. New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. Proc R Soc B. 2012;279(1741):3367–75.
Martin M, Del Cacho E, Codina C, Tuset M, De Lazzari E, Mallolas J, et al. Relationship between adherence level, type of the antiretroviral regimen, and plasma HIV type 1 RNA viral load: a prospective cohort study. AIDS Res Human Retrovir. 2008;24(10):1263–8.
Maynard-Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974;23 (1):23–35.
McDonald SM, Nelson MI, Turner PE, Patton JT. Reassortment in segmented RNA viruses: mechanisms and outcomes. Nat Rev Microbiol. 2016;14(7):448–60.
McGowan JE Jr. Economic impact of antimicrobial resistance. Emerg Infect Dis. 2001;7(2):286.
McHardy AC, Adams B. The role of genomics in tracking the evolution of influenza A virus. PLoS Pathog. 2009;5(10):e1000566.
Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol. 2013;28(11):659–69.
Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, et al. Zika virus evolution and spread in the Americas. Nature. 2017;546:411–5.
Miralles R, Gerrish PJ, Moya A, Elena SF. Clonal interference and the evolution of RNA viruses. Science. 1999;285:1745–7.
Muller HJ. Some genetic aspects of sex. Am Nat. 1932;66:118–38.
Murrell B, De Oliveira T, Seebregts C, Kosakovsky Pond SL, Scheffler K, Southern African Treatment and Resistance Network (SATuRN) Consortium. Modeling HIV-1 drug resistance as episodic directional selection. PLoS Comput Biol. 2012;8(5):e1002507.
Negroni M, Buc H. Copy-choice recombination by reverse transcriptases: reshuffling of genetic markers mediated by RNA chaperones. Proc Natl Acad Sci U S A. 2000;97(12):6385–90.
Neogi U, Shet A, Sahoo PN, Bontell I, Ekstrand ML, Banerjea AC, Sonnerborg A. Human APOBEC3G-mediated hypermutation is associated with antiretroviral therapy failure in HIV-1 subtype C-infected individuals. J Int AIDS Soc. 2013;16(1):18472.
Neverov AD, Kryazhimskiy S, Plotkin JB, Bazykin GA. Coordinated evolution of influenza A surface proteins. PLoS Genet. 2015;11(8):e1005404.
Noguera-Julian M, Cozzi-Lepri A, Di Giallonardo F, Schuurman R, Däumer M, Aitken S, et al. Contribution of APOBEC3G/F activity to the development of low-abundance drug-resistant human immunodeficiency virus type 1 variants. Clin Microbiol Infect. 2016;22(2):191–200.
Nomikou K, Hughes J, Wash R, Kellam P, Breard E, Zientara S, et al. Widespread reassortment shapes the evolution and epidemiology of bluetongue virus following European invasion. PLoS Pathog. 2015;11(8):e1005056.
Patiño Galindo JA, González-Candelas F. Comparative analysis of variation and selection in the HCV genome. Infect Genet Evol. 2017;49:104–10.

Pennings PS, Kryazhimskiy S, Wakeley J. Loss and recovery of genetic diversity in adapting populations of HIV. PLoS Genet. 2014;10(1):e1004000.
Phillips PC. Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet. 2008;9(11):855–67.
Pybus OG, Rambaut A. Modelling: evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet. 2009;10:540–50.
Pybus OG, Tatem AJ, Lemey P. Virus evolution and transmission in an ever more connected world. Proc Biol Sci. 2015;282(1821):20142878.
Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. The genomic and epidemiological dynamics of human influenza A virus. Nature. 2008;453(7195):615.
Ramirez BC, Simon-Loriere E, Galetto R, Negroni M. Implications of recombination for HIV diversity. Virus Res. 2008;134(1):64–73.
Renzette N, Caffrey DR, Zeldovich KB, Liu P, Gallagher GR, Aiello D, et al. Evolution of the influenza A virus genome during development of oseltamivir resistance in vitro. J Virol. 2014;88(1):272–81.
Ridenhour B, Kowalik JM, Shay DK. Unraveling R0: considerations for public health applications. Am J Public Health. 2014;104(2):e32–41.
Rogozin I, Spiridonov A, Sorokin A, Wolf Y, Jordan I, Tatusov R, et al. Purifying and directional selection in overlapping prokaryotic genes. Trends Genet. 2002;18(5):228–32.
Ross HA, Rodrigo AG. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. J Virol. 2002;76(22):11715–20.
Rothenberger S, Torriani G, Johansson MU, Kunz S, Engler O. Conserved endonuclease function of hantavirus L polymerase. Viruses. 2016;8(5):108.
Sanjuán R, Bordería AV. Interplay between RNA structure and protein evolution in HIV-1. Mol Biol Evol. 2011;28(4):1333–8.
Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. J Virol. 2010;84 (19):9733–48.
Shen S, Shi J, Wang J, Tang S, Wang H, Hu Z, Deng F. Phylogenetic analysis revealed the central roles of two African countries in the evolution and worldwide spread of Zika virus. Virol Sin. 2016;31(2):118–30.
Simmonds P, Tuplin A, Evans DJ. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: implications for virus evolution and host persistence. RNA. 2004;10(9):1337–51.
Simon-Loriere E, Holmes EC. Why do RNA viruses recombine? Nat Rev Microbiol. 2011;9 (8):617–26.
Simon-Loriere E, Holmes EC, Pagán I. The effect of gene overlapping on the rate of RNA virus evolution. Mol Biol Evol. 2013;30(8):1916–28.
Smyth RP, Davenport MP, Mak J. The origin of genetic diversity in HIV-1. Virus Res. 2012;169 (2):415–29.
Snoeck J, Fellay J, Bartha I, Douek D, Telenti A. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. Retrovirology. 2011;8(1):87.
Sobel Leonard A, McClain MT, Smith GJD, Wentworth DE, Halpin RA, Lin X, et al. Deep sequencing of influenza A virus from a human challenge study reveals a selective bottleneck and only limited intrahost genetic diversification. J Virol. 2016;90(24):11247–58.
Sobel Leonard A, McClain MT, Smith GJD, Wentworth DE, Halpin RA, Lin X, et al. The effective rate of influenza reassortment is limited during human infection. PLoS Pathog. 2017;13(2): e1006203.
Steel J, Lowen AC. Influenza A virus reassortment. In: Influenza pathogenesis and control – volume I. Cham: Springer; 2014. p. 377–401.
Thurner C, Witwer C, Hofacker IL, Stadler PF. Conserved RNA secondary structures in Flaviviridae genomes. J Gen Virol. 2004;85(5):1113–24.
Van Valen L. A new evolutionary law. Evol Theory. 1973;1:1–30.

Veeramachaneni V, Makalowski W, Galdzicki M, Sood R, Makalowska I. Mammalian overlapping genes: the comparative perspective. Genome Res. 2004;14(2):280–6.
Vrancken B, Baele G, Vandamme AM, Van Laethem K, Suchard MA, Lemey P. Disentangling the impact of within-host evolution and transmission dynamics on the tempo of HIV-1 evolution. AIDS. 2015;29(12):1549–56.
Vuilleumier S, Bonhoeffer S. Contribution of recombination to the evolutionary history of HIV. Curr Opin HIV AIDS. 2015;10(2):84–9.
Wang W, Zhang X, Xu Y, Weinstock GM, Di Bisceglie AM, Fan X. High-resolution quantification of hepatitis C virus genome-wide mutation load and its correlation with the outcome of peginterferon-alpha2a and ribavirin combination therapy. PLoS One. 2014;9(6):e100131.
Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW Jr, Swanstrom R, et al. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature. 2009;460(7256):711–6.
WHO. Emerging zoonoses. 2017. http://www.who.int/zoonoses/emerging_zoonoses/en.
WHO Scientific Working Group. Antimicrobial resistance. Bull World Health Organ. 1983;61 (3):383–94.
Wilson BA, Garud NR, Feder AF, Assaf ZJ, Pennings PS. The population genetics of drug resistance evolution in natural populations of viral, bacterial and eukaryotic pathogens. Mol Ecol. 2016;25(1):42–66.
Woelk CH, Holmes EC. Reduced positive selection in vector-borne RNA viruses. Mol Biol Evol. 2002;19(12):2333–6.
Worobey M. Molecular mapping of Zika spread. Nature. 2017;546:355–7.
Wright JK, Brumme ZL, Carlson JM, Heckerman D, Kadie CM, Brumme CJ, et al. Gag-protease-mediated replication capacity in HIV-1 subtype C chronic infection: associations with HLA type and clinical parameters. J Virol. 2010;84(20):10820–31.
Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, et al. Population genomics of intrapatient HIV-1 evolution. Elife. 2015;4:e11282.

Population Genomics of Bacteriophages

Harald Brüssow

To the memory of the late Roger Hendrix and Hans Ackermann, who dedicated their scientific life to bacteriophage research.

Abstract Due to their small genome size, an abundance equaling or surpassing that of bacteria, and an obligatory dependence on their host bacteria, bacteriophages are an ideal study object for population genomics. However, due to a certain research neglect, less than 2,700 phage genomes were deposited in the NCBI database, far less than the 90,000 prokaryotic genomes. Large and ecologically representative phage genome sequencing projects have so far only conducted for a small number of phage systems. Phages of dairy bacteria belong to this group since they were systematically collected and extensively sequenced due to their negative impact on industrial milk fermentation. More than ten different phage species were defined for *Lactococcus lactis* and four for *Streptococcus thermophilus*, the two most important starter bacteria in cheese and yogurt production, respectively. The genetic interrelationship between the phages infecting the same host species and between phages infecting phylogenetically (*L. lactis* vs. *L. garvieae* and *S. thermophilus* vs. *S. salivarius* phages) or ecologically closely related host bacteria (*L. lactis* vs. *S. thermophilus* dairy phages) is here reviewed. Dairy phages allowed the study of population genomics as a function of time, geography, and distinct fermentation technologies. The elucidation of the CRISPR-Cas antiviral defense system in *S. thermophilus* provided first insights into the phage-bacterium arms race at

H. Brüssow (✉)
Division of Animal and Human Health Engineering, Laboratory of Gene Technology, University of Leuven, Leuven, Belgium
e-mail: harald.bruessow@kuleuven.be

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_16,

the level of phage and bacterial population genomics. Phages studied by applied microbiologists thus became important study objects for fundamental questions of biology.

Keywords Bacteriophages · Cheese · Dairy · Lactic acid bacteria · *Lactococcus* · Milk fermentation · Phylogeny · Population genomics · *Streptococcus* · Taxonomy

1 Introduction

Population genomics is an extension of population genetics into the genomic sequencing era. By large-scale, whole genome comparisons of DNA sequences, population genomics tries to understand the phylogenetic history, the "phylogeography," "demography," and occupation of ecological niches by a defined population of organisms. Bacteriophages (bacterial viruses or, in short, phages) are a particularly interesting study object for population genomics for a number of reasons. The majority of described phages are from the order of *Caudovirales*, which are the tailed phages (Ackermann 2003). They consist of relatively small genome sizes (10–300 kbp) packaged into a viral capsid, whose size is closely linked to genome length. At the genome sequence, phages are specific to the bacterial genus level (Grose and Casjens 2014). With respect to infection, phages are mostly specific to the species or even subspecies level of host bacteria (Bourdin et al. 2014), and such specificity is largely mediated through tail fibers that adsorb to surface receptors. Tailed phages are further classified into one of three families based on their tail length: *Siphoviridae* (long, noncontractile tail), *Podoviridae* (short tail), or *Myoviridae* (long, contractile tail) (Maniloff et al. 1999; Ackermann 2006). Phages are either virulent or temperate (Campbell 2006). Both have a lytic cycle of replication and host lysis, while temperate phages have an additional lifestyle that permits them to integrate into and subsequently exist dormant as part of the host bacterial chromosome until a later activation (Gottesman and Oppenheim 1999; Little 2006). Virulent and temperate phages can be distinguished by genome analysis (possession of an integrase gene mediating prophage formation and a characteristic genetic switch region) (Lucchini et al. 1999b). Phage genomes are organized into modules of genes for replication, structural proteins, and host lysis, although not in a set order (Desiere et al. 2001a). These basic properties of phage biology seem simplistic in an isolated examination, yet they become quite complex when analyzed at a population level in a natural environment.

The viral nature of phages itself is of interest for population genomics. While phages are not considered living (micro)organisms, they do contain nucleic acids as their genetic material (Villarreal 2004) and are therefore subject to genome evolution (Brüssow and Desiere 2006). The biological status of viruses has been controversially discussed, with some researchers considering them as genetic material that became secondarily independent from cellular life, while others considered viruses as remnants of biological precursors of cellular life (Brüssow 2009). Whatever the answer, the split of phages from bacteria goes back to the root of the universal tree of life, if it does not represent an independent evolutionary tree itself.

Furthermore, the Darwinian concept of species does not apply to phages and different modes of evolution for phages than for sexual organisms for which the term species was defined can therefore be anticipated. Indeed, a distinct modular concept of evolution was proposed for bacteriophages in 1980 where it was recognized that recombination and horizontal gene transfer shaped phage genomes more than the sequential accumulation of point mutations over time on which selection forces worked (Botstein 1980). This special status of phages was later lost when the role of horizontal gene transfer also became obvious for bacteria (Doolittle 1999), which are true living microorganisms, but still not species in the original Darwinian sense.

It is indeed difficult to discuss phages and bacteria separately as phages represent a major element of mobile DNA for bacteria and are therefore important drivers of bacterial evolution. When temperate phages integrate their DNA into the bacterial genome as a prophage and become part of the bacterial genome, they are subject to selection forces working on the host bacterium, therefore resulting in genetic cooperation between phages and bacteria (Canchaya et al. 2003). In a number of important human pathogens, phage-encoded genes are responsible for the virulence of the pathogen (e.g., *Vibrio cholerae*, *Escherichia coli*, *Streptococcus pyogenes*, *Staphylococcus aureus*). Phages have thus impacts ranging from bacterial pathogenicity (Brüssow et al. 2004) to biogeochemical cycles (Roux et al. 2016).

Bacteria additionally evolve to develop mechanisms against phage infection. This is accomplished through a variety of different strategies, including receptor modification, restriction-modification systems, CRISPR-Cas systems, etc. Phages in turn must evolve to overcome the resistance mechanism, creating an ongoing evolutionary "arms race" between a predator and a prey.

Bacterial species in turn are infected by a variety of phages. Since many phages have a host range limited to a single bacterial species, frequently only to a limited number of strains within a single species, this could mean that the number of different phage types is one order of magnitude higher than the number of bacterial species (Rohwer 2003), which concurs with recent estimates from metagenome data (Paez-Espino et al. 2016). Despite the fact that phage genomes are smaller than those of bacteria, the phage sequence space might thus still be of comparable size to that of bacteria. The large number of phage genes with no homology to known genes agrees with this estimate of a large phage DNA sequence space.

Lastly, many environments contain a significant number of phages. Estimates vary from a ratio of 10:1 to 1:1 for phage to bacterial numbers in marine environments, and it is suggested that 20% of the bacterial biomass in the oceans is lysed by phage infection daily. Phages are thus also major ecological players and one of the most numerous biological entities in the biosphere (Wommack and Colwell 2000).

Phage research has historically been a springboard for molecular biology, thanks to their relative simplicity, small size, and ease and speed of experimental manipulation (Cairns et al. 1966). These properties remain assets even in contemporary biology (Brüssow and Hendrix 2002), including for population genomics. In view of these considerations, phage population genomics is a subject of substantial theoretical and practical interest in microbiology, if not for biology in general.

2 Database Limitations

When stressing the theoretical importance of phage population genomics, there are several considerable practical problems associated with the subject. While many large bacterial strain collections exist, only one moderately large phage collection is maintained in Canada (www.phage.ulaval.ca/). In addition, there are phage collections comprising phage isolates from a single or few related bacterial species of specific research interest to academic and industrial laboratories, which are not part of international collections. Researchers interested in phage population genomics can't therefore rely on existing collections, but must frequently first isolate their study objects from the field. For several decades, phage has been absent from the limelight of biological research, and the number of sequenced phage genomes deposited in the NCBI database reflects this neglect. Despite the scientific attractiveness of phage and the convenience of sequencing small genomes (Brüssow and Hendrix 2002), less than 2,700 phage genomes were deposited in the NCBI database at the beginning of 2017, far less than the 90,000 prokaryotic genomes (Hayes et al. 2017). In addition, viral taxonomists officially recognized only 441 species of tailed phages in 2016 (*Caudovirales*) (Krupovic et al. 2016), which most likely represent a serious underestimate of their true number.

Phage population genomics as an emerging research branch is thus still seriously data limited; a meaningful analysis can currently only be done for a handful of well-documented phage systems. Historically, phages have been intensively studied for *Escherichia coli*. Following the reductionist approach of Max Delbrück, coliphage research was historically focused on a detailed genetic analysis of a few selected model phages. Thanks to efforts initiated by Roger Hendrix for lambdoid coliphages (Hendrix et al. 1999) and Henry Krisch for T4-like phages (Filée et al. 2005; Petrov et al. 2010), knowledge of these two coliphage groups has been extended to phylogenetic and ecological aspects. A few other phage systems have been investigated with substantial genome sequencing efforts, providing also valuable material for phage population genomics. Of note here are the efforts of the "Phage Hunters" from Graham Hatfull's lab who sequenced a large number of mycobacteriophages in student and high school courses (Pedulla et al. 2003; Pope et al. 2015), marine phage surveys (Roux et al. 2016), and the constitution of large phage collections as a basis for phage therapy approaches (Kwan et al. 2005; Sarker et al. 2012). However, currently we owe the best sequencing datasets for phage population genomics to applied dairy microbiologists, who collect, characterize, and sequence phages from lactic acid bacteria, which are used as starter bacteria in industrial milk fermentation.

In cheese and yogurt production, a major cause of fermentation failure is phage attack in both the industrial and artisanal settings. Dairy fermentation is largely reliant upon *Lactococcus lactis* and *Streptococcus thermophilus* starter cultures. For the control of the fermentation process, dairy industries regularly screen the factories for phages, put the isolated phages into collections, and classify these phages. This activity provides an interesting study material for phage population

genomics, although it is restricted to a man-made industrial environment. A single large dairy uses half a million liters of milk per day. Dairies represent an important part of the food industry from Western societies, which have now been running for more than half a century (Brüssow 2001). Dairies and their phages represent therefore an upscaled version of the "evolution machine" constructed by Manfred Eigen and therefore promise important insights for evolutionary biologists and geneticists. The present review will concentrate on phage population genomic aspects in the dairy environment, which reflects the background of the reviewer, but also the unique possibilities of this man-made environment for experimental phage population genomics studies.

3 General Aspects of Dairy Phages

Approximately a dozen distinct phage types have been characterized in *L. lactis*. With an average genome size of 40 kb and limited nucleotide sequence sharing between the different phage groups, the minimal sequence space of lactococcal phages is 12×40 kb $= 0.48$ Mb compared to 2 Mb for the host genome. Both estimates under project the true size of the respective sequence space. The pangenome of *L. lactis* converges toward 6,000 genes (Kelleher et al. 2017). The average lactococcal gene is 1 kb, which would mean that the pangenome of the lactococcal phage is about one tenth the size of its host. However, lactococcal phages belonging to the same phage species still differ by many genes, and each newly discovered *L. lactis* phage presents many genes without homologies to known genes; the pangenome for lactococcal phages is therefore certainly greater and will come close to that of its host pangenome. For many well-investigated phage-host systems (coliphages, mycobacteriophages), the viral DNA sphere seems to match that of its bacterial host, if not to exceed it since many phage pangenomes are still far from saturation.

The close relatedness of *L. lactis* and *S. thermophilus* brings into questions the genomic relatedness of their phages. The pangenomes of these two major dairy starter bacteria can be in first approximation considered to be additive. If lactococcal phages were able to also infect the streptococcal starter and vice versa, the combined pangenome for lactococcal and streptococcal phages would be less than additive and possibly far from it. In practice, there are clear cross-species barriers for phage infection between these two dairy starters: *L. lactis* phages in general do not infect *S. thermophilus* and vice versa. If cross-species infections occur as suggested by transduction experiments, they are rare and unusual and do not necessarily involve a fully completed infection (Szymczak et al. 2017). In addition, genetic relationships between lactococcal and streptococcal phages extend only in exceptional cases to the nucleotide level and then only over short segments of the genome. The data do not exclude the exchange of phage genes between these two phage systems, but they exclude a model where phages are ecologically shared between species. There is thus no reason to discount the pangenome of lactococcal phages for overlaps with

streptococcal or other phages. Data from dairy phages thus support a model where their pangenomes come close to that of their bacterial hosts. Obviously, the large size of the phage sequence space and its relative isolation from the bacterial sequence space raise interesting questions about the origin of phage genes.

If the lactococcal and streptococcal phages and their host bacteria present mostly distinct genes with respect to DNA sequence, this does not mean that they are totally unrelated. Several lactococcal and streptococcal phage types demonstrate relatedness at the protein sequence level. Comparative genomics with dairy phages has demonstrated that several phage types infecting different bacterial species and even bacterial genera can be traced back to a hypothetical common ancestor phage (Lucchini et al. 1998; Brüssow and Desiere 2001). However, comparative genomics in coliphages and dairy phages has shown that phages are not the unit of evolution. Due to a pervasive role of horizontal gene transfer in phage evolution, evolutionary histories can only be established for individual phage modules, i.e., units of interacting phage genes, like the genes involved in DNA packaging and phage head construction.

When considering phage evolution, one might ask whether relatedness of phages reflects more the phylogenetic or the ecological relatedness of their host bacteria. The phylogenetic model anticipates relatedness by descent over time scales of bacterial species and genus formation, entailing that few trans-species genetic exchanges would occur once species have split. The ecological model anticipates more frequent genetic exchanges across species barriers for bacteria inhabiting the same ecological niche. Current data do not allow a clear conclusion on these alternatives: *L. lactis* phages are indeed more closely related to *S. thermophilus* phages than to phages from a pathogenic *Lactococcus* species of fish, therefore suggesting that ecological relatedness dominates over evolutionary relatedness. However, the database for such comparison of phages from different *Lactococcus* species is still small. With the possible exception of the most recently described *S. thermophilus* phage group 987 (McDonnell et al. 2016), the dairy phages from *S. thermophilus* are more similar to those of *S. salivarius*, a human oral commensal, than to *L. lactis*, indicating dominance of evolutionary relationships. The graded relatedness of streptococcal phages with other phages that range from nucleotide over protein sequence similarity to sharing a gene order without sequence similarity and that mirror the phylogenetic relatedness of their hosts suggests elements of coevolution of phages with their bacterial hosts. However, a clear separation of both models cannot be expected. Phages are mobile DNA; recombination between phage DNA can be expected between two phages infecting the same cell or between a phage infecting a cell that contains phage as an integrated prophage. The difference between both models therefore concerns mostly the question of how frequent are cross-species infections. If they occur very rarely, phylogenetic relationships of the hosts will dominate for phage relatedness; if they occur more frequently, niche relatedness will dominate the similarity of phages from different species.

However, comparative phage genomics is not phage population genomics. The first analysis addresses mainly problems of evolution, while the second looks for ecological aspects. Dairy phages have here provided some answers. Overall, dairy

phages are cosmopolites: only weak signals for a geographical fragmentation of dairy phages were detected. Other phage systems have also shown closely related phages isolated on different continents. However, in the case of dairy phages, the dairy factory is a man-made environment where a limited number of defined bacterial strains are worldwide distributed by starter companies, precluding geographical specialization of dairy phages. Time series of phage isolation do not reveal clear trends probably because dairy phages cover too short time periods (decades). In addition, dairy phage genomes were not systematically sequenced from the same place over time. Finally, dairies represent a constrained area for phage evolution since the starter is frequently changed and the dairy is cleaned between each new fermentation cycle.

With the number of sequenced dairy phages, investigations into genome variations within a given phage type reveal a type of population dynamics at the genomics level. This type of analysis allowed the differentiation of a core genome (not necessarily a contiguous segment of genes, but more a set of shared genes, which can be scattered across the phage genome) from non-shared genes. The latter must not necessarily be nonessential genes, but can represent different alleles for the same gene function. The non-shared genes probably also contain genes that mediate ecological adaptation to the specific niche. Since most of the non-shared genes in defined phage types lack annotations, an ecology-oriented population genomics is not (yet) possible. However, intragroup phage genome comparisons allow inferences about the mechanisms of genome reshuffling in dairy phages. Exchanges of entire modules and even to a larger extent single gene exchanges punctuate dairy phage genomes.

Phages have a couple of easily measurable phenotypes, e.g., the burst size = number of phages produced per infected cell; latent period = time between infection and appearance of progeny phage. Burst size and latent period cannot be read from the genome, but must be measured with isolated phages in vitro. Dairy microbiologists reported short latent periods in phages from industrial and long latent adaptation periods in phages from artisanal cheese production (Samson and Moineau 2010). Short latent periods and high burst size were interpreted as an adaptation to the industrial environment. Also the lactococcal phage types differed between industrial and artisanal cheese production. The latter are probably closer to lactococcal phages in their natural environment (plant leaves). However, phage genomes are still not easily "readable" to the researcher such that ecological adaptations can be gleaned from genome analysis.

Phage population genomics has already delivered important insights for the evolution and ecological adaptation of important human pathogens. This subject has been extensively reviewed in the past (Brüssow 2007, 2008). Many of the temperate phages conferring virulence genes to human bacterial pathogens carry these genes in a specific genome region (lysogenic conversion module). *S. thermophilus* phage isolates show relatively few temperate phages, which is explained ecologically by the abundance of target cells in the dairy. In contrast, lactococcal phages particularly from the P335 group represent many temperate phages. However, very few have been described, but they are rather abundant in

lactococcal genomes. The temperate lifestyle is believed to represent an adaptation to rare target cells in nature when producing progeny virions would be a futile effort. However, some phages from dairy bacteria are very similarly organized as these prophages from human pathogens. In addition, they contain genes in a genome region corresponding to this lysogenic conversion module. It was speculated that these phage genes are candidates for fitness factors for human commensals and by extension also to dairy bacteria containing prophages. While representing an attractive model, no proof for this hypothesis has been provided because all genes from the putative lysogenic conversion module lacked so far annotations and phenotypes except for the *sie* (superinfection exclusion) genes in lactococcal and streptococcal phages (McGrath et al. 2002).

Phage population genomics can also provide interesting insights into the arms race between phages and their host bacteria. Dairy phages respond to the introduction of phage resistance mechanisms into dairy starters with countermeasures circumventing these infection inhibitors. Since dairy microbiologists follow carefully the factories for the appearance of such escape phages after introduction of a phage-resistant starter cell, the effect of strong selective forces on phage genome adaptation can be studied in detail for dairy phages by phage genome sequencing. In the man-made industrial dairy environment, this arms race can be studied experimentally and, at large scale, not disturbed by the myriad of confounding factors met in natural environments. Dairy microbiology and phage population genomics are likely two cross-fertilizing fields in the near future.

The next sections will provide a detailed review on the population genomics of dairy phages with literature references for readers interested in the state of current knowledge in that specific field.

4 Population Genomics of Lactococcal Phages

4.1 Early Taxonomy of Lactococcal Phages

Lactococcus lactis belongs to the handful of bacteria domesticated by humans shortly after the domestication of milk-producing animals (Passerini et al. 2010). For millennia, milk fermentation was an empirical small-scale activity of dairy farmers, and industrial milk fermentation was only developed in the twentieth century. *L. lactis* is now extensively used as a bacterial starter for many fermented dairy products. Because of the worldwide industrial and financial consequences of phage attack, particularly in cheese production, *L. lactis* phages are among the most commonly isolated phages. By 2007, over 700 lactococcal phage isolates had been reported in the literature (Ackermann and Kropinski 2007). Over the last decade, this number has probably passed beyond 1,000 with a concomitant increase in complete lactococcal phage genomes, which tallied 84 sequenced phage genomes in a census from 2016 (Murphy et al. 2016) and will have surpassed 100 by now (Mahony et al. 2017b). Early on, dairy microbiologists tried to establish

a taxonomic differentiation scheme for lactococcal phages based on phage morphology, host range determination, restriction analysis, and especially DNA-DNA hybridization techniques (Jarvis 1984; Braun et al. 1989; Prevots et al. 1990). Electron microscopic analysis revealed that all isolated lactococcal phages belonged to the group of tailed phages (*Caudovirales*). Tailed phages come in three different phage families: *Siphoviridae* (phages with long, flexible, noncontractile tails), *Podoviridae* (phages with short tails), and *Myoviridae* (phages with contractile tails). The vast majority of lactococcal phages belong to the family of *Siphoviridae* (Deveau et al. 2006). Morphologically, the lactococcal *Siphoviridae* can be distinguished by capsid diameter ranging from 45 to 70 nm (the capsid size is roughly proportional to genome size), capsid form (isometric to slightly elongated), and tail lengths, which cover sizes from 93 to 490 nm (the latter corresponds to half the length of a small-sized bacterium). Baseplate structures are likewise variable, as well as the collar structures. Only two *Podoviridae* were detected within the lactococcal phage isolates. The lactococcal *Podoviridae* showed either a slightly or extensively elongated, cigar-shaped phage, both with small tails (19 and 32 nm, respectively). Strikingly, not a single myovirus has so far been identified in *L. lactis*. Based on morphology and DNA-DNA hybridization analysis, ten phage groups (sometimes also called phage species) were distinguished (Deveau et al. 2006). The species concept was used for phage taxonomy purposes (Ackermann et al. 1992), but one should be aware that it represents a polythetic species concept. Polythetic means sharing a number of characteristics, which occur commonly in members of a group, but none is essential for membership in that group. The polythetic species concept for phages is a direct reflection of the modular mode of phage evolution (Botstein 1980) and is of central importance for the understanding of phage population genomics.

4.2 Lactococcal Phage Sequencing and Database

At least one representative from each of the ten lactococcal phage species was sequenced, allowing a comprehensive comparative genomic analysis of lactococcal phages. Of interest for phage population genomics, multiple phage isolates from the "936 species" were sequenced, including 28 isolates from a survey in 8 Australian cheese factories conducted between 1994 and 2001 (Castro-Nallar et al. 2012) and 35 isolates sampled between 2009 and 2013 from 4 Dutch dairy factories (Murphy et al. 2016). Several genomes were also deposited for the "P335 species": 15 phage and 8 prophage genomes were reported (Oliveira et al. 2016), and recently 17 novel genomes were added (Mahony et al. 2017b). The "949 species" is represented with 12 complete phage genomes, the "P087 species" with 7 genomes (Mahony et al. 2017a), and the "1706 species" with 5 (Kot et al. 2014), while only 1 or 2 genomes are available for the remaining lactococcal phage species (Mahony and van Sinderen 2014). The overrepresentation of genome sequences for lactococcal phages from the 936 and P335 species reflects the fact that these two

lactococcal phage species also represent the main causes of milk fermentation failures worldwide (Josephsen et al. 1994; Moineau et al. 1992), while the other species are rare isolates from dairy factories, sometimes even representing unique isolates, which are thus of fundamental interest for defining the breadth of lactococcal phage genome variety, but of limited interest for practical phage control in the factories. However, this preponderance of lactococcal phages from the 936 and P335 species depends on the ecological context. While these isolates dominate in industrial cheese environments, the "rare" lactococcal species 949 and P087 are prevalent in whey from artisanal cheese making (Mahony et al. 2017b). The major difference between both systems is the fact that large dairies work with a genetically restricted set of industrial starter strains provided by specialized commercial suppliers, while artisanal cheese makers use complex, variable, and undefined starter mixtures. The streamlining of the *L. lactis* starters in the industry has thus also narrowed the diversity of lactococcal phages encountered in industrial context. The natural habitat of *L. lactis* is decaying plant material where this bacterium ferments plant cell wall material. The dairy strains of *L. lactis* living in the nutritionally richer milk environment are derived by a process of genome erosion from plant strains of *L. lactis* (Siezen et al. 2008). It will be interesting to investigate the ecological specialization of *L. lactis* phages following their host from plant association to artisanal and from there to industrial milk fermentation, but phage isolates from the original niche occupation of *L. lactis* in plants are completely lacking for such studies and are still scarce for artisanal cheese production.

4.3 The 936 Species (Now: **Sk1virus***)*

The prototype of the "936 species" is phage sk1 for which a detailed transcription map was established, differentiating blocks of early, middle, and late transcription genes (Chandry et al. 1994). The overall organization of this 28 kb genome resembled that of a size-reduced *E. coli* phage λ genome. The left half of the sk1 genome contains late structural genes, all transcribed rightward, and intriguingly the gene order closely resembles the structural gene order of phage λ, without, however, sharing any sequence similarity at nucleotide or protein level with phage λ. Only the gene map and the order of the prospective structural genes were very similar between both phages. A similar genome organization was also found in a number of phages from Gram-negative and high GC-content and low GC-content Gram-positive bacteria, even in viruses infecting one subgroup of *Archaea* (Brüssow and Desiere 2001). This structural gene module seems to represent an old evolutionary constellation that was conserved from a hypothetical ancestor phage into modern descendants of this evolutionary viral lineage when adapting to a wide range of prokaryotes. The conservation of this gene order has fascinated biologists, and several explanations were proposed. Some researchers noted a parallel between the phage particle's morphology and the gene map (DNA packaging, head, head to tail, tail fiber, lysis genes) and suggested that this order would assure a temporal

sequence of expression parallel to their requirement during morphogenesis (Casjens and Hendrix 1974). Others have argued that a fixed linkage arrangement has a selective advantage since it minimizes recombination events that lead to nonviable phages when the same cell is infected with two different phages (Stahl and Murray 1966). This concept underlies the Botstein hypothesis of modular phage evolution (Botstein 1980). Modular evolution does not exclude recombination within a functional module between unrelated phages, but since they bring together incompatible genetic elements, no viable phage is formed and consequently counter-selected (Juhala et al. 2000). Selection thus favors phages with a conserved gene order within a given module. Other researchers suggested that short DNA repeats within phage genomes (Blatny et al. 2004) or promotor sequences demarcate preferred sites for recombination enzymes.

The 28 Australian 936-like phages indeed showed an extremely conserved structural gene module (Castro-Nallar et al. 2012). The only variation was found in the length of the neck passage structural gene and a small adjacent gene, which allowed the separation of the Australian 936 phages into two subgroups. The subgroups correlated with host tropism, but not geographical location of the factories. The neck passage protein and the receptor-binding protein represented peak regions in the compared 936 genomes with respect to genetic diversity and recombination rate. The 35 Dutch 936-like phages also showed a conserved structural gene module, but they displayed more variations than the Australian phages (Murphy et al. 2016). One additional site of diversity was the region between the small and large terminase genes involved in DNA packaging which were in some isolates separated by one to three extra genes. This additional extra DNA comprised a homing endonuclease and a methyltransferase gene, which suggest invasion of the phage genome by another mobile DNA element (Foley et al. 2000). Another site of diversity is located in the major tail gene, which is in some phages not preceded by a neck passage structural gene and followed in others by a tail extension gene. Variations in that genome region correlated with morphological changes in the neck of the phages displaying either whiskers, a double disc, or no extra structure. Another tail morphology variant, a spiral structure around the tail, was associated with a translational frameshift between the major tail and the tail extension genes. Also, the Dutch 936-like phages showed sequence diversity over the receptor-binding gene, which correlated with the cell wall polysaccharide biosynthesis type of the *Lactococcus* host cell, suggestive of positive selection for host adaptation.

The right genome half of the 936-like phages is nearly entirely occupied by numerous, small-sized, early transcribed genes; all are transcribed in opposite direction to the structural genes. The few annotated genes suggest DNA replication function. The Australian phages showed highly conserved replication modules within each of the two 936 subgroups (Castro-Nallar et al. 2012). The early gene modules from the Dutch phages displayed substantial differences between the sequenced isolates suggesting gene gain and loss. Some variation was tentatively associated with genes circumventing host cell-encoded antiphage systems of abortive infection (Abi) systems (Murphy et al. 2016) resulting from the arms race between bacterial phage defense systems and phage anti-defense countermeasures. Demographic analysis revealed a constant population size and genetic diversity of

the phages through time, which could mean that there was a stable number of outbreaks through time over the time period covered by phage isolates (Castro-Nallar et al. 2012) and that the introduction of Abi systems met countermeasures of phages (Labrie et al. 2010).

Middle transcripts of 936-like phages are derived from a small genome region at the right end of their genomes. These genes are transcribed in opposite direction to the early genes. The Australian phages differed by the presence or absence of tRNA genes, while the Dutch phages showed more gene variation in this genomic region.

Notably, the Australian phages clustered into the same two distinct groups whether based on late, early, or middle gene clusters, therefore suggesting that the reshuffling of phage genomes between modules – as predicted by Botstein – did not occur during the last decades. Molecular clock arguments indicate time frames between 60 and 110 years for the differentiation of the two subgroups, a time frame compatible with the industrialization of dairy activities in Australia, but probably too short for the Botstein type of phage genome evolution to occur. This argument limits also the use of dairy factories as proxies for an evolutionary machine as an alternative to large-scale laboratory evolution experiments (Lenski 2017). Loss and gain of single genes under selective pressure of Abi systems were by these population genomics data identified as rapid processes, occurring within the considered time frame. This concurs with the observation of resistant mutants created by natural recombination after biotechnological introduction of Abi systems into starter strains (Labrie et al. 2010).

When phages were sorted according to a presence or absence matrix of protein families, some clustering by country was evident. However, frequent exceptions were seen and attributed to the global movement of industrial starter cultures, which distorted any geographical clustering via natural means of phage dispersal (Murphy et al. 2016).

4.4 The P335 Quasi-Species

The second major group of phages isolated from cheese fermentation failures is named after a phage isolated in 1979 from a German dairy (Labrie et al. 2008). It showed an isometric small head, a relatively short tail, and a complex baseplate. Its genome is 33.6 kb long and showed a clear modular organization. A replication/transcription module is followed by a morphogenesis and a lysis module. With lactococcal phages TP901-1 and Tuc2009, better investigated members of the P335 group, it shares not only a closely related gene map, but also sequence identity over the morphogenesis module. The alignment was, however, patch-wise: regions with protein sequence identity >95% alternated with regions of lesser or no sequence identity. Differences in the tail tape and baseplate genes correlated with differences in tail length and baseplate structures. Over the replication/transcription module, P335 shared only 4 out of 22 genes with TP901-1, but the shared genes were not adjacent, therefore excluding a single modular transfer event. Phage P335 lacked the entire lysogeny module located at the left genome end of the temperate

phages TP901-2 and Tuc2009. Between the transcription and morphogenesis module, P335 possesses a transcribed four-gene insert, not known in other lactococcal phages but found in a prophage from a streptococcal pathogen.

Comparative genomics in the P335 phage species thus demonstrates a different mode of genome evolution than seen for the 996 species. Quite extensive gene exchanges have apparently shaped the P335-like phage genomes. Small blocks of genome segments rather than entire modules were the target for genetic recombination. This conclusion was not only derived from genome comparisons but also seen in natural mutants from another lactococcal phage from the P335 group, namely, phage ul36 (Bouchard and Moineau 2000). Phage ul36 acquired DNA sequences from a prophage of its *L. lactis* host. The recombinant phage had the selective advantage to be insensitive to AbiK abortive infection system. The exchange was mediated by homologous recombination over a stretch of 23 bp with 100% bp identity between phage and prophage. This was not a singular event: DNA acquisition by phages from the bacterial chromosome (Moineau et al. 1994) and from a plasmid (Hill et al. 1991) has previously been described. Also, phage LC3, a member of another subgroup of the P335 phage species, contained 14 noncoding regions larger than 50 bp that shared sequence identity with similar positioned intergenic regions of other P335 phages and might thus represent potential recombination sites for gene exchange. Two of these regions were indeed located at mosaic borders (Blatny et al. 2004).

With such a prominent role of recombination for phage genome evolution, it is obvious that the definition of a phage species will meet with severe difficulties. Indeed, the P335 species was subdivided into four subgroups, I–IV. The common denominator was literally a single gene, a dUTPase, different from the chromosomal *dut* gene but strangely conserved across the phages of the P335 species. The function of this phage gene is still unknown, but its conservation was used by dairy microbiologists to develop a diagnostic PCR assay for P335 phages. However, exceptions have been found (Kelly et al. 2013), questioning both the essential role of this gene for phages of the P335 group and removing the last shared denominator in the P335 species. The P335 species was therefore referred to as a polythetic species (Labrie et al. 2008) or as a quasi-species (Labrie and Moineau 2007; Chopin et al. 2001). The lactococcal phages from the P335 species are thus a good illustration for the distinct mode of genome evolution in bacteriophages in general. The need for collecting and characterizing large numbers of phages for the practical needs of the dairy industry has provided interesting raw material for phage population genomics studies. After the recent sequencing of 17 novel P335 phage genomes from 8 different countries (Mahony et al. 2017a), 27 complete genomes can now be analyzed.

4.4.1 P335: Subgroup I (e.g., BK5-T)

The prototype of this phage group is the temperate phage BK5-T with a 40 kb-long genome. BK5-T shares >60% aa sequence identity with two other members of this subgroup (phages 4268, bIL286), but the sequence relatedness is restricted to the

morphogenesis module and excludes the tail appendages (Deveau et al. 2006). Only two genes from the remainder of the genome are shared with phage 4268. With lactococcal phages from the other P335 subgroups, BK5-T shares only a few isolated nonstructural genes. BK5-T displayed similarity at the protein sequence level over a 15 kb-long part of the morphogenesis module with *S. thermophilus* phage Sfi21. Notably, a gradient of relatedness ranging from nucleotide sequence to protein sequence similarity to gene map similarity without sequence relatedness was seen with phages from Gram-positive bacteria. Since the degree of relatedness was correlated with the evolutionary distance separating their bacterial hosts, these observations were interpreted as elements of vertical evolution for the structural gene cluster of this phage group (Desiere et al. 2001a, b). Protein sequence similarity was also observed between several genes of the DNA replication module from BK5-T and Sfi21 (Desiere et al. 1997).

4.4.2 P335: Subgroup II (e.g., TP901-1)

When ten lactococcal phages from subgroup II, representing isolates from four continents, were compared, a complex pattern of genome sharing was observed. Phages independently isolated in Australia, the UK, and the USA shared more than 90% aa identity over the entire genome except for two tail proteins (TMP, Tal), while phage isolates from the USA and Canada aligned essentially across the morphogenesis module only, excluding the receptor-binding protein-encoding gene (Mahony et al. 2017a). The alignments of the structural module showed a patchwork of related and unrelated genes for some subgroup II phages, suggesting genetic exchange of single genes or small gene groups. Apparently, single genes or gene groups exist in multiple alleles that can be relatively freely assorted, leading to genetic mosaics for these phage genomes (Labrie and Moineau 2007). No significant sequence identity was seen between the structural gene modules from subgroup II phages and those of the other subgroup P335 lactococcal phages. One might therefore ask whether subgroup II phages are not better defined as a distinct phage species particularly since the different subtypes of the P335 phage species can also be clearly distinguished by morphologically distinct distal tail regions (Mahony et al. 2017a). Since this would further increase the already large number ($n = 11$) lactococcal phage species, dairy microbiologists have (so far) opted against this solution.

4.4.3 P335: Subgroup III (e.g., LC3)

The dot-plot analysis of the subgroup III lactococcal phages LC3 (Blatny et al. 2004) and r1t (van Sinderen et al. 1996) showed an alignment over the entire genome, punctuated by smaller and larger alignment gaps. The gaps mostly respect gene borders, suggesting that single genes or groups of adjacent genes were the units of genetic exchange. Interestingly, these gene groups were partially flanked by direct repeats, which may represent target regions for recombination enzymes and thus

demarcate predestined breakpoints (Blatny et al. 2004). When comparative genomics was extended to further subgroup III phages, substantial variation over the nonstructural genes was observed. Some phages shared only a small number of isolated genes (Mahony et al. 2017a), less than between some inter-subgroup comparisons within P335 phages (Blatny et al. 2004). In contrast, the structural modules were shown to be better conserved (Mahony et al. 2017a).

4.4.4 P335: Subgroup IV (e.g., Q33)

Subgroup IV phage isolates from three continents shared a highly conserved structural gene module, while no links to structural genes from other lactococcal phages were detected. Since Q33 structural genes shared similarity with prophage from diverse colonizers of the gastrointestinal tract (*Enterococcus*, *Bifidobacterium*), it was suspected that Q33-like phages infected initially another host before acquiring the appropriate machinery to infect *Lactococcus*. In support of this hypothesis, subgroup IV phages shared substantial gene identity over nonstructural genes with subgroup III P335 phages (Mahony et al. 2013, 2017a).

4.5 The c2 Species (Now: C2virus*)*

Together with phages from the 936 and P335 group, phages from the c2 group are consistently found in industrial milk fermentation failures that use *L. lactis* as starter. *C2virus*, which is recognized by the International Committee on Taxonomy of Viruses (ICTV), shows a prolate head in contrast to the isometric head of the 936 and P335 phages and displays a smaller 22 kb genome. Two subgroups were distinguished in this exclusively virulent phage species: c2- and bIL67-like phages. Both genomes showed an absolute synteny of their genome maps, nearly bp sequence identity over the leftmost 17 kb of their genomes, but differed for a group of three structural genes at the right end of the genome, which displayed only about 40% aa sequence identity (Lubbers et al. 1995). Sequencing of a larger set of *C2virus* isolates confirmed this separation into two subgroups and linked the difference observed with the three genes at the right genome end (l14–l16 in c2) with the recognition of two alternative phage receptor proteins Pip and YjaE on *L. lactis* host cells (Millen and Romero 2016). Hybrids between c2 and bIL67 phages could be created in vivo, demonstrating that host-determinant specificity serves as a strong selective pressure for phage evolution by recombination (Millen and Romero 2016). By using only host strain rotation as selection pressure, phage-phage recombinants could be created, which associated further genome regions with host range determinants at the level of DNA entry (*l10* in c2) and *cos*-end ligation (1.5kb right terminal end) (Rakonjac et al. 2005). Starter strain rotation, used in the dairy industry to keep phage titers low, might ironically stimulate horizontal gene transfer contributing to the observed mosaicism of dairy phages.

4.6 The Not So Rare Lactococcal Phage Species 949 and P087 from Artisanal Cheese

More than half a dozen further *L. lactis* phage groups have so far been identified. Due to their low industrial prevalence, they represent sometimes only single, sequenced isolates. However, it must be remembered that their rarity refers to a man-made environment, while their prevalence is sometimes higher in a natural environment (Mahony et al. 2017b). This was demonstrated for two such rare lactococcal phage groups, repetitively isolated and sequenced from artisanal cheese production. Of course this observation contradicts their rarity and shows that the frequency of isolation depends on the ecological niche investigated. Isolation of phages is difficult from artisanal cheese since starter cultures are undefined and complex. When using a test panel of 25 indicator strains for phage detection in traditional Sicilian cheese whey, 59 phage isolates were obtained. With a multiplex PCR approach, 51 isolates could not be typed, pointing to a high prevalence of variant phage groups, not covered by this PCR. The artisanal cheese whey yielded representatives of the previously defined "rare" phage groups 949 and P087.

Phage 949 is, in many respects, unusual: morphologically it represents a siphovirus with a large isometric head of 70 nm diameter. It is the largest known lactococcal phage genome with 115 kb of DNA, containing terminal redundancy suggesting a *pac*-site phage. Phage 949 presents a 500 nm-long tail, which is larger than previously reported for any siphovirus and renders it more sensitive to thermal inactivation (Mahony et al. 2017b) practiced in industrial cheese making. Its genome comprises 154 ORFs that are subdivided into 4 segments of opposite orientation. Half of the ORFs have no significant identity with entries from the NCBI database. Matches were mostly with *Firmicutes* and their phages at an aa identity level below 50%. The genome organization was unusual and characterized by a number of similarities with T4 phage (e.g., the presence of ribonucleotide reductases, group I introns, many tRNA genes) (Samson and Moineau 2010). When 12 further 949-like genomes were included into the comparison, only 94 gene families were identified for the core genome (where members of a given core gene family are defined as those whose deduced products share aa identity above 50%, while all analyzed genomes should contain at least 1 member of that family). The core genome extends only across half of the genome map indicating substantial variability within this phage species (Mahony et al. 2017b). A further unusual feature of the 949 phage group was its broad host range on lactococci and its wide ecological distribution: 949-like phages were also found in rennet from the stomach of young ruminants. However, these phages may have originated from the whey of the artisanal fermentation that may have been used as feed for piglets and calves. In fact, except for a few scattered, gene-poor regions, the cheese phage isolate WRP3 aligned at the nucleotide level across the entire genome length with the sewage phage L47 (Mahony et al. 2015), isolated on a grass-associated *L. lactis* strain (Cavanagh et al. 2014). As mentioned above, *L. lactis* is believed to have evolved from plant strains. The domestication of this organism to the milk environment is

associated with genome reduction and gene decay and the acquisition of specific genes involved in protein and lactose utilization by horizontal gene transfer (Cavanagh et al. 2015). The case of the 949-like phages indicates that the ecological distribution and genetic diversity of *L. lactis* phages might be substantially greater than currently appreciated.

Phage genes conferring adaptation to a specific niche beyond receptor-anti-receptor interactions in host recognition (Mahony et al. 2017a, b) are less well investigated for lactococcal phages. For example, phage 949 has a latency period of 70 min, which is longer than latency periods commonly found in dairy phages from industrial environments, which can be as short as 20 min (Samson and Moineau 2010) and most likely represent an adaptation to the fast-growing industrial *L. lactis* starters. The predominance of phages from the 936, P335, and c2 group in dairies might present adaptations to highly domesticated *L. lactis* strains, which went through two selection processes: first, the selection for milk-adapted lactobacilli in prehistoric time of early dairy activities documented in archaeological sherds from Turkey (Evershed et al. 2008) and, second, a selection for efficient industrial starters during the twentieth century. *L. lactis* not undergoing this domestication process have certainly survived in their natural habitat on plant leaves and phages adapted to this environment. As there is no applied incentive to study lactococcal phages from this environment, our knowledge about these natural lactococcal phages is extremely limited. The distinct nature of phage isolates from artisanal cheese allows a fascinating glimpse into the world of preindustrial lactococcal phages.

A similar situation applies to lactococcal phage P087, a siphovirus with an isometric head and a 60 kbp-long, circularly permutated genome with an uncommon organization. All genes are oriented in one direction. The left genome half encodes nonstructural genes involved in the early phage infection cycle. The right half encodes structural proteins, which display distant sequence similarity with a prophage from a clinical *Enterococcus faecalis* isolate (Villion et al. 2009). *E. faecalis* overlaps the ecological niche of *L. lactis* since it is also isolated from dairy foods, but the level of sequence identity is mostly below 35%, which argues against direct horizontal gene transfer between *Enterococcus* and *Lactococcus*. Six further P087-like phages were isolated from artisanal Sicilian cheese whey. Their sequencing revealed a conserved P087 core genome that comprised 80% of the predicted genome. Despite this high degree of genome conservation, phylogenetic tree analysis split P087 phages into three separate branches (Mahony et al. 2017b).

4.7 Raw Milk-Associated Lactococcal Phages: The Case of the 1706 Species

Another interesting case is represented by phage 1706. It was isolated 20 years ago from a failed French soft cheese production, but rarely encountered in dairies since. Phage 1706 is a siphovirus with an isometric head, a relatively long tail (276 nm),

and a 56 kb-long *cos*-site containing DNA genome, with an unusual constellation of the lysis cassette (Garneau et al. 2008). A large, contiguous segment of the morphogenesis module and several regions of nonstructural genes share a common genome organization and up to 60% aa sequence identity with a prophage from *Clostridium leptum*, a constituent of the human fecal microbiota, and *Ruminococcus torques*, found in the rumen of cattle and sheep. This finding recalls the isolation of 949 phages from rennet of young ruminants with the caveat cited above. However, a recent horizontal transfer from clostridia is unlikely: the GC content of phage 1706 is typical for lactococci (34 vs. 36%), while the *C. leptum* prophage and its host have a much higher GC content (45 vs. 50%) (Garneau et al. 2008). If phage 1706 originated in clostridia, enough time has elapsed since its cross-species infection to adapt its genome to the new low GC-content host. Interestingly, phage 1706 carries a clear signature of adaptation to lactococci in possessing a receptor-binding gene that shares 71% aa identity with a *L. lactis* prophage. A similar prophage had contributed genes to *L. lactis* phages infecting a host containing two different phage resistance mechanisms, AbiK and AbiT (Labrie and Moineau 2007). Apparently, under selective pressure of resistance genes, prophage elements provide the genetic material to superinfection phages to escape from control. Prophage-phage interaction thus drives the evolution of lytic phages in *L. lactis*. Phage 1706 might therefore represent the outcome of a cross-species phage infection where the resulting phage adapted to the new host by modifying its GC content and acquiring a *L. lactis*-specific receptor-binding gene. Acquisition of only the receptor-binding gene was apparently not enough to change the host specificity, since this gene was preceded by two tail proteins also related to *L. lactis*, more specifically phage bIL286 (P335 subgroup I). Apparently, these tail proteins are also needed to connect the receptor-binding protein to the remainder of the tail from phage 1706.

Four lactococcal phages isolated 40 years ago from raw milk were recovered from a freezer and sequenced. They turned out to be closely related, sharing 90% bp identity between each other. They also shared 45% overall bp identity with phage 1706, but regions of similarity were scattered across 19 non-contiguous genomic segments, and none was larger than 1 kb (Kot et al. 2014), pointing to a complicated evolutionary history of phage 1706, which might be the result of multiple, sequential recombination events.

4.8 Rare Lactococcal Siphoviridae *Isolates: Q54 and 1358*

For the dairy technologist, the rare, sometimes unique lactococcal phage isolate represents curiosities of little practical interest. For the population geneticist, such isolates provide important additional information on the overall natural diversity of *L. lactis* phages and could shed light on their origins, evolution, and relationships with other phages.

Phage Q54 is such a rare lactococcal phage isolate. With its prolate head, this siphovirus closely resembles *C2virus*. However, DNA-DNA hybridization failed to reveal any homology with the known lactococcal phage species. Q54 has a narrow

host range, a property which it shares notably with the rare phage isolates 1706 and 949. Its 27 kb-long *cos*-site DNA genome contains early and late morphogenesis modules on the upper strand and a smaller segment containing further early genes on the lower strand. The morphogenesis module shares a nearly collinear gene map and up to 34% aa identity for six genes with *C2virus*. Five isolated nonstructural genes show up to 59% aa identity with the lactococcal phage sk1 (Fortier et al. 2006). The genome analysis suggests recombination events between c2-like and 936-like lactococcal phages. The presence of three separate genes, related to P335, indicates a complicated recombination and evolutionary history of phage Q54. Genes related to P335-like phages are not surprising since this lactococcal phage group contains temperate phages, which reside as prophages in the bacterial genome (Chopin et al. 2001). Prophage DNA is genetically accessible for recombination to phages superinfecting such lysogens.

Phage 1358 was isolated in 1981 in New Zealand. With respect to size, it is the smallest of the lactococcal *Siphoviridae*. The left half of the 37 kb-long dsDNA genome encodes the morphogenesis module, the right half nonstructural genes; all genes are transcribed in the same direction (Dupuis and Moineau 2010). Several features make this lactococcal phage unusual. The genes of the morphogenesis module share 25–49% aa identity with *Listeria monocytogenes* phages P40 and P35, and two segments of the DNA transaction genes resemble putative prophage genes from *Listeria*. Not a single best gene match was with a lactococcal phage. Even more unusual was the high GC content of 51%, far above all other lactococcal and *Listeria* phages, rendering the origin of this phage enigmatic. Problems with the codon usage might explain the long latent period of 90 min and the rare isolation of this phage from industrial cheese making.

4.9 Rare Lactococcal Podoviridae *Isolates: P034 and KSY1*

Lactococcal phages from the P034 species display a prolate head, whiskers, and a short tail characteristic for *Podoviridae*. While they have been repetitively isolated in the dairy industry, they represent less than 1% of the phage isolates (Braun et al. 1989). Their 19 kb-long dsDNA genome is in several aspects unusual for lactococcal phages: it contains 0.6 kb inverted terminal repeats, a phage-encoded DNA polymerase, and a terminal protein (Kotsonis et al. 2008). The genome contains nonstructural and structural genes in opposing transcriptional orientations. Despite a distinct overall genome map organization, several similarities including weak sequence similarity place this lactococcal phage close to *Bacillus subtilis* phage ϕ21 and *Streptococcus pneumoniae* phage Cp-1. However, two putative tail genes show 53% aa identity with lactococcal phages, which were possibly needed to interact with the *L. lactis* host. DNA-DNA hybridization between different members of the lactococcal P034 species suggested some genomic variability.

Lactococcal phage KSY1 is another unusual and so far unique isolate. It shows a 220 nm large cigar-shaped head structure, an elaborate base plate, and a short tail, defining another *Podoviridae*. In contrast to the lactococcal *Siphoviridae*, which

show distantly related λ-like genome maps for the structural module, KSY1 displays a T7-like transcription system, including an RNA polymerase. Several genes show low-level sequence similarity with diverse *E. coli*, *Bacillus*, *Lactobacillus*, *Streptococcus*, and *Staphylococcus* phages. Notable is a >80% bp identity over a 5 kb segment encoding putative baseplate and adjacent tail genes from *Lactococcus* phages of the P335 group (Chopin et al. 2007). This observation suggests that genes enabling interaction with the *L. lactis* cell surface were possibly acquired from lactococcal prophages by an alien phage following a cross-species infection. This scheme seems to be a common motive in rare lactococcal phage types.

4.10 An Outgroup: L. garvieae *Phages*

A close phylogenetic relative of *L. lactis* is *L. garvieae*, an important fish pathogen, which has also been isolated from raw milk, sewage water, and vegetables. Several *L. garvieae* phages were characterized, e.g., the soil isolate GE1 (Eraclio et al. 2015). GE1 displays a morphology that closely resembles *L. lactis* phage c2. In addition, GE1 shares a comparable genome map despite a slightly larger genome size (25 vs. 22 kb). Thirteen of the forty-eight GE1 ORFS showed aa identity up to 58% with phage c2, but at most three adjacent genes demonstrated this sequence identity suggesting several modular exchanges between both phage systems in a distant past.

L. garvieae phage WP-2, isolated from water of a rainbow trout farm, presents a 19 kb genome with two opposing nonstructural and structural gene modules including a DNA polymerase. It is a new member of the *Ahjdlikevirus* genus of *Podoviridae* with numerous matches to *Staphylococcus* phages at 30–40% protein sequence identity; only two genes shared similarity with other *L. garvieae* phages (Ghasemi et al. 2014).

Subsequently a prophage was induced from a *L. garvieae* strain of marine fish. The siphovirus PLgT-1 possesses a 40 kb genome with a genome organization that is typical of many temperate dairy phages. A short lysogeny module is divergently transcribed from the remainder of the genome comprising DNA replication/morphogenesis/lysis modules. It shared with *L. lactis* phage TP712, a member of the P335 species, the overall genome organization and high sequence identity over the central part of the genome that encodes transcriptional regulation, DNA packaging, and head and major tail genes (Hoai et al. 2016). However, homology was frequently closest with an *Enterococcus* phage genome. Recently, eight prophages from *L. garvieae* isolated from dairy products and fish were characterized. Prophage PLg-TB25 with a 38 kb genome, isolated from a cheese strain, displayed homology with other *L. garvieae* phages over nonstructural genes, while the structural genes resembled mostly various phages from other genera of low GC-content *Firmicutes* (however, protein sequence identity remained <55%). Another prophage from a dairy *L. garvieae* strain was shown to display >90% and 50% protein sequence identity with head and tail proteins from *L. lactis* phage ul36.k1 (a subgroup II P335

species). The same strain yielded another prophage which showed a similar pattern of identity with *L. lactis* phage r1t (Eraclio et al. 2017). Further *L. garvieae* prophages showed a similar genome organization, but no significant identity with other phage genomes. Overall, *L. garvieae* phages seem not to be in active gene exchange with *L. lactis* phages prevented either by ecological or species barrier effects.

5 *Streptococcus thermophilus* Phages

S. thermophilus is an interesting reference for comparative phage genomics with lactococcal phages for several reasons. *Lactococcus* was initially referred to the genus *Streptococcus*. *Lactococcus* has in the meanwhile been attributed to a distinct genus. *S. thermophilus* is, after the other *Lactococcus* species, the closest phylogenetic relative of *L. lactis*. Both species share the same ecological niche, the dairy environment, and a similar evolutionary pathway. Like *L. lactis*, *S. thermophilus* adapted mainly through loss of function to milk as a habitat (Hols et al. 2005). Today, *S. thermophilus* is, after *L. lactis*, the second most important starter in the dairy industry. Both species have ample opportunity for genetic exchange since both starter bacteria are used in the same vat in many industrial fermentation processes. It is thus interesting to compare the bacteriophages from both species to understand forces that drive genome evolution of phages in this ecological niche.

Two basic observations have struck dairy microbiologists when comparing both phage-host systems. Firstly, *S. thermophilus* phages are highly uniform compared to the diversity of *L. lactis* phages (Mahony and van Sinderen 2014). Secondly, *L. lactis* mounts a multitude of frequently plasmid-encoded abortive infection defense mechanisms (Labrie et al. 2010), while plasmids are rare in *S. thermophiles.* This suggests a fundamentally different phage-host interaction, as recently underlined by the discovery of the CRISPR-Cas defense system in *S. thermophilus* (Barrangou et al. 2007).

5.1 *Two Phage Lineages in* Streptococcus thermophilus*:* cos-*Site Sfi21 and* pac-*Site Sfi11 Phages*

Compared to the great diversity of lactococcal phage species, the taxonomy of *S. thermophilus* phages was, until recently, quite simple. Over decades of factory surveys, the collected phages revealed morphologically uniform phages: *Siphoviridae* with 65 nm isometric heads, 260 nm-long noncontractile tails, and a long tail fiber (e.g., Brüssow et al. 1994a; for reviews see Brüssow 2001; Mahony and van Sinderen 2014). Most phages represented virulent isolates, but the less numerous temperate *S. thermophilus* phages showed close genetic relationship with the lytic phages (Brüssow and Bruttin 1995). In fact, by serial passage of the temperate *S. thermophilus* phage Sfi21 in the laboratory, spontaneous lytic phage

mutants were derived that had showed deletions at identical nucleotide positions, which suggests a site-specific recombination system transforming temperate into lytic *S. thermophilus* phages (Bruttin and Brüssow 1996). A faulty side reaction of the phage integrase may have been responsible for part of the spontaneous deletions leading to lytic derivative phages (Bruttin et al. 1997b).

The early partial sequencing of *S. thermophilus* phage genomes had revealed a highly conserved DNA replication module (Desiere et al. 1997). The conservation was twofold: first, this module was widely shared between three quarters of all *S. thermophilus* phages from dairy factories (Brüssow et al. 1994a, b; Le Marrec et al. 1997). Second, the degree of nucleotide sequence identity over the DNA replication module between many independent isolates differed frequently by <0.1% (Brüssow et al. 1994b). At a protein sequence level, the genes of the conserved DNA replication module matched genes from other phages in a patchwise fashion, including coliphages. Most prominent were, however, similarities with lactococcal phages of the P335 group (Desiere et al. 1997). *S. thermophilus* phage 7201 displayed an alternative DNA replication module: it shared up to 40% protein sequence identity with replication proteins from widely distributed plasmids in *Firmicutes* and with *Bacillus*, *Lactococcus*, and *E. coli* phages (Stanley et al. 2000).

Subsequently, *S. thermophilus* phages were classified into two groups based on the presence or absence of cohesive genome ends. This difference in DNA packaging co-segregated with differences in the virion protein composition: *cos*-site phages showed two major structural proteins, while *pac*-site phages had three (Le Marrec et al. 1997). Neither the host range nor the DNA replication modules were correlated with these differences in DNA packaging and virion structure. In line with the modular theory of phage evolution developed 40 years ago from comparisons of lambdoid coliphages (Botstein 1980), *S. thermophilus* phages are composed by either an Sfi21-like *cos*-site structural gene cluster or an Sfi11-like *pac*-site structural gene cluster that associates independently with either of two nonstructural gene clusters, represented by Sfi21 or 7201 phages (Lucchini et al. 1999a). In streptococcal phages, large modular DNA exchanges between phages were further modified by the accumulation of point mutations and the subsequent acquisition of small deletions and insertions (Desiere et al. 1998). Each module has thus a complex evolutionary history. The structural module of the Sfi11-like phages resembled *Lactococcus* phages (BK5-T, r1t, TP901-1), and to a lesser degree *Bacillus* phages (SPP1), mycobacteriophages, and *E. coli* phage λ, defining a gradient of relatedness (Lucchini et al. 1998). In contrast, the structural gene module from *S. thermophilus* phage Sfi21 shared a close relationship with *L. lactis* phage BK5-T and moderate protein sequence identity with various phages of low GC-content Gram-positive bacteria and even a gene constellation and distant protein relatedness with *E. coli* phage HK97 (Desiere et al. 1999). Notably, *E. coli* phages λ and HK97 represent two distinct modes of head assembly of virions. It thus seems that the dichotomy of Sfi21- versus Sfi11-like structural modules seen in *S. thermophilus* phages represents two distinct and very ancient modes of virion assembly that were invented by ancestor phages and transmitted through bacterial evolution to extant phages. Comparative phage genomics can thus retrace part of the evolutionary history of

phage modules, like capsid building or the establishment of the lysogenic state. A gradient of relatedness is seen with various degrees of nucleotide sequence identity between phages infecting bacteria that currently exchange DNA or exchanged DNA in a relatively recent past. Phage modules that share only protein sequence identity represent more distant relatives that were in genetic separation for sufficient time to lose nucleotide sequence identity. Phage modules separated by even greater evolutionary distances share only a common genome organization (synteny of genes) without any sequence identity. Additional evidence for an evolutionary relationship beyond any detectable sequence relationship is the discovery of a specific viral head protein fold motif named after the prototype *E.* coli phage HK97, which was found in viruses from all three domains of life by crystallographic structural analysis (Pietilä et al. 2013). Tailed phages are thus the result of both vertical and horizontal evolution, which explains the weblike phylogenies of phages (Brüssow and Desiere 2001). Structural genes are particularly well conserved, but synteny of the nonstructural genes in the lysogeny module was also seen across phages from diverse bacterial hosts (Lucchini et al. 1999b).

In this context, it is notable that dairy phages from both *L. lactis* and *S. thermophilus* share substantial protein sequence identity with phages from an exclusive human pathogen, *Streptococcus pyogenes* (Desiere et al. 2001a). Apparently, with respect to the genetic relationship between these phages, phylogenetic relationships between their host bacteria are more important than their current ecological separation. However, it is likewise significant that the genome *S. thermophilus* phage Sfi21 still exhibits a considerable degree of DNA sequence identity with that of *L. lactis* phage BK5-T: over half of the DNA packaging and head morphogenesis module (Desiere et al. 2001b). This was not a singular case: also *S. thermophilus* phage 7201 does still share a measurable level of DNA sequence similarity with *L. lactis* phage bIL286 (Proux et al. 2002). Due to the close phylogenetic relationship between their bacterial hosts, lactococcal and lactic streptococcal phages have shared ancestor phages or have exchanged DNA modules in a relatively recent past such that DNA sequence similarity was not entirely eroded by the relentless accumulation of point mutations. Interestingly, the relationship between Sfi21 and BK5-T phages (or 7201 and bIL286) is closer than between currently known *L. lactis* and *L. garvieae* phages. This observation suggests that sharing a common ecological niche allows substantial phage gene exchanges across bacterial genus barriers.

5.2 Phage 5093

When phage isolates from mozzarella cheese whey samples were screened with a DNA probe for a highly conserved *S. thermophilus* phage anti-receptor gene (Binetti et al. 2005), the nonreactive *S. thermophilus* phage 5093 was identified. Except for a distinct tail tip structure (distinct baseplate with flexible globular appendices instead of a long tail fiber), it showed the usual siphovirus morphology and shared the

overall genome organization of *S. thermophilus* phages. With only three gene replacements, the right genome half comprising the lysis, lysogeny, and DNA replication modules displayed a close relatedness with known *S. thermophilus* phages. In contrast, over the structural gene modules, phage 5093 shared relatedness with neither Sfi21- nor Sfi11-like *S. thermophilus* phages, but with phages isolated from evolutionary related species (*S. pneumoniae*, *S. pyogenes*, *S. gordonii*), which inhabit distinct ecological niches, namely, the nasopharynx, pharynx, and oral cavity of humans. Since these habitats are quite different from dairy environments, the role of recombination events within an interbreeding phage population is unlikely in phage 5093 evolution (Mills et al. 2011). A similar phage was isolated from a different collection (Szymczak et al. 2017).

5.3 The 987 Phage Group

When two groups applied a modified PCR test detecting *cos*- and *pac*-site *S. thermophilus* phages (Quiberoni et al. 2006) to large phage collections, only few nonreactive phage isolates were identified (McDonnell et al. 2016). Four variant phages were identified, which differed from Sfi21-, Sfi11-, and 5093-like phages by a shorter genome size of 33 kb and a substantially shorter tail length (140 vs. 250 nm). These so-called 987-like phages were closely related across their structural gene modules despite their distinct geographical origin, which was probably explained by the fact that they were all isolated from the same industrial starter. Over the structural modules, the best matches were with subgroup II phages of the lactococcal phage species P335 and notably not with other *S. thermophilus* phages. Over the nonstructural gene modules, similarities with S. thermophilus phage 7201 were seen. However, for these genes, substantial variations were observed between the individual 987-like phage isolates. Another group found two *S. thermophilus* phages that showed an even closer relatedness to subgroup II P335 lactococcal phages over the structural modules reaching >80% bp identity over a long, contiguous stretch of DNA packaging, head, and tail genes, excluding anti-receptor and tail appendage genes (Szymczak et al. 2017). Over the nonstructural genes, one of these variant phages aligned with the typical *S. thermophilus* phage DNA replication module, while the second variant phage displayed unattributed genes. 987-like streptococcal and subgroup II P335 lactococcal phages were able to absorb with reduced efficiency to the reciprocal heterologous host but were unable to infect them, demonstrating a strong species barrier effect for a direct genetic exchange even for such close cross-species infections (McDonnell et al. 2016; Szymczak et al. 2017). While cross-species infections apparently occur and punctuate the evolution of dairy phages, the events are of low frequency, and lactococcal and streptococcal phages do not represent a genetically interbreeding population even when sharing the same dairy environment.

Phylogenetic tree analysis based on whole genome nucleotide comparisons reveals currently four clusters of *S. thermophilus* phages. The majority belongs to

either the Sfi21- or the Sfi11-like phage group, while rare isolates constitute the 5093-like and 987-like phage groups, respectively. The 987-like phage group represents a side branch of the more numerous subgroup II P335 lactococcal phage cluster (McDonnell et al. 2016; Szymczak et al. 2017).

5.4 *An Outgroup:* S. salivarius *Phages*

Streptococcus salivarius was initially classified together with *S. thermophilus* in the same species, and both were considered as two subspecies. In the meanwhile, bacterial taxonomists have separated them into two different, although phylogenetically closely related, species. However, the ecological niches are totally different, therefore suggesting substantial niche adaptation. *S. thermophilus* has raw milk as habitat, while *S. salivarius* is a major inhabitant of the oral cavity and the small intestine from humans. It is thus not only interesting to compare phages from *L. lactis* and *S. thermophilus*, which share the same niche and are separated by a larger evolutionary distance, but also phages from *S. thermophilus* and *S. salivarius*, which share a closer phylogenetic relationship, but differ for niche specialization. Such a comparison is now possible with the report of one temperate *S. salivarius* phage and three *S. salivarius* prophages (Chou et al. 2017). The answer is clear-cut: with few exceptions of small hypothetical proteins, *S. salivarius* phage YMC-2011 has the closest homology along its entire genome with *S. thermophilus* phages. A phylogenetic tree analysis attributes YMC-2011 together with *S. salivarius* prophage JIM8777 to the Sfi21 branch of *S. thermophilus* phages, while two other *S. salivarius* prophages were attributed the 5093-like *S. thermophilus* phage branch.

6 Limitations and Opportunities for Population Genomics Work with Dairy Phages

As for *L. lactis* phages, a word of caution should be mentioned for the relative prevalence of *S. thermophilus* phages. Phages have mostly been isolated from industrial fermentation failures using relatively few defined *S. thermophilus* starters that were distributed worldwide to dairy factories by commercial starter strain companies. However, mozzarella cheese in Italy is still produced by the “madre naturale” technique where complex, undefined mixtures of thermophilic starters are used. The starters are obtained by controlled heating of raw milk from cows from spring pastures. These “starters” contain mostly, but not exclusively, *S. thermophilus* isolates. Titers greater than 10^6 phages/ml cheese whey have regularly been observed in mozzarella factories employing these starters and open vat fermentation.

After starter changes (rotation), new phages entered into the factory derived from raw milk since some phages survived the pasteurization process (Bruttin et al.1997a). Many of the cheese factory *S. thermophilus* phage isolates could not be typed as either Sfi21- or Sfi11-like phages (Brüssow et al. 1994a, b), suggesting that under these artisanal fermentation conditions, many more variant phages remain to be described. As raw milk is the ecological niche of *S. thermophiles*, such surveys are more likely to represent the natural variability of these phages.

The cheese factory phages were isolated on indicator strains or occasionally on *S. thermophilus* colonies isolated directly from the raw milk. Due to the rather narrow host range of *S. thermophilus* phages, few of the many different phages are identified by cultivation techniques. The introduction of metagenome sequencing to cheese whey samples is likely to change the situation since it allows phage detection by a culture-independent method and potentially allows a true population genomic analysis of phages (Muhammed et al. 2017). Population genetics studies with *S. thermophilus* phages have so far been limited to the sequencing of small genome segments. With that limited resolution, it could be demonstrated that the source of new phages invading a cheese factory was raw milk phages. In addition, it was observed that phages isolated from a single cheese factory over a 2-year survey period showed the same degree of sequencing differences as *S. thermophilus* phages isolated over 30 years from dairy factories located in different countries (Bruttin et al. 1997a). Extending the sequencing analysis to the whole genome will enable population genomics insights not possible with these older population genetics studies.

The broad background knowledge gained about dairy phages over the last decades, combined with the small size of dairy phage genomes, makes the dairy environment an ideal system for studying phage population genomics, even if this system represents an artificial, man-made industrial environment. With the discovery of the CRISPR-Cas system in dairy phages and their industrial starters, applied food microbiologists have already yielded fundamental insights for basic biology (see next section). Phage-host interaction at the genome level has already started with the analysis of phage genomes after the introduction of abortive infection plasmids in *L. lactis* (Labrie and Moineau 2007; Labrie et al. 2012) and with spacer acquisition in *S. thermophilus* after phage challenge (Achigar et al. 2017). The continuous struggle between phages and their hosts, between bacterial resistance and viral anti-resistance, and on evaluations of the cost of resistance has already started (Vale et al. 2015). Extending this analysis to the phage genome level will provide new insights. Phage population genomics under the strong selective pressure of newly introduced resistance mechanisms in dairy starters (Labrie and Moineau 2007; Labrie et al. 2012) is likely to become a fruitful research area in phage population genomics and possibly beyond.

7 *Streptococcus thermophilus* Phages and the CRISPR-Cas System

It is not for the first and probably not for the last time that phage research laid the foundation for the next revolution in molecular biology. Research on *E. coli* phages led to the discovery of restriction-modification enzymes, which became instrumental for molecular cloning. Research on *S. thermophilus* phages prepared the ground for an even more versatile instrument of targeted genetic manipulation potentially revolutionizing human gene therapy.

In 2005, clustered regularly interspaced short palindromic repeats (CRISPRs), composed of 25–50 bp repeats separated by unique sequence spacers of similar length, were found in *S. thermophilus*, located in the vicinity of *cas* (*C*RISPR-*as*sociated) genes encoding RNA-guided DNA endonuclease enzymes. The spacers showed homology with sequences from phages. Phage resistance of different strains correlated with the number of spacers in the CRISPR locus, which led to the hypothesis that the spacer elements are the traces of past invasions by extrachromosomal elements and represent a type of cell immunity against phage infection (Bolotin et al. 2005). Indeed, 2 years later, scientists demonstrated that after viral challenge, bacteria integrated new spacers derived from phage genomic sequences. Removal or addition of particular spacers modified the phage-resistant phenotype of the cell. The resistance specificity was determined by spacer-phage sequence similarity (Barrangou et al. 2007). When 124 *S. thermophilus* strains were studied, 109 unique spacer arrangements were observed across the 3 CRISPR loci. Most showed identity to phage sequences (77%), but identity with plasmid sequences (16%) was also found, while only few matched bacterial sequences. CRISPR loci evolved both via polarized addition of novel spacers after exposure to foreign genetic elements and via internal deletion of spacers (Horvath et al. 2008). Each CRISPR contains a ~100–500 bp leader element that typically includes a transcription promoter, followed by an array of captured ~35 bp sequences (spacers) sandwiched between copies of an identical ~35 bp direct repeat sequence. In general new spacers are added immediately downstream of the leader.

Interference is based on small RNAs carrying a spacer sequence. These RNAs guide the defense apparatus to foreign molecules carrying sequences that match the spacers. Soon it was demonstrated by in vivo experiments that the CRISPR1/Cas system specifically cleaves plasmid and bacteriophage double-stranded DNA within the proto-spacer at specific sites (Garneau et al. 2010). The conservation of proto-spacer adjacent motifs (PAMs) was a common theme for the most diverse CRISPR systems (Mojica et al. 2009). *S. thermophilus* CRISPR3/Cas system could be transferred into *Escherichia coli* and provided there heterologous protection against plasmid transformation and phage infection. The interference was sequence-specific, and mutations in the vicinity or within the proto-spacer adjacent motif (PAM) allowed plasmids to escape CRISPR-encoded immunity; in these experiments, *cas9* was the only *cas* gene necessary for CRISPR-encoded interference (Sapranauskas et al. 2011). The silencing of invading nucleic acids is executed by ribonucleoprotein complexes preloaded with small, interfering CRISPR RNAs

(crRNAs) that act as guides for targeting and degradation of foreign nucleic acid. The *S. thermophilus* CRISPR3/Cas system introduced a double-strand break at a specific site in DNA containing a sequence complementary to crRNA. DNA cleavage is executed by Cas9, which uses two distinct active sites to generate site-specific nicks on opposite DNA strands (Gasiunas et al. 2012). Cas9 co-purifies with an additional RNA molecule, tracrRNA (trans-activating CRISPR RNA), and it is the ternary Cas9-crRNA-tracrRNA complex that cleaves DNA (Karvelis et al. 2013). The CRISPR-Cas9 nuclease has been engineered by biotechnologists, and a *cas9* gene from the related bacterium *Streptococcus pyogenes* has now been repurposed for hyper-accurate genome editing in human cells (Chen et al. 2017).

The CRISPR-Cas systems have been categorized into three major types (I–III). Type I and II systems provide immunity against invading DNA. The type IIIA system in *S. thermophilus* type (StCsm) restricts the MS2 RNA phage and cuts RNA in vitro. Upon phage infection, crRNA-guided StCsm binds to the emerging transcript and recruits Cas10 DNase to the actively transcribed phage DNA, resulting in degradation of both the transcript and phage DNA, but not the host DNA (Kazlauskiene et al. 2016). The molecular details have recently been deciphered. Target RNA binding by the Csm effector complex of *S. thermophilus* triggers Cas10 to synthesize cyclic oligoadenylates. Acting as signaling molecules, cyclic oligoadenylates bind Csm6 to activate its nonspecific RNA degradation by allosteric activation (Kazlauskiene et al. 2017; Niewoehner et al. 2017). The CRISPR-Cas system reveals here striking conceptual similarity to oligoadenylate signaling in mammalian innate immunity. CRISPR-Cas also shows astonishing resemblance with the adaptive immune system by the acquisition of the spacer sequences specific for the infecting phages. The parallels with the adaptive immune go even further. Bacterial cells can acquire spacers not only from infectious phages (which would mostly kill the target bacterium before it had time enough to mount a CRISPR-Cas-based resistance system) but also from defective phages at a rate directly proportional to the quantity of replication-deficient phages to which the cells are exposed (Hynes et al. 2014). This process reminded the researchers of immunization in humans by vaccination with inactivated viruses.

Beyond its enormous impact on biology and biotechnology in general, CRISPR-Cas systems have also delivered important insights for dairy microbiologists and for phage population genomics. Bacteriophage-insensitive mutants (BIMs) of a *S. thermophilus* yogurt starter were generated with the same phage in different phage challenge experiments. Each BIM acquired unique spacer regions ranging between one and four new spacers in CRISPR1. Formation of second-generation BIMs did not lead to increases in spacer numbers, but to alterations in spacer regions (Mills et al. 2010). In another study of 23 spontaneous BIMs, all of them had acquired at least 1 new spacer in their CRISPR1 array. While 14 BIMs had acquired spacer at the 5′-end of the array, 9 other BIMs acquired a spacer within the array (Achigar et al. 2017), challenging the concept of preferential spacer insertion at the 5′- end. The diversification and host-phage coevolution in a population derived from a single colony were also characterized after 1 week of co-culturing using metagenome sequencing approaches. The acquisition of new spacers led to a genetically diverse population with multiple subdominant strain lineages.

Phage mutations that circumvented the interference were localized in or near the proto-spacer adjacent motif (PAM) indicating a strong selection force on these phage regions (Sun et al. 2013). A strong and reproducible bias in the phage genome locations from which spacers derive was also observed in a further report. Spacers that target the host chromosome are infrequent and strongly selected against, suggesting autoimmunity is lethal. The researchers observed early dominance by a few spacer subpopulations and rapid oscillations in subpopulation abundances (Paez-Espino et al. 2013). Strains that acquired a single spacer showed only an incomplete resistance phenotype (Levin et al. 2013). Increased resistance can also be obtained by combining different resistance systems: restriction-modification (R-M) and CRISPR-Cas systems are compatible and act together to increase the overall phage resistance. Specifically, methylation of phage DNA does not impair CRISPR-Cas acquisition or interference activities (Dupuis et al. 2013).

Mathematical models revealed a highly complex coevolutionary dynamics in the virus-host arms race, with viruses escaping resistance and hosts reacquiring it through the capture of new spacers, when taking fitness cost of CRISPR-Cas systems into account (Koonin and Wolf 2015). Others developed an eco-evolutionary model called distributed immunity – the coexistence of multiple, equally fit immune alleles among individuals in a microbial population – and how it emerges and fluctuates in multi-strain communities of hosts and viruses. Distributed immunity promoted in this model sustained diversity and stability in host communities and decreased viral population density that could lead to viral extinction (Childs et al. 2014).

The fitness costs of two type II functional CRISPR-Cas systems were experimentally measured in *S. thermophilus* with growth assays in isolation or in pairwise competition. Cas protein expression was particularly costly, as Cas-deficient mutants achieved higher competitive abilities than the wild-type strain with functional Cas proteins. Increasing immune memory by acquiring more than one and up to four phage-derived spacers was not associated with fitness costs, while the activation of the CRISPR-Cas system during phage exposure induced a significant, but small, fitness costs (Vale et al. 2015). Long-term *S. thermophilus*-phage coevolution experiments followed by massive deep sequencing demonstrated that CRISPR immunity drives fixation of single nucleotide polymorphisms that accumulated exclusively in phage genome regions targeted by CRISPR. The presence of multiple phages increased phage persistence by enabling recombination-based formation of chimeric phage genomes in which sequences heavily targeted by CRISPR were replaced. These observations identify CRISPR as one of the fundamental drivers of phage evolution (Paez-Espino et al. 2015).

8 Future Perspectives

In the following outlook, we take a quick bird's eye view on contemporary biological research and suggest phage research as a suitable approach to bring mechanistically oriented molecular genetics and phylogeny- and ecology-oriented genomics together.

In physics, natural phenomena are conceptually reduced to basic principles described by a few fundamental laws with the ultimate aim to be summarized in a universally applicable "world formula." Such an approach is not meaningful in biology. Biological phenomena – as we know them today – lack universality since they are limited to the special conditions observed on a single planet. However, despite this drastic limitation of biological phenomena in physical space, biology is characterized by an exuberant diversity of phenomena manifested as many life forms beyond the grasp of any single biologist. There are millions of eukaryotic species and a currently still unchartered number of prokaryotic "species." Even more disturbing, the extant organisms are only a small fraction of the species, which have populated the planet through evolutionary time periods. Biology has thus elements of a historical science with the uniqueness of a given historical situation which upon a replay would result in a different constellation. The history of organisms can only indirectly be deduced from the study of fossils and retro-projections of genome analyses.

The only unifying theory of biology is currently that of evolution formulated by Charles Darwin and his modern followers. Yet, when reducing biology to abstract principles, the most interesting and defining aspects of biology are neglected, namely, the filling of all habitable niches on earth by different life forms. Biology strives to define the position of a constantly evolving organism in its given ecological niche, which is highly variable due to changing physicochemical conditions of the niche over time and varying under the pressure exerted by a myriad of other competing biological organisms, which are also in a constant flux.

To come research-wise to grip with this complexity, physicists have introduced the reductionist principle into biological research by taking out a few living elements from the environment and by studying them in the splendid isolation of simple, constant, and defined laboratory conditions. The most spectacular and most successful of these reductionist approaches in biology was the study of a handful of coliphages infecting a few *E. coli* strains in a defined broth culture or on the Petri dish. We owe to this conceptional approach the molecular biology revolution, which has fundamentally changed biology. With the current and still ongoing analytical technology revolution in biology, which started with the sequencing and omics technologies, biology has entered another era, which addresses increasingly complex systems. By sequencing techniques, huge datasets are now created that aim to describe entire ecosystems. Some biologists have expressed the concern that these datasets will outstrip our intellectual capacity to interpret them. With this type of research, frequently we can't see the wood for the trees. Or as expressed more poetically by Goethe in his "Faust" play, "Wer will was Lebendigs erkennen und beschreiben/ sucht erst den Geist heraus zu treiben/ Dann hat er die Teile in seiner Hand/ Fehlt leider! Nur das geistige Band" (to docket living things past any doubt/ You cancel first the living spirit out/The parts lie in the hollow of your hand/You only lack the living link you banned). Clearly, we need new overarching concepts which allow us to reduce the apparent complexity that prevents us from seeing the "wood." Some proposals are emerging; for example, it was suggested that the global pool of available metabolic functions, rather than the distribution of functions among

organisms, drives community assembly (Coles et al. 2017). This concept could rationalize the apparent randomness and fluctuation in the microbial species composition in the ocean or in the gut microbiota. According to this concept, the entire ecosystem at least in model calculations is evolving in an understandable way, while randomness, if not chaos, might blur our view when we analyze these systems from the perspective of individual species constituting the ecosystem (Sarker et al. 2017). Another possible solution could be found in a combination of a reductionist principle with the current trend for ecological complexity. Phages, due to their position at the lowest level of biological complexity, might permit such a compromise when being investigated at the population genomics level in natural environments like the ocean or the intestine of humans or animals. Time series of virome metagenome sequences from the gut of individual animals or humans could represent meaningful first steps for future phage ecology studies, which reach beyond the artificial environment of the man-made dairy factory, opening new vistas for phage population genomics. Such longitudinal phage population genomics studies might reveal the dynamics of such systems much better than cross-sectional studies. It might even be desirable to trigger microbiota changes by targeted interventions to induce changes in phage composition, allowing dynamic studies of phages. Recently, we have, for example, observed that systemic antibiotic application in malnourished diarrhea patients leads to an outgrowth of *E. coli* in their gut. In some patients this microbiota change is mirrored by an outgrowth of two specific coliphage types (Kieser et al. 2018). The microbiota changes seen in these patients showed striking parallels to observations in a mouse model of *Salmonella* infection (Faber et al. 2016). Clinical observations can thus be integrated into the large body of knowledge acquired for coliphage *E. coli* host interaction in vitro and in animal models. Perhaps coliphage population genomics studies in such patients could become appropriate bridge points into the interpretation of the large and complex datasets of contemporary metagenome research.

Acknowledgment The author thanks Douwe van Sinderen (University College Cork, Ireland) and Shawna McCallin (University of Lausanne, Switzerland) for their critical reading of the manuscript and many useful comments.

References

Achigar R, Magadán AH, Tremblay DM, Julia Pianzzola M, Moineau S. Phage-host interactions in Streptococcus thermophilus: genome analysis of phages isolated in Uruguay and ectopic spacer acquisition in CRISPR array. Sci Rep. 2017;7:43438.

Ackermann HW. Bacteriophage observations and evolution. Res Microbiol. 2003;154(4):245–51.

Ackermann HW. Classification of bacteriophages. In: Calendar R, editor. The bacteriophages. Oxford: Oxford University Press; 2006. p. 8–16.

Ackermann HW, Kropinski AM. Curated list of prokaryote viruses with fully sequenced genomes. Res Microbiol. 2007;158(7):555–66.

Ackermann HW, DuBow MS, Jarvis AW, Jones LA, Krylov VN, Maniloff J, Rocourt J, Safferman RS, Schneider J, Seldin L. The species concept and its application to tailed phages. Arch Virol. 1992;124(1–2):69–82.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007;315(5819):1709–12.

Binetti AG, Del Río B, Martín MC, Alvarez MA. Detection and characterization of Streptococcus thermophilus bacteriophages by use of the antireceptor gene sequence. Appl Environ Microbiol. 2005;71(10):6096–103.

Blatny JM, Godager L, Lunde M, Nes IF. Complete genome sequence of the Lactococcus lactis temperate phage phiLC3: comparative analysis of phiLC3 and its relatives in lactococci and streptococci. Virology. 2004;318(1):231–44.

Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology. 2005;151(Pt 8):2551–61.

Botstein D. A theory of modular evolution for bacteriophages. Ann N Y Acad Sci. 1980;354:484–90.

Bouchard JD, Moineau S. Homologous recombination between a lactococcal bacteriophage and the chromosome of its host strain. Virology. 2000;270(1):65–75.

Bourdin G, Navarro A, Sarker SA, Pittet AC, Qadri F, Sultana S, Cravioto A, Talukder KA, Reuteler G, Brüssow H. Coverage of diarrhoea-associated Escherichia coli isolates from different origins with two types of phage cocktails. Microb Biotechnol. 2014;7(2):165–76.

Braun V, Hertwig S, Neve H, Geis A, Teuber M. Taxonomic differentiation of bacteriophages of Lactococcus lactis by electron microscopy, DNA-DNA hybridization, and protein profiles. J Gen Microbial. 1989;135:2551–60.

Brüssow H. Phages of dairy bacteria. Annu Rev Microbiol. 2001;55:283–303.

Brüssow H. The impact of phages on the evolution of bacterial pathogenicity. In: Pallen M, Nelson KE, Preston GM, editors. Bacterial pathogenomics. Washington: ASM Press; 2007. p. 267–300.

Brüssow H. Phage-bacterium co-evolution and its implication for bacterial pathogenesis. In: Hensel M, Schmidt H, editors. Horizontal gene transfer in the evolution of pathogenesis. New York: Cambridge University Press; 2008. p. 49–77.

Brüssow H. The not so universal tree of life or the place of viruses in the living world. Philos Trans R Soc Lond Ser B Biol Sci. 2009;364(1527):2263–74.

Brüssow H, Bruttin A. Characterization of a temperate Streptococcus thermophilus bacteriophage and its genetic relationship with lytic phages. Virology. 1995;212(2):632–40.

Brüssow H, Desiere F. Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages. Mol Microbiol. 2001;39(2):213–22.

Brüssow H, Desiere F. Evolution of tailed phages: insights from comparative phage genomics. In: Calendar R, editor. The bacteriophages. Oxford: Oxford University Press; 2006. p. 26–36.

Brüssow H, Hendrix RW. Phage genomics: small is beautiful. Cell. 2002;108(1):13–6.

Brüssow H, Fremont M, Bruttin A, Sidoti J, Constable A, Fryder V. Detection and classification of Streptococcus thermophilus bacteriophages isolated from industrial milk fermentation. Appl Environ Microbiol. 1994a;60(12):4537–43.

Brüssow H, Probst A, Frémont M, Sidoti J. Distinct Streptococcus thermophilus bacteriophages share an extremely conserved DNA fragment. Virology. 1994b;200(2):854–7.

Brüssow H, Canchaya C, Hardt WD. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev. 2004;68(3):560–602.

Bruttin A, Brüssow H. Site-specific spontaneous deletions in three genome regions of a temperate Streptococcus thermophilus phage. Virology. 1996;219(1):96–104.

Bruttin A, Desiere F, d'Amico N, Guérin JP, Sidoti J, Huni B, Lucchini S, Brüssow H. Molecular ecology of Streptococcus thermophilus bacteriophage infections in a cheese factory. Appl Environ Microbiol. 1997a;63(8):3144–50.

Bruttin A, Foley S, Brüssow H. The site-specific integration system of the temperate Streptococcus thermophilus bacteriophage phiSfi21. Virology. 1997b;237(1):148–58.

Cairns J, Stent GS, Watson JD, editors. Phage and the origins of molecular biology. New York: Cold Spring Harbor Laboratory of Quantitative Biology; 1966.

Campbell A. General aspects of lysogeny. In: Calendar R, editor. The bacteriophages. Oxford: Oxford University Press; 2006. p. 66–73.

Canchaya C, Proux C, Fournous G, Bruttin A, Brüssow H. Prophage genomics. Microbiol Mol Biol Rev. 2003;67(2):238–76.
Casjens S, Hendrix R. Comments on the arrangement of the morphogenetic genes of bacteriophage lambda. J Mol Biol. 1974;90(1):20–5.
Castro-Nallar E, Chen H, Gladman S, Moore SC, Seemann T, Powell IB, Hillier A, Crandall KA, Chandry PS. Population genomics and phylogeography of an Australian dairy factory derived lytic bacteriophage. Genome Biol Evol. 2012;4(3):382–93.
Cavanagh D, Guinane CM, Neve H, Coffey A, Ross RP, Fitzgerald GF, McAuliffe O. Phages of non-dairy lactococci: isolation and characterization of ΦL47, a phage infecting the grass isolate Lactococcus lactis ssp. cremoris DPC6860. Front Microbiol. 2014;4:417.
Cavanagh D, Fitzgerald GF, McAuliffe O. From field to fermentation: the origins of Lactococcus lactis and its domestication to the dairy environment. Food Microbiol. 2015;47:45–61.
Chandry PS, Davidson BE, Hillier AJ. Temporal transcription map of the Lactococcus lactis bacteriophage sk1. Microbiology. 1994;140(Pt 9):2251–61.
Chen JS, Dagdas YS, Kleinstiver BP, Welch MM, Sousa AA, Harrington LB, Sternberg SH, Joung JK, Yildiz A, Doudna JA. Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. Nature. 2017;550(7676):407–10.
Childs LM, England WE, Young MJ, Weitz JS, Whitaker RJ. CRISPR-induced distributed immunity in microbial populations. PLoS One. 2014;9(7):e101710.
Chopin A, Bolotin A, Sorokin A, Ehrlich SD, Chopin M. Analysis of six prophages in Lactococcus lactis IL1403: different genetic structure of temperate and virulent phage populations. Nucleic Acids Res. 2001;29(3):644–51.
Chopin A, Deveau H, Ehrlich SD, Moineau S, Chopin MC. KSY1, a lactococcal phage with a T7-like transcription. Virology. 2007;365(1):1–9.
Chou WC, Huang SC, Chiu CH, Chen YM. YMC-2011, a temperate phage of streptococcus salivarius 57.I. Appl Environ Microbiol. 2017;83(6):e03186–16.
Coles VJ, Stukel MR, Brooks MT, Burd A, Crump BC, et al. Ocean biogeochemistry modeled with emergent trait-based genomics. Science. 2017;358(6367):1149–54.
Desiere F, Lucchini S, Bruttin A, Zwahlen MC, Brüssow H. A highly conserved DNA replication module from Streptococcus thermophilus phages is similar in sequence and topology to a module from Lactococcus lactis phages. Virology. 1997;234(2):372–82.
Desiere F, Lucchini S, Brüssow H. Evolution of Streptococcus thermophilus bacteriophage genomes by modular exchanges followed by point mutations and small deletions and insertions. Virology. 1998;241(2):345–56.
Desiere F, Lucchini S, Brüssow H. Comparative sequence analysis of the DNA packaging, head, and tail morphogenesis modules in the temperate cos-site Streptococcus thermophilus bacteriophage Sfi21. Virology. 1999;260(2):244–53.
Desiere F, McShan WM, van Sinderen D, Ferretti JJ, Brüssow H. Comparative genomics reveals close genetic relationships between phages from dairy bacteria and pathogenic streptococci: evolutionary implications for prophage-host interactions. Virology. 2001a;288(2):325–41.
Desiere F, Mahanivong C, Hillier AJ, Chandry PS, Davidson BE, Brüssow H. Comparative genomics of lactococcal phages: insight from the complete genome sequence of Lactococcus lactis phage BK5-T. Virology. 2001b;283(2):240–52.
Deveau H, Labrie SJ, Chopin MC, Moineau S. Biodiversity and classification of lactococcal phages. Appl Environ Microbiol. 2006;72(6):4338–46.
Doolittle WF. Phylogenetic classification and the universal tree. Science. 1999;284(5423):2124–9.
Dupuis ME, Moineau S. Genome organization and characterization of the virulent lactococcal phage 1358 and its similarities to Listeria phages. Appl Environ Microbiol. 2010;76(5):1623–32.
Dupuis MÈ, Villion M, Magadán AH, Moineau S. CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. Nat Commun. 2013;4:2087.
Eraclio G, Tremblay DM, Lacelle-Côté A, Labrie SJ, Fortina MG, Moineau S. A virulent phage infecting Lactococcus garvieae, with homology to Lactococcus lactis phages. Appl Environ Microbiol. 2015;81(24):8358–65.

Eraclio G, Fortina MG, Labrie SJ, Tremblay DM, Moineau S. Characterization of prophages of Lactococcus garvieae. Sci Rep. 2017;7(1):1856.

Evershed RP, Payne S, Sherratt AG, Copley MS, Coolidge J, Urem-Kotsu D, Kotsakis K, Ozdoğan M, Ozdoğan AE, Nieuwenhuyse O, Akkermans PM, Bailey D, Andeescu RR, Campbell S, Farid S, Hodder I, Yalman N, Ozbaşaran M, Biçakci E, Garfinkel Y, Levy T, Burton MM. Earliest date for milk use in the near east and southeastern Europe linked to cattle herding. Nature. 2008;455(7212):528–31.

Faber F, Tran L, Byndloss MX, Lopez CA, Velazquez EM, et al. Host-mediated sugar oxidation promotes post-antibiotic pathogen expansion. Nature. 2016;534(7609):697–9.

Filée J, Tétart F, Suttle CA, Krisch HM. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. Proc Natl Acad Sci U S A. 2005;102(35):12471–6.

Foley S, Bruttin A, Brüssow H. Widespread distribution of a group I intron and its three deletion derivatives in the lysin gene of Streptococcus thermophilus bacteriophages. J Virol. 2000;74(2):611–8.

Fortier LC, Bransi A, Moineau S. Genome sequence and global gene expression of Q54, a new phage species linking the 936 and c2 phage species of Lactococcus lactis. J Bacteriol. 2006;188(17):6101–14.

Garneau JE, Tremblay DM, Moineau S. Characterization of 1706, a virulent phage from Lactococcus lactis with similarities to prophages from other Firmicutes. Virology. 2008;373(2):298–309.

Garneau JE, Dupuis MÈ, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadán AH, Moineau S. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010;468(7320):67–71.

Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A. 2012;109(39):E2579–86.

Ghasemi SM, Bouzari M, Yoon BH, Chang HI. Comparative genomic analysis of Lactococcus garvieae phage WP-2, a new member of Picovirinae subfamily of Podoviridae. Gene. 2014;551(2):222–9.

Gottesman M, Oppenheim A. Lysogeny and prophage. In: Granoff A, Webster RG, editors. Encyclopedia of virology. 2nd ed. San Diego: Academic Press; 1999. p. 925–33.

Grose JH, Casjens SR. Understanding the enormous diversity of bacteriophages: the tailed phages that infect the bacterial family Enterobacteriaceae. Virology. 2014;468–470:421–43.

Hayes S, Mahony J, Nauta A, van Sinderen D. Metagenomic approaches to assess bacteriophages in various environmental niches. Virus. 2017;9(6):E127.

Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci U S A. 1999;96(5):2192–7.

Hill C, Miller LA, Klaenhammer TR. In vivo genetic exchange of a functional domain from a type II A methylase between lactococcal plasmid pTR2030 and a virulent bacteriophage. J Bacteriol. 1991;173(14):4363–70.

Hoai TD, Nishiki I, Yoshida T. Properties and genomic analysis of Lactococcus garvieae lysogenic bacteriophage PLgT-1, a new member of Siphoviridae, with homology to Lactococcus lactis phages. Virus Res. 2016;222:13–23.

Hols P, Hancy F, Fontaine L, Grossiord B, Prozzi D, Leblond-Bourget N, Decaris B, Bolotin A, Delorme C, Dusko Ehrlich S, Guédon E, Monnet V, Renault P, Kleerebezem M. New insights in the molecular biology and physiology of Streptococcus thermophilus revealed by comparative genomics. FEMS Microbiol Rev. 2005;29(3):435–63.

Horvath P, Romero DA, Coûté-Monvoisin AC, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R. Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. J Bacteriol. 2008;190(4):1401–12.

Hynes AP, Villion M, Moineau S. Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages. Nat Commun. 2014;5:4399.

Jarvis AW. Differentiation of lactic streptococcal phages into phage species by DNA-DNA homology. Appl Environ Microbiol. 1984;47(2):343–9.

Josephsen J, Andersen N, Behrndt E, Brandsborg E, Christinasen G, Hansen MB, Hansen S, Nielsen EW, Vogensen FK. An ecological study of lytic bacteriophages of Lactococcus lactis subsp. Cremoris isolated in a cheese plant over a five year period. Int Dairy J. 1994;4:123–40.

Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J Mol Biol. 2000;299(1):27–51.

Karvelis T, Gasiunas G, Miksys A, Barrangou R, Horvath P, Siksnys V. crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus. RNA Biol. 2013;10(5):841–51.

Kazlauskiene M, Tamulaitis G, Kostiuk G, Venclovas Č, Siksnys V. Spatiotemporal control of type III-A CRISPR-Cas immunity: coupling DNA degradation with the target RNA recognition. Mol Cell. 2016;62(2):295–306.

Kazlauskiene M, Kostiuk G, Venclovas Č, Tamulaitis G, Siksnys V. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. Science. 2017;357(6351):605–9.

Kelleher P, Bottacini F, Mahony J, Kilcawley KN, van Sinderen D. Comparative and functional genomics of the Lactococcus lactis taxon; insights into evolution and niche adaptation. BMC Genomics. 2017;18(1):267.

Kelly WJ, Altermann E, Lambie SC, Leahy SC. Interaction between the genomes of Lactococcus lactis and phages of the P335 species. Front Microbiol. 2013;4:257.

Kieser S, Sarker SA, Berger B, Sultana S, Chisti MJ, et al. Antibiotic treatment leads to fecal Escherichia coli and coliphage expansion in severely malnourished diarrhea patients. Cell Mol Gastroenterol Hepatol. 2018. https://doi.org/10.1016/j.jcmgh.2017.11.014. (in press).

Koonin EV, Wolf YI. Evolution of the CRISPR-Cas adaptive immunity systems in prokaryotes: models and observations on virus-host coevolution. Mol BioSyst. 2015;11(1):20–7.

Kot W, Neve H, Vogensen FK, Heller KJ, Sørensen SJ, Hansen LH. Complete genome sequences of four novel Lactococcus lactis phages distantly related to the rare 1706 phage species. Genome Announc. 2014;2(4):e00265–14.

Kotsonis SE, Powell IB, Pillidge CJ, Limsowtin GK, Hillier AJ, Davidson BE. Characterization and genomic analysis of phage asccphi28, a phage of the family Podoviridae infecting Lactococcus lactis. Appl Environ Microbiol. 2008;74(11):3453–60.

Krupovic M, Dutilh BE, Adriaenssens EM, Wittmann J, Vogensen FK, Sullivan MB, Rumnieks J, Prangishvili D, Lavigne R, Kropinski AM, Klumpp J, Gillis A, Enault F, Edwards RA, Duffy S, Clokie MR, Barylski J, Ackermann HW, Kuhn JH. Taxonomy of prokaryotic viruses: update from the ICTV bacterial and archaeal viruses subcommittee. Arch Virol. 2016;161(4):1095–9.

Kwan T, Liu J, DuBow M, Gros P, Pelletier J. The complete genomes and proteomes of 27 Staphylococcus aureus bacteriophages. Proc Natl Acad Sci U S A. 2005;102(14):5174–9.

Labrie SJ, Josephsen J, Neve H, Vogensen FK, Moineau S. Morphology, genome sequence, and structural proteome of type phage P335 from Lactococcus lactis. Appl Environ Microbiol. 2008;74(15):4636–44.

Labrie SJ, Moineau S. Abortive infection mechanisms and prophage sequences significantly influence the genetic makeup of emerging lytic lactococcal phages. J Bacteriol. 2007;189(4):1482–7.

Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol. 2010;8(5):317–27.

Labrie SJ, Tremblay DM, Moisan M, Villion M, Magadán AH, Campanacci V, Cambillau C, Moineau S. Involvement of the major capsid protein and two early-expressed phage genes in the activity of the lactococcal abortive infection mechanism AbiT. Appl Environ Microbiol. 2012;78(19):6890–9.

Le Marrec C, van Sinderen D, Walsh L, Stanley E, Vlegels E, Moineau S, Heinze P, Fitzgerald G, Fayard B. Two groups of bacteriophages infecting Streptococcus thermophilus can be distinguished on the basis of mode of packaging and genetic determinants for major structural proteins. Appl Environ Microbiol. 1997;63(8):3246–53.

Lenski RE. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. ISME J. 2017;11(10):2181–94.

Levin BR, Moineau S, Bushman M, Barrangou R. The population and evolutionary dynamics of phage and bacteria with CRISPR-mediated immunity. PLoS Genet. 2013;9(3):e1003312.

Little JW. Gene regulatory circuitry of phage λ. In: Calendar R, editor. The bacteriophages. Oxford: Oxford University Press; 2006. p. 74–82.
Lubbers MW, Waterfield NR, Beresford TP, Le Page RW, Jarvis AW. Sequencing and analysis of the prolate-headed lactococcal bacteriophage c2 genome and identification of the structural genes. Appl Environ Microbiol. 1995;61(12):4348–56.
Lucchini S, Desiere F, Brüssow H. The structural gene module in Streptococcus thermophilus bacteriophage phi Sfi11 shows a hierarchy of relatedness to Siphoviridae from a wide range of bacterial hosts. Virology. 1998;246(1):63–73.
Lucchini S, Desiere F, Brüssow H. Comparative genomics of Streptococcus thermophilus phage species supports a modular evolution theory. J Virol. 1999a;73(10):8647–56.
Lucchini S, Desiere F, Brüssow H. Similarly organized lysogeny modules in temperate Siphoviridae from low GC content gram-positive bacteria. Virology. 1999b;263(2):427–35.
Mahony J, Martel B, Tremblay DM, Neve H, Heller KJ, Moineau S, van Sinderen D. Identification of a new P335 subgroup through molecular analysis of lactococcal phages Q33 and BM13. Appl Environ Microbiol. 2013;79(14):4401–9.
Mahony J, van Sinderen D. Current taxonomy of phages infecting lactic acid bacteria. Front Microbiol. 2014;5:7.
Mahony J, Randazzo W, Neve H, Settanni L, van Sinderen D. Lactococcal 949 group phages recognize a carbohydrate receptor on the host cell surface. Appl Environ Microbiol. 2015;81(10):3299–305.
Mahony J, Cambillau C, van Sinderen D. Host recognition by lactic acid bacterial phages. FEMS Microbiol Rev. 2017a;41(Supp 1):S16–26.
Mahony J, Moscarelli A, Kelleher P, Lugli GA, Ventura M, Settanni L, van Sinderen D. Phage biodiversity in artisanal cheese wheys reflects the complexity of the fermentation process. Virus. 2017b;9(3):E45.
Maniloff J, Ackermann HW, Jarvis A. Phage taxonomy and classification. In: Granoff A, Webster RG, editors. Encyclopedia of virology. 2nd ed. San Diego: Academic Press; 1999. p. 1221–8.
McDonnell B, Mahony J, Neve H, Hanemaaijer L, Noben JP, Kouwen T, van Sinderen D. Identification and analysis of a novel group of bacteriophages infecting the lactic acid bacterium *Streptococcus thermophilus*. Appl Environ Microbiol. 2016;82(17):5153–65.
McGrath S, Fitzgerald GF, van Sinderen D. Identification and characterization of phage-resistance genes in temperate lactococcal bacteriophages. Mol Microbiol. 2002;43(2):509–20.
Millen AM, Romero DA. Genetic determinants of lactococcal C2viruses for host infection and their role in phage evolution. J Gen Virol. 2016;97(8):1998–2007.
Mills S, Griffin C, Coffey A, Meijer WC, Hafkamp B, Ross RP. CRISPR analysis of bacteriophage-insensitive mutants (BIMs) of industrial Streptococcus thermophilus – implications for starter design. J Appl Microbiol. 2010;108(3):945–55.
Mills S, Griffin C, O'Sullivan O, Coffey A, McAuliffe OE, Meijer WC, Serrano LM, Ross RP. A new phage on the 'Mozzarella' block: bacteriophage 5093 shares a low level of homology with other *Streptococcus thermophilus* phages. Int Dairy J. 2011;21:963–9.
Moineau S, Fortier J, Ackermann HW, Pandian S. Characterization of lactococcal bacteriophages from Quebec cheese plants. Can J Microbiol. 1992;38:875–82.
Moineau S, Pandian S, Klaenhammer TR. Evolution of a Lytic Bacteriophage via DNA Acquisition from the Lactococcus lactis Chromosome. Appl Environ Microbiol. 1994;60(6):1832–41.
Mojica FJ, Díez-Villaseñor C, García-Martínez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology. 2009;155(Pt 3):733–40.
Muhammed MK, Kot W, Neve H, Mahony J, Castro-Mejía JL, Krych L, Hansen LH, Nielsen DS, Sørensen SJ, Heller KJ, van Sinderen D, Vogensen FK. Metagenomic analysis of dairy bacteriophages: extraction method and pilot study on whey samples derived from using undefined and defined mesophilic starter cultures. Appl Environ Microbiol. 2017;83(19):e00888–17.
Murphy J, Bottacini F, Mahony J, Kelleher P, Neve H, Zomer A, Nauta A, van Sinderen D. Comparative genomics and functional analysis of the 936 group of lactococcal Siphoviridae phages. Sci Rep. 2016;6:21345.
Niewoehner O, Garcia-Doval C, Rostøl JT, Berk C, Schwede F, Bigler L, Hall J, Marraffini LA, Jinek M. Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. Nature. 2017;548(7669):543–8.

Oliveira J, Mahony J, Lugli GA, Hanemaaijer L, Kouwen T, Ventura M, van Sinderen D. Genome sequences of eight prophages isolated from Lactococcus lactis dairy strains. Genome Announc. 2016;4(6):e00906–16.

Paez-Espino D, Morovic W, Sun CL, Thomas BC, Ueda K, Stahl B, Barrangou R, Banfield JF. Strong bias in the bacterial CRISPR elements that confer immunity to phage. Nat Commun. 2013;4:1430.

Paez-Espino D, Sharon I, Morovic W, Stahl B, Thomas BC, Barrangou R, Banfield JF. CRISPR immunity drives rapid phage genome evolution in Streptococcus thermophilus. MBio. 2015;6(2):e00262–15.

Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. Uncovering earth's virome. Nature. 2016;536(7617):425–30.

Passerini D, Beltramo C, Coddeville M, Quentin Y, Ritzenthaler P, Daveran-Mingot M-L, Le Bourgeois P. Genes but not genomes reveal bacterial domestication of Lactococcus lactis. PLoS One. 2010;5(12):e15306.

Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WR Jr, Hendrix RW, Hatfull GF. Origins of highly mosaic mycobacteriophage genomes. Cell. 2003;113(2):171–82.

Petrov VM, Ratnayaka S, Nolan JM, Miller ES, Karam JD. Genomes of the T4-related bacteriophages as windows on microbial genome evolution. Virol J. 2010;7:292.

Pietilä MK, Laurinmäki P, Russell DA, Ko CC, Jacobs-Sera D, Hendrix RW, Bamford DH, Butcher SJ. Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story. Proc Natl Acad Sci U S A. 2013;110(26):10604–9.

Pope WH, Bowman CA, Russell DA, Jacobs-Sera D, Asai DJ, Cresawn SG, Jacobs WR, Hendrix RW, Lawrence JG, Hatfull GF, Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science, Phage Hunters Integrating Research and Education, Mycobacterial Genetics Course. Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. elife. 2015;4:e06416.

Prevots F, Mata M, Ritzenthaler P. Taxonomic differentiation of 101 lactococcal bacteriophages and characterization of bacteriophages with unusually large genomes. Appl Environ Microbiol. 1990;56(7):2180–5.

Proux C, van Sinderen D, Suarez J, Garcia P, Ladero V, Fitzgerald GF, Desiere F, Brüssow H. The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like Siphoviridae in lactic acid bacteria. J Bacteriol. 2002;184(21):6026–36.

Quiberoni A, Tremblay D, Ackermann HW, Moineau S, Reinheimer JA. Diversity of Streptococcus thermophilus phages in a large-production cheese factory in Argentina. J Dairy Sci. 2006;89(10):3791–9.

Rakonjac J, O'Toole PW, Lubbers M. Isolation of lactococcal prolate phage-phage recombinants by an enrichment strategy reveals two novel host range determinants. J Bacteriol. 2005;187(9):3110–21.

Rohwer F. Global phage diversity. Cell. 2003;113(2):141.

Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E, Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M, Searson S, Cruaud C, Alberti A, Duarte CM, Gasol JM, Vaqué D, Tara Oceans Coordinators, Bork P, Acinas SG, Wincker P, Sullivan MB. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature. 2016;537(7622):689–93.

Samson JE, Moineau S. Characterization of Lactococcus lactis phage 949 and comparison with other lactococcal phages. Appl Environ Microbiol. 2010;76(20):6843–52.

Sapranauskas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res. 2011;39(21):9275–82.

Sarker SA, McCallin S, Barretto C, Berger B, Pittet AC, Sultana S, Krause L, Huq S, Bibiloni R, Bruttin A, Reuteler G, Brüssow H. Oral T4-like phage cocktail application to healthy adult volunteers from Bangladesh. Virology. 2012;434(2):222–32.

Sarker SA, Berger B, Deng Y, Kieser S, Foata F, et al. Oral application of Escherichia coli bacteriophage: safety tests in healthy and diarrheal children from Bangladesh. Environ Microbiol. 2017;19(1):237–50.

Siezen RJ, Starrenburg MJC, Boekhorst J, Renckens B, Molenaar D, van Hylckama Vlieg JET. Genome-scale genotype-phenotype matching of two Lactococcus lactis isolates from plants identifies mechanisms of adaptation to the plant niche. Appl Environ Microbiol. 2008;74(2):424–36.

Stahl FW, Murray NE. The evolution of gene clusters and genetic circularity in microorganisms. Genetics. 1966;53(3):569–76.

Stanley E, Walsh L, van der Zwet A, Fitzgerald GF, van Sinderen D. Identification of four loci isolated from two Streptococcus thermophilus phage genomes responsible for mediating bacteriophage resistance. FEMS Microbiol Lett. 2000;182(2):271–7.

Sun CL, Barrangou R, Thomas BC, Horvath P, Fremaux C, Banfield JF. Phage mutations in response to CRISPR diversification in a bacterial population. Environ Microbiol. 2013;15(2):463–70.

Szymczak P, Janzen T, Neves AR, Kot W, Hansen LH, Lametsch R, Neve H, Franz CM, Vogensen FK. Novel variants of *Streptococcus thermophilus* bacteriophages are indicative of genetic recombination among phages from different bacterial species. Appl Environ Microbiol. 2017;83(5):e02748–16.

Vale PF, Lafforgue G, Gatchitch F, Gardan R, Moineau S, Gandon S. Costs of CRISPR-Cas-mediated resistance in Streptococcus thermophilus. Proc Biol Sci. 2015;282(1812):20151270.

van Sinderen D, Karsens H, Kok J, Terpstra P, Ruiters MH, Venema G, Nauta A. Sequence analysis and molecular characterization of the temperate lactococcal bacteriophage r1t. Mol Microbiol. 1996;19(6):1343–55.

Villarreal LP. Are viruses alive? Sci Am. 2004;291(6):100–5.

Villion M, Chopin MC, Deveau H, Ehrlich SD, Moineau S, Chopin A. P087, a lactococcal phage with a morphogenesis module similar to an enterococcus faecalis prophage. Virology. 2009;388(1):49–56.

Wommack KE, Colwell RR. Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev. 2000;64(1):69–114.

Index

M. F. Polz, O. P. Rajora (eds.), *Population Genomics: Microorganisms*,
Population Genomics, https://doi.org/10.1007/978-3-030-04756-6