Kirk E. Lohmueller
Rasmus Nielsen *Editors*

Human Population Genomics

Introduction to Essential Concepts and Applications

Springer

Human Population Genomics

Kirk E. Lohmueller • Rasmus Nielsen
Editors

Human Population Genomics

Introduction to Essential Concepts and Applications

Springer

Editors

Kirk E. Lohmueller
Ecology and Evolutionary Biology
University of California
Los Angeles, CA, USA

Rasmus Nielsen
Integrative Biology
University of California
Berkeley, CA, USA

ISBN 978-3-030-61644-1 ISBN 978-3-030-61646-5 (eBook)
https://doi.org/10.1007/978-3-030-61646-5

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

To Walter and James
And
To Hannah and Milla

Preface

Human population genomics is an exciting and rapidly changing field. Over the past decade, there has been a deluge of genetic variation data from the entire genome of individuals from many populations. These data have allowed us an unprecedented look at human history and how natural selection has impacted humans during this journey. Simultaneously, there have been increased efforts to determine how genetic variation affects complex traits in humans. Due to technological and methodological advances, progress has been made at determining the architecture of complex traits.

Because the field is so rapidly evolving, it can become difficult for newcomers to gain footing in the morass of technical concepts of population genetics, complex trait genetics, and data analysis. This is unfortunate as human population genomics brings together researchers from disparate areas of expertise, including computer science, statistics, medicine, genetics, evolution, and anthropology.

This book represents our modest attempt to make the highly technical and contemporary aspects of the field of human population genomics more accessible to researchers in all these groups. We wanted to start with the basics, but then include more advanced and current research. Thus, we hope that this book can serve as a gateway to modern human population genetics research for those new to the field. We hope that seasoned practitioners will also find it as a useful reference on their shelf.

The book has three parts. The first part provides an introduction to essential concepts in population genetics. These chapters will be useful for learning some of the more arcane, but important, concepts in population genetics, which are relevant for any organism. The second part covers the genetics of complex traits in humans. The third part of the book focuses on applying these techniques and concepts to genetic variation data to learn about demographic history and natural selection in humans.

In Chap. 1, Wakeley provides an introduction to one of the most useful tools in population genetics, the coalescent. The coalescent is used ubiquitously throughout population genetics, and an understanding of it will help develop intuition about how evolutionary forces impact genetic variation. In Chap. 2, Slatkin provides a discussion of linkage disequilibrium. While linkage disequilibrium may have originally been an esoteric theoretical topic with little practical value, it has become an essential phenomenon for associating genetic variation with complex traits. In Chap. 3, Skoglund and colleagues discuss tools used to measure and quantify human population structure. While understanding human population structure is valuable in itself, it is vital to properly design association studies for complex traits. In Chap. 4, Enard provides an introduction to natural selection in humans and contemporary statistical methods used to detect positive selection.

Chapters 5–7 discuss medical genetic and association studies between genetic variants and complex traits. In Chap. 5, Graff and Witte discuss statistical methods used to associate genetic variants with complex phenotypes. This chapter provides an introduction to genome-wide association studies (GWAS) which have become a mainstay in human genetics. In Chap. 6, Thompson describes an exciting approach of using identity by descent (IBD) mapping for implicating genetic variants in disease risk. In Chap. 7, Voight provides a summary of some of the lessons that the human genetics community has learned from the early GWAS.

Lastly, Chaps. 8 and 9 provide an overview of human evolutionary genetics. In Chap. 8, Wall summarizes what we have learned about human history for the analysis of genome-wide genetic variation data. In Chap. 9, Emery and Akey provide an overview of what we have learned about how natural selection has shaped genetic variation across the human genome. They discuss some specific examples of human adaptations and what these selection scans tell us about human evolution.

We wish to thank the authors of these chapters as well as the editors at Springer for their continued patience with us. This project took nearly a decade to complete, and we are grateful for their support and helping us eventually complete the book.

Los Angeles, CA
Berkeley, CA
May 2020

Kirk E. Lohmueller
Rasmus Nielsen

Contents

Part I Population Genetic Theory

Part II Association Studies and Medical Genetics

Part III Human Evolutionary Population Genetics

Part I

Population Genetic Theory

Coalescent Models

1

John Wakeley

Abstract

The standard neutral coalescent model and its extensions to include changes in population size over time and population structure are reviewed. Gene genealogies are shown to provide the hidden structure behind patterns of genetic variation. Expressions for expected levels of genetic variation are presented and explained, and tests of the standard neutral model based on the frequencies of mutations at single-nucleotide sites (aka "site frequencies") are outlined. Several examples of deviations from the standard model are discussed, and their effects on expected site frequencies are illustrated. Some attention is given to the fact that coalescent theory has not fully grappled with the existence of underlying population pedigrees.

1.1 Aims and Clarifications

The goal of the coalescent theory is the same as that of population genetics, namely, to understand the forces which produce and maintain variation. The models presented in this chapter support this endeavor. They are abstract and idealized tools applicable to many different kinds of organisms or species. Due to the persistence of racism, studies of human diversity call for a great deal of sensitivity. It is not enough just to agree with Hochman (2019) that we have "changed the topic from 'race' to 'population.'" We have to clarify that we are not just switching words. Rather, if "race" is used to refer to a group of people, that group is not a "population" in the population-genetic sense.

J. Wakeley (✉)
Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA
e-mail: wakeley@fas.harvard.edu

K. E. Lohmueller, R. Nielsen (eds.), *Human Population Genomics*,
https://doi.org/10.1007/978-3-030-61646-5_1

Winther et al. (2015) discuss three common uses of "population": in mathematical models, in the laboratory, and in the wild. The populations in this chapter are of the first kind. They are theoretical constructs to be applied for the sake of better understanding. Their application may lead either to further advances in modeling or to new hypotheses about populations of actual organisms including ourselves. The framework is statistical and involves sampling from populations. It should be borne in mind that "the sample" in what follows means genetic data taken from a number of individuals.

As a matter of perspective, it is important to recognize the surprising overall truth about human genetic variation—that there is very little of it compared to what is found in most species (Leffler et al. 2012). Further, the degree of substructure among humans is remarkably low (Rosenberg et al. 2005). As a first approximation, it is not uncommon or unreasonable to compare global patterns of human genetic variation to the predictions for a single, well-mixed population.

1.2 Introduction: Gene Genealogies Within a Population or Species

Population-genetic datasets, with their typically complex and interesting patterns of polymorphism among the DNA sequences, haplotypes or genotypes in the sample, are the result of an equally complex and interesting set of ancestral genetic processes. Each single-nucleotide polymorphism (SNP) reflects the specific patterns of descent from the ancestors of the sample and the mutation(s) at that nucleotide site during genetic transmission. Patterns of descent from ancestors are influenced by the random processes of genetic transmission and a host of demographic processes which may include natural selection, population growth, and population structure. The data consist only of patterns of polymorphism, and the challenge is to use these to make whatever inferences we can about the underlying processes.

Forgetting mutations for the moment, the term gene genealogy refers to the pattern of genetic ancestry among the members of a sample at a single-nucleotide site or a genetic locus made up of a non-recombining sequence of sites. If there is intra-locus recombination, then gene genealogies at different sites may be different (see Chap. 2). In this chapter, gene genealogies are considered without intra-locus recombination. Under mild restrictions on the sample size, the population size, and the demography of the species, gene genealogies may be depicted accurately as rooted, bifurcating trees, with the samples at the tips and the most recent common ancestor, or MRCA, of the sample at the root. The branches represent the genetic lineages ancestral to the sample.

Figure 1.1 shows a hypothetical dataset and a corresponding gene genealogy. For a real dataset, the gene genealogy would be unknown, but it is clear from Fig. 1.1 that the structure of the gene genealogy is a very strong determinant of the patterns of mutations (e.g., the frequencies) in the sample. Fig. 1.1 depicts the simple case in which every mutation in the ancestry of the sample occurs at a different nucleotide

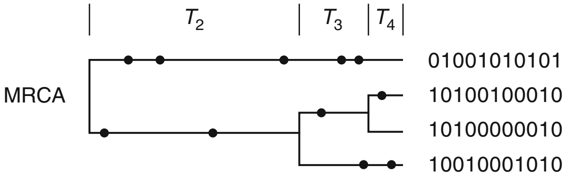


Fig. 1.1 A hypothetical gene genealogy of $n = 4$ sequences from the standard neutral coalescent model without recombination and assuming infinite-sites mutation. The coalescent intervals T_2, T_3, and T_4 are drawn in proportion to their expected values. Hypothetical DNA sequence data (or haplotypes) are coded such that the ancestral base at each site is denoted 0 and the mutant base is denoted 1

site. Thus, each dot (mutation) on the tree in Fig. 1.1 corresponds to exactly one SNP.

It is due to their position as intermediaries between patterns of polymorphism and population-level demographic processes that gene genealogies became important objects of study in the early 1980s. Hudson (1983) and Tajima (1983) initiated the study of gene genealogies in population genetics, on the stage set previously by Ewens (1972, 1974) and Watterson (1975). Together, these publications anticipated the current abundance of genetic data and laid the foundations for modern computational approaches to data analysis, which often make explicit use of gene genealogies and typically treat them as unknown "nuisance" parameters or hidden variables.

Work on gene genealogies ushered in a new way of thinking in population genetics, in which the classical models were turned around and viewed backward in time (Ewens 1990). The subfield of population genetics that treats gene genealogies is called a coalescent theory. For reviews, see Hein et al. (2005) and Wakeley (2009). The word coalescent captures the idea that the ancestral genetic lineages of a sample are imagined to join together in common ancestors (i.e., they coalesce) as they travel backward in time. Kingman (1982a, b, c) gave the formal mathematical proof of the existence of the standard neutral coalescent process, which is the same process Hudson (1983) and Tajima (1983) considered from a biological point of view.

The fruit of the study of gene genealogies may be seen, for example, in the work of Li and Durbin (2011), who modeled the distribution of SNPs across the genome in a sample of two (haploid) human genomes, taking into account the fact that recombination occurs across the genome. In standard neutral coalescent models, the shape of this distribution depends on the distribution of pairwise times to common ancestry across the genome, which in turn depends on the size of the population in each past generation. Li and Durbin (2011) applied a simulation-based method of inference, specifically a hidden Markov model (HMM) of times to common ancestry, to make detailed estimates of past human population sizes. Spence et al. (2018) review the development of such HMMs following Li and Durbin (2011).

The purpose of this chapter is to provide an intuitive introduction to the mathematics of coalescent theory. All the basic results of the standard neutral model for a single locus without recombination are presented and explained. The presentation begins with the idea that gene genealogies are embedded within organismal pedigrees. Although this is not a controversial idea, it is not, in fact, how gene genealogies are modeled in standard coalescent derivations.

1.2.1 Organismal Pedigrees and Gene Genealogies

Within any sexually reproducing species, such as humans, there exists a pattern of ancestry and descent which we may call the *population pedigree*. If it were known, the population pedigree would be a record of all reproduction events, connecting parents with their offspring and extending from the distant past to the present day. It would reflect the movement of individuals across the globe, changes in local population sizes over time, and events such as the selective sweeps of advantageous alleles through the population. Like the gene genealogy, the population pedigree is an unknown but important outcome of the population-level processes which affect genetic variation. Further, the gene genealogy at a locus without intra-locus recombination is simply the result of Mendelian segregation within the parts of the population pedigree relating to the sampled individuals.

The derivations of coalescent theory average over the unknown population pedigree within each generation, with details depending on the reproduction form assumed. For example, consider the probability that two gene copies or alleles at an autosomal locus, obtained by randomly sampling two individuals from the population without replacement, are descended from a common ancestor (i.e., they "coalesce") in the immediately previous generation. This quantity, which we may call c_N following Möhle (1998a) is fundamental in coalescent theory because it sets the timescale of the coalescent process. Under the diploid, dioecious Wright–Fisher model (Fisher 1930; Wright 1931) with random mating between the two sexes, we have

$$c_N = \frac{1}{8}\left(\frac{1}{N_\mathrm{f}} + \frac{1}{N_\mathrm{m}}\right)$$

in which N_f and N_m are the numbers of females and males in the population. In other words, coalescence occurs when the sampled individuals share a female parent ($1/N_f$) or a male parent ($1/N_\mathrm{m}$), and both samples come from that shared parent (1/4), and they descend from the same copy in that parent (1/2). The first two probabilities, $1/N_\mathrm{f}$ and $1/N_\mathrm{m}$, follow from the assumption of random mating, and the second two probabilities, 1/4 and 1/2, follow from the process of Mendelian segregation.

In the case that $N_\mathrm{f} = N_\mathrm{m} = N/2$, then we have $c_N = 1/(2N)$, which is identical to the result for the diploid, monoecious Wright–Fisher model. Although the details of the coalescence probability c_N depend on the details of reproduction, in general, c_N will depend inversely on the size of the population as it does in this example. Now, if we apply this same probability in every generation in the past, the number of generations back to the common ancestor of the sample is geometrically distributed:

$$P(g) = c_N(1 - c_N)^{g-1}. \tag{1.1}$$

Then on average, looking backward in time, it will take $1/c_N$ generations for the pair of genetic lineages ancestral to the sample to coalesce. Under the diploid, monoecious Wright–Fisher model, this would be $2N$ generations. This is what sets the timescale of the coalescent process. Time is usually measured in units proportional to N generations, e.g., $2N$ in this case of the diploid, monoecious Wright–Fisher model.

The derivation of Eq. (1.1) is incorrect for two reasons. First, it is exact only in the diploid, monoecious model if "random mating" includes the possibility of reproduction by selfing. When there are two sexes or if selfing is not possible, it is wrong to apply the same probability, c_N, in every generation because when the two lineages ancestral to the sample are in the same individual and they are distinct, they necessarily descend separately from the two parents. Consequently, the probability of coalescence in the immediately previous generation is equal to zero. In spite of this, the geometric distribution is still approximately correct as long as the population size is large (Möhle 1998b, c), because the two ancestral lineages will only be in the same individual a small number of times, while the number of generations back to their coalescence would be large, proportional to N generations on average.

The second reason has to do with how the population pedigree is treated in the derivation. Here it is important to recognize that coalescent theory is typically used to describe patterns of genetic variation across the genome, for example, as in Li and Durbin (2011) mentioned above. Because this chapter does not treat recombination, we may imagine a genomic dataset made up of sequences at a number of genetic loci within which there is no recombination but between which there is either completely independent assortment, as with different chromosomes, or effectively independent assortment, as with loci that far enough apart on the same chromosome. It is conceptually wrong to use results such as Eq. (1.1) because, in averaging over the process of reproduction, they do not capture the actual patterns of relatedness among the sampled individuals which are encoded in the population pedigree.

Figure 1.2, which is adapted from Fig. 3 in Wakeley et al. (2016), shows part of a larger pedigree for the Spanish Habsburg royal family reported in Alvarez et al. (2009). At a single genetic locus, two sequences sampled from Mary of Portugal and King Philip II would have zero chance of being descended from a common ancestral sequence in the immediately previous generation. In the grand-parental generation, however, the probability of coalescence is a substantial 1/8 due to the special relatedness of Mary and Philip as double first cousins. Since Mary and Philip also share one pair of great grandparents, there is a 1/32 chance of coalescence in the third generation in the past.

These probabilities are calculated by tracing each ancestral lineage back to the mother or the father of each individual with a 50:50 chance and letting two lineages coalesce with probability 1/2 whenever they trace back to the same individual. If

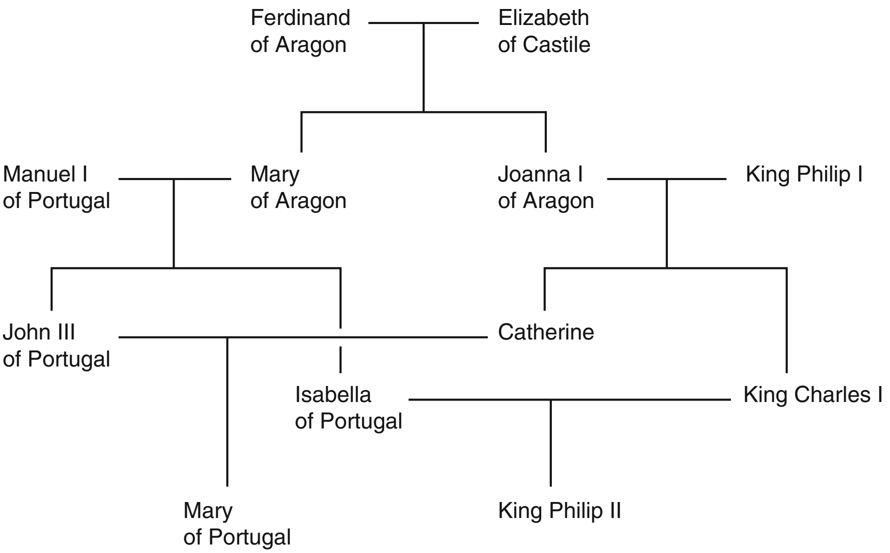


Fig. 1.2 Part of the pedigree of the Spanish Habsburg royal family, extracted from Fig. 1 in Alvarez et al. (2009) and redrawn

two lineages are in the same individual but are distinct, another uniform random choice decides which is maternal and which is paternal. Calculations like this have a long history and numerous uses in human genetics (Ott 1999) and can be used to compute any probability of interest on arbitrarily complicated pedigrees (Cannings et al. 1978).

Following these rules, which are just Mendel's laws viewed backward in time, also allows for the straightforward simulation of gene genealogies within population pedigrees. In the particular case of Fig. 1.2, if the genetic ancestry of a large number of loci were simulated beginning with one sample from Mary of Portugal and one from King Philip II, the results would show that no loci would have their MCRA in past generation one, 1/8 of loci would have their MCRA in past generation two, and 1/32 of loci would have their MCRA in past generation three.

The chances of common ancestry in each generation for the pedigree in Fig. 1.2 are markedly different from the predictions of the standard neutral coalescent model, with its constant probability of coalescence, c_N, in every generation. Averaging over the process of reproduction or equivalently over pedigrees is conceptually wrong because for any given species, there is in fact just one population pedigree. Sample ancestries might include relationships like those in Fig. 1.2, which standard coalescent models ignore. Again, the primary application of coalescent theory is to model the distribution of genetic variation across the genome within a sample. As this distribution is the outcome of transmission within a single, fixed population pedigree, coalescent theory should ideally model the distribution of gene genealogies in this way too.

Fortunately, it appears that inferences based on the standard coalescent model may often involve little error because the process of coalescence on a fixed pedigree is practically indistinguishable from the standard neutral coalescent process (Wakeley et al. 2012). This result comes from simulations of large well-mixed populations, so it is important to note that it may not hold for all possible extensions of coalescent modeling. For example, in those finely divided, structured populations, often referred to as "meta-populations" (Hanski and Gaggiotti 2004), the sizes of local subpopulations may be small, and averaging over pedigrees could give highly inaccurate results.

Standard coalescent models become accurate for large well-mixed populations because the ancestries of all present-day individuals overlap broadly (Chang 1999; Rohde et al. 2003). If all ancestors are distinct, then every individual will have 2^g pedigree ancestors in generation g in the past. Thus, just 40 generations, or perhaps 1000 years ago, we should each have more than one trillion ancestors. However, according to Fig. 1 of Keinan and Clark (2012), the number of people alive 1000 years ago was only about 100 million. For each of us, our $>10^{12}$ expected pedigree ancestors must all map onto 10^8 actual pedigree ancestors. This causes a huge degree of overlap of our ancestries. For a Wright–Fisher population of large constant size N, Chang (1999) found that by $1.77\log_2 N$ generations ago the population is divided neatly into two groups: a fraction (~0.7698) who are ancestors of every present-day individual and a fraction (~0.2302) who have no descendants today. For perspective, $1.77\log_2 N$ is roughly 35 generations for a population of size $N = 10^6$ and 47 generations for $N = 10^8$.

Roughly speaking, it is because of this broad overlap of pedigree ancestries in the relatively recent past that coalescent models based on the incorrect assumption of homogeneity of coalescent probabilities over time actually make reasonable predictions about the distribution of gene genealogies within fixed population pedigrees, at least for large well-mixed populations. Of course, the distribution of gene genealogies within fixed population pedigrees is not identical to the distribution of gene genealogies under the standard neutral coalescent. But the differences are primarily restricted to the past $\log_2 N$ generations, at which point there is a rapid transition to the type of homogeneous, essentially pedigree-independent behavior found in the standard model (Wakeley et al. 2012).

Random samples from large well-mixed populations are very unlikely to include closely related individuals, so the typical effect of the population pedigree is to bar coalescence in the very recent past. But since $\log_2 N$ generations is much less than the timescale of the coalescent process, i.e., N generations, the effects of population pedigrees on gene genealogies will often be negligible. This provides some justification for the common practice of discarding individuals when high levels of relatedness are detected in population-genetic data (Rosenberg 2006). However, it may be preferable to account for recent pedigrees explicitly, particularly in structured populations (Wilton et al. 2017). In cases where the pedigree itself is of interest, Ko and Nielsen (2019) describe how it can be estimated from genetic data.

1.3 The Standard Neutral Model: The Kingman Coalescent

The standard neutral coalescent model, also called the Kingman coalescent, begins with single-generation probabilities like c_N above and then takes advantage of the fact that a relatively simple model of the ancestral genetic process for a sample of size $n \geq 2$ holds for many different kinds of reproduction as long as the population size (N) is large and the sample size n is much smaller than N. Mathematically, this involves rescaling time so that it is measured in units of $1/c_N$ generations ($2N$ for the diploid, monoecious Wright–Fisher model) and then taking the limit $N \to \infty$. The standard coalescent is a backward-time dual process (see Möhle 1999) of the standard forward-time diffusion model of population genetics, as both models use this procedure of rescaling time and taking the limit $N \to \infty$ (Ewens 2004).

Kingman assumed a general family of haploid models of reproduction introduced by Cannings (1974) which includes the diploid, monoecious Wright–Fisher model. By studying all possible events in the immediate ancestry of a sample of size n, Kingman found—see Eq. 4.3 in Kingman (1982a)—that the most likely event is a coalescent event between a pair of lineages. Considering all possible pairs of lineages and averaging over the process of reproduction, the probability of a coalescent event is

$$\binom{n}{2}\frac{\sigma^2}{N} + O\left(\frac{1}{N^2}\right) \tag{1.2}$$

where n is the sample size, σ^2 is the variance of the number of offspring of a single (haploid) individual under the model or reproduction, and $\binom{n}{2}$ $=n(n-1)/2$ is the number of possible pairs of lineages. Equation (1.2) is written with large populations in mind. The exact probability is not captured entirely by the first term; $O(1/N^2)$ represents all remaining terms in a power series expansion of the coalescence probability, the largest of which is proportional to $1/N^2$. The other possible events, which involve more than two ancestral lineages coalescing in a single generation, have probabilities proportional to$1/N^2$ or smaller. As $N \to \infty$, all events and terms of order $1/N^2$ or smaller become negligible compared to the first term in Eq. (1.2).

Like the standard diffusion model, the standard coalescent process is a limiting model which is meant to capture the essential behavior of large populations. Its timescale is set by the probability of coalescence for a sample of size two. When time is rescaled in Eq. (1.1), so that it is measured in units of $1/c_N$ generations (which is, again, proportional to N generations) and then the limit $N \to \infty$ is taken, the geometric distribution $P(g)$ converges to an exponential distribution $f(t) = e^{-t}$. Thus, on the new timescale, coalescence occurs with a rate equal to 1 between the two ancestral lineages. Similarly, using the probability in Eq. (1.2) for a sample of arbitrary size n, the rate of coalescence becomes equal to $n(n-1)/2$. That is, each of the $n(n-1)/2$ pairs of lineages coalesces with a rate equal to 1 independently in the coalescent limit.

When the Kingman coalescent is the limiting ancestral process, it is useful to refer to the *coalescent effective population size* N_e (Sjödin et al. 2005), which is given by N/σ^2 in Kingman's derivation from the general Cannings' model, by the familiar $2N$ in the diploid Wright–Fisher model, under both the monoecious model and the dioecious model with equal numbers of males and females, and by $1/c_N$ in general.

Note that the statement $N \to \infty$ does not refer to changes in the size of the population. The population size N is assumed to be constant over time in the standard neutral coalescent model (later, one may relax this assumption). The limit simply means that we consider a series of such (constant-size) populations, with the aim of identifying the dominant behavior of the ancestral process when N is very large.

The standard neutral coalescent has been shown to be robust to many deviations from Kingman's initial assumptions (Möhle 1998a). It applies when generations are overlapping and to populations of diploid, biparental organisms. The latter case requires mathematical formalism beyond what Kingman used. This was developed in a pair of papers by Möhle which treated partial selfing (Möhle 1998b) and diploid, biparental inheritance (Möhle 1998c). In all these cases, the derivation of the coalescent begins with the description of an expected, single-generation process, which is the average over all possible outcomes of reproduction or over the pedigree.

1.3.1 The Sampling Structure of Coalescent Gene Genealogies

The end product of these calculations is a continuous-time model of the ancestral genetic process which begins with the n genetic lineages of the present-day sample and proceeds back into the past. Each pair of lineages coalesces independently with a rate equal to one, so that the total rate is $i(i-1)/2$ when there are i ancestral lineages. Again, $i(i-1)/2$ is the total number of pairs of lineages that can coalesce. Thus, the total rate of coalescence is higher when there are more lineages available to coalesce with each other. Coalescent events occur between randomly chosen pairs of lineages at randomly (exponentially) distributed times. The process is stopped when the last two ancestral lineages coalesce into a single lineage, the MRCA of the sample.

One run of this process produces a random-joining tree with associated branch lengths determined by the series of exponentially distributed coalescence times, which is taken to represent a single gene genealogy sampled from the distribution of all possible gene genealogies under the model. Multiple independent runs are used to represent collections of gene genealogies at multiple unlinked or effectively unlinked loci. Gene genealogies vary quite dramatically, in both branching structure and coalescence times (reflected in the heights of the genealogies). This is shown in Fig. 1.3 which displays ten randomly generated gene genealogies for a sample of size $n = 20$ under the standard neutral coalescent. A key purpose of coalescent theory is to model variation in gene genealogies, as in Fig. 1.3, reflecting the randomness of the evolutionary process.

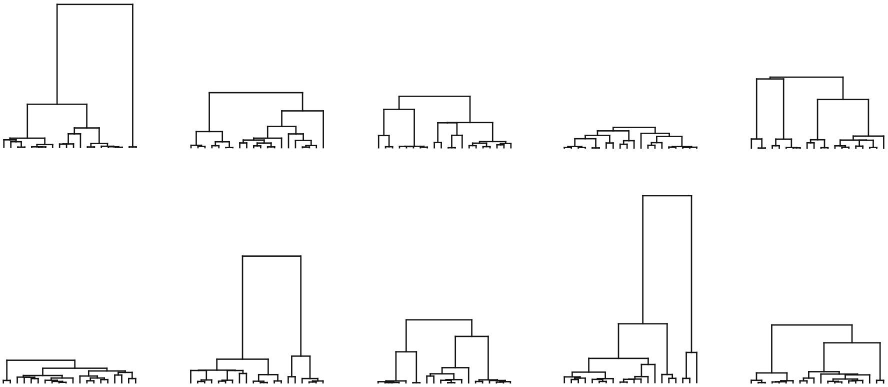


Fig. 1.3 Ten independently generated gene genealogies for a sample of size $n = 20$, produced using a Mathematica Demonstrations Project "Coalescent Gene Genealogies" written by John Hawks

The standard neutral coalescent provides a prior distribution of gene genealogies which can be invoked (logically before a sample is taken) to make predictions about expected patterns of genetic variation or for purposes of statistical inference from data. For example, Huff et al. (2010) used a simple result from coalescent theory, due to Tajima (1983), to identify loci in a pair of human genomes that had twice the average coalescence time of randomly chosen loci, and employed these older loci to make inferences about ancient human effective population sizes. General methods of statistical inference for larger samples, such as those mentioned in Sect. 1.4.3, treat gene genealogies as missing data and average over them using the coalescent prior.

1.3.2 Including Mutations in the Coalescent

The lineages of the gene genealogy represent all the opportunity for mutations in the ancestry of the sample: any polymorphisms in the data must be the result of mutations that occurred along the branches of the gene-genealogical tree. Predictions about genetic variation and inferences from genetic data cannot be made unless mutations are included in the model. Fortunately, this is straightforward in the standard neutral coalescent. By definition, neutral genetic variation does not affect the probabilities of reproduction or the distribution of the number of offspring per individual, so mutation and coalescence can be treated separately. In particular, conditional on the gene-genealogical tree, mutations occur independently along each branch.

Because the timescale of coalescence is in units of N generations, each branch in the tree represents a huge number of opportunities for a mutation to occur. Then, because the probability of mutation per generation is very small, mutation may be modeled quite accurately as a continuous-time Markov process or sometimes simply as a Poisson counting process. A four-state Markov process is appropriate for a mutation in DNA, with its four nucleotides. When a Poisson counting process is used as an approximation, it is also often assumed that at most one mutation can occur per site. This is known as the infinite-sites (or infinitely-many-sites) mutation model.

The mutation rate for a genetic locus is typically denoted $\theta/2$, the mutation parameter θ being proportional to the product of the population size, N, and neutral mutation rate per generation, u. In general, $\theta = 2N_e u$, with $\theta = 4Nu$ in the diploid Wright–Fisher model. Technically, θ is assumed to exist in the limit $N \to \infty$, but less formally, the model is valid when N is large and u is small. With θ defined this way, the number of mutations on a branch or branches of total length t follows a Poisson distribution with expected value $t\theta/2$. The critical feature of the infinite-sites model is that each mutation creates a unique polymorphic site. Thus, for a given nucleotide site in the genome, at most, one mutation can have occurred in the history of the sample. This chapter will focus exclusively on this mutation model, which is due in this form (i.e., without recombination) to Watterson (1975). The infinite-sites model is a reasonable starting approximation for human autosomal genetic diversity, because only about 1/1000 nucleotide sites are polymorphic when two human genomes are compared (Cargill et al. 1999; Stephens et al. 2001) and only about 1/500 SNPs show more than two bases segregating (Hodgkinson and Eyre-Walker 2010).

A large number of four-state models have been put forward to represent DNA mutations, the HKY85 model being one of the most commonly used (Hasegawa et al. 1985). Models of "stepwise" mutation have also been added to the coalescent in order to account for variation in repeat sequences, such as microsatellite loci (Valdes et al. 1993). In general, mutation is a time-inhomogeneous process and must be modeled separately along each branch, forward in time starting from the MRCA or root of the tree. A number of simpler, "parent-independent" mutation models have been employed as approximations; for example, see Stephens and Donnelly (2003) and Fearnhead (2006). The infinite-sites model considered here and the infinite-alleles model used by Ewens (1972) are special cases of parent-independent mutation.

1.4 Fundamental Predictions for Single Loci in Well-Mixed Populations

The mathematical convenience of the standard neutral coalescent, the ease with which it may be applied, and all of the detailed predictions one can make using it follow from three key properties of standard neutral gene genealogies:

- The branching structure of a coalescent tree is determined by randomly joining pairs of ancestral lineages until the MRCA is reached

- The time during which there are i lineages ancestral to the sample, denoted T_i, follows an exponential distribution with parameter $i(i - 1)/2$
- T_i and T_j are statistically independent for $i \neq j$

For reference, the gene genealogy in Fig. 1.1 is drawn with the lengths of coalescent intervals T_2, T_3, and T_4 equal to their expected values from the exponential distribution, $E[T_i] = 2/(i(i - 1))$. In the case of parent-independent mutation, such as in the infinite-sites model considered here, we may include a fourth property:

- Mutations occur with rate $\theta/2$ along each branch of the coalescent tree

Then, conditional on the tree, the number of mutations over a length t of the tree follows a Poisson distribution with parameter $t\theta/2$. For example, the 11 mutations on the gene genealogy in Fig. 1.1 are exactly the number expected if $\theta = 6$ because the total length of the gene genealogy in Fig. 1.1 is 11/3 (see $E[T_{\text{Total}}]$ below).

1.4.1 The Size and Shape of a Gene Genealogy

Considering just the gene genealogy, without mutations, a great deal can be gleaned from the first three properties listed above. Because lineages always coalesce in pairs in the standard neutral coalescent, every gene genealogy includes exactly $n - 1$ coalescent events. Let T_{MRCA} represent the time to the most recent common ancestor of the sample and let T_{Total} represent the total length of the gene genealogy, or the sum of the lengths of all the branches in the tree. These two measures have been used extensively to characterize the sampling properties of gene genealogies. Knowledge of T_{MRCA} may be of direct biological interest, while knowledge of T_{Total} is important because T_{Total} quantifies the total opportunity for mutations to occur in the ancestry of the sample.

From their definitions, T_{MRCA} is simply the sum of the individual times T_i, or

$$T_{\text{MRCA}} = \sum_{i=2}^{n} T_i$$

and T_{Total} is obtained similarly, except that each time is weighted by the number of lineages that existed during the interval, so that

$$T_{\text{Total}} = \sum_{i=2}^{n} i\,T_i$$

The $n - 1$ times, T_n, $T_{n-2}, \ldots,$ T_2, are called *coalescence intervals* here to avoid confusion with T_{MRCA}, which is often referred to as *the* coalescence time.

Using the properties of the exponential distribution, namely, that $E[T_i] = 2/(i(i - 1))$ and $\text{Var}[T_i] = 4/(i(i - 1))^2$ and the fact that the $n - 1$ coalescent intervals are statistically independent, so that $\text{Cov}[T_i, T_j] = 0$ for $i \neq j$, one obtains the fundamental predictions about the time to the most recent common ancestor,

$$E\left[T_{\mathrm{MRCA}}\right] = \sum_{i=2}^{n} \frac{2}{i\,(i-1)} = 2\left(1 - \frac{1}{n}\right) \approx 2$$

and

$$\mathrm{Var}\left[T_{\mathrm{MRCA}}\right] = \sum_{i=2}^{n} \left(\frac{2}{i\,(i-1)}\right)^2 \approx \frac{4}{3}\left(\pi^2 - 9\right)$$

in which the approximations are for large n. Recall that time here is measured in units of N_e generations, so that 2 in the first equation above corresponds to $4N$ generation in the diploid Wright–Fisher model. The significance of this equation is that T_{MRCA} does not grow indefinitely with increasing sample size n, but converges to a constant value. This is because when i is large, T_i tends to be extremely short. For example, $E[T_{100}] = 0.0002$, whereas $E[T_2] = 1$. Again, the coalescence intervals in Fig. 1.1 are drawn to scale according to their expected values: $[T_2] = 1$, $E[T_3] = 1/3$, and $E[T_4] = 1/6$.

Even if the sample size is large, the statistical properties of gene genealogies are strongly affected by a relatively small number of coalescent intervals, those deep in the past for which the number of ancestral lineages, i, is small. It is in large part because of this that inferences based on genetic variation at a single locus tend to be poor. For example, note that $\mathrm{Var}[T_{\mathrm{MRCA}}]$ does not decrease to zero as n increases but converges to a constant value (~1.16). Increasing the sample size n in population genetics does not induce the kinds of “law of large numbers” behaviors one finds in standard statistical scenarios where samples are independent.

For the total length of the gene genealogy, one finds

$$E\left[T_{\mathrm{Total}}\right] = 2\sum_{i=1}^{n-1} \frac{1}{i} \approx 2\left(\ln n + \gamma\right)$$

in which $\gamma = 0.577216$ is Euler’s constant, and

$$\mathrm{Var}\left[T_{\mathrm{Total}}\right] = 4\sum_{i=1}^{n-1} \frac{1}{i^2} \approx \frac{2\pi^2}{3}$$

Again, the approximations are for large n. In this case, there is a somewhat greater effect of increasing the sample size n, but the effect is weak. For example, the coefficient of variation of T_{Total}—defined as the standard deviation divided by the mean—does tend to zero as n tends to infinity, but it decreases very slowly, in proportion to one over the natural logarithm of n.

It is possible to obtain explicit expressions for the full distributions of T_{MRCA} and T_{Total} and also for measures of genetic variation such as S in the next section, but the expressions are cumbersome. Interested readers should consult the review by Tavaré (1984) or the textbooks by Hein et al. (2005) and Wakeley (2009).

1.4.2 Levels and Patterns of Genetic Variation

The introduction of the coalescent was revolutionary in population genetics because it provided a way to compute the probability of a dataset. This probability, also known as the likelihood, is the foundation of rigorous statistical inference. Population-genetic data are complicated, with nested patterns of variation among subsets of the sample as in Fig. 1.1, and no general analytical results are available for the likelihood. Inference has proceeded by the numerical computation of likelihoods, using simulations to sample gene genealogies from the coalescent prior distribution. Even this is quite complicated due to the enormity of the sample space of gene genealogies.

Two ways of accounting for the unknown gene genealogy were developed in the 1990s: importance sampling and Markov chain Monte Carlo (MCMC). If we knew the order of mutation events and coalescent events, as in Fig. 1.1, but not the times, we could compute the likelihood by multiplying the probabilities of the ordered events. For example, under the infinite-sites model, all likelihoods include the familiar probability of identity by descent, $1/(\theta + 1)$, originally due to Malécot (1946), because for all gene genealogies, the final event is that two ancestral lineages coalesce at the MRCA before either of them mutates. Importance-sampling methods average these products of probabilities over possible orderings of events (Griffiths and Tavaré 1994, 1996; Stephens and Donnelly 2000; Wu 2010). Alternatively, if we knew the tree structure and the coalescence times, we could model the process of mutation along the branches of the tree in computing the likelihood. MCMC methods do this and average over the underlying trees and times (Kuhner et al. 1995; Kuhner 2006; Beerli 2006; Hey and Nielsen 2004, 2007; Drummond et al. 2012).

There has been a growing trend to make inferences based on "summary statistics" rather than the full data, often within the framework of approximate Bayesian computation (Beaumont 2010; Alvarado-Serrano and Hickerson 2016). The summary-statistic approach to inference reduces the dimensionality of the data, ideally to a small set of simpler measures of genetic variation which are highly informative with regard to a set of parameters or phenomena of interest. As with importance sampling and MCMC, summary-statistic approaches use coalescent models to average over gene genealogies. Coalescent theory is also used to make predictions about summary statistics, a number of which (e.g., heterozygosity) have also been important historically in population genetics.

Three kinds of summary statistics have been well studied in a variety of settings. These are segregating sites, average pairwise differences, and site frequencies, which are defined as follows for a sample from a single population. The number of segregating sites, S, is the number of polymorphisms, e.g., SNPs in a dataset of DNA sequences. The average number of pairwise differences, which is denoted

π, is found by comparing each sampled sequence with every other, counting the number of differences between them, and taking the average over all pairs. The site frequencies, ξ_i, are obtained by counting the number of SNPs in the data at which the mutant base is found in i copies in the sample, for i from 1 to $n-1$. The ancestral state at each SNP is generally ascertained as the state in a sequence from a closely related species. If this information is not available or is unreliable, site frequencies just count the less frequent base, so that $1 \leq i \leq n/2$.

Under the infinite-sites model, the number of segregating sites is equal to the number of mutations on the coalescent tree. Then, by conditioning on T_{Total}, one obtains

$$E\left[S\right] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

and

$$\text{Var}\left[S\right] = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2}$$

(Watterson 1975). The properties of S are similar to the properties of T_{Total}. In particular, the expected number of segregating sites increases very slowly—like ln$n$ as more and more sequences are sampled. Further, the quality of estimates of θ based on S improves rather slowly with increasing sample size, again like $1/\ln n$ rather than the usual $1/n$ that holds in standard statistical applications, because the samples are not independent due to their shared gene genealogy.

For the average number of pairwise sequence differences, one obtains

$$E\left[\pi\right] = \theta,$$

which follows directly from $E[S]$ because the marginal expectation for each pair of sequences in a sample is identical to the expectation for a single pair. Further,

$$\text{Var}\left[\pi\right] = \frac{n+1}{3\left(n-1\right)}\theta + \frac{2\left(n^2+n+3\right)}{9n\left(n-1\right)}\theta^2$$

(Tajima 1983). Estimates of θ based on π are unbiased but inconsistent in the statistical sense because Var$[\pi]$ does not decrease to zero as the sample size n tends to infinity. This is due to the fact that the ancestries of different pairs of sequences in the sample share many of the same branches of the gene genealogy, causing some mutations to be counted more than once in the computation of the average number of pairwise differences π.

Beyond the general statement that population-genetic samples are not independent due to the underlying gene genealogy, the relatively poor statistical properties of estimates of θ are further explained by the probabilistic structure of gene genealogies. For example, Rauch and Bar-Yam (2004) studied the distribution of the genetic "uniqueness" of a sample, defined as the length of the branch connecting that sample to the rest of the gene genealogy. This distribution has an extremely long tail because there is a chance the $(n + 1)$th sample will establish a new MRCA and thus be markedly unique. Forward in time, the loss of such markedly unique lineages causes neutral substitutions to accumulate in (non-Poisson) bursts (Watterson 1982; Pfaffelhuber and Wakolbinger 2005). Greater statistical power to estimate θ may be achieved by sampling more loci rather than increasing the sample size at a single locus (Pluzhnikov and Donnelly 1996; Felsenstein 2006).

Predictions about site frequencies are obtained by considering mutations that occur on branches in the coalescent tree that have i descendants in the sample. Fu (1995) derived expressions for the expected values, variances, and covariances of the site-frequency counts. Here, we will focus on the expected values, which are

$$E\left[\xi_i\right] = \frac{\theta}{i}$$

for i from 1 to $n - 1$. This simple relationship, for a sample of size $n = 20$, is graphed as a "site-frequency spectrum" in Fig. 1.4a, which means that site frequencies are plotted as expected proportions of all segregating sites. Figure 1.4b–d displays a range of site-frequency spectra for three different models discussed in Sect. 1.5. When depicted in this way, as expected proportions, the site-frequency spectrum gives the probability of observing each type of polymorphism at a randomly chosen SNP.

If the ancestral states at the sites of SNPs are not known, it is only possible to discern the "folded" site-frequency patterns, which have expected values

$$E\left[\eta_i\right] = \theta\left(\frac{1}{i} + \frac{1}{n-i}\right)\frac{1}{\delta_{i,n-i}}$$

for $1 \leq i \leq n/2$. In the last term, $\delta_{i,\,n-i}$, is the Kronecker delta, so this term is a correction for the case $i = n - i$, in which only one kind of SNP contributes to η_i. The full site-frequency spectrum is referred to as the "unfolded" site-frequency spectrum.

None of these three measures of genetic variation depends on how variation is arrayed along the sequences in the sample. All of them can be computed by considering each SNP in isolation from all other SNPs. Patterns of linkage disequilibrium between sites and the process of recombination that produces them are the subject of Chap. 2. Although here our focus is on single loci without recombination, it is important to note that all of the expected values given above hold regardless of recombination. This is because the marginal coalescent process at each site is the same as the corresponding single-locus coalescent process. However, the variances given above hold only for single loci without recombination. Generally speaking, recombination acts to decrease these variances because it introduces a level of independence among sites.

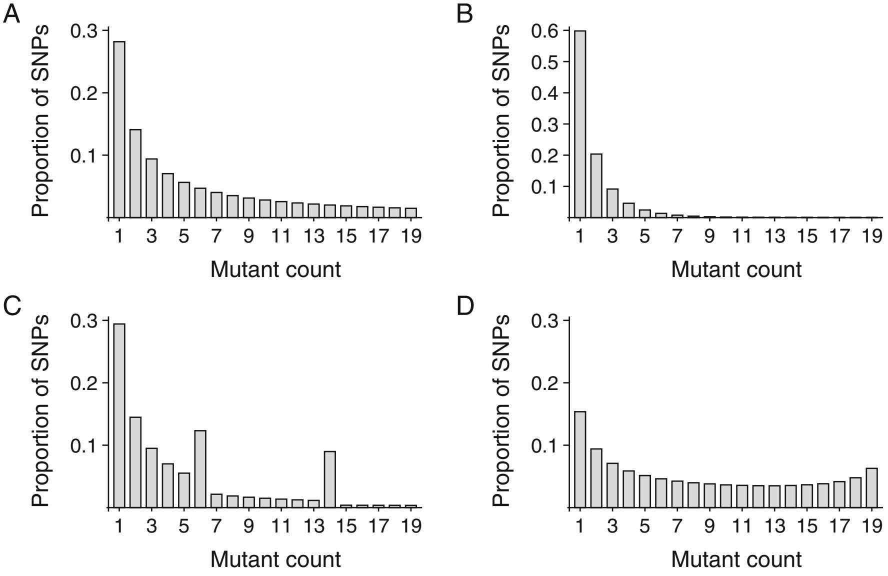


Fig. 1.4 Four site-frequency spectra illustrating the range of possible predictions of neutral population-genetic models. For a given SNP, the mutant count is the number of sequences that carry the mutant base out of a total sample of $n = 20$ sequences. The heights of bars give the proportion of all SNPs that have each mutant count. The four panels show results for samples from (**a**) a standard neutral population, (**b**) a population that recently grew 100-fold, (**c**) two isolated populations, and (**d**) a single deme in a subdivided population with migration. Details and parameters for each case are given in the text. The values in panels (**b**) and (**c**) were computed using Eqs. (20) and (22) in Wakeley and Hey (1997). The values in panel (**d**) were generated using simulations described in Wakeley (1999)

1.4.3 Tests of the Standard Neutral Coalescent Based on Site Frequencies

Section 1.5 introduces deviations from the simple Kingman coalescent, motivated by the desire to apply coalescent models more broadly. It is also of interest to test the simple Kingman coalescent, and this can be done using the three measures of genetic variation considered in the previous section. A large number of test statistics have been proposed, modeled after Tajima's (1989) initial suggestion of the statistic

$$D = \frac{\pi - S/a_1}{\sqrt{\mathrm{Var}\,(\pi - S/a_1)}},$$

in which

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

The numerator of Tajima's D compares two unbiased estimates of θ, one from pairwise differences and one from segregating sites. Tajima's D thus has an expected value very nearly equal to zero (it is not exactly equal to zero because of the denominator), and significant deviations either in the positive or the negative direction warrant rejection of the standard neutral coalescent. The denominator is a normalization factor which decreases the sensitivity of the sample size and requires an estimate of the variance of the numerator, typically made from the same data.

Because S and π are linear functions of the site frequencies (Tajima 1997), Tajima's D may be viewed as a measure of goodness-of-fit of the prediction displayed in Fig. 1.4a (for $n = 20$). Actually, Tajima's D depends only on the folded site-frequency spectrum because sites that contribute to ξ_i and ξ_{n-i} are weighted equally, proportional to $i(n - i)$, in the calculation of π, and all sites are weighted equally in the calculation of S. Deviations in the positive direction indicate an excess of middle-frequency SNPs (i around $n/2$) and deviations in the negative direction indicate either an excess of low-frequency SNPs (i close to 1) or an excess of high-frequency SNPs (i close to $n - 1$). Tajima's D is sometimes portrayed as a test for selection (see Chap. 4), but in fact, it is sensitive to a whole battery of nonselective deviations from the standard neutral model, including population structure and changes in population size over time.

The distribution of Tajima's D takes on a variety of shapes depending on the sample size, the mutation rate, and other factors. Because it is computed from the site-frequency counts, ξ_i, Tajima's D is a discrete random variable. Figure 1.5 shows two distributions of Tajima's D, illustrating the range of shapes it can assume. Figure 1.5a is for a sample of $n = 20$ at a hypothetical locus with $\theta = 10$ under the standard neutral coalescent, and Fig. 1.5b is for the same number of sequences all sampled from a single subpopulation in the migration model discussed in Sect. 1.5.2.

Fu and Li (1993) and Fu (1997) introduced a large number of related statistics, including many that test deviations from the folded site-frequency spectrum and

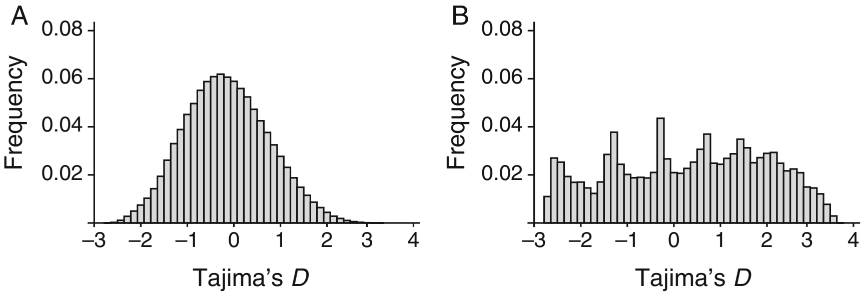


Fig. 1.5 (**a**) The distribution of Tajima's D among 10^6 data sets simulated under the standard neutral coalescent model at a hypothetical locus with $n = 20$ and $= 10$. (**b**) The corresponding distribution for a sample from a single subpopulation under the island migration model with $M = 1$ and with other parameters set so the expected number of pairwise differences in the sample is equal to 10, as it is in (**a**). The data in (**a**) were generated using the algorithm in Hudson (1990), and the data in (**b**) were generated using the algorithm in Wakeley (1999)

others that test deviations from the unfolded site-frequency spectrum. Simonsen et al. (1995) outlined a method of assessing the significance of such statistics which accounts for the fact that P-values depend on an estimate of θ from the data. Fay and Wu (2000) adapted one of Fu's (1997) statistics as a test specifically for positive selection. Their statistic, H, is sensitive to an excess of high-frequency SNPs, and positive selection is one of a small number of deviations from the standard neutral model that can cause such an excess. Achaz (2009) advanced the theory of devising optimal statistics based on site frequencies, and Ferretti et al. (2010) extended this approach to design optimal statistics for specific deviations from the standard neutral model. Recently, Sainudiin and Véber (2018) described a way to compute the likelihood of the full site-frequency spectrum of a sample at a locus without recombination.

1.5 Extensions of the Standard Model

The strong simplifying assumptions of the standard neutral model—no selection, constant population size over time, and no population structure—are appropriate if the aim is to establish a null model to be tested. If one wishes to make more detailed inferences about a range of biological phenomena, then coalescent models must be extended to include those phenomena and their key parameters. Many such extensions have been made, significantly broadening coalescent theory beyond the standard neutral case. In this section, we will encounter examples of how two important deviations from the Kingman coalescent, namely, changes in population size over time and geographic population structure, can affect the site-frequency spectrum and thus might be detected, for example, using Tajima's D or related statistics.

Figure 1.6 displays hypothetical gene genealogies for four different population models, showing how the size and shape of gene genealogies depend on the details of population structure and history. The standard neutral model (Fig. 1.6a), for which expected site-frequency results are given in Fig. 1.4a, is compared to a model of population growth (Fig. 1.6b), and two models of population structure: divergence in isolation (Fig. 1.6c) and subdivision and migration (Fig. 1.6d). These three types of deviations from the standard model are considered in detail below, with expected site-frequency results given in the corresponding panels of Fig. 1.4.

A consideration of how natural selection affects site frequencies is taken up in Chap. 4. Note that in some species, though probably not humans, extreme differences in offspring numbers among individuals can cause site-frequency patterns that closely mimic those produced by natural selection. When the variance σ^2 of the offspring-number distribution is very large, the Kingman coalescent may not hold. Instead, gene genealogies may include multiple mergers of ancestral genetic lineages (Möhle and Sagitov 2001). None of the predictions listed in Sect. 1.4 may hold, and there may be a dramatic excess of high-frequency SNPs; e.g., see Sargsyan and Wakeley (2008). Strong natural selection induces a very similar phenomena near the locus at which selection acts (Durrett and Schweinsberg 2004; Etheridge et al.

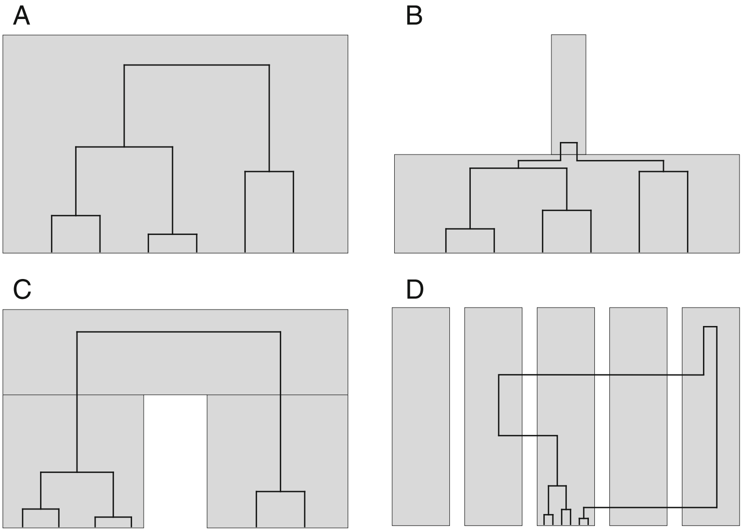


Fig. 1.6 Cartoon depictions of the four types of neutral population-genetic models for which results for expected site-frequency distributions are given in Fig. 1.4. Shaded blocks represent populations over time, with the present at the bottom and the past at the top and with widths in proportion to relative population size. Hypothetical gene genealogies are shown within these populations, constrained by their structure, and with coalescent times in roughly inverse proportion to relative population sizes. The four panels show (**a**) a standard neutral population of constant size, (**b**) a population that recently grew tenfold, (**c**) two isolated populations descended from a common ancestral population, and (**d**) a subdivided population in which migration can occur among five local populations

2006) similarly because a few individuals leave very many descendants in a short period of time.

1.5.1 Fluctuations in Population Size over Time

Due to the inverse dependence of the probability of coalescence on N, for example, in Eq. (1.2), changes in population size lead to changes in the rate of coalescence. Comparing two populations which are otherwise identical, if one population is twice the size of the other, then its gene genealogies will be twice as long on average. In a single population, with time rescaled by the current population size, then at a time in the past when the population size was twice as large, the rate of coalescence will be half what it is now. To make this precise, under arbitrary changes in population size, if $\lambda(t)$ is the size of the population at time t relative to what it is today, then by defining

$$\Lambda(t) = \int_0^t \frac{1}{\lambda(s)} \mathrm{d}s$$

the occurrence of a coalescent event in the interval $(0, t)$ may be modeled using the first two of the key properties listed in Sect. 1.4 but over the corresponding, rescaled interval $(0, \Lambda(t))$ (Donnelly and Tavaré 1995).

Equivalently, one can imagine taking a standard gene genealogy, such as the one in Fig. 1.1, then stretching or shrinking its coalescent intervals accordingly, so that they become proportionately longer (or, respectively, shorter) when the population size was larger (respectively, shorter). Alternatively, one may model changes in population size as proportional changes in the mutation parameter θ over time. Based on these considerations, for simple types of changes in population size, it is possible to obtain analytical expressions for some quantities of interest (Slatkin and Hudson 1991; Polanski and Kimmel 2003; Wakeley and Hey 1997).

Figure 1.4b shows the expected site-frequency spectrum for a sample of size $n = 20$ from a population which was much smaller in the past than it is now. Specifically, the population grew 100-fold instantaneously at time $t = 0.2$ in the past, measured on the coalescent timescale based on the current population size. In terms of the scaled mutation rate, between the present and time $t = 0.2$, the mutation parameter was $\theta = 1$, while before time $t = 0.2$ in the past, the mutation parameter was $\theta = 0.01$.

In this situation, only a small fraction of mutations will occur during the more ancient coalescent intervals of the gene genealogy. These are the mutations that would have produced high-frequency SNPs. For example, on average for $n = 20$, there will be seven ancestral genetic lineages at time $t = 0.2$. These more ancient intervals, with from seven down to two ancestral genetic lineages, are the only ones in which mutations can create SNPs contributing to site frequencies ξ_{14} through ξ_{19}. Figure 1.6b shows a similar scenario for the gene genealogy of a sample of size $n = 6$, illustrating the dramatic compression of ancient coalescent intervals under population growth.

Because more ancient mutations are disproportionately the source of high-frequency SNPs, population growth causes an excess of low-frequency SNPs (Fig. 1.4b) compared to the standard neutral coalescent (Fig. 1.4a). Population decline causes an excess of high-frequency SNPs (not shown). However, as long as the branching structure of the gene genealogy is determined by randomly joining pairs of ancestral lineages, the site-frequency spectrum will always be a convex decreasing function of the mutant count (Sargsyan and Wakeley 2008). Thus, the most extreme excess of high-frequency SNPs that population decline or any series of changes in population size can produce under the coalescent model is a flat site-frequency spectrum.

1.5.2 Population Subdivision and Migration

The great simplicity of the standard neutral coalescent follows from the exchangeability of the genetic lineages ancestral to the sample (Kingman 1982c) which holds only under neutrality for well-mixed populations. Whenever lineages carry labels, such as allelic types when there is selection or locations when there is geographic structure, and these labels affect the rates of coalescence, then the lineages are not exchangeable, and modeling gene genealogies become more complicated. In the case of population subdivision, either with or without migration, the chance of coalescence is greater for pairs of lineages in the same subpopulation than for pairs of lineages in different subpopulations. This can only be modeled by explicitly keeping track of the locations of ancestral lineages as they are followed backward in time.

Other complications arise because structured populations may contain many subpopulations, which may be of different sizes and between which any number of complex patterns of migration might exist. A general model of D subpopulations, or "demes" as they are often called, would have D^2 parameters: D deme sizes and $D(D-1)$ migration rates. In addition, it is not clear what sort of simplified limiting models should be developed for structured populations. Some populations might comprise a small number of very large demes, while others might comprise a large number of small demes. It could be that the very idea of demes/subpopulations is inapplicable, rather than that the population is continuously distributed across its habitat.

Accordingly, a number of different coalescent models of geographic structure have been developed—these are reviewed in Hein et al. (2005) and Wakeley (2009)—and the choice of model must depend on the species being studied. Wright (1931) introduced the island model of population subdivision with migration, which has been the source of a great number of other models and methods of data analysis. Herbots (1997) and Notohara (1990) described the general mathematical coalescent approach to these discrete-deme models, following Takahata (1988). In these models, the deme sizes are assumed to be large, like the population size in the Kingman coalescent ($N \to \infty$). The migration rates are assumed to be small. They are treated in the same manner that mutation is treated in the limit leading to the Kingman coalescent.

The resulting structured coalescent model allows the straightforward derivation of useful expressions concerning genetic variation. For instance, consider a simple version of the island model in which all D demes are of the same (haploid) size N, migration between all pairs of demes occurs with the same per-generation probability m, and reproduction occurs by haploid Wright–Fisher sampling. Then, if π_w and π_b are the average number of differences between pairs of sequences from the same deme (i.e., "within") and the average number of differences between pairs of sequences from two different demes (i.e., "between"), it can be shown that

$$E[\pi_w] = \theta D$$

and

$$E[\pi_b] = \theta D \left(1 + \frac{1}{M}\right)$$

where the parameter $M = 2Nm$ is the scaled migration rate. This simple model can be extended to include other modes of reproduction or diploidy, as in the Kingman coalescent, by replacing N with N_e in both θ and M.

The expression for $E[\pi_w]$ says that the expected level of genetic variation within demes is identical to the expected level in a single population of the same total size, ND. The same is not true of the variance (not shown), which depends inversely on M. The expression for $E[\pi_b]$ says that the expected level of genetic variation between demes is increased by an amount inversely proportional to the migration parameter M. When M is small, the population is expected to contain high levels of variation, and the difference between π_b and π_w is expected to be great. This is the basis of F_{ST} as a measure of the degree of population subdivision (Slatkin 1991). It is important to note that the scaled migration rate M may be large, so that little evidence of subdivision is apparent, even if the per-generation probability of migration is small.

These results had been known previously (Li 1976; Slatkin 1987; Strobeck 1987), but the introduction of the structured coalescent greatly facilitated the development of sophisticated methods of inference for structured populations, akin to those mentioned in Sect. 1.4.3, where the model is employed to average over gene genealogies in the computation of the likelihood (de Iorio et al. 2005; Beerli 2006). Similar methods have been proposed for cases of nonequilibrium migration, in which two or more populations descend from a single ancestral population (Hey and Nielsen 2004, 2007; Wilkinson-Herbots 2008; Hey 2010).

Population subdivision can have a dramatic effect on site frequencies. Figure 1.4C shows the site-frequency spectrum for a sample of size $n = 20$ for a hypothetical case of hidden population structure. Specifically, the sample contains $n_1 = 6$ sequences from one population and $n_2 = 14$ from the other population under the isolation model of Takahata and Nei (1985) in which two populations split from a common ancestral population at some time in the past and after that exchanged no migrants. The same mutation rate $\theta = 1$ was used for all three populations, and the split time was assumed to be $t = 1$, measured on the coalescent timescale. In this case, there is a tendency for the gene genealogy to be composed of two subtrees with 6 and 14 tips connected by a long internal branch, with mutations on this branch contributing to ξ_6 and ξ_{14}. Figure 1.6c depicts a similar scenario for a sample of total size six.

Figure 1.4d shows the site-frequency spectrum for a sample of size $n = 20$ taken from a single deme in the island model with many demes and a migration rate of $M = 1$. Here, in the recent past, ancestral genetic lineages not only coalesce within the deme from which they were sampled but also migrate to other demes. When all remaining ancestral lineages are in separate demes, the process of coalescence is dependent on migration events that bring ancestral lineages together into the same deme, where they have a chance to coalesce. Both the branching pattern and the lengths of coalescent intervals differ from those of standard coalescent gene genealogies. For $M = 1$, this results in the slightly U-shaped site-frequency spectrum shown in Fig. 1.4d.

We can understand the pattern in Fig. 1.4d by imagining a mixture of patterns like the one in Fig. 1.4c, depending on the specific outcome of migration and coalescence in the recent ancestry of the sample. For example, if one ancestral lineage migrates out of the sampled deme and the other 19 ancestral lineages coalesce within it, then the gene genealogy will resemble one for which two demes were sampled with $n_1 = 1$ and $n_2 = 19$. This would cause mutants counts ξ_1 and ξ_{19} to be inflated. Fig. 1.6d shows an analogous scenario for a sample of size six in a five-deme model, in which the counts ξ_2 and ξ_4 would be inflated.

Figure 1.5 illustrates that such deviations can be detected using Tajima's D. For example, if we adopt the lower 2.5% critical value of -1.803 and the upper 97.5% critical value of 2.001 for $n = 20$ from Table 2 in Tajima (1989), the one million pseudo-datasets from the standard neutral coalescent that yielded Fig. 1.5a reject the null model 2.0% of the time in the negative direction and 1.1% in the positive direction. In contrast, the pseudo-datasets that yielded Fig. 1.5b, which were generated under the same kind of many-demes model that gave the site-frequency spectrum in Fig. 1.4d (but with $E[\pi_w] = 10$ to allow comparison with Fig. 1.5a) reject the null model 11.4% of the time in the negative direction and 18.8% in the positive direction.

1.6 Conclusion: Current Challenges of Big Data

Now is a particularly exciting time in human population genetics. The field is awash in data, with major efforts such as the Simons Genome Diversity Project (Mallick et al. 2016), the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015), and the UK Biobank (Bycro et al. 2018) promising that, soon, many millions of genomes will be available for study. The continued relevance of the models presented here may be seen in the recent papers by Kelleher et al. (2019) and Speidel et al. (2019). These present new methods for the population-genetic analysis of very large numbers of genomes. With the caveat that at the genomic scale, it is crucial to include recombination (see Chap. 2), parts of the analyses in both Kelleher et al. (2019) and Speidel et al. (2019) rely on the standard neutral coalescent model. The aim in both works is to infer the ordered series of mutation events and coalescent events at loci across the human genome (recall the importance-sampling methods described in Sect. 1.4.2). In doing so, both works use the techniques of Li and Stephens (2003), which extend the importance-sampling method of Stephens and Donnelly (2000) to account for recombination. The results of Kelleher et al. (2019) and Speidel et al. (2019) provide first-pass estimates of ancient relationships and associated mutations among humans across the genome (Harris 2019).

The aim of this chapter has been to describe the foundational models of coalescent theory. They are simplified models which capture the effects of neutral mutation, reproduction, and genetic transmission in shaping distributions of genetic diversity. The simplest model, the standard neutral coalescent, assumes a single well-mixed population of constant size, but extensions to include changes in population size over time and idealized kinds of population structure were also described. These models aid in the interpretation of genetic data and express our understanding of how evolutionary forces produce and maintain variation at a single locus without recombination.

References

Achaz G (2009) Frequency spectrum neutrality tests: one for all and all for one. Genetics 183:249–258

Alvarado-Serrano DF, Hickerson MJ (2016) Spatially explicit summary statistics for historical population genetic inference. Methods Ecol Evol 7:418–427

Alvarez G, Ceballos FC, Quinteiro C (2009) The role of inbreeding in the extinction of a European royal dynasty. PLoS One 4(4):e5174

Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. Annu Rev Ecol Evol Syst 41:379–406

Beerli P (2006) Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics 22:341–345

Bycro C et al (2018) The UK biobank resource with deep phenotyping and genomic data. Nature 562:203–209

Cannings C (1974) The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. Adv Appl Probab 6:260–290

Cannings C, Thompson EA, Skolnick MH (1978) Probability functions on complex pedigrees. Adv Appl Probab 10:26–61

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daly GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231–237

Chang JT (1999) Recent common ancestors of all present-day individuals. Adv Appl Probab 31:1002–1026

de Iorio M, Griffiths RC, Leblois R, Rousset F (2005) Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. Theoret Pop Biol 68:41–53

Donnelly P, Tavaré S (1995) Coalescents and genealogical structure under neutrality. Annu Rev Genet 29:401–421

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973

Durrett R, Schweinsberg J (2004) Approximating selective sweeps. Theoret Pop Biol 66:129–138

Etheridge AM, Pfaffelhuber P, Wakolbinger A (2006) An approximate sampling formula under genetic hitchhiking. Ann Appl Probab 16:685–729

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theoret Pop Biol 3:87–112

Ewens WJ (1974) A note on the sampling theory for infinite alleles and infinite sites models. Theoret Pop Biol 6:143–148

Ewens WJ (1990) Population genetics theory—the past and the future. In: Lessard S (ed) Mathematical and statistical developments of evolutionary theory. Kluwer Academic, Amsterdam, pp 177–227

Ewens WJ (2004) Mathematical population genetics, vol I: theoretical foundations. Springer, Berlin

Fay JC, Wu C-I (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Fearnhead P (2006) Perfect simulation from nonneutral population genetic models: variable population size and population subsdivision. Genetics 174:1397–1406

Felsenstein J (2006) Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? Mol Biol Evol 23:691–700

Ferretti L, Perez-Enciso M, Ramos-Onsins S (2010) Optimal neutrality tests based on the frequency spectrum. Genetics 186:353–365

Fisher RA (1930) The genetical theory of natural selection. Clarendon, Oxford

Fu Y-X (1995) Statistical properties of segregating sites. Theoret Pop Biol 48:172–197

Fu Y-X (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147:915–925

Fu Y-X, Li W-H (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709

Griffiths RC, Tavaré S (1994) Simulating probability distributions in the coalescent. Theoret Pop Biol 46:131–159

Griffiths RC, Tavaré S (1996) Monte Carlo inference methods in population genetics. Math Comput Modelling 23:141–158

Hanski I, Gaggiotti OE (2004) Ecology, genetics, and evolution of metapopulations. Elsevier Academic, London

Harris K (2019) From a database of genomes to a forest of evolutionary trees. Nat Genet 51:1304–1307

Hasegawa M, Kishino H, Yano H (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Hawks J. "Coalescent Gene Genealogies" from the Wolfram Demonstrations Project. http://demonstrations.wolfram.com/CoalescentGeneGenealogies/

Hein J, Schierup MH, Wiuf C (2005) Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford University Press, Oxford

Herbots HM (1997) The structured coalescent. In: Donnelly P, Tavaré S (eds) Progress in population genetics and human evolution, IMA volumes in mathematics and its applications, vol 87. Springer, New York, pp 231–255

Hey J (2010) Isolation with migration models for more than two populations. Mol Biol Evol 27:905–920

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167:747–760

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci U S A 104:2785–2790

Hochman A (2019) Race and reference. Biology & Philosophy 34:32

Hodgkinson A, Eyre-Walker A (2010) Human triallelic sites: evidence for a new mutational mechanism? Genetics 184:233–241

Hudson RR (1983) Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203–217

Hudson RR (1990) Gene genealogies and the coalescent process. In: Futuyma DJ, Antonovics J (eds) Oxford surveys in evolutionary biology, vol 7. Oxford University Press, Oxford, pp 1–44

Huff CD, Xing J, Rogers AR, Witherspoon D, Jorde LB (2010) Mobile elements reveal small population size in the ancient ancestors of *Homo sapiens*. Proc Natl Acad Sci USA 107:2147–2152

Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336:740–743

Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G (2019) Inferring whole-genome histories in large population datasets. Nat Genet 51:1330–1338

Kingman JFC (1982a) On the genealogy of large populations. J Appl Probab 19A:27–43

Kingman JFC (1982b) The coalescent. Stoch Process Appl 13:235–248

Kingman JFC (1982c) Exchangeability and the evolution of large populations. In: Koch G, Spizzichino F (eds) Exchangeability in probability and statistics. North-Holland, Amsterdam, pp 97–112

Ko A, Nielsen R (2019) Joint estimation of pedigrees and effective population size using Markov chain Monte Carlo. Genetics 212:855–868

Kuhner MK (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22:768–770

Kuhner MK, Yamato J, Felsenstein J (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 140:1421–1430

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P, Przeworski M (2012) Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol 10(9):e1001388

Li W-H (1976) Distribution of nucleotide difference between two randomly chosen cistrons in a subdivided population: the finite island model. Theoret Pop Biol 10:303–308

Li H, Durbin R (2011) Inference of population history from individual whole-genome sequences. Nature 475:493–496

Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–2233

Malécot G (1946) La consaguinite dans une population limitee. Comp Rendus Acad Sci Paris 222:841–843

Mallick S et al (2016) The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature 538:201–206

Möhle M (1998a) Robustness results for the coalescent. J Appl Probab 35:438–447

Möhle M (1998b) A convergence theorem for Markov chains arising in population genetics and the coalescent with partial selfing. Adv Appl Probab 30:493–512

Möhle M (1998c) Coalescent results for two-sex population models. Adv Appl Probab 30:513–520

Möhle M (1999) The concept of duality and applications to Markov processes arising in neutral population genetics models. Bernoulli 5:761–777

Möhle M, Sagitov S (2001) A classification of coalescent processes for haploid exchangeable population models. Ann Appl Probab 29:1547–1562

Notohara M (1990) The coalescent and the genealogical process in geographically structured population. J Math Biol 9:59–75

Ott J (1999) Analysis of human genetic linkage, 3rd edn. Johns Hopkins University Press, Baltimore

Pfaffelhuber P, Wakolbinger A (2005) The process of most recent common ancestors in an evolving coalescent. Stoch Proc App 116:1836–1859

Pluzhnikov A, Donnelly P (1996) Optimal sequencing strategies for surveying molecular genetic diversity. Genetics 144:1247–1262

Polanski A, Kimmel M (2003) New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics 165:427–436

Rauch EM, Bar-Yam Y (2004) Theory predicts the uneven distribution of genetic diversity within species. Nature 431:449–452

Rohde DLT, Olsen S, Chang JT (2003) Modeling the recent common ancestry of all living humans. Nature 425:798–804

Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70:841–847

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW (2005) Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1:e70

Sainudiin R, Véber A (2018) Full likelihood inference from the site frequency spectrum based on the optimal tree resolution. Theoret Pop Biol 124:1–15

Sargsyan O, Wakeley J (2008) A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theoret Pop Biol 74:104–114

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141:413–429

Sjödin P, Kaj I, Krone S, Lascoux M, Nordborg M (2005) On the meaning and existence of an effective population size. Genetics 169:1061–1070

Slatkin M (1987) The average number of sites separating DNA sequences drawn from a subdivided population. Theoret Pop Biol 32:42–49

Slatkin M (1991) Inbreeding coefficients and coalescence times. Genet Res Camb 58:167–175

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562

Speidel L, Forest M, Sinan S, Myers SR (2019) A method for genome-wide genealogy estimation for thousands of samples. Nat Genet 51:1321–1329

Spence JP, Steinrücken M, Terhorst J, Song YS (2018) Inference of population history using coalescent HMMs: review and outlook. Curr Op Genet Devel 53:70–76

Stephens M, Donnelly P (2000) Inference in molecular population genetics. J R Stat Soc Ser B 62:605–655

Stephens M, Donnelly P (2003) Ancestral inference in population genetics models with selection. Aust N Z J Stat 45:395–430

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han J-H, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell W, Windemuth A, Xu C, Kalbfleisch TS, Shaner SL, Arnold K, Schultz V, Drysdale CM, Nandabalan K, Judson RS, Ruaño G, Vovis GF (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489–493

Strobeck C (1987) Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics 117:149–153

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA. Genetics 123:585–595

Tajima F (1997) Estimation of the amount of DNA polymorphism and statistical tests of the neutral mutation hypothesis based on DNA polymorphism. In: Donnelly P, Tavaré S (eds) Progress in population genetics and human evolution. Springer, New York, pp 149–164

Takahata N (1988) The coalescent in two partially isolated diffusion populations. Genet Res Camb 53:213–222

Takahata N, Nei M (1985) Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344

Tavaré S (1984) Lines-of-descent and genealogical processes, and their application in population genetic models. Theor Popul Biol 26:119–164

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–74

Valdes AM, Slatkin M, Freimer NB (1993) Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133:737–749

Wakeley J (1999) Non-equilibrium migration in human history. Genetics 153:1863–1871

Wakeley J (2009) Coalescent theory: an introduction. Macmillan Learning, Macmillan, New York

Wakeley J, Hey J (1997) Estimating ancestral population parameters. Genetics 145:847–855

Wakeley J, King L, Low BS, Ramachandran S (2012) Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. Genetics 190:1433–1445

Wakeley J, King L, Wilton P (2016) Effects of the population pedigree on genetic signatures of historical demographic events. Proc Natl Acad Sci USA 113:7994–8001

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theoret Pop Biol 7:256–276

Watterson GA (1982) Mutant substitutions at linked nucleotide sites. Adv Appl Probab 14:166–205

Wilkinson-Herbots HM (2008) The distribution of the coalescence time and the number of pairwise nucleotide differences in the "isolation with migration" model. Theoret Pop Biol 73:277–288

Wilton PR, Baduel P, Landon MM, Wakeley J (2017) Population structure and coalescence in pedigrees: comparisons to the structured coalescent and a framework for inference. Theoret Pop Biol 115:1–12

Winther GW, Giordano R, Edge MD, Nieslen R (2015) The mind, the lab, and the field: three kinds of populations in scientific practice. Stud Hist Phil Biol Biomed Sci 52:12–21

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

Wu Y (2010) Exact computation of coalescent likelihood for panmictic and subdivided populations under the infinite sites model. IEEE/ACM Trans Comput Biol Bioinform 7:611–618

Linkage Disequilibrium

2

Montgomery Slatkin

Abstract

Linkage disequilibrium (LD) is the nonrandom association between alleles at closely linked loci. LD is created by genetic drift and natural selection, and it decays exponentially with time at a rate proportional to the recombination rate. This chapter reviews the theory of LD between pairs of loci and the use of LD for detecting past episodes of selection and for gene mapping.

2.1 Introduction

Although population genetics largely focuses on one locus at a time, much is to be learned from considering two or more loci together. The reason is that alleles at different loci are transmitted together, creating the opportunity for correlations that reflect their common history. This correlation is important for gene mapping, where the goal is to identify loci that affect a trait, and when considering the effects of natural selection. Loci that are selected influence nearby neutral loci. Therefore, the study of sets of loci together can provide more insight into evolutionary processes and give additional information about gene action than can be obtained by focusing on each locus separately.

For simplicity, I will start by presenting results for two loci. Assume that the two loci are on the same pair of homologous chromosomes. If there are two alleles at each of the two loci (*A/a* and *B/b*), there are four combinations, called haplotypes, on a chromosome, *AB*, *Ab*, *aB*, and *ab*. These haplotypes can be thought of as the four kinds of gametes that can be produced by an individual.

M. Slatkin (✉)
Department of Integrative Biology, University of California, Berkeley, CA, USA
e-mail: slatkin@berkeley.edu

K. E. Lohmueller, R. Nielsen (eds.), *Human Population Genomics*,
https://doi.org/10.1007/978-3-030-61646-5_2

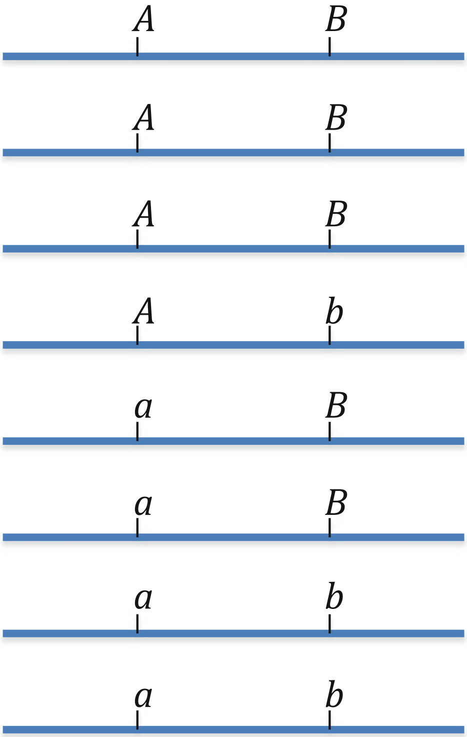
A population is characterized by the frequencies of the four haplotypes, f_{AB}, f_{Ab}, f_{aB}, and f_{ab}. Allele frequencies can be recovered from the haplotype frequencies:$f_A = f_{AB} + f_{Ab}$, etc. Given the haplotype frequencies, we can determine whether an allele's presence in a haplotype is independent of the allele present at the other locus. If they are independent, then the haplotype frequency is the product of the allele frequencies. For example, $f_{AB} = f_A f_B$. If that is the case, the two loci are said to be in linkage equilibrium. If not, they are in linkage disequilibrium, often abbreviated LD. If they are in LD, the extent of LD is quantified by the difference between the actual haplotype frequency and the frequency expected at linkage equilibrium:

$$D = f_{AB} - f_A f_B. \tag{2.1}$$

This calculation is illustrated in Fig. 2.1 in a sample of eight chromosomes. $f_A = 1/2, f_B = 5/8$, and $f_{AB} = 3/8$, which gives $D = 1/16$.

The quantity D is called the coefficient of linkage disequilibrium. $D = 0$ implies there is linkage equilibrium. D can be regarded as a covariance in the allelic state at the two loci.

Fig. 2.1 Illustration of haplotype counts in a hypothetical sample of 8 chromosomes



Although there are four haplotypes, there is only a single D needed to describe the extent of LD. That is clear when the relationship between haplotype and allele frequencies is used. For example, if D is defined by Eq. (2.1), then

$$f_{Ab} - f_A f_b = f_A - f_{AB} - f_A f_b = f_A (1 - f_b) - f_{AB} = f_A f_B - f_{AB} = -D.$$

It is possible, then, to express all the haplotype frequencies in terms of the allele frequencies and D:

$$\begin{aligned} f_{AB} &= f_A f_B + D \\ f_{Ab} &= f_A f_b - D \\ f_{aB} &= f_a f_B - D \\ f_{ab} &= f_a f_b + D. \end{aligned} \tag{2.2}$$

Notice that, if there are more AB haplotypes than expected at linkage equilibrium ($D > 0$), there have to be more ab haplotypes and fewer Ab and aB haplotypes.

Although D is defined by Eq. (2.1), there is an equivalent but different expression:

$$\begin{aligned} D &= f_{AB} - f_A f_B = f_{AB} - (f_{AB} + f_{Ab})(f_{AB} + f_{aB}) \\ &= f_{AB}(1 - f_{AB} - f_{Ab} - f_{aB}) - f_{Ab} f_{aB} = f_{AB} f_{ab} - f_{aB} f_{Ab} \end{aligned} \tag{2.3}$$

Equation (2.2) tells us that, because none of the haplotype frequencies can be negative, there is a limit on the magnitude of D imposed by the allele frequencies. If $D > 0$, D must be no greater than the smaller of $f_A f_b$ and $f_a f_B$, and if $D < 0$, D must be larger than $-f_A f_B$ and $-f_a f_b$. That is,

$$-\min(f_a f_b, f_A f_B) \leq D \leq \min(f_A f_b, f_a f_B). \tag{2.4}$$

One question that arises when computing D for different pairs of loci is whether a particular value is large or small. For example, does $D = 0.006$ indicate a small or large amount of LD for a pair of loci? Inequality (2.4) tells us that the answer depends on the allele frequencies and suggests that it is useful to express D relative to its maximum or minimum possible value. We can do this by defining

$$\begin{aligned} D' &= \frac{D}{\min(f_A f_b, f_a f_B)} \text{ if } D > 0 \\ &= \frac{D}{-\min(f_a f_b, f_A f_B)} \text{ if } D < 0 \end{aligned} \tag{2.5}$$

which indicates how close D is to its maximum or minimum value (Lewontin 1964). In the example with $D = 0.006$, if $f_A = 0.4$ and $f_B = 0.01$, then $D^{'} = 1$, while if $f_A = 0.4$ and $f_B = 0.3$, then $D^{'} = 0.0333$.

2.2 Tests of whether $D = 0$

The quantity D' is, as will be seen below, useful for some purposes, but its value alone does not tell us whether there is statistically significant LD between a pair of loci, that is, whether the hypothesis that $D = 0$ can be rejected. Two tests are commonly used. One is the standard χ^2 test of significance for a 2×2 contingency table whose entries are the numbers of each of the four haplotypes. Closely related to the χ^2 test is the r^2 statistic,

$$r^2 = \frac{D^2}{f_A f_a f_B f_b} \tag{2.6}$$

which provides another way to quantify the extent of LD. r^2 is formally a correlation coefficient. Although the upper bound of r^2 is 1, it does in general not take the value 1 even when $D' = 1$. It is convenient to use r^2 because, when testing for significance in the contingency table, $\chi^2 = nr^2$ where n is the number of haplotypes sampled. This result follows from the fact that D is the difference between the observed haplotype frequency and the haplotype frequency expected if alleles at the two loci are randomly associated. Such a difference arises naturally when doing a χ^2 test of statistical significance in a 2×2 contingency table. When sample sizes are large and neither allele is rare, the χ^2 test is powerful and easy to use. When sample sizes are small or at least one of the haplotypes is rare (<5 in count), then Fisher's exact test is preferable (Weir 1996).

For most practical purposes, D' and r^2 are equally useful, and in real data sets, their values are highly correlated. One important feature of D' is that when its value is 1, at least one of the four haplotype frequencies is 0. That situation is particularly important because, when a new mutation arises at a previously monomorphic locus, $D' = 1$ with all polymorphic loci on the same chromosome. The single copy of the new mutant arises on only one genetic background. After this time, D' between the mutant and another locus becomes less than 1 only if there is recombination between the two loci.

2.3 More than Two Alleles per Locus

Describing LD for a pair of biallelic loci is relatively simple because a single quantity D, combined with the allele frequencies, provides a complete characterization. If there are more than two alleles at either or both loci, more coefficients of LD are needed, one for the difference between the frequency of each haplotype and the frequency expected under random association. If the alleles at one locus are A_1, A_2, A_3 ... and those at the other are B_1, B_2, B_3 ..., then

$$D_{ij} = f_{A_i B_j} - f_{A_i} f_{B_j}. \tag{2.7}$$

The D_{ij} are not independent because of the requirement that allele frequencies at each locus sum to 1. If there are n_1 alleles at the first locus and n_2 at the second, there are $(n_1 - 1)(n_2 - 1)$ independent values of the D_{ij}.

The theory of LD for multiple alleles has been important, particularly in the application to the major histocompatibility complex (MHC) region in humans and other vertebrates. MHC loci often have dozens and even hundreds of alleles, and there is abundant LD among most of the loci (Hedrick et al. 1986). Furthermore, there is evidence of balancing selection which could contribute to the extent of LD. For genomic data, nearly all single-nucleotide polymorphisms (SNPs) and insertion/deletion polymorphisms are biallelic, so a single value of D is sufficient.

2.4 More than Two Loci: Haplotype Blocks and the HapMap Project

When more than two polymorphic loci are analyzed together, as is always the case with genomic data, the analysis can become quite complicated. There are analogs of D defined for three or more loci that represent higher-order linkage disequilibria. For example, with three diallelic loci (*A/a, B/b*, *C/c*), the third-order disequilibrium coefficient is defined to be

$$D_{ABC} = f_{ABC} + f_A D_{BC} + f_B D_{AC} + f_C D_{AB} - f_A f_B f_C \tag{2.8}$$

where D_{AB}, D_{AC}, and D_{BC} are the pairwise coefficients of LD defined above (Geiringer 1944). The higher-order coefficients of LD are analogous to higher-order interaction terms in the analysis of contingency tables with more than two dimensions.

These higher-order coefficients are well defined, and their theoretical properties have been studied extensively, but they are difficult to estimate and interpret. Furthermore, there are many of them because the number of higher-order coefficients grows as an exponential function of the number of loci. Higher-order coefficients of LD have been used primarily in the study of the human MHC loci (Robinson et al. 1991) because there are strong multilocus patterns of LD that are of clinical significance.

The approach much more commonly taken to analyzing numerous polymorphic loci is to compute D' and r^2 for all pairs of polymorphic sites. In the human genome, it is often found that relatively large values of D' and r^2 are found between sets of closely linked sites (Daly et al. 2001). Sets of closely linked sites in strong LD are called haplotype blocks, and they have played an important role in human genetics in the past several years. The discovery that much of the human genome is made up of haplotype blocks was part of the impetus for the HapMap project, which had the goal, now achieved, of finding nearly all common SNPs in several human populations (Consortium 2003, 2005, 2007; HapMap3 2010). The emphasis on common SNPs, i.e., those with allele frequencies in the range (0.05, 0.95) was motivated by the hypothesis that alleles in this frequency range are largely responsible for the genetic basis of complex inherited diseases. For complex diseases, such as most cancers, most forms of heart disease, and many psychiatric disorders and autoimmune diseases, disease risk in close relatives of an affected individual is higher than the average risk in a population, which strongly suggests there is a genetic basis, yet the genetic basis is not attributable to single Mendelian loci.

2.5 Dynamics of *D*

The term "linkage disequilibrium" is unfortunate for two reasons. First, it does not necessarily tell us something about the linkage of two loci. Two loci on different chromosomes might be in linkage disequilibrium, while two closely linked loci might be in linkage equilibrium. Second, the term implies that it describes a dynamic process, but it does not. Instead, D, $D^{'}$, and r^2 quantify the relationship between haplotype and allele frequencies in a population at a given time. We can understand the dynamics of LD by determining how D changes under the influence of various forces, including random mating, natural selection, recombination, and genetic drift.

We will begin with random mating and recombination. We assume that zygotes are formed by randomly combining haplotypes and that the two loci have a recombination rate c between them. We also assume the haplotype frequencies in generation t are f_{AB}, f_{Ab}, f_{aB}, and f_{ab}. We can compute the haplotype frequencies in the next generation ($t + 1$) by assuming gametes are randomly combined into genotypes. Table 2.1 shows the genotypes, the genotype frequencies, and the frequencies of gametes produced by each genotype. It will be necessary to distinguish haplotypes of the two parental gametes of doubly heterozygous individuals, meaning those with genotype *AaBb*, because the gametes produced depend on whether they are doubly heterozygous because their parental gametes are *AB* and *ab* or *Ab* and *aB*.

Table 2.1 Two-locus genotypes and their frequencies in a randomly mating population, along with the frequencies of gametes produced by each genotype

		Gametes produced			
Genotype	Frequency	*AB*	*Ab*	*aB*	*ab*
AB/AB	f_{AB}^2	1	0	0	0
AB/Ab	$2f_{AB}f_{Ab}$	1/2	1/2	0	0
AB/aB	$2f_{AB}f_{aB}$	1/2	0	1/2	0
AB/ab	$2f_{AB}f_{ab}$	$(1–c)/2$	$c/2$	$c/2$	$(1–c)/2$
Ab/ab	f_{Ab}^2	0	1	0	0
Ab/aB	$2f_{Ab}f_{aB}$	$c/2$	$(1–c)/2$	$(1–c)/2$	$c/2$
Ab/ab	$2f_{Ab}f_{ab}$	0	1/2	0	1/2
aB/aB	f_{aB}^2	0	0	1	0
aB/ab	$2f_{aB}f_{ab}$	0	0	1/2	1/2
Ab/ab	f_{ab}^2	0	0	0	1

From Table 2.1, it is straightforward to compute the haplotype frequencies in the next generation. For example,

$$\begin{aligned} f_{AB}\left(t+1\right) &= f_{AB}^2 + 2f_{AB}f_{Ab}\left(1/2\right) + 2f_{AB}f_{aB}\left(1/2\right) \\ &\quad +2f_{AB}f_{ab}\left(1-c\right)/2 + 2f_{Ab}f_{aB}\left(c/2\right) \\ &= f_{AB}^2 + f_{AB}f_{Ab} + f_{AB}f_{aB} + f_{AB}f_{ab} - c\left(f_{AB}f_{ab} - f_{Ab}f_{aB}\right) \\ &= f_{AB} - c\left(f_{AB}f_{ab} - f_{Ab}f_{aB}\right) \\ &= f_{AB} - cD. \end{aligned} \tag{2.9}$$

where the last step uses Eq. (2.3).

We conclude that under random mating, haplotype frequencies change each generation unless $D = 0$. That is in contrast to what happens at each locus separately. The Hardy-Weinberg law tells us that random mating does not change allele frequencies. It is also important that the extent of change in the haplotype frequencies depends on the recombination rate between the two loci. We can find the recursion equation for D alone by using the fact that $f_{AB}(t+1) = f_A(t+1)f_B(t+1) + D(t+1)$, $f_A(t+1) = f_A$, and $f_B(t+1) = f_B$ to obtain

$$f_Af_B + D\left(t+1\right) = f_Af_B + D(t) - cD(t) \tag{2.10}$$

or

$$D\left(t+1\right) = \left(1-c\right)D(t). \tag{2.11}$$

Equation (2.11) tells us that D decreases by a factor of (1–c) after one generation of random mating in a very large population. Because the decrease is by the same factor each generation, D decreases exponentially with time from its initial value:

$$D(t) = (1-c)^t D(0). \tag{2.12}$$

The rate of decrease is determined by the recombination rate between the two loci. If c is small, then $(1-c)^t \approx e^{-ct}$, and we conclude that D will decrease to roughly 37% of its initial value after 1/c generations of random mating.

It is important that D does not go to zero after one generation of random mating. Even between unlinked loci, for which $c = 1/2$, D decreases only by a factor of 1/2 each generation. That behavior is in marked contrast to what happens to genotype frequencies after one generation of random mating. The foundation of population genetics is that the Hardy–Weinberg (HW) genotype frequencies are established in one generation of random mating regardless of the initial genotype frequencies. That D between unlinked loci does not go to zero in a single generation of random mating is surprising because both HW genotype frequencies and linkage equilibrium indicate statistical independence. At the HW frequencies, the presence of an allele on one homologue is independent of the presence of an allele on the other. At linkage equilibrium, the presence of an allele at one locus in a haplotype is independent of the presence of an allele at the other locus. Yet the above results show that independence between homologues is established in one generation, but the independence of loci on different chromosomes is established more slowly.

These results can be related to human populations in a way that illustrates the reason that linkage disequilibrium has become such an important part of human population genetics. There are approximately 24,000 coding genes in the human genome, which has a total recombination length of 30 Morgans or 3000 cM (Lander 2001). Therefore, the average recombination distance between adjacent coding genes is 1/8 cM or $c = 0.00125$. If D is initially nonzero between adjacent coding genes, then D will decrease to 37% of its initial value in $1/0.00125 = 800$ generations. The generation time in humans is about 25 years, so 800 generations represent about 20,000 years. In other words, the extent of linkage disequilibrium between adjacent coding genes in the human genome is expected to decay on a timescale comparable to major events in the history of modern humans, i.e., the colonization of North America and the arrival of agriculture in Europe or the domestication of horses, sheep, and cattle.

Another implication of the preceding theory gives us an additional reason for being concerned with LD. We first recall that Eq. (2.5) tells us that $D' = 1$ when D takes either its maximum positive value or its minimum negative value. In that case, the set of equations in (2.2) implies that at least one of the haplotypes has 0 frequency. The reverse is also true, namely, that if only three of four haplotypes are present in a population, then necessarily $D' = 1$. (Understanding this fact allows a few students in population genetics courses to quickly answer examination questions that occupy the rest of their classmates for a considerable time.)

Now consider what happens when a mutation occurs at a locus that was previously monomorphic and is linked to another locus that is polymorphic. To be specific, suppose the *A/a* locus is polymorphic and the *B/b* locus is initially fixed for *b* (i.e., all individuals in the population carry the *b* allele). In a particular generation, *B* appears in one copy as a new mutant. When *B* appears, it does so on a chromosome that initially carries an *A* or an *a*, but it cannot appear on both types of chromosomes. Suppose it appears on an *A*-bearing chromosome. In that generation, there are then only three haplotypes, *AB*, *Ab*, and *ab*. The fourth haplotype (*aB*) is not present. Therefore, when *B* appears as a new mutant allele, we know that $D' = 1$ regardless of the frequency of *A* and regardless of the recombination distance between *A* and *B*. That is, when a new mutant allele appears, $D' = 1$ between it and every polymorphic locus on the same chromosome.

What happens to D' after *B* appears depends on c. For loci sufficiently far apart on the same chromosome that they are effectively unlinked ($c = 1/2$), D and hence D' decrease rapidly to zero. For more closely linked loci ($c < 1/2$), D and D' decrease more slowly, with the rate of decrease being slowest for very closely linked loci. Thus, each new mutant will be expected to be in strong LD with alleles at closely linked polymorphic loci for a long time, roughly $1/c$ generations. That prediction is valid for every new mutant. Because mutants are appearing every generation, the overall pattern of LD we expect is one in which D' is large between closely linked loci and then decreases with increasing recombination distance between loci.

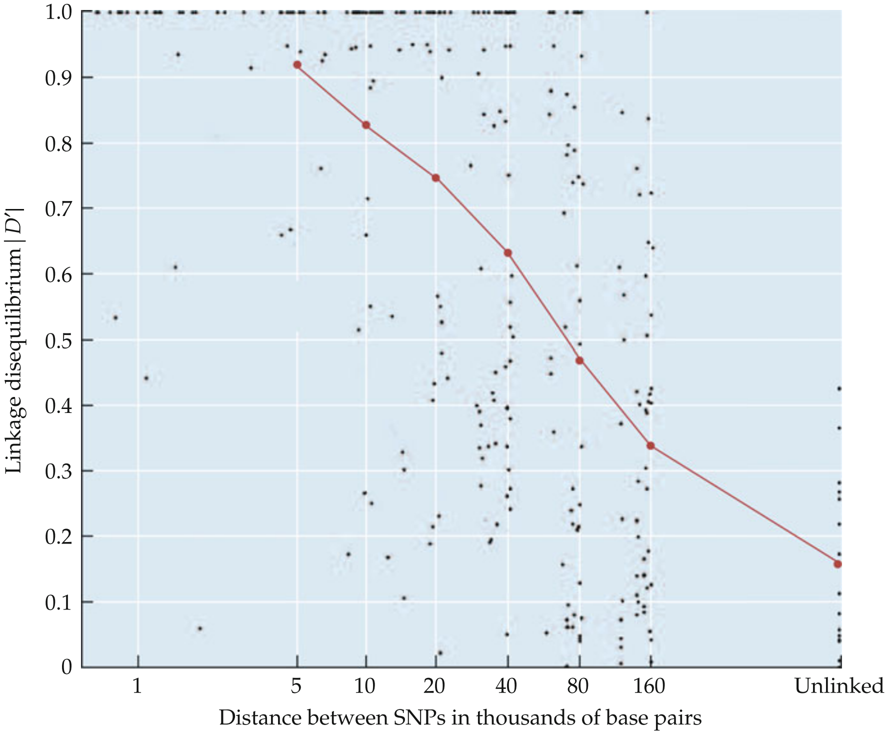


Fig. 2.2 D′ plotted against distance separating SNPs in the human genome. The line indicates the average values of D'. (Reproduced with permission from Nielsen and Slatkin, 2013, *An Introduction to Population Genetics: Theory and Applications,* p. 121, Oxford University Press)

This prediction ignores other population genetic forces, particularly genetic drift and natural selection, which also affect LD, but it reflects the combined effects of recombination and random mating which affect the whole genome.

This prediction is consistent with many observations of LD in the human and other genomes. In humans, significant LD is usually found between polymorphic nucleotide positions that are separated by 50 kb or less but usually less so between sites separated by 100 kb or more (Reich et al. 2001). There is, however, considerable variation in D' values even between sites separated by the same distance (Fig. 2.2), something that is not predicted by the simple theory presented so far.

2.6 Genetic Drift and LD

The preceding theory assumes a population of effectively infinite size. That is what allowed us to assume that allele frequencies do not change from generation to generation. Real populations are of finite size, and that implies that allele frequencies change from generation to generation because of genetic drift (Wright 1931). Genetic drift also affects the extent of LD. The mathematical theory that predicts the effect of drift on D is too complicated to be presented here, but the main conclusions from the theory are relatively simple (Hill and Robertson 1968; Ohta and Kimura 1969). At an equilibrium under drift, random mating, and recombination, genetic drift will maintain some LD between closely linked sites. Although the expected value of D is 0, the expectation of D^2 is nonzero and decreases roughly with $1/c$ as c increases.

2.7 Genealogical Interpretation of LD

There is a close relationship between the gene genealogies of two loci and the extent of LD between them. Refer to Chap. 1 for a discussion of gene genealogies for a single locus. First, consider the case in which there is no recombination between the A and B loci. Because there is no recombination, the gene genealogies of the two loci are the same, as shown in Fig. 2.3. If there is only one mutation at each of the two loci, as shown, no more than three of the four possible haplotypes will be present in the sample. Which haplotype is missing depends on where on the genealogy the two mutations occur. As shown in Fig. 2.3, AB is missing, but if B instead arose on one of the descendent branches carrying A, then aB would be missing. And if A and B happened to have arisen on the same branch, then only ab and AB haplotypes would be present. Therefore, in the absence of recombination and recurrent mutation, $D' = 1$ necessarily, as noted above. It follows that, if $D' < 1$, either recombination or recurrent mutation occurred. At the level of individual nucleotides, recurrent mutation is unlikely, which implies that observing $D' < 1$ for two loci indicates

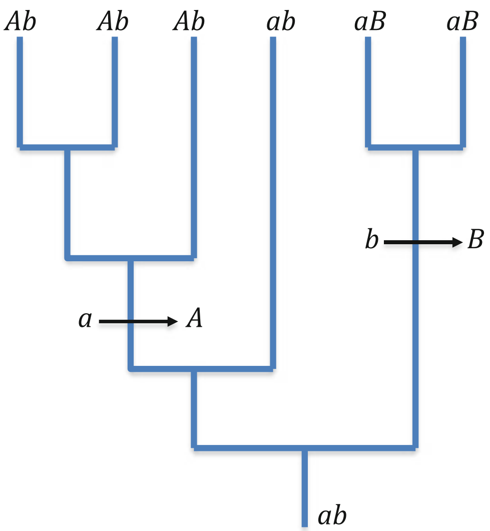


Fig. 2.3 Illustration of the gene genealogy of two completely linked loci, showing the generation of haplotypes by mutation

that recombination occurred between them. When there is recombination, the gene genealogies are no longer the same. Recombination has the effect of breaking the gene genealogy of one of the loci and attaching somewhere else on the genealogy of the other locus. When that occurs, the relationship between D or D' and the recombination rate is no longer simple, and the genealogical approach does not in general lead to tractable analytic results. The similarity of the genealogies at the two loci will be determined, in part, by the recombination rate between the two loci. Loci separated by smaller genetic distances will have genealogies that are more correlated with one another than loci which are farther apart.

2.8 Natural Selection and LD

If the genotypes at two linked loci affect survival and reproduction, the resulting natural selection can increase the extent of linkage disequilibrium under some conditions and even maintain permanent disequilibrium in the face of recombination (Lewontin and Kojima 1960; Felsenstein 1965; Karlin and Feldman 1970). The effect of selection on LD is weak, however, because it depends not on the selection coefficients themselves but on the degree of epistasis, which is necessarily smaller. To illustrate in a simple context, assume there is haploid selection on two linked biallelic loci (A/a and B/b). Let the relative fitnesses of the four haplotypes be w_{AB}, w_{Ab}, w_{aB}, and w_{ab}. Selection of this type will tend to increase D if $R = (w_{AB}w_{ab})/(w_{Ab}w_{aB}) > 1$ (Felsenstein 1965). In other words, R is greater than 1 when the increase in fitness from having A and B together (w_{AB}/w_{ab}) exceeds the product of the gains from having A or B separately (w_{Ab}/w_{ab} and w_{aB}/w_{ab}).

For diploid populations, more complicated but similar conditions have been derived that show when selection can overcome the effects of recombination and random mating and maintain permanent LD. Roughly speaking, permanent LD can be maintained under restrictive conditions, namely, that there has to be overdominance in fitness at each locus and c has to be less than a quantity that summarizes the extent of epistasis in fitness (Karlin and Feldman 1970). It is currently unclear whether epistasis in fitness among closely linked loci contributes to observable patterns of LD.

2.9 Genetic Hitchhiking

Natural selection at a locus affects the frequencies of neutral alleles closely linked to it, a process termed "genetic hitchhiking." (Maynard Smith and Haigh 1974) The idea is simple. As described above, when a new mutant arises, it is in complete LD ($D' = 1$) with alleles at linked polymorphic loci. If that mutant increases rapidly in frequency because it confers a selective advantage to carriers, then neutral alleles on the same chromosome will increase in frequency also. For example, if B arises on an A-bearing chromosome and subsequently increases in frequency, A also will increase in frequency. The result will be an excess of AB chromosomes.

Recombination between the *A* and *B* loci will reduce that excess. Simple theory shows that, if the selection coefficient in favor of *A* is s, then substantial LD will be created by hitchhiking at neutral loci for which $c < s$. That is, neutral loci that are very closely linked to the selected locus will remain in substantial LD with the advantageous allele, while more distant neutral loci will not.

These theoretical results have useful practical applications. If there is substantial LD in a region surrounding a functionally important allele, it is likely that the allele has increased in frequency recently because of positive selection. For example, the *A*– allele of the *G6PD* gene in a west African population has a frequency of 11%. Loci as far away as 700 kb are in significant LD with the *A*– allele, a distance much larger than the normal scale of LD in the human genome (Saunders et al. 2005). Data of this type not only indicate that the *A*– allele increased because of positive selection but also make it possible to infer that the selection coefficient in favor of *A*– was at least 0.05 and that it arose by mutation between 3000 and 6000 years ago (Slatkin 2008).

2.10 Population Subdivision

Population subdivision creates LD when there are local differences in allele frequencies. We can see why by considering a simple example. Suppose that two populations are fixed for different alleles at each of two loci, population 1 is fixed for *A* and *B*, while population 2 is fixed for *a* and *b*. In this case, every individual in both populations is doubly homozygous, either *AABB* or *aabb.* Next, suppose that a researcher who is concerned with LD at these two loci samples individuals from both populations. If the researcher does not realize that there are in fact two distinct populations, individuals from both would be combined into a single sample. The resulting sample would be a mixture of *AABB* or *aabb* individuals. In this sample, only *AB* and *ab* haplotypes would be present, so there is apparently perfect LD between these two loci ($D' = 1$). Yet, that conclusion is obviously an artifact of mixing individuals from two populations with quite different allele frequencies.

Although this example is an extreme case that can be understood without doing any calculations, the conclusion is quite general. If allele frequencies at two loci differ at all between two or more populations and if samples from those two populations are combined, there will in general more LD in the mixture than in the separate populations (Mitton et al. 1973; Nei and Li 1973). This effect is called the "two-locus Wahlund effect" because of its similarity to the classic Wahlund effect, which is the decrease in heterozygosity in a mixture of two or more populations. In the simple example, there are no heterozygous individuals at either locus, which is an extreme case of the Wahlund effect.

It is usually possible to distinguish the two-locus Wahlund effect from selection as a cause of LD because the Wahlund effect affects all pairs of loci at which allele frequencies differ among subpopulations, while selection will probably affect only one genomic region. Still, the two-locus Wahlund effect is important for the design and interpretations of genome-wide association studies (GWAS). GWAS are discussed in greater detail in Chap. 5. Because a GWAS is designed to detect significant LD between alleles that cause a complex disease and SNP markers, the two-locus Wahlund effect can create a spurious signal of association if individuals from different subpopulations are mixed together in the cases and controls. The term “population stratification” is used in this context. It is difficult to completely eliminate the effects of population stratification even if care is taken not to combine individuals from different ethnic groups. The problem is that the actual extent of variation among subpopulations in the frequencies of causative alleles is unknown, and hence, it is not clear how narrowly defined a subpopulation has to be in order to eliminate the effect of population stratification. For example, in carrying out a GWAS in people of European ancestry, is it appropriate to include people of both northern and southern European ancestry in the same study or not? Including both would increase the sample size and hence increase the statistical power to detect significant associations but at the risk of inducing spurious false-positive associations. This trade-off is especially problematic for rarer complex diseases for which the total number of affected individuals might be small. One resolution of the problem is to allow for some population stratification by using overall genomic averages of LD, called “genomic controls,” to infer the overall extent of LD created by subtle population stratification (Devlin et al. 2001).

Gene flow among populations that have diverged can maintain LD in each subpopulation separately. When there is gene flow, the organisms themselves do the mixing and create LD between all pairs of loci that differ in allele frequency among the subpopulations. Mathematical analysis shows that substantial LD between closely linked loci can be maintained by this mechanism (Mitton et al. 1973; Nei and Li 1973).

2.11 Conclusion

When the term linkage disequilibrium was introduced by Lewontin and Kojima (Lewontin and Kojima 1960) in 1960, it was in the context of an abstract mathematical model developed to understand the combined effects of natural selection and recombination in an infinitely large population at equilibrium. Extensive further mathematical studies of LD were carried out in the 1960s and 1970s, but there was almost no attempt to relate the theory to data because almost no information about closely linked loci was available. This was the era during which genetic variation was studied by detecting differences in electrophoretic mobility of proteins (Hubby and Lewontin 1966). Polymorphic protein-coding loci that could be studied with electrophoresis were not usually closely enough linked for LD to be detectable. Linkage disequilibrium remained a somewhat arcane and mathematically difficult part of population genetics, known and appreciated by only a few specialists.

That situation changed with the development of direct means of assessing polymorphisms at the DNA sequence level—first restriction fragment length polymorphisms (RFLPs), then microsatellite loci, and finally SNPs. Instead of being obscure, linkage disequilibrium became well-known, then fashionable, and finally essential. By 1999, that situation changed drastically. In the program for the 1999 annual meeting of the American Society of Human Genetics, 17% of the paper titles or abstracts contained the term "linkage disequilibrium." Since then, LD has only increased in importance in human genetics and is the foundation of GWAS, discussed in Chap. 5. As genomic tools become more widely used in plant and animal populations, LD will assume equal prominence in evolutionary biology.

References

Consortium H (2003) The international HapMap project. Nature 426:789–796

Consortium IH (2005) A haplotype map of the human genome. Nature 437:1299–1320

Consortium IH (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nature 29:229–232

Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theo Popul Biol 60:155–166

Felsenstein J (1965) The effect of linkage on directional selection. Genetics 52:349–363

Geiringer H (1944) On the probability theory of linkage in Mendelian heredity. Ann Math Stat 15:25–57

HapMap3 (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58

Hedrick PW, Thomson G, Klitz W (1986) Evolutionary genetics: HLA as an exemplary system. Academic, New York

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

Hubby JL, Lewontin RC (1966) A molecular approach to study of genic heterozygosity in natural populations. I. Number of alleles at different loci in Drosophila Pseudoobscura. Genetics 54:577–594

Karlin S, Feldman MW (1970) Linkage and selection: two locus symmetric viability model. Theo Popul Biol 1:39–71

Lander ES (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49–67

Lewontin RC, Kojima K (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23–35

Mitton JB, Koehn RK, Prout T (1973) Population genetics of marine pelecypods. 3. Epistasis between functionally related isoenzymes of Mytilus-Edulis. Genetics 73:487–496

Nei M, Li W (1973) Linkage disequilibrium in subdivided populations. Genetics 75:213–219

Ohta T, Kimura M (1969) Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics 63:229–238

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R et al (2001) Linkage disequilibrium in the human genome. Nature 411:199–204

Robinson WP, Cambon-Thomsen A, Borot N, Klitz W, Thomson G (1991) Selection, hitchhiking and disequilibrium analysis at three linked loci with application to HLA data. Genetics 129:931–948

Saunders MA, Slatkin M, Garner C, Hammer MF, Nachman MW (2005) The span of linkage disequilibrium caused by selection on G6PD in humans. Genetics 171:1219–1229

Slatkin M (2008) A Bayesian method for jointly estimating allele age and selection intensity. Genet Res 90:129–137

Weir BS (1996) Genetic data analysis II, Sunderland, Sinauer

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

Analysis of Population Structure

3

Per Sjödin, Lucie Gattepaille, Pontus Skoglund, Carina Schlebusch, and Mattias Jakobsson

Abstract

For humans, like any sexually reproducing diploid organism, mating may be random in the sense that individuals are equally likely to mate and produce offspring. Such a view of a population has been important in population genetics as a basis for modeling and analysis. Population structure denotes deviation from this panmixia, regardless of the cause. In this chapter, we will briefly discuss random mating, populations, population structure, and various methods and practices to infer population structure among individuals from empirical genome-wide data.

3.1 What Is Population Structure?

A loose definition of a "panmictic population" is any collection of randomly mating individuals living at a specific point in time. Strictly speaking, a population is "randomly mating" if the probability to produce offspring is larger than zero

P. Sjödin · L. Gattepaille ·
Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden

P. Skoglund
Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden

The Francis Crick Institute, London, UK

C. Schlebusch · M. Jakobsson (✉)
Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden

Science for Life Laboratory, Uppsala University, Uppsala, Sweden

Department of Anthropology and Development Studies, Centre for Anthropological Research, University of Johannesburg, Auckland Park, South Africa
e-mail: mattias.jakobsson@ebc.uu.se

K. E. Lohmueller, R. Nielsen (eds.), *Human Population Genomics*,
https://doi.org/10.1007/978-3-030-61646-5_3

and equals for all possible pairs of individuals drawn from the population (often, depending on the species, given that a pair consists of a female and a male).

Random mating in this sense is, however, rarely an accurate representation of reality since most populations have a spatial distribution that affects the mating probability of a random pair of individuals. Hence, if by "structured population," we refer to a population that is not randomly mating, then essentially all populations are structured, and a dual categorization of "structured populations" vs "unstructured populations" is not very useful. Instead, it may be better to think of any population as structured and attempt to quantify the degree of structure. The level of population structure can also (often) be accounted for in downstream analyses. In practice, there are often groups of individuals for which no structure is detectable (with the data and methods at hand), and these groups can be regarded as unstructured for most intents and purposes.

It should be noted that this way of thinking about populations does not reflect common practice in which the definition of populations is typically subjective, based on, for example, linguistic, cultural, ethnic, and/or the geographic location of sampled individuals. Moreover, almost all population structure analyses are necessarily based on samples, and detection of stratification within the sample does not necessarily reflect biological populations. For instance (following Pritchard et al. (2000)), imagine a species that lives on a continuous plane but has a low dispersal rate, so allele frequencies vary continuously across the plane. A few clustered sampling points will result in a signal of (a few) clustered genetic groups, which does not give an accurate description of the biological reality. This example illustrates that studies of populations are always indirect and limited to information contained in samples and sensitive to sampling biases.

Accounting for population structure is often crucial in order to reduce both type I and type II errors in statistical analyses of genetic data. For instance, it has been shown that not accounting for population structure can result in spurious signals in association mapping studies and will thus invalidate standard tests (Ewens and Spielman 1995; Pritchard et al. 2000). It is also important to account for population structure in forensic applications like DNA fingerprinting to estimate the probability of random individuals matching a particular profile (Balding and Nichols 1994, 1995; Foreman et al. 1997; Roeder et al. 1998).

Since the first historical opportunity of quantifying molecular genetic variation until today's (almost) complete genome sequencing, numerous molecular techniques have been used to genotype individuals, which in turn can be used to investigate population structure. Early strategies involved typing the human blood groups (Landsteiner and Weiner 1940), followed by the development of allozymes (Lewontin and Hubby 1966), and various forms of DNA fragment length assays, including microsatellites that are abundant in eukaryote genomes (Katti et al. 2001). The sequencing of the human genome (Venter et al. 2001) led to access to a large number of human single-nucleotide polymorphisms (SNPs) as well as the human genome sequence. Various studies have employed novel molecular techniques to investigate human population structure, including classical markers (Cavalli-Sforza et al. 1994), mitochondrial genome (Cann et al. 1987), microsatellites

(Rosenberg et al. 2002), SNPs (Jakobsson et al. 2008; Li et al. 2008), and complete genomes (1000 Genomes, Mallick et al. 2016). Although these different types of data have different properties, the general study of population structure follows straightforward principles of genetic variation, and the difference in datatype can be accounted for by using different assumptions on the mutation model (see e.g., Veeramah and Hammer 2014).

We will review some methods to quantify structure within a population. We will (primarily) assume biallelic markers, where only two states exist that give rise to three possible genotypes. These methods are often part of an initial exploratory step in order to get a better picture of how and to what extent the sample and the population have been influenced by population structure. First, we will present methods where all individuals are treated equally and no a priori information is used for clustering individuals or parts of individual genomes. We will then discuss some methods to contrast clusters of individuals and how this can be used in more explicit demographic modeling. To illustrate concepts and methods, we will analyze a subset of the HGDP data, which has been thoroughly investigated in previous studies (Cann et al. 2002; Li et al. 2008; Jakobsson et al. 2008). This example dataset consists of individuals from Africa—the West African Yoruba, the Southern African San, and the North African Mozabite—and from Europe (France).

3.2 Individual-Based and Unsupervised Methods for Inferring Population Structure

3.2.1 Tree Construction Methods at the Individual Level

Genetic distance is a traditional measure of differentiation in population genetics. Distances are calculated between each pair of individuals (can also be calculated between groups of individuals; see below) and are represented by a pairwise distance matrix. Distance matrices can be visualized by various approaches such as the multidimensional statistics discussed in the next section or in the form of graphs/trees.

A common distance measure between a pair of individuals is the identity by state (IBS) measure. IBS examines biallelic SNPs between two individuals and puts them into one of three categories: identical $= 1$ (e.g., for the genotypes AA and AA, BB and BB, and AB and AB, where A and B denote the two alleles), one-allele-shared $= 0.5$ (i.e., AA and AB; AB and BB), and no-allele-shared $= 0$ (i.e., AA and BB). The state-values are then averaged over all loci to provide genome-wide pairwise IBS similarity values between 1 and 0 for all individual pairwise comparisons. This is summarized in an individual similarity matrix of which 1-IBS will give the distance matrix.

Distance measures based on substitution models for DNA and protein sequence evolution have also been developed. The evolutionary distance between a pair of sequences is measured by the number of nucleotide (or amino acid) substitutions occurring between them. The p-distance is the simplest model and is based on the proportion (p) of nucleotide sites at which two sequences being compared are different. The proportion is obtained by dividing the number of nucleotide differences by the total number of nucleotides compared. It does not make any correction for multiple substitutions at the same site, substitution rate biases (for example, differences in the transition and transversion rates), or differences in evolutionary rates among sites. More specialized measures (e.g., Jukes-Cantor, Kimura 2-parameter, Tamura 3-parameter, Tamura-Nei) take some/more of these complexities into account. The pairwise distances between diploid individuals are then obtained by averaging over the distances obtained for all pairwise comparisons of sequences between the two individuals.

There are several tree construction methods for distance data, two of the most common methods are unweighted pair group method with arithmetic mean (UPGMA) and neighbor joining (NJ) (Saitou and Nei 1987). UPGMA is a simple hierarchical clustering method that combines the nearest two units (individuals or grouped individuals) in a distance matrix into a higher-level cluster. The distance between any two units is the average of all distances between each element of each unit. UPGMA assumes a constant molecular clock model and produces an ultrametric tree (a tree where all the path lengths from the root to the tips are of equal length). Similar to UPGMA, neighbor joining is also a bottom-up clustering method; however, compared with UPGMA, neighbor joining has the advantage that it does not assume that all lineages evolve at the same rate. Both NJ and UPGMA are fast-clustering (tree-building) algorithms, but since only two elements of the distance matrix are considered at a time, they have no optimization criterion to fit the best solution (or tree) over all the data. An optimal criterion method that is commonly applied to distance data is minimum evolution (ME), which accepts the tree with the shortest sum of branch lengths, and thus minimizes the total amount of evolution assumed. Tree-building methods applicable to discrete characters such as nucleotides are also available (e.g., maximum parsimony and maximum likelihood), but these are more applicable to phylogenetic purposes and fall outside the scope of this chapter. Note that inferred trees based on non-recombining chromosomes, such as the mitochondrial genome or the Y chromosome, represent estimates of the genealogy of a specific chromosome (see Chap. 1 for a review of gene genealogies) that may poorly capture an individual's or a population's evolutionary history or structure. Inferred trees based on genome-wide data represent averages over the genealogical process across the genome, and such summary trees capture population structure in a more accurate way.

After an initial tree is produced from the distance matrix, a confidence measure can be calculated making use of procedures such as jackknifing or bootstrapping. The most widely used tool for confidence inference is a version of bootstrapping introduced by Felsenstein (1983). Each bootstrap sample consists of the same number of markers resampled (with replacement) from the original data set and is then subjected to the same distance calculation and tree reconstruction. From these trees produced by bootstrapping, a consensus tree can be constructed in which the confidence of the tree is noted on the nodes as a bootstrap value (the percentage of times the bootstrap procedure supported the specific node). Jackknife is a similar resampling procedure, but in this case, the estimate is systematically recomputed by leaving out one or more observations at a time from the sample set. The bootstrap and jackknife assume near-independence of the markers used, and so the linkage between nearby markers can be accounted for by performing the bootstrap or jackknife procedure over larger blocks of the genome or whole chromosomes.

Various software packages that can handle different types of genetic data are able to calculate many of the distance measures discussed above and give a pairwise distance matrix as output. The distance matrix can then be used for clustering using, e.g., a tree construction algorithm. In certain software, both matrix calculation and tree construction algorithms are available. Some examples of commonly used software are Mega (Kumar et al. 2016), Arlequin (Excoffier and Lischer 2010), and Plink (Purcell et al. 2007).

An example tree of West African Yoruba, European French, North African Mozabites, and Southern African San is given in Fig. 3.1. To construct the tree,

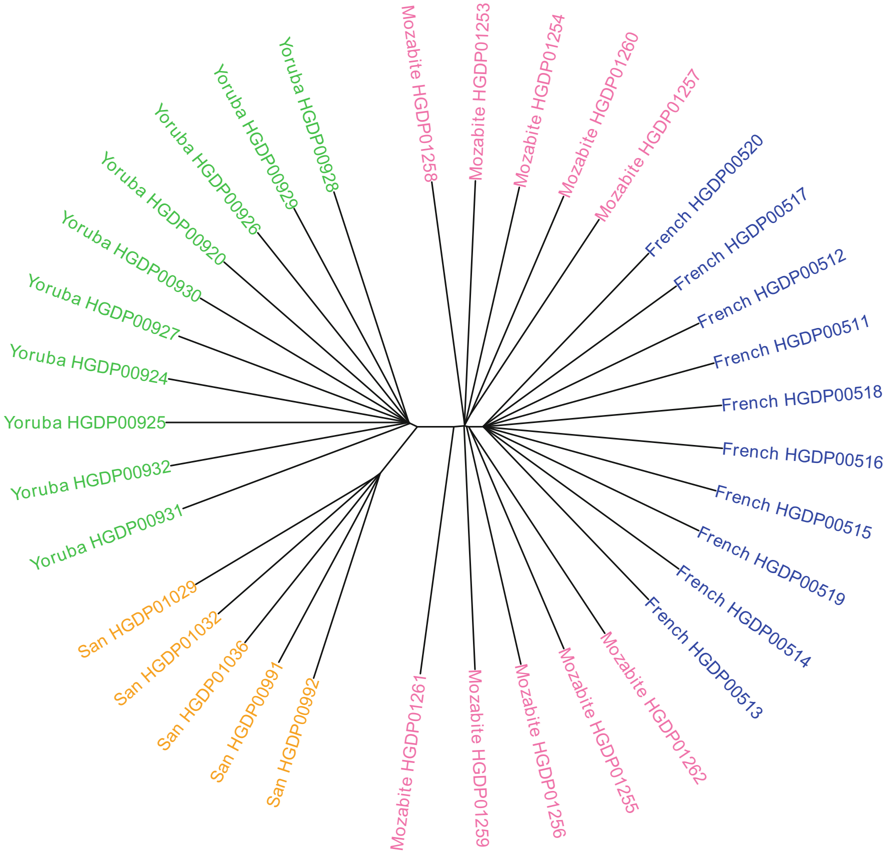


Fig. 3.1 An example neighbor-joining tree of individuals from West Africa (Yoruba), Europe (French), North Africa (Mozabite), and Southern African (San). The NJ tree was built from an IBS distance matrix (computed in Plink) and the R-package *ape*. The individuals cluster according to their sample locations, except for the Mozabites that contain the French sample as a subset

we created an IBS distance matrix in Plink and used the R-package *ape* to construct an NJ tree. We see that the individuals cluster quite well according to their sample locations, except for the Mozabites that contain the French sample as a subset.

3.2.2 Principal Component Analysis and Related Approaches

A single individual can be represented by a position in a multidimensional space where each locus characterizes one dimension. The number of dimensions is therefore very large if we have access to information from many sites; oftentimes, however, many of these dimensions are correlated. A number of methods based on linear algebra aim at finding the best way to summarize and visualize the data in a reduced number of dimensions that capture the greatest axes of variation. These methods, which are typically agnostic to the underlying model of genetic variation, can potentially reveal inherent population structure in a set of sampled individuals. We will describe one of the most widely used method for initial data exploration—principal component analysis (PCA)—and then give a brief overview of a few other related approaches.

The principle of PCA is straightforward: finding and ordering orthogonal axes (or principal components, PCs) that capture the variation of the sample so that the first PC represents the axis of greatest variation in the data, the second PC represents the axis of greatest remaining variation when the data is projected orthogonally to the first PC, and so on, down to the last PC where all variation has been taken into account. Consider, for example, n SNP-loci. Each individual is represented by a vector of n values, with 0 if homozygous for the reference allele, 1 if homozygous for the alternative allele, and 0.5 if heterozygous. PCA performs a rotation of the original n-dimensional orthogonal base, where each of the n loci represents one dimension, into a new orthogonal base, formed by linear combinations of the loci, and defining directions called principal components. The first PC is the direction that maximally explains the variance among individuals when projected into a 1-dimensional space. Together with the first PC, the second PC defines the plane that maximizes the variance of the individuals when projected into a 2-dimensional orthogonal space, and more generally, together with the first k-1 PCs, PC k defines an orthogonal space that maximizes the variance of the individuals when projected into a k-dimensional space. Each PC explains a proportion of the total variance, with the first PC explaining the most variance and the last PC explaining the least.

Applied to our example data of four populations, PCA reveals differences between the groups (Fig. 3.2). We see that the first and second PCs explain 10.8% and 5.5% of the total sample variance, respectively. The first PC separates the sub-Saharan African populations from the European and the North African population; the second PC separates the Southern African San and the West African Yoruba, with the French and the Mozabite in-between. The first two PCs together show the Mozabite individuals in-between the European population and the West African population, suggesting historical models for the ancestry of the Mozabite such as (1) a history of admixture between European and West African groups, (2) shared

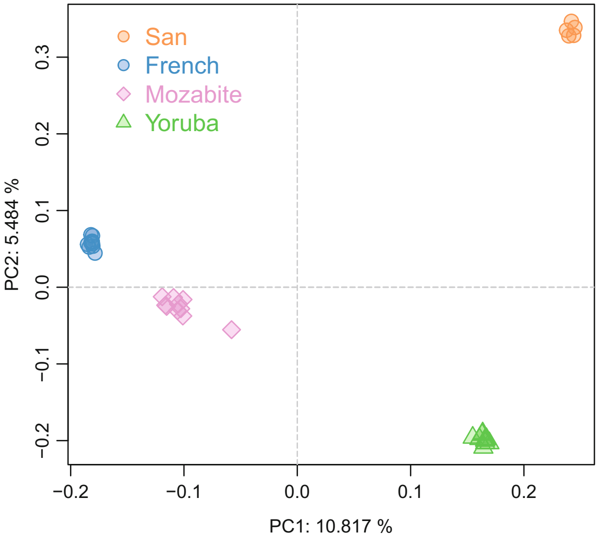


Fig. 3.2 Principal component analysis of our example HGDP data of four populations, with the first two PCs displayed, computed using EIGENSOFT. The first PC explains 10.8% of the total sample genetic variance, and the second PC explains 5.5%. The first PC separates the sub-Saharan African populations from the European and the North African population, and the second PC separates the Southern African San and the West African Yoruba, with the French and the Mozabite in-between. The first two PCs together show the Mozabite individuals distributed between the European population and the West African population, consistent with the Mozabite being an admixed group with a European and a West African source population

ancestry in a treelike population model (without direct admixture) with both these groups, or (3) a combination of the aforementioned (1) and (2).

Investigating outliers, which are easily identifiable using, e.g., PCA, is an important step in many applications. In this particular example, there are no obvious outlier individuals. Outliers are easily identified as individuals "far away" from any other cluster of individuals in PC space. Outliers can be due to low genotype quality (for particular individuals), recent migrants from unsampled populations, or displaying some, potentially unknown, level of population structure in the sample.

The number of PCs to visualize and investigate is arbitrary, but some rule of thumb has been utilized in past studies. Sometimes, this number can be decided by a threshold of variance to be explained (for example, investigate the K PCs that explain at least a fixed percentage, X, of the variance). However, for genome-wide data, the variance explained by each PC, including the first few ones, is typically small due to the high dimensionality (see e.g., Fig. 3.2). Another way to choose the number of PCs to investigate would be to use a "scree plot." A property of PCA is that each PC represents an eigenvector of the variance-covariance matrix of the data (or the correlation matrix if the data is scaled) and is associated with an eigenvalue. The first PC is associated with the highest eigenvalue λ_1, the second PC with the

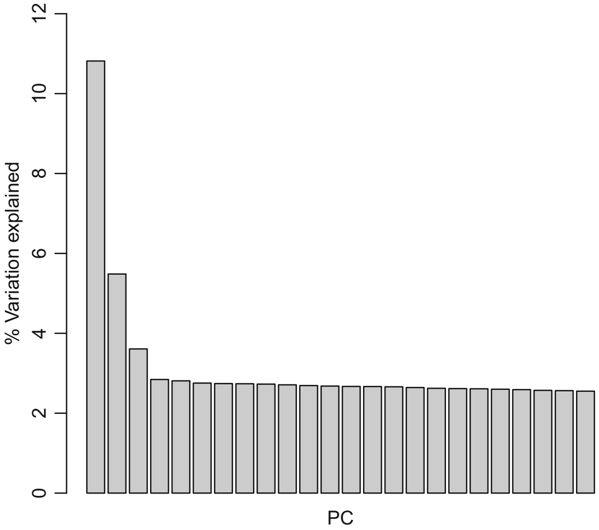


Fig. 3.3 The eigenvalues associated with each PC from our example PCA on HGDP data, sorted decreasingly. In a PCA, the eigenvalues of the variance-covariance matrix (or correlation matrix if the data is scaled) are directly linked to the variance explained by the PCs: the first PC is associated with the highest eigenvalue, the second PC to the second-highest eigenvalue, and so on

second-highest eigenvalue λ_2, and so on. The scree plot is the graph that displays the eigenvalues, sorted decreasingly. In a PCA, the eigenvalues of the variance-covariance matrix (or correlation matrix if the data is scaled) are directly linked to the variance explained by the PCs: λ_i divided by the sum of all eigenvalues gives the proportion of variance explained by the *i*th PC. In our HGDP example, there is a clear flattening of the difference between consecutive PCs from PC4 and onward suggesting that the most interesting patterns can be seen using the first three PCs (see Fig. 3.3). An alternative approach is to test each PC if there is significant evidence for structure (Patterson et al. 2006).

PCA is an important tool for data exploration. For a richer mathematical description and illustrations in different contexts, see Jolliffe (2005). For population genetics, McVean (2009) showed that expected pairwise coalescent times is what determines the primary PCs in a PCA implying that it is impossible to distinguish models with the same expected coalescent times using a PCA approach. He also demonstrated how PCA can, under some models, be used to estimate divergence time between populations, as well as admixture proportions within individuals (McVean 2009). There are several software packages that compute PCA including the R prcomp package and Eigensoft (Patterson et al. 2006).

Multidimensional scaling (MDS) is a group of methods that use a matrix of dissimilarities between individuals and represent the individuals in a smaller number of dimensions, so that the pairwise distances between individuals in the plotting space are good approximations of the original dissimilarities (see Quinn and Keough

2002, for a review on some of these methods). The dissimilarity measure of the individuals in the original data is the scientist's choice, and some examples of useful dissimilarity measures were presented in the previous section. The number of dimensions in the plotting space is chosen in advance and is usually low, to facilitate observation and interpretation of the data.

3.2.3 Ancestry Component Estimation with Few Model Assumptions

The program STRUCTURE (Pritchard et al. 2000) implements a method that makes very few assumptions about the data, and it was one of the first tools that utilized the power of multiple markers for inference. This approach and the many ensuing approaches have become standard in population structure investigations and population-genetic studies in general. STRUCTURE-like methods infer a predefined number of ancestry components (K) among individuals, based on genotype frequencies. Each individual's genotype is assigned to one of K number of clusters with a certain probability. In the first implementation, STRUCTURE searched for the assignment of individuals that minimizes deviation from Hardy-Weinberg equilibrium (see Box 3.1) and linkage equilibrium in each of the K clusters and allowed individuals to be admixed and to have membership proportions to more than one of the K clusters (Pritchard et al. 2000). Population structure is then visible in the dataset as individuals that are closely related having a greater proportion of their genome assigned to the same cluster/s than individuals that are not. This approach analyzed single markers separately and then added up the information to produce a global estimate for each individual. Information about the relative positions of markers to each other was not used and was considered to be segregating independently. This approach works well for low-density marker sets but is less suitable for the high density and full genome datasets that are available today. In the 2003 update of the STRUCTURE algorithm (Falush et al. 2003), sites/markers are not required to be independent, and correlations between subsequent markers due to admixture events are explicitly modeled. This allowed for individual ancestry estimates, known as local ancestry estimates, where the ancestry of chromosomal chunks can be traced along the chromosomes. It also introduced a simplistic model (the F-model, originally described in Nicholson et al. (2002)) to account for correlations of allele frequencies between populations. Although a clearly unrealistic model, it improved the performance of the algorithm considerably (Falush et al. 2003).

Box 3.1 Hardy–Weinberg Equilibrium (HWE)

An assumption of random mating is that the probability to produce viable offspring is equal for all possible pairs of individuals drawn from the

(continued)

population (given that a pair consists of a female and a male). A consequence of this is that the probability of an allele contributed by the mother being of type A is equal to the population frequency of allele type A. The same is true for alleles contributed by the father. Given a population frequency p of allele A and $1-p$ of alleles of type a, the probability of an offspring being of type AA, Aa, and aa are p^2, $2p(1-p)$, and $(1-p)^2$, respectively. When this situation is true, the population is said to be in Hardy-Weinberg equilibrium (HWE).

STRUCTURE has been very popular for population structure inference. However, with the ever-increasing density of genome-wide markers, meeting the computational demands of the algorithm has become a challenge. The Markov chain Monte Carlo method that STRUCTURE employs places a high burden on computer resources for large datasets. This has led to the recent development of alternative approaches, using fast maximum-likelihood-based estimations (FRAPPE (Tang et al. 2005) and ADMIXTURE (Alexander et al. 2009)).

HAPMIX (Price et al. 2009) extends the local ancestry method implemented in the second version of STRUCTURE. It is based on the Li and Stephens (2003) model for patterns of linkage disequilibrium (Li and Stephens 2003) between markers and infers local ancestry estimates of unphased admixed individuals based on the phased haplotype data of exactly two populations. Modeling the full demographic process with recombination and mutation is a notoriously difficult and computationally intractable problem. The Li and Stephens (2003) approach models the $k+1$ haplotype by (imperfectly) copying from the first k "parental" haplotypes where recombination events correspond to changing parental haplotype from which to copy from. Correlations of genealogies (due to linkage) across the sequence is surprisingly well captured by this approach, and importantly, it is sufficiently simple to permit even full genome analyses.

Recent developments along this line have led to a method (CHROMOPAINTER, Lawson et al. (2011)) that does not need discretely defined admixed and parental populations. Instead, each individual in a sample is considered, in turn, both as a recipient and a donor, and chromosomes are reconstructed using blocks of DNA donated by the individuals to each other. Each individual's chromosome is thus "painted" by markers donated by donor individuals in any number of other populations or within the same population. These "painted" chromosomes can be summarized as a co-ancestry matrix, which is proposed to fully capture the information provided by PCA and STRUCTURE-like methods (also for nonindependent sites). In addition, consecutive markers that are in linkage disequilibrium are combined into haplotypes, which increases the ability of the method to observe subtle population structure. A downstream model-based extension (fineSTRUCTURE, Lawson et al. (2011)) is then used to identify discrete populations using the inferred co-ancestry matrix.

A subclass of population admixture models can utilize the spatial coordinates of sampled individuals (BAPS (Corander et al. 2003), TESS (Chen et al. 2007), GENELAND (Guillot et al. 2005)). While in STRUCTURE-like methods, the assignment to a population is independent and identical among all individuals in the dataset, this class of methods takes into account (a priori) the spatial distribution of individuals and aims to detect genetic discontinuities in space. Since geographic spatial correlation is often present among individuals and populations, it can be useful to incorporate spatial coordinates into the population structure analysis. Recent attempts to incorporate geographic information into population structure estimations involve approaches that use Wishart distributions (a generalization of multidimensional gamma distributions) to model genetic similarity as a function of spatial distance. In uniform isolation-by-distance scenarios, genetic distances visualized in two dimensions should mirror the samples/individuals in geographic space. Migration and admixture and hinders to gene flow would disturb this correlation. In one approach, implemented in the software SpaceMix, a covariance model of genetic data is used to build maps of the geographic positions of the populations, but distances are distorted according to inferred rates of gene flow (Bradburd et al. 2016). Barriers to gene flow result in larger distances between groups, while migration and admixture can be identified as abnormal strong covariances over long distances. The inferred admixture is then estimated and represented as "arrows," on a generated map, from the source population to the recipient population. Another approach based on the Wishart distribution, EEMS (Petkova et al. 2016), uses pairwise genetic similarities among populations and estimates a surface map of effective migration rates. The effective migration rates are scaled by effective population sizes under an equilibrium model. These methods that incorporate spatial information may highlight important features of population structure that might have remained undetected using other, spatially "blind," methods for inferring population structure. Two other methods that use F_{ST} measures between populations to identify violations of isolation-by-distance patterns have also been developed (Duforet-Frebourg and Blum 2014; Jay et al. 2013).

For our example dataset, we ran 10 iterations in the program ADMIXTURE at $K = 2$ to $K = 5$. The iterations at each value of K were then compared to detect different clustering solutions using the program CLUMPP (Jakobsson and Rosenberg 2007). For $K = 2$ to $K = 4$, all 10 iterations arrived at very similar solutions, and the combined output is shown in Fig. 3.4 (visualized with the program DISTRUCT (Rosenberg 2004)). The analysis shows clustering of sub-Saharan Africans (orange component) and clustering of Europeans and North Africans (blue component) at $K = 2$, although Yoruba individuals show a small fraction of ancestry from the blue component, and the Mozabites show a small ancestry fraction from the orange component. At $K = 3$, the Yoruba obtains its own cluster (green), while the Mozabite still clusters with the French but showing some ancestry fraction from the green component. The Mozabite forms its own group at $K = 4$, but some individuals showed shared ancestry with Yoruba and French. For $K = 5$, there was no common solution, and each of the 10 iterations had a different solution (this pattern can be seen clearly from the similarity matrix output of CLUMPP). The lack of a "common

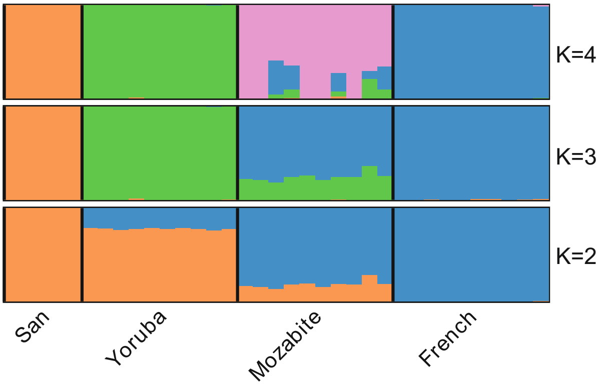


Fig. 3.4 Combined output from 10 iterations of running ADMIXTURE with our example HGDP data analysis for $K = 2$ to $K = 4$ visualized by DISTRUCT. For $K = 5$, there was no common solution, and each of the 10 iterations had a different solution. The analysis show clustering of sub-Saharan Africans (orange component) and clustering of Europeans and North Africans (blue component) at $K = 2$, although Yoruba individuals show a small fraction of ancestry from the blue component and the Mozabite show a small ancestry fraction from the orange component. At $K = 3$, the Yoruba obtains its own cluster (green), while the Mozabite still clusters with the French but showing some ancestry fraction from the green component. The Mozabite forms its own group at $K = 4$, but some individuals showed shared ancestry with Yoruba and French. Note that these algorithms, like the ADMIXTURE algorithm, is typically set up to only utilize the genetic information and to be agnostic to all other information (and other information like self-identified ancestry/ethnicity, geographic sample location, and/or language can be added onto the results for visualization purposes)

mode" in the iterations at $K = 5$ is an indication that there is no additional level of structure to reveal by dividing the individuals' genomes into additional ancestry components. In summary, we note that these three choices of assumed number of clusters ($K = 2$, 3, and 4) all reveal interesting patterns of population structure that are related to a hierarchical ancestry relationship among the four populations (e.g., Jakobsson et al. 2008; Schlebusch et al. 2012). We further note the interesting pattern for the Mozabite that display ancestry components related to Europeans and West Africans at $K = 3$ but that form their own cluster (to a large extent) at $K = 4$—a pattern consistent with a population with mainly Eurasian ancestry, followed by some level of admixture with West Africans. This admixture likely happened some time ago since the Mozabites make up their own cluster at $K = 4$, which is consistent with subsequent genetic drift in the Mozabites since the admixture.

3.3 Population-Based and Supervised Methods

Once an overview of the signals in the data has been obtained using PCA, STRUCTURE-like analyses, and tree construction at the individual level, a natural next step is to assign individuals to predefined populations. We may then want to quantify the amount of structure among groups in order to learn something about the demographic past of the full collection of individuals. Although these populations are ideally identified with the aid of the previously presented methods, it is not uncommon to have assessments based on geography-only defined populations. It may still be of value to contrast these predefined populations in order to affirm that they do correspond to separate biological populations.

3.3.1 Genetic Differentiation at the Population Level

3.3.1.1 F_{ST}

Introduced by Sewall Wright (Wright 1949), F_{ST} is one of the first measures of genetic differentiation (sometimes referred to as "genetic distance") among or between populations. There are many variations on the original definition, and the usefulness of F_{ST} and relatives is still a debated topic (e.g., Holsinger and Weir 2009; Rousset 2013; Jost 2008; Ryman and Leimar 2009). F_{ST} was originally defined as the correlation between gametes chosen randomly from within the same subpopulation relative to the entire population or, equivalently, as the departure of genotype frequencies from Hardy–Weinberg expectations relative to the entire population (see Holsinger and Weir 2009 for a thorough review of F_{ST}). A common definition for more practical purposes is

$$F_{\mathrm{ST}} = \frac{\mathrm{Var}\left(p_i\right)}{E\left[p_i\right]\left(1 - E\left[p_i\right]\right)},$$

(e.g., Holsinger and Weir 2009) where Var(p_i) is the variance in allele frequencies across subpopulations and $E[p_i]$ is the expected allele frequency. There are other formulations, including (Nei 1973) in terms of heterozygosity,

$$G_{\mathrm{ST}} = \frac{H_{\mathrm{T}} - H_{\mathrm{S}}}{H_{\mathrm{T}}},$$

where H_{T} is the total (pooled) heterozygosity and H_{S} is the mean heterozygosity across subpopulations. Note that these definitions are all coined in terms of genetic variation. If demography is the primary interest, these definitions may not be ideal since the distribution of genetic variation depends on the mutational process as well as demography—via the genealogical process. Slatkin (1995) isolated the purely genealogical aspect of F_{ST} by studying the limit as the mutation rate approached zero and showed that F_{ST} can be expressed in terms of expected coalescent times

$$F_{\mathrm{ST}} = \frac{t_s - t_w}{t_s},$$

where t_s is the average time for two randomly picked genes—from the whole population—to find a common ancestor and t_w is the average time for two genes from the same subpopulation to find a common ancestor. The intuition for this definition is that it measures the relatively longer time it takes for two genes situated in individuals from different subpopulations to find a common ancestor compared to when they are situated in individuals from the same subpopulation. From this definition, it is clear that if the expected coalescent time for two genes does not depend on which population the genes are drawn from ($t_s = t_w$), then $F_{\text{ST}} = 0$. In contrast, if genes from different populations take much longer to find a common ancestor than genes from the same population, F_{ST} tends toward 1.

In order to estimate F_{ST}, we need genetic variation from individuals drawn from predefined populations. Depending on the type of genetic data, different assumptions of the mutational process can be used. For sequence data, the mutational process is well approximated by the infinite sites model, for which at most one mutation is allowed for each site. The mutation rate per site is typically very low and the influence of branch-specific, novel mutations when estimating F_{ST} will be assumed to be negligible compared to demographic factors that affect sites polymorphic in the ancestral population to the predefined populations. Alternatively, an outgroup can be utilized to delimit the data to SNPs that were polymorphic prior to the time period of interest. In order to account for the sample variance (due to limited sample sizes), Weir and Cockerham (1984) developed a robust (and commonly used) estimator for F_{ST} (see also Weir 1996; Bhatia et al. 2013).

Model-specific demographic parameters such as migration rate and/or divergence time can often be directly related to F_{ST}, although caution is warranted for directly equating an F_{ST} estimate with a specific demographic parameter as there are many different factors that influence estimates of F_{ST}. For instance, in a two-population divergence model, the relationship between F_{ST} and the divergence time t is (Slatkin 1995):

$$F_{\text{ST}} = \frac{t}{t + 8N_e},$$

while in an infinite island model, the migration rate m is related to F_{ST} as (see Fig. 3.5 and Box3.2)

$$F_{\text{ST}} = \frac{1}{1 + 4N_e m}.$$

Note that estimates of F_{ST}, like many other population genetic parameters, depend on genetic drift and include the term of effective population size. Hence, estimates of F_{ST} transformed into estimates of other population genetic parameters, such as divergence time or migration rate, are typically estimates of the scaled (in terms of N_e) parameter.

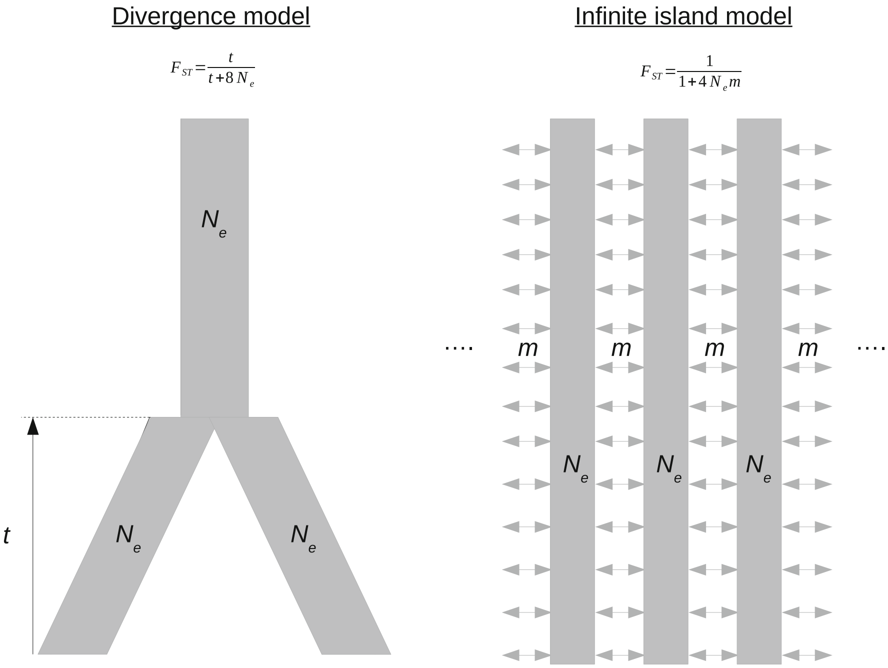


Fig. 3.5 Two simple demographic models and their relation to F_{ST} (in the limit as the mutation rate approaches 0 (Slatkin 1995)). The left model is a two-population split model with no migration, and the scaled divergence time $t/2N_e$ under this model can be estimated by $4F_{ST}/(1-F_{ST})$. The right model illustrates the infinite island model with an infinite number of equally sized subpopulations with a constant migration rate between neighboring subpopulations. Under this model, the scaled migration rate $4N_em$ is estimated by $(1-F_{ST})/F_{ST}$

Box 3.2 F_{ST}

Applied to the example data, the pairwise F_{ST} values, estimated using Weir and Cockerham's (1984) estimator, are displayed in the table below. The largest estimate of F_{ST} is found between the San and the French leading to the largest estimated divergence time (scaled by effective population sizes) or, alternatively, the smallest estimated migration rate, between these two populations. Note that because San ancestors may have diverged earlier than any of the other populations (Schlebusch et al. 2012), we may expect that the divergence time between San and any of the other population would be equal, but because time is scaled in N_e, this is not necessarily the case. The large divergence between San and French is, for instance, likely to be a consequence of a relatively small N_e for the French. Interestingly, F_{ST} is markedly smaller between the French and the Mozabite than between the Mozabite and the

(continued)

Yoruba. Under a pure migration model, this would suggest more gene flow between the Mozabite and the French compared to between the Mozabite and the Yoruba. An alternative interpretation (under a simple divergence model) would state that the first population split was between Yoruba and the ancestor population to the French and the Mozabite. As both these models are highly unlikely to be a good approximation to human demographic history in this particular case, caution is warranted about such direct interpretations when the underlying model is largely unknown. From the PCA and STRUCTURE analyses, we see indications of more complicated demographic scenarios that incorporate both gene flow and population divergence, which can be reconciled with these pairwise F_{ST} values.

	F_{st}
San vs. French	0.105
San vs. Mozabite	0.091
Yoruba vs. French	0.0905
Yoruba vs. Mozabite	0.073
San vs. Yoruba	0.0511
French vs. Mozabite	0.0185

3.3.1.2 Other Measures of Genetic Distance

There are many alternative measures of genetic distance between populations. The simplest distance measure is the Euclidian distance between two points in a multidimensional space. Many variations of this distance are available such as Rogers's (1972) scaled Euclidian distance, Prevosti et al.'s (1975) distance, Cavalli-Sforza and Edwards' (1967) chord distance, and Nei et al.'s (1983) D_{A} distance. These are all geometric distances and do not involve any evolutionary models.

Other, more model-based, distance measures are Cavalli-Sforza's chord distance (1969); Reynolds, Weir, and Cockerham's θ_{W} (1983), and Nei's (1972) D_{s}. The first two measures utilize existing variation (without modeling the possibility of new mutations), while Nei's D_{s} includes the possibility of new mutations occurring in an infinite allele mutation model.

Distance measures have also been designed to handle the stepwise mutation model (SMM) that can be useful for microsatellite data. Goldstein et al.'s (1995) $(\delta\mu)^2$ distance is commonly used, as is the closely related average square distance (ASD) (Slatkin 1995). Two other commonly employed distances for microsatellites are Shriver et al.'s (1995) distance and the shared allele distance D_{SA} (Chakraborty and Jin 1993).

3.3.2 Formal Tests for Admixture Under a Population Tree-Model

Once we have some proposed demographic model/s, we can start estimating demographic parameters in these models. An alternative, but not mutually exclusive, way to proceed is to construct formal tests of these models. The proposed model is used as a null model to make predictions, and then these predictions are compared to the observed data in order to test if the data is consistent with the model. A recent suite of tests along these lines is the 3-population test, the 4-population test (Reich et al. 2009), and the D-test (Green et al. 2010). These have been successfully applied to test for admixture among human populations as well as for identifying a significant level of admixture from archaic humans (Neandertals and Denisovans) among human populations (Green et al. 2010; Reich et al. 2010). These methods, collectively referred to as f-statistics (in contrast to Wright's F-statistics), relate the expected covariances in allele frequencies between not only 2 but also 3 and 4 populations in a bifurcating population phylogeny with the possibility of punctual admixture events.

The f_3 statistic, or 3-population test, is computed as the product $(p_{\mathrm{X}} - p_A)(p_{\mathrm{X}} - p_{\mathrm{B}})$, where p_{X}, p_{A}, and p_{B} are the allele frequencies at each locus in population A, B, and X. The expected value of this product is positive under a tree model, but the estimate from data can be negative under certain admixture scenarios (which violate the tree model), and negative f_3-statistics can only occur due to admixture events.

The f_4 statistic, or 4-population test, is computed as the product $(p_{\mathrm{A}} - p_{\mathrm{B}})(p_{\mathrm{X}} - p_{\mathrm{Y}})$, where p_{A}, p_{B}, p_{X}, and p_{Y} are the allele frequencies at each locus in population A, B, X, and Y. This product is expected to be 0 if the 4 populations are related by an unrooted phylogeny of the form (A, B), (X, Y) without admixture. Violations of this assumption can create (significantly) positive or negative values where the sign of the statistic contains information on the direction of the admixture.

The D-test is a version of the f_4 statistic with a denominator that includes a term for heterozygosity. Jackknife or bootstrap permutation tests of chromosomes or blocks of the genome can be used to assess statistical uncertainty and perform hypothesis tests using these statistics (Reich et al. 2009).

We illustrate these methods by performing the D-test on our data (Table 3.1). Using San as the outgroup, we see that the single-tree hypothesis with the smallest deviation from $D = 0$ has the Mozabite and the French as the closest related populations. However, the negative D-value for this tree suggests gene flow from the Yoruba into the Mozabite. This result is consistent with the Mozabite having ancestry related to both Yoruba and the French with more gene flow from the French than from the Yoruba. This result closely mirrors the analysis based on F_{ST}, and it is (supposedly) robust to effects of genetic drift (e.g., from different effective population sizes in the different populations), which could impact F_{ST} results. However, we cannot rule out alternative models without a more detailed model of genetic drift in the population history model.

Table 3.1 *D*-test for different topologies (W, (X, (Y, Z)))

W	X	Y	Z	D-stat	Z-score
San	Yoruba	Mozabite	French	−0.0089	−8.018
San	Mozabite	Yoruba	French	0.1358	67.627
San	French	Yoruba	Mozabite	0.1445	76.227

A large absolute value of the Z-score indicates a poor fit of the observed data and the proposed topology. If gene flow involving the outgroup lineage can be ignored, a negative D value suggests gene flow between X and Y, while a positive D value indicates gene flow between X and Z

3.3.3 More Advanced Modeling

3.3.3.1 Population Graph Fitting

The f-statistic framework in Reich et al. (2009) where the 3- and 4-population tests were introduced also in a more complex model fitting framework where a multi-population model can be fitted so that population topology, admixture events, and genetic drift along lineages are fitted to the observed f-statistics. This approach has been implemented in the package qpgraph (Patterson et al. 2012) and in MixMapper (Lipson et al. 2013). A similar approach of fitting ancestry graphs to genetic data was attained by Pickrell and Pritchard (2012). Their method, implemented in the software TREEMIX, finds the tree structure with potential admixture events between populations that best explains the observed matrix of allele frequency covariances between populations.

3.3.3.2 Isolation-Migration Models

Several methods that attempt to co-estimate effective population sizes, divergence time, and migration rates in a 2-population "isolation-migration" (IM) model setting have been developed. In one line of approaches, the estimates are based on haplotype information using a Bayesian framework (Nielsen and Wakeley 2001; Hey and Nielsen 2004, 2007). Here the haplotypes are assumed to be known and that there is no intra-locus recombination. The latter assumption has been relaxed in the implementation of MIMAR (Becquet and Przeworski 2007). These methods are usually too computationally intensive to be applied to full genomic data. Alternatively, loci are assumed to be independent, and the full joint frequency spectrum is utilized in a composite likelihood approach to estimate the migration rates and divergence time (Gutenkunst et al. 2009). ABC methods (see below) have also been developed specifically for the IM model (Lopes et al. 2009; see also Tellier et al. (2011)).

3.3.3.3 Approximate Bayesian Computation

Approximate Bayesian computation (ABC) is a powerful and extremely flexible approach to fit and compare models to real data that does not rely on calculating the full likelihood of the data given a model (see, for instance, Beaumont et al. 2002; Csilléry et al. 2010). Instead, some (well-chosen) summary statistics calculated for simulated data are compared to the values of these summary statistics observed from the real data. The simulations are performed by drawing model parameters from prior distributions and then choosing those simulations that best mimic the real data. The distribution of the model parameters in this chosen set of best fitting simulations can then be used to estimate the model parameters. Different models can then be contrasted using Bayes factors. The ABC approach has proven very flexible for inferring model parameters, and there is an active community developing novel and faster algorithms (e.g., Pudlo et al. 2016; Csilléry et al. 2012).

3.4 Summary and Guidelines

Good practice when investigating a population-genetic dataset for population structure is to start by visualizing the data in a way that reveals the inherent characteristics of the data. To get an indication of whether or not the data contains different groups; PCA, simple tree-building methods, and "STRUCTURE-like" approaches are all good tools for initial data exploration. Once some overview of the data is obtained, we can start building simple models to investigate additional hierarchical patterns and characteristics of the underlying demographic history of the sample and population/s. More detailed hypotheses can subsequently be scrutinized using more explicit and advanced models. The more accurate models of demography we can infer, the better we understand the underlying processes shaping the population genetic patterns of variation. Ultimately, this understanding may allow in-depth analysis of the genetic architecture of traits and patterns of selection impacting the genome (Li et al. 2012), after controlling for patterns caused by demographic history manifesting as population structure.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664

Balding DJ, Nichols RA (1994) DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 64:125–140

Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96:3–12

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162:2025–2035

Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. Genome Res 17:1505–1519

Bhatia G, Patterson N, Sankararaman S, Price AL (2013) Estimating and interpreting FST: the impact of rare variants. Genome Res 23:1514–1521

Bradburd GS, Ralph PL, Coop GM (2016) A spatial framework for understanding population structure and admixture. PLoS Genet 12:e1005703

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V et al (2002) A human genome diversity cell line panel. Science 296:261–262

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis -models and estimation procedures. Am J Hum Gen 19:233–257

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The History and Geography of Human Genes. Princeton University Press, Princeton, NJ

Chakraborty R, Jin L (1993) A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances. In: DNA fingerprinting: state of the science. Birkhäuser, Basel, pp 153–175

Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Mol Ecol Notes 7:747–756

Corander J, Waldmann P, Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. Genetics 163:367–374

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian computation in practice. Trends Ecol Evol 25:410–418

Csilléry K, François O, Blum MGB (2012) abc: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol 3:475–479

Duforet-Frebourg N, Blum MGB (2014) Nonstationary patterns of isolation-by-distance: inferring measure of local genetic differentiation with Bayesian kriging. Evolution 68:1110–1123

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57:455–464

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Res 10:564–567

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Felsenstein, J (1983) Parsimony in systematics: biological and statistical issues. Annu Rev Ecol Syst 14:313–333

Foreman L, Smith A, Evett I (1997) Bayesian analysis of DNA profiling data in forensic identification applications. J R Stat Soc A 160:429–469

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 92:6723–6727

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U et al (2010) A draft sequence of the Neandertal genome. Science 328:710–722

Guillot G, Estoup A, Mortier F, Cosson JF (2005) A spatial statistical model for landscape genetics. Genetics 170:1261–1280

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167:747–760

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci USA 104:2785–2790

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet 10:639–650

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM et al (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998–1003

Jay F, Sjödin P, Jakobsson M, Blum MGM (2013) Anisotropic isolation by distance: the main orientations of human genetic differentiation. Mol Biol Evol 30:513–525

Jolliffe I (2005) Principal component analysis. Wiley, New York

Jost L (2008) G(ST) and its relatives do not measure differentiation. Mol Ecol 17:4015–4026

Katti MV, Rajekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161–1167

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874

Landsteiner K, Weiner AS (1940) An agglutinable factor in human blood recognized by immune sera for rhesus blood. Proc Soc Exp Biol NY 43:223

Lawson DJ, Hellenthal G, Myers S, Falush D (2011) Inference of population structure using dense haplotype data. PLoS Genet 8:e1002453

Lewontin RC, Hubby JL (1966) A molecular approach to the study of genetic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics 54:595–609

Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–2233

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM et al (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104

Li J, Li H, Jakobsson M, Li S, Sjödin P, Lascoux M (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? Mol Ecol 21:28–44

Lipson M, Loh PR, Levin A, Reich D, Patterson N, Berger B (2013) Efficient moment-based inference of population admixture parameters and sources of gene flow. Mol Biol Evol 30:1788–1802

Lopes JS, Balding D, Beaumont MA (2009) PopABC: a program to infer historical demographic parameters. Bioinformatics 25:2747–2749

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M et al (2016) The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature 538:201–206

McVean G (2009) A genealogical interpretation of principal components analysis M. PLoS Genetics 5:e1000686

Nei M (1972) Genetic distance between populations. Am Nat 106:283–292

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. II Gene frequency data. J Mol Evol 19:153–170

Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánsson K, Donnelly P (2002) Assessing population differentiation and isolation from single nucleotide polymorphism data. J R Stat Soc B 64:695–715

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158:885–896

Patterson N, Price AL, Reich D (2006) Population structure and eigen analysis. PLoS Genetics 2:e190

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N et al (2012) Ancient admixture in human history. Genetics 192:1065–1093

Petkova D, Novembre J, Stephens M (2016) Visualizing spatial population structure with estimated effective migration surfaces. Nat Genet 48:94–100

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

Prevosti A, Ocana J, Alonzo G (1975) Distances between populations for Drosophila subobscura based on chromosome arrangement frequencies. Theor Appl Genet 45:231–241

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N et al (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet 5:e1000519

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP (2016) Reliable ABC model choice via random forests. Bioinformatics 32:859–866

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR et al (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am Hum Genet 81:559–575

Quinn GP, Keough MJ (2002) Experimental design and data analysis for biologists. Cambridge University Press, Cambridge

Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. Nature 461:489–494

Reich D, Green RE, Kircher M, Krause J, Patterson N et al (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060

Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767–779

Roeder K, Escobar M, Kadane JB, Balazs I (1998) Measuring heterogeneity in forensic databases using hierarchical Bayes models. Biometrika 85:269–287

Rogers JS (1972) Measures of similarity and genetic distance. In: Studies in genetics VII. University of Texas Publication 7213. Austin, Texas, pp 145−153

Rosenberg NA (2004) Distruct: a program for the graphical display of population structure. Mol Ecol Notes 4:137–138

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK et al (2002) Genetic structure of human populations. Science 298:2381–2385

Rousset F (2013) Exegeses on maximum genetic differentiation. Genetics 194:557–559

Ryman N, Leimar O (2009) G(ST) is still a useful measure of genetic differentiation – a comment on Jost's D. Mol Ecol 18:2084–2087

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D et al (2012) Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science 338:374–379

Shriver M, Jin L, Boerwinkle E, Deka R, Ferrell RE et al (1995) A novel measure of genetic distance for highly polymorphic tandem repeat loci. Mol Biol Evol 12:914–920

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462

Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol 28:289–301

Tellier A, Pfaffelhuber P, Haubold B, Naduvilezhath L, Rose LE et al (2011) Estimating parameters of speciation models based on refined summaries of the joint site-frequency spectrum. PLoS One 6:e18155

Veeramah KR, Hammer MF (2014) The impact of whole-genome sequencing on the reconstruction of human population history. Nat Rev Genet 15:149–162

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al (2001) The sequence of the human genome. Science 291(5507):1304–51. https://doi.org/10.1126/science.1058040. Erratum in: Science 292(5523):1838 (2001). PMID: 11181995.

Weir BS (1996) Genetic data analysis II. Sinauer Associates, Sunderland

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370

Wright S (1949) The genetical structure of populations. Ann Hum Gen 15:323–354

Types of Natural Selection and Tests of Selection

4

David Enard

Abstract

Natural selection leaves distinct signals at loci in the genome that have experienced deleterious or advantageous mutations during their evolution. This chapter provides an overview of the main signatures left by natural selection both at the level of genetic variation within populations and at the level of differences between species. The chapter also details a number of classic summary statistics that exploit these signals to detect and quantify natural selection and the rationale behind these statistics. As such, it offers a basic introduction.

4.1 Types of Selection and Their Effect on Linked Neutral Sites

4.1.1 Types of Selection

Different types of natural selection are distinguished based on the dynamics of the selected alleles. There are three broad categories: positive directional selection, balancing selection, and purifying selection.

Alleles under positive directional selection are advantageous alleles (here advantageous means higher reproductive success) where individuals carrying the allele have a reproductive advantage. In the case of directional selection, assuming the mutation becomes established in the population, the greater selective advantage quickly drives the frequency of advantageous alleles from close to zero when they first appear to 100% after a number of generations.

D. Enard (✉)
Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA
e-mail: denard@email.arizona.edu

K. E. Lohmueller, R. Nielsen (eds.), *Human Population Genomics*,
https://doi.org/10.1007/978-3-030-61646-5_4

Alleles under balancing selection are alleles where heterozygous individuals have a greater selective advantage than homozygous individuals (heterozygote advantage) or where the selective advantage of either heterozygous individuals carrying the allele on one chromosome or homozygous individuals carrying the allele on both chromosomes depends on the frequency of the allele (frequency-dependent selection). In both cases of heterozygote advantage and frequency-dependent selection, the effect of selection is to maintain the frequency of the selected alleles at intermediate frequencies, for a number of generations greater than expected under genetic drift alone.

Alleles under purifying selection, also known as negative selection, are alleles with deleterious consequences for their carriers. Deleterious alleles reduce the reproductive success of individuals. By reducing reproductive success, deleterious alleles are removed from populations because individuals carrying them reproduce less than individuals free of such alleles. Strongly deleterious alleles are eliminated from populations rapidly enough that they never make it to appreciable frequencies. Weakly deleterious alleles can initially reach appreciable frequencies due to genetic drift and sometimes, in small populations, can become fixed by drift in spite of selection (Ohta and Gillespie 1996).

For further discussion of the mathematical details of the basics of natural selection, we recommend that readers consult several excellent textbooks on population genetics (Hartl and Clark 2007; Hamilton 2009; Charlesworth and Charlesworth 2010).

4.1.2 The Effect of Selection on Linked Neutral Sites

Natural selection affects not only the frequency of the selected variants themselves but also the linked neutral genetic variation. Neutral variants, as opposed to selected ones, have no effect on an individual's reproductive success. In the absence of selection, their evolution is driven by random genetic drift, where the frequency of a variant fluctuates over generations as a result of the random sampling of gametes at each generation. The probability that a new neutral mutation becomes fixed is only one over twice the effective population size. In other words, the vast majority of neutral variants never reach fixation. The crucial difference between neutral and selected variants is that when selection is strong enough, it can change the frequency of linked neutral variants much faster than genetic drift alone. Only those neutral variants that are genetically linked to one or several selected variants also experience faster changes in their frequencies. In the case of positive directional selection, neutral variants are said to hitchhike with the selected variants. It is therefore possible to localize and/or quantify natural selection in a genome by detecting the effect of selected variants on linked neutral variation. Regions of the genome where the frequency of neutral variants has changed the most are also the ones that are more likely to be influenced by linkage to sites affected by natural selection.

There are three forms of natural selection that influence patterns of linked neutral diversity. The first is positive directional selection and results in a phenomenon called genetic hitchhiking; as a directionally selected variant increases rapidly to a high frequency, genetically linked neutral variants are also driven to high frequencies. Balancing selection is the second form of selection that influences nearby neutral variants. The third form of selection is called background selection. Unlike directional selection and balancing selection that involve advantageous mutations, background selection results from the removal of deleterious mutations. We will first describe the effect of hitchhiking on linked neutral variation (part A). Second, we will describe how balancing selection shapes neighboring neutral diversity (part B). Finally, we will focus on background selection (part C).

(A) Hitchhiking

Genetic hitchhiking occurs when linked neutral variants are driven to higher frequencies together with a positively selected, advantageous mutation that is on the way to fixation. Hitchhiking leaves several signatures in neutral diversity that can be used to detect and quantify positive selection. One of the most important signatures is a local decrease of neutral diversity around the selected mutation (Maynard Smith and Haigh 1974; Kaplan et al. 1989). As the selected mutation increases in frequency, at neighboring neutral sites, linked alleles increase in frequency, and unlinked alleles decrease in frequency. If they do not recombine with the selected mutation, alleles unlinked to the selected mutation disappear from the population. Following a hitchhiking event, the loss of neutral diversity is most severe near the selected mutation and becomes gradually less severe as one gets further and further away from the selected mutation (Fig. 4.1). This is because very few recombination events occurred between the selected and the nearest neutral variants that could rescue the latter from disappearing. As the distance from the selected mutation increases, more and more recombination events occur, and as a result, more neutral variants survive. The resulting pattern is a local dip in the level of diversity called a selective sweep (Fig. 4.1). Diversity has been swept away during the process of hitchhiking.

The loss of genetic variants is only one of several of the signatures left by hitchhiking on patterns of diversity. Hitchhiking also affects both the site frequency spectrum (SFS) and the structure of haplotypes in a population. The site frequency spectrum is the distribution of the frequencies of variants. In an equilibrium population, most neutral variants arose recently and segregate at low frequencies. Most of these low-frequency variants ultimately get lost. Indeed, being at low frequency means that under drift these variants are much more likely to be lost than fixed, since it is far easier for them to reach a proportion of 0 than 100%. Only a small minority of low-frequency variants make it to intermediate or high frequencies, and the SFS is skewed toward low-frequency variants. In population genetics, two types of SFS are used, the folded SFS and the unfolded SFS. The unfolded SFS is the distribution of frequencies of derived alleles. The derived allele at a particular position is the most recent (i.e. mutant) allele, as opposed to the ancestral, preexisting allele. Determining which allele is derived and which allele is ancestral is called polarization. It requires the use of a closely related species, close enough that it is very likely that both ancestral allele and allele found in the

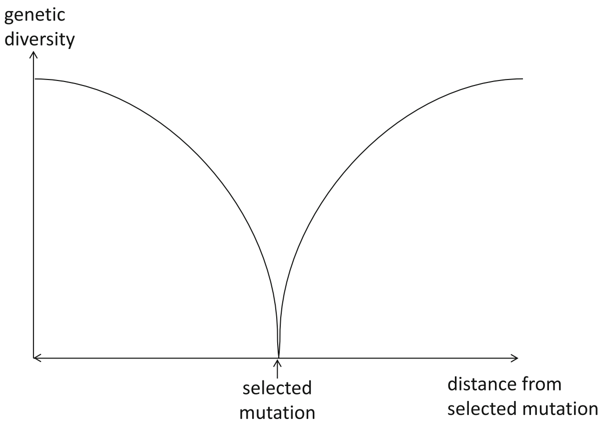


Fig. 4.1 Selective sweep around a selected mutation. The graph represents the loss of genetic diversity or selective sweep due to hitchhiking near a selected mutation that has reached fixation. The loss of diversity is complete right next to the selected mutation and gradually recovers as the distance from the selected mutation increases

close species still correspond to the allele in the common ancestor. Polarization of variants in humans is usually performed using chimpanzees and other primates for comparison (Hernandez et al. 2007). When it is not possible or not necessary to know which allele is derived and which is ancestral, the folded SFS is used. The folded SFS is the distribution of minor allele frequencies. Below we discuss how hitchhiking affects the SFS.

Hitchhiking affects the SFS in systematic ways that can be detected using a range of statistical methods. One of the characteristic signatures of hitchhiking on the SFS is an excess of rare alleles (Nielsen et al. 2007). One may consider the simplest case of an advantageous de novo mutation increasing in frequency to fixation. As the frequency of the advantageous mutation increases, the frequency of unlinked alleles decreases until they are lost. At fixation, many old variants that were previously at intermediate frequencies have been lost. The new variants that appear after fixation are all at low frequency. These two things combine to create a local excess of rare alleles that can be detected using either the folded or the unfolded SFS. The stronger the selection, the shorter the time to fixation, the less time for recombination to rescue unlinked variants from loss, and the stronger the excess of rare alleles (Fig. 4.2).

The second effect of hitchhiking on the SFS is to create an excess of high-frequency derived alleles. Because this pattern is based on derived alleles, the unfolded SFS is required. When a de novo advantageous mutation occurs, it is genetically linked to several derived, low-frequency alleles. In the absence of selection, the vast majority of these derived alleles disappear. However, linked derived alleles hitchhike to higher frequencies together with the advantageous mutation. As a result, soon before fixation, there is an unusually high number of high-frequency derived alleles (Fig. 4.2). When the advantageous mutation fixes, or

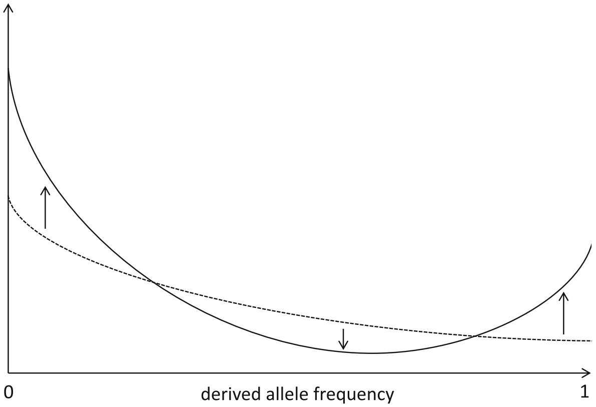


Fig. 4.2 Distortion of the site frequency spectrum by positive selection. The dashed line represents the distribution of derived allele frequencies under neutrality without selection. The full line represents the distribution of derived allele frequencies under selection soon after the fixation of a selected mutation. Compared to the neutral case, there is an excess of rare and high-frequency derived alleles and reciprocally a lack of derived alleles at intermediate frequencies

soon thereafter, most of these alleles become fixed, and the excess of high-frequency derived alleles disappears quickly (Przeworski 2002). It is a very transient signal useful for detecting selective sweeps close to fixation (Zeng et al. 2006). Too soon before fixation, the derived alleles have not reached frequencies high enough that they can be considered unusual. Too long after fixation, the derived alleles have fixed.

Both excess of rare alleles and excess of high-frequency derived variants are signals that are best suited for detecting hitchhiking from de novo advantageous mutations, as opposed to selection on standing genetic variation (Hermisson and Pennings 2005). Selected variants from standing genetic variation differ from the selection on de novo mutations in that in the case of standing variation, the selected sites had time to recombine with several genetic backgrounds before selection started acting upon them. This means that a far greater proportion of neutral variants are linked to the advantageous standing variant. As a result, far fewer variants are lost during the hitchhiking process, and the impact on the SFS is much more subtle. These factors make it less likely to detect selection on standing variation than on neutral mutations using the SFS (Hermisson and Pennings 2005; Przeworski et al. 2005).

In addition to affecting the SFS, hitchhiking can also have profound effects on the structure of haplotypes in a population. A haplotype is a combination of neighboring alleles that are genetically linked in a number of individuals of a population. Longer haplotypes can be defined in genomic regions that experience low rates of recombination, since in this case, it is less likely that recombination events would break down any specific combination of alleles. Hitchhiking specifically creates unusually long and unusually high-frequency haplotypes (Fig. 4.3) (Sabeti et al.

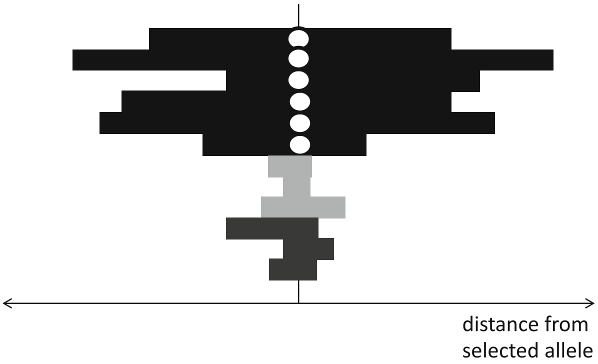


Fig. 4.3 Selected alleles are associated with large, frequent haplotypes. The selected allele (white circle) is associated with a specific haplotype (dark) that is more frequent and extends much further away from the allele compared to haplotypes unlinked to it (gray haplotypes). The dark haplotype is more or less extended in different individuals because of different recombination events in the ancestral lineages of different individuals

2002; Bersaglieri et al. 2004; Voight et al. 2006). In the absence of recombination, a de novo advantageous mutation would drive an entire chromosome to fixation. As a de novo advantageous mutation increases in frequency, linked neutral variants forming a specific haplotype also increase in frequency. As one gets further and further away from the advantageous mutation, it becomes more and more likely that recombination events will break down the specific combination of variants that form the haplotype carrying this mutation. The faster the increase in frequency, the fewer recombination events have time to break the advantageous mutation-bearing haplotype, and the longer this haplotype is as a result. The presence of a single, unusually long, and high-frequency haplotype is therefore a good potential signal of hitchhiking driven by a de novo mutation.

Hitchhiking on standing variation leaves a different pattern on the linked haplotypes. During the initial phase when the selected standing variant is not yet advantageous, it has time to recombine with several haplotypes. When the standing variant becomes advantageous, all the haplotypes recombined with increase in frequency, and the signature of selection on standing variation is the presence of multiple long, high-frequency haplotypes as opposed to only one haplotype in the case of selection on a de novo mutation (Hermisson and Pennings 2005; Messer and Petrov 2013b). Haplotype signatures of hitchhiking are short-lived as recombination rapidly breaks down haplotypes after fixation. For this reason, the haplotype structure is used to detect ongoing selective sweeps or sweeps completed very recently.

The impact of hitchhiking on neutral diversity depends on the intensity of selection and on the amount of recombination. The intensity of selection is defined as the product of the selection coefficient and population size. In the case of hitchhiking, it defines how fast an allele goes to fixation and how systematic the advantageous mutation will increase in frequency due to selection in spite of genetic drift. In a small population, genetic drift is strong, and even mutations with high selection coefficients can experience many ups and downs in frequency before reaching fixation. In a large population, genetic drift is weaker, and for the same selection coefficient, there are systematically more ups than downs on the road to fixation. In other words, as the intensity of selection increases, the path to fixation becomes more deterministic and shorter. The faster the fixation, the fewer recombination events and the larger the region around the selected mutation with reduced diversity and skewed SFS and haplotype structure. Selective sweeps leave a larger footprint in regions of the genome with lower recombination. For this reason, if recombination rates are well known, the size of the genomic region affected by hitchhiking can be used to estimate the intensity of selection.

(B) Balancing Selection

Balancing selection occurs when selection maintains polymorphism at a specific frequency in the population. Balancing selection happens in three different situations: in the case of heterozygous advantage, in the case of frequency-dependent selection, and in the case of fluctuating selection. In heterozygous advantage, heterozygous individuals have, on average, higher reproductive success than homozygous ones. At the selected site, allelic variants tend to oscillate around an equilibrium frequency. In frequency-dependent selection, the selective advantage of an initially uncommon allele depends on its frequency and results in balancing selection if the equilibrium is at intermediate frequencies between the different alleles at the selected site. In the case of fluctuating selection, the direction of selection can fluctuate in time and space in such a way that advantageous alleles are maintained for extended periods of time in a population.

For both heterozygote advantage and frequency-dependent selection, the initial phase is very similar to the case of hitchhiking (Charlesworth 2006). A de novo mutation spreads quickly to the equilibrium frequency (the equilibrium frequency for heterozygote advantage). Linked neutral variants or haplotypes are also driven to intermediate frequencies, much like in the initial phase of a hitchhiking event. In other words, balancing selection is first characterized by a short phase of hitchhiking that results in a partial sweep. Such partial sweeps leave an excess of intermediate allele frequencies in the SFS and also result in larger than usual intermediate-frequency haplotypes. The second phase of balancing selection is usually much more extended in time. Once the equilibrium frequency has been reached, this equilibrium is maintained for a large number of generations. The stronger the selection, the longer the equilibrium is maintained. The neutral variants that initially hitchhiked together with the new mutation are also maintained to the equilibrium frequency as long as they are not dissociated by recombination. Soon after the establishment of the equilibrium, there is an excess of intermediate-frequency neutral alleles in a large region surrounding the balanced site. With time, however, recombination breaks down the linkage between the balanced site and neutral variants, and the region with an excess of intermediate variants shrinks a bit more every generation. Ultimately, neutral variation is affected by balancing selection only in a small region right next to the balanced site, where the maintenance of balanced alleles increases the observed diversity. After a long time, it is possible to detect balancing selection as balanced polymorphisms that can be detected as cases of trans-specific polymorphisms (Leffler et al. 2013). The effects of fluctuating selection on patterns of linked neutral diversity have been less extensively studied and are expected to strongly depend on the type and pace of fluctuations (Bergland et al. 2014). Fluctuating selection includes seasonal adaptations where allele frequencies of selected alleles fluctuate across seasons for extended periods of time, as in the case of *Drosophila* populations in Northern America (Bergland et al. 2014).

(C) Background Selection

Unlike hitchhiking and balancing selection, background selection is driven not by advantageous mutations but by deleterious mutations (Charlesworth et al. 1993). Deleterious mutations are mutations that decrease an individual's reproductive success. They are far more common than advantageous mutations (Boyko et al. 2008). Unlike advantageous mutations that happen very intermittently, there is a constant input of deleterious mutations at any functionally important locus in the genome. Because they are negatively selected, deleterious mutations are unlikely to fix, even the weakly deleterious ones. The important differences between weakly and strongly deleterious mutations are the time these mutations can segregate in a population before they are lost and the frequencies they can reach. Strongly deleterious mutations always segregate at low frequencies and are removed very quickly from a population. Weakly deleterious mutations can segregate at higher frequencies and during a greater number of generations before they are removed by purifying selection.

The effect of background selection on linked neutral variation depends on whether it is driven by strongly or by weakly deleterious mutations. Strongly deleterious mutations remain at very low frequencies and disappear as fast as they appeared. They only take away from the population of the chromosomes where they initially appeared. By doing so, they reduce the overall diversity but have very little impact on the SFS (Charlesworth et al. 1993). In this respect, background selection due to strongly deleterious mutations results in a decrease in genetic diversity. In other words, the level of neutral genetic variation would look as though the population size was smaller.

Weakly deleterious mutations also decrease the overall diversity but in addition also affect the SFS (Charlesworth et al. 1993). Weakly deleterious mutations can rapidly reach higher frequencies than strongly deleterious ones through genetic drift. Intermediate frequency, older alleles unlinked to such weakly deleterious mutations can get fixed faster in the process. Population size is being constant, and at the same time, there is always the same input of new neutral mutations and the same number of recent, low-frequency neutral alleles. Background selection due to weakly deleterious mutations therefore creates an excess of rare alleles relative to higher-frequency ones in the SFS (Desai et al. 2012). Note however that locally the effect is not as drastic as in the case of strong hitchhiking where the skew toward rare alleles can be much more pronounced, especially soon before or soon after the fixation of a beneficial mutation. Unlike beneficial mutations, weakly deleterious mutations do not create an excess of high-frequency derived alleles. Unlike hitchhiking, background selection does not drive long haplotypes to high frequencies (Enard et al. 2014), but it does increase overall linkage disequilibrium (Zeng and Charlesworth 2011).

4.2 Tests of Selection

As previously described, hitchhiking, balancing selection, and background selection all have a specific range of effects on patterns of diversity: on overall levels of diversity, on the SFS, and on the structure of haplotypes. There is considerable interest in testing the presence of natural selection at specific loci (Akey 2009) in a genome or genome-wide (Hernandez et al. 2011; Lohmueller et al. 2011; Enard et al. 2014). Most frequently evolutionary biologists are interested in detecting local signals of hitchhiking. Balancing selection has received less attention in the human genome (Andres et al. 2009; Leffler et al. 2013; DeGiorgio et al. 2014). There has been little interest in detecting background selection at specific loci, and background selection is usually quantified at a global genome-wide scale, not a local one (Hernandez et al. 2011; Lohmueller et al. 2011; Enard et al. 2014). Here we will therefore focus on tests aimed at detecting hitchhiking. However, we will also mention those cases where tests of positive selection are confounded by background selection or when a specific test can also be used to detect balancing selection.

We will first describe tests of positive selection based on patterns of genetic diversity. There are two flavors of these tests: those that use the site frequency spectrum and those that reply on haplotype patterns. We will then describe one particular test based on the combination of both genetic diversity and divergence, the MacDonald–Kreitman test. The tests described here are the most commonly used: Tajima's *D* (Tajima 1989), Fay and Wu's *H* (Fay and Wu 2000) and likelihood ratio-based tests (Kim and Stephan 2002; Nielsen et al. 2005) for the SFS, extended haplotype homozygosity (EHH) (Sabeti et al. 2002), integrated haplotype score (*iHS*) (Voight et al. 2006) and cross-population extended haplotype score (XP-EHH) (Sabeti et al. 2007) for haplotype structures, and the McDonald–Kreitman test (McDonald and Kreitman 1991) for approaches based both on diversity and divergence. Other tests are often variations of the former tests. We have voluntarily limited our description to those tests that were most extensively used to analyze selection in the human genome.

4.2.1 Tests of Selection Based on the Site Frequency Spectrum

4.2.1.1 Tajima's *D*

Fumio Tajima proposed his test in 1989 (Tajima 1989), based on the summary statistic *D*. Although Tajima's *D* has been widely used to detect hitchhiking, Tajima initially introduced it in order to detect loci with an excess of rare deleterious mutations. However, because of the popularity of genome scans for positive selection, it has been used more widely to detect selective sweeps. The principle of the Tajima's *D* test is to contrast two well-known estimators of genetic diversity, namely, π and *S* on the folded SFS to detect a departure from neutrality. π is the average of differences between each pair of sequences in the sample studied. *S* is the number of segregating variants. When a selective sweep is very near or right after fixation, variation at perfectly linked sites right next to the selected mutation has been wiped out. Further from the selected mutation, recombination maintains variation, but many preexisting older intermediate-frequency alleles have been swept away from the population. At the same time, there is a constant input of new, low-frequency alleles. This results in an excess of young, rare alleles compared to the neutral SFS. The strength of the excess of rare variants reflects the strength and speed of the selective episode. The missing intermediate-frequency alleles do not affect *S* strongly as they represented only a small proportion of all segregating variants in the first place. Unlike *S*, π is strongly affected by the lack of intermediate-frequency alleles. As a reminder, π is the average of differences between each pair of sequences in the sample studied. Intermediate-frequency alleles in the folded SFS result in differences between many pairs of sequences in the sample, whereas rare alleles result in differences between very few pairs of sequences. π is therefore far more sensitive to intermediate-frequency variants than *S* is. Tajima's *D* is the difference between π and an estimate of the mutation rate based on *S*, normalized by the standard deviation of the difference. Under perfectly neutral conditions, the difference is expected to be null. When a sweep is right before or right after fixation, Tajima's *D* is negative: π is strongly reduced because of the missing intermediate frequency alleles, while *S* is less affected. It is important to note that any demographic expansion or bottleneck that induces a local excess of rare alleles can create spurious detections of sweeps when using Tajima's *D*. Tajima's *D* is also sensitive to the excess of rare alleles induced by background selection due to weakly deleterious mutations. Tajima's *D* can also be used to detect recent balancing selection, when there is an excess of intermediate-frequency alleles across a wide region surrounding the balanced allele not yet broken down by recombination. In this case, Tajima's *D* is strongly positive.

4.2.1.2 Fay and Wu's *H*

Tajima's *D* inspired a number of other statistics similarly based on the SFS (Achaz 2009). In an attempt to create a statistic more sensitive to positive selection at the exclusion of other confounding processes, Fay and Wu created the statistic *H* (Fay and Wu 2000), usually called Fay and Wu's *H*. *H* uses the unfolded SFS, which means an out-group has to be used to polarize alleles. The rationale is to detect selective sweeps using intermediate- and high-frequency alleles of the unfolded SFS, at the exclusion of rare alleles. Indeed, the main effect of population expansions, bottlenecks, and background selection on the SFS is to create an excess of young rare alleles. By focusing on intermediate- and high-frequency variants, *H* is expected to be more robust to these confounding factors. In brief, *H* is the difference between π and θ_H where π is sensitive to intermediate-frequency alleles as explained in the description of Tajima's D, and θ_H is a measure of diversity that is sensitive to high-frequency alleles. Right before or after fixation, many intermediate frequency alleles have been fixed, and π is low. At the same time, young, derived alleles linked to the selected mutation were rapidly driven to high frequencies. There is an excess of young, high-frequency derived alleles, and θ_H is high. The difference $H = \pi - \theta_H$ is therefore strongly negative soon before and after the fixation of a selected mutation. Because the effect of population expansions is to increase the supply of new, rare alleles, population expansions have no effect on H. Conversely, population bottlenecks can affect H. The rapid rise of a specific haplotype to a high frequency can occur by chance during a bottleneck. This mimics a selective sweep and creates both an excess of high-frequency derived alleles and a paucity of intermediate alleles on a local scale. Fay and Wu's H is robust to background selection driven by strongly deleterious mutations (Zeng et al. 2006). The robustness of the statistic to background selection driven by weakly deleterious mutations remains to be evaluated.

4.2.1.3 Likelihood Ratio Tests

In order to achieve greater statistical power to detect selective sweeps, several authors have developed Likelihood Ratio Tests (LRTs) based on the SFS. The advantage of such tests is to consider the entire SFS rather than just using parts of it as for Tajima's D or Fay and Wu's H. Further, LRTs compare the likelihoods of the observed SFS under a selective sweep model and a neutral model, with the likelihood of the observed SFS under a neutral model. As it provides an explicit test of selection, it should be more powerful and robust than simple tests of neutrality. Under both sweep and neutral models, it is possible to estimate the probability of the observed frequency for each allele of the tested locus. The likelihood of the SFS under each model is then calculated by multiplying all the individual probabilities for all the sites present in the tested locus. Once the likelihoods are known both with and without sweeps, the likelihood ratio can be used to decide whether or not the neutral hypothesis can be rejected. There are two distinct approaches to estimate the likelihood of the SFS under the neutral hypothesis. The first approach consists of using a classical neutral model under panmictic assumptions (Kim and Stephan 2002; Li and Stephan 2005). A major issue with this neutral model is that it does not take demographic perturbations into account. Consequently, this approach is very sensitive to demographic events such as population expansions or bottlenecks (Jensen et al. 2005). The second approach, called CLR for composite likelihood ratio (Nielsen et al. 2005), is to estimate the likelihood of the SFS at a local scale based on the global, genome-wide SFS as the neutral model. Compared to the previous classical neutral model, this empirical neutral model is a good approximation of the average expected SFS given the past demographic history of the population studied. Although this approach does integrate the systematic, average influence of demography in the neutral model, it still does not account for the increased variance in the local SFS along chromosomes expected after bottlenecks, and false positives due to demography are still an issue (Huber et al. 2014).

4.2.2 Tests of Selection Based on Haplotypes

4.2.2.1 Extended Haplotype Homozygosity (EHH)

The EHH (extended haplotype homozygosity) test was introduced in 2002 by Sabeti et al. (2002) to test the hypothesis of positive selection of two variants protecting against malaria at the *G6PD* and *CD40L* loci. Both alleles decrease the risk of malaria by approximately 50% in heterozygous individuals. In African regions where malaria is endemic, *G6PD*-202A is at a frequency of 20% despite the fact that it causes mild G6PD deficiency. This clearly suggests that increased resistance to malaria has given *G6PD*-202A heterozygous individuals a strong selective advantage despite the adverse effects of the allele. If *G6PD*-202A has driven a selective sweep, it is a partial ongoing selective sweep. Statistics aimed at detecting complete or near-complete sweeps such as Tajima's D or Fay and Wu's H have no power in such a case. If selection for *G6PD*-202A has been strong enough, the allele should be associated with an unusually long haplotype block that has not been broken down by recombination. The EHH statistic was designed to detect such long haplotypes. It is built by first defining core haplotypes. Core haplotypes are combinations of nearby alleles that are in perfect linkage (i.e., no recombination has occurred between the alleles). For all possible pairs of chromosomes with the same core haplotype, EHH is calculated as the proportion of pairs with perfect homozygosity over a physical distance x from the core haplotype. The distance x is fixed and largely defined by the size of the locus that was sequenced around the candidate allele. A relative EHH is then calculated as the ratio of EHH for the core haplotype of interest to the EHH for all other core haplotypes grouped together. The relative EHH is meant to normalize for recombination. Note however that there is always more power to detect selection in regions with low recombination rates and a greater risk of false positives (O'Reilly et al. 2008; Ferrer-Admetlla et al. 2014). The observed relative EHH can finally be compared with a distribution of simulated neutral relative EHH. Sabeti et al. (2002) successfully used the EHH test to detect strong partial sweeps at the G6PD and CD40L loci. To this date, the G6PD locus is one of the best examples of an ongoing, partial sweep in the human genome.

4.2.2.2 Integrated Haplotype Score (iHS)

A major limitation of the EHH test is that it uses an arbitrary distance to the core haplotype. One can imagine the curve of EHH as a function of the distance to the core haplotype. The area under the curve captures more information about haplotype structure and represents a less arbitrary statistic than EHH measured for a fixed distance to the core haplotype. In this respect, the integrated Haplotype Score (*iHS*) (Voight et al. 2006) is an improved, integrated, and standardized version of EHH. Unlike the classic EHH, *iHS* is oriented based on the derived or ancestral status of alleles. Single ancestral or derived alleles are used as core haplotypes, and *iHS* detects selection by comparing the haplotype structure around derived alleles with the haplotype structure around ancestral alleles. The integral of EHH is noted *iHH* and is calculated for both the ancestral (iHH_A) and the derived allele (iHH_D). The unstandardized *iHS* is then calculated as:

$$unst\text{-}iHS = \ln\left(iHH_A/iHH_D\right).$$

The logarithm is used so that *unst-iHS* is normally distributed. The *iHS* is finally calculated by standardizing *unst-iHS* with the genome-wide average and standard deviation for alleles with a similar frequency. Unlike the classic EHH, *iHS* is calculated using genetic distances instead of physical distances. Genetic distances are more appropriate than physical distances since the decay of EHH is a function of the amount of recombination between the tested allele and the position in question. *iHS* has maximal power to detect ongoing partial sweeps where the selected allele has a frequency between 60% and 90%. One of the strongest *iHS* signals in the human genome is found at the lactase locus in the European population (Voight et al. 2006). Methods similar to *iHS* such as *nSL* have since been developed that are more robust to confounding factors such as low recombination (Ferrer-Admetlla et al. 2014).

4.2.2.3 Cross-Population EHH (XP-EHH)

The EHH and *iHS* tests were designed to detect ongoing, partial sweeps using the haplotype structure information from a single population. The *XP-EHH* statistic (Sabeti et al. 2007) is essentially a cross-population *iHS* designed to detect nearly complete or complete sweeps. Where *iHS* compares the haplotype structure around ancestral and derived alleles within a single population, *XP-EHH* compares the haplotype structure around the same allele in two different populations. *XP-EHH* has maximal power to detect sweeps with selected allele frequencies over 90%, a frequency range where *iHS* performs very poorly. The idea behind *XP-EHH* of comparing haplotypes between different populations is similar to the idea behind the classic *F*st approaches where selection is detected not with changes in haplotype structure but with relative changes in allele frequencies between populations.

4.2.3 Tests Based on both Diversity and Divergence: The McDonald–Kreitman Test

The tests we have described so far only make use of genetic diversity to detect single episodes of selection at a very transient time scale in evolution. There has also been considerable interest in quantifying not only recent adaptation but also adaptation that has occurred since divergence from chimpanzee. The most commonly asked questions are (1) which loci have experienced recurrent episodes of positive selection since divergence with chimpanzees and (2) how much positive selection occurred genome-wide since divergence with chimpanzees? These questions have been addressed using the McDonald–Kreitman test (McDonald and Kreitman

1991). The aim of the test is to quantify the rate of adaptive substitutions with a potential functional impact, like amino acid substitutions in a coding sequence or substitutions affecting a regulatory segment. The test can be conducted from scales ranging from a single gene to the entire genome. The idea is to compare functional diversity and divergence with nonfunctional diversity and divergence to measure an excess of functional divergence. The test has been most frequently used to quantify the rate of amino acid changes in coding sequences (Bustamante et al. 2005; Boyko et al. 2008), where it requires to know the numbers of non-synonymous and synonymous polymorphic sites P_n and P_s and the number of non-synonymous and synonymous divergent sites with a closely related species D_n and D_s. Recurrent adaptive amino acid changes are expected to increase D_n while leaving P_n unaffected. Indeed, adaptive mutations get fixed rapidly and do not contribute to polymorphism. Recurrent adaptive amino acid changes thus make D_n higher than expected given P_n. Variations in the mutation rate are then controlled by comparing the ratios D_n/D_s and P_n/P_s instead of just comparing D_n and P_n. The rate of adaptive amino acid changes can then be easily computed as a function of D_n, D_s, P_n, and P_s. Note that the category "non-synonymous" can be replaced with any other functional category and that "synonymous" can also be replaced by any other neutral category.

The McDonald–Kreitman test has been adapted for quantifying adaptation genome-wide (Boyko et al. 2008) or at single genes (Bustamante et al. 2005). Although the approach is elegant and looks simple at first, it suffers from many biases. The most important one is deleterious mutations (Eyre-Walker and Keightley 2009; Messer and Petrov 2013a). Recessive deleterious mutations can reach relatively high frequencies before they get eliminated from a population. This means that they have a significant contribution to polymorphism, but no contribution to divergence. In coding sequences, non-synonymous deleterious mutations inflate P_n while leaving D_n unaffected, which biases estimates of the rate of adaptive substitutions downward. This problem is especially severe after a long population bottleneck (Eyre-Walker 2002). Indeed, reduced population size means that purifying selection is less efficient at weeding out deleterious mutations, which increases P_n. Fluctuations in selective constraint are also an issue, as D_n represents the integration of selective constraint over a long period of time while P_n is affected only by the recent level of constraint. Despite these limitations, the McDonald–Kreitman approach has been remarkably useful to quantify rates of adaptation in the human genome.

4.3 How to Choose a Specific Test of Selection?

The choice of a test of positive selection depends on a number of parameters:

- *The type of selection*. Different types of selection often require different tests, although the same tests can sometimes be used to study both hitchhiking and balancing selection. For example, SFS-based tests like Tajima's *D* test (Tajima

1989) (see above) can be used both to detect close to fixation or recently fixed advantageous mutations or recently established balancing selection.

- *The timing of selection*. Different tests of positive selection aim at detecting either hitchhiking or balancing selection at different times relative to the start and end of the selective events. Ongoing selective sweeps not yet close to fixation or balancing selection in its initial hitchhiking phase are best detected with tests using haplotype structure in a single population. Recently completed selective sweeps are best detected using either SFS-based tests or haplotype-based tests using multiple populations (see above). Selection over extended time periods involving multiple amino acid sites may best be detected using the McDonald–Kreitman type test.
- *The nature of the data*. Tests of selection have strong requirements regarding the data they need to be performed. Here we will discuss only the case where genome-wide variation is available. The data used to test selection in the genome-wide study usually consists either of sequencing data or of genome-wide genotype calls obtained with a genotyping chip. Regardless of the kind of data used, a fundamental parameter for selection tests is the number of individuals in the dataset. Most tests need a minimum of 20 chromosomes or 10 individuals to have the power to detect selection. The more individuals, the higher the power to detect selection. Sequencing can be performed for each individual separately, or for all individuals grouped together during the sequencing process. The latter is called pooled sequencing.

  Compared to sequencing, genome-wide genotyping is still much cheaper, but there has recently been a clear movement toward adapting approaches to full genome sequencing. Indeed, sequencing has become far more affordable. Importantly, genome-wide genotyping data suffers from the so-called ascertainment bias (Clark et al. 2005). Ascertainment bias comes from the fact that only those variants that were previously known are genotyped. This biases the genotyped diversity in several ways. First, already known variants tend to be high-frequency ones. This is a serious issue when using the SFS to detect selection, since the SFS will be skewed toward intermediate frequencies. Such a skew is not trivial to correct, since it varies along a chromosome with some loci being more affected than others (Ramirez-Soriano and Nielsen 2009). Second, genotyped positions are chosen so that they are regularly spaced along the genome. This regular spacing removes the information on selection that could be extracted from the local amount of variable sites (Clark et al. 2005). It also means that important selected variants can be completely missed. A great amount of effort has been done in the past 10 years to deal with ascertainment bias, but it is still a strong limitation that explains why full sequencing is now favored over genotyping.

  Sequencing is not free of local biases either. The main bias of sequencing when trying to detect selection corresponds to the local variations in sequencing depth (Abecasis et al. 2012). The lower the sequencing depth, the lower the probability that all chromosomes have been sequenced in all individuals of the studied samples. Incomplete sequencing affects the SFS by reducing the accuracy of the estimated frequencies. At a given position the frequency of a variant in the sequencing data may not accurately represent the true frequency in the sample if sequencing depth is too low. Sequencing depth is more convenient to handle than ascertainment bias. Indeed, sequencing depth unlike ascertainment bias can be readily measured and is therefore much easier to take into account. The process of imputation further makes it possible to infer alleles in regions of the genome with insufficient sequencing depth (Li et al. 2009).

– *The confounding factors of selection*. Confident detection of positive selection is always challenging due to the multiple confounding factors that create the same signatures in diversity as positive selection. The choice of a specific test of selection thus depends on which confounding factors are acceptable and which are not in a particular study context.

Demographic history represents the most important confounding factor of selection. Indeed both decreases and increases in population size can create local signatures that are undistinguishable from the signatures left by a local event of positive selection. Increases in population size or population expansions affect both the SFS and haplotype structure. Population expansion creates an increased influx of new, low-frequency mutations. The main effect of a population expansion is therefore to skew the SFS toward rare alleles, just like hitchhiking. Decreases in population size or bottlenecks can also severely confound tests of selection. Bottlenecks can either result in an excess of rare alleles, due to all lineages coalescing during the bottleneck, or lead to an excess of intermediate-frequency alleles, if only a few lineages make it through the bottleneck into the larger ancestral population. Either of these effects can occur depending on the specific parameter (e.g., timing and severity) of the bottleneck. Thus, there could be depletion in genetic variation or an excess of rare alleles that can be mistaken with selection. Specific haplotypes can reach high frequency quickly during a bottleneck, which can also mimic selection.

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65

Achaz G (2009) Frequency spectrum neutrality tests: one for all and all for one. Genetics 183(1):249–258

Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19(5):711–722

Andres AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, Bustamante CD et al (2009) Targets of balancing selection in the human genome. Mol Biol Evol 26(12):2755–2764

Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA (2014) Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. PLoS Genet 10(11):e1004775

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74(6):1111–1120

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR et al (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4(5):e1000083

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD et al (2005) Natural selection on protein-coding genes in the human genome. Nature 437(7062):1153–1157

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2(4):e64

Charlesworth B, Charlesworth D (2010) Elements of evolutionary genetics. Roberts and Company Publishers, Greenwood Village

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134(4):1289–1303

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15(11):1496–1502

DeGiorgio M, Lohmueller KE, Nielsen R (2014) A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS Genet 10(8):e1004561

Desai MM, Nicolaisen LE, Walczak AM, Plotkin JB (2012) The structure of allelic diversity in the presence of purifying selection. Theor Popul Biol 81(2):144–157

Enard D, Messer PW, Petrov DA (2014) Genome-wide signals of positive selection in human evolution. Genome Res 24(6):885–895

Eyre-Walker A (2002) Changing effective population size and the McDonald-Kreitman test. Genetics 162(4):2017–2024

Eyre-Walker A, Keightley PD (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol 26(9):2097–2108

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155(3):1405–1413

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014) On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol 31(5):1275–1291

Hamilton M (2009) Population genetics. Wiley-Blackwell, Hoboken

Hartl DL, Clark AG (2007) Principles of population genetics. Sinauer, Sunderland

Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169(4):2335–2352

Hernandez RD, Williamson SH, Bustamante CD (2007) Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol 24(8):1792–1800

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M (2011) Classic selective sweeps were rare in recent human evolution. Science 331(6019):920–924

Huber CD, Nordborg M, Hermisson J, Hellmann I (2014) Keeping it local: evidence for positive selection in Swedish *Arabidopsis thaliana*. Mol Biol Evol 31(11):3026–3039

Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD (2005) Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170(3):1401–1410

Kaplan NL, Hudson RR, Langley CH (1989) The “hitchhiking effect” revisited. Genetics 123(4):887–899

Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160(2):765–777

Leffler EM, Gao Z, Pfeifer S, Segurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G et al (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 339(6127):1578–1582

Li H, Stephan W (2005) Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. Genetics 171(1):377–384

Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10:387–406

Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, Tian G, Huerta-Sanchez E, Feder AF, Grarup N et al (2011) Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet 7(10):e1002326

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23(1):23–35

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351(6328):652–654

Messer PW, Petrov DA (2013a) Frequent adaptation and the McDonald-Kreitman test. Proc Natl Acad Sci USA 110(21):8615–8620

Messer PW, Petrov DA (2013b) Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol 28(11):659–669

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15(11):1566–1575

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8(11):857–868

Ohta T, Gillespie JH (1996) Development of neutral and nearly neutral theories. Theor Popul Biol 49(2):128–142

O'Reilly PF, Birney E, Balding DJ (2008) Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome Res 18(8):1304–1313

Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 160(3):1179–1189

Przeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. Evolution 59(11):2312–2323

Ramirez-Soriano A, Nielsen R (2009) Correcting estimators of theta and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process. Genetics 181(2):701–710

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ et al (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832–837

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R et al (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164):913–918

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4(3):e72

Zeng K, Charlesworth B (2011) The joint effects of background selection and genetic recombination on local gene genealogies. Genetics 189(1):251–266

Zeng K, Fu YX, Shi S, Wu CI (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174(3):1431–1439

Part II

Association Studies and Medical Genetics

Methods for Association Studies

5

Rebecca E. Graff, Caroline G. Tai, Linda Kachuri, and John S. Witte

Abstract

Association studies are a key approach to evaluating the relationship between genetic factors and phenotypes or traits. This chapter presents general methods for genetic association studies in unrelated humans. Topics covered include types of association studies, study design considerations, measurement of genetic information, and analytical techniques. This material provides readers with background for interpreting results from association studies and for undertaking their own studies.

5.1 Introduction

Since the first sequences of base pairs were published in the late 1960s and early 1970s (Gilbert and Maxam 1973; Wu and Kaiser 1968; Wu and Taylor 1971), our ability to investigate the human genome has advanced immensely. Genetic epidemiology largely aims to identify genetic factors that are associated with a particular phenotype or disease state. To evaluate these relationships, one essential approach used by researchers is the *genetic association study*. These studies relate germline genetic variants—or other sources of genetic variation—to some measure of phenotype, disease status, progression, and/or mortality.

Before association studies became pervasive in the journey toward deciphering the genetic basis of complex disease, *linkage analysis* was a common method for detecting genes with a major effect on phenotype (Claussnitzer et al. 2020). In the 1980s and early 1990s, many researchers undertook genetic studies that

R. E. Graff (✉) · C. G. Tai · L. Kachuri · J. S. Witte
Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA
e-mail: Rebecca.Graff@ucsf.edu

K. E. Lohmueller, R. Nielsen (eds.), *Human Population Genomics*,
https://doi.org/10.1007/978-3-030-61646-5_5

utilized family structures ranging from sibling pairs to large multiplex pedigrees. Such studies use families with numerous disease-affected individuals to evaluate markers spaced widely across the genome, at intervals of up to 20 million base pairs, and to examine how these markers segregate with the disease phenotype across multiple families (Botstein et al. 1980). Linkage analyses are often successful in the evaluation of rare and/or monogenetic disorders but are generally underpowered to detect genetic factors with subtle effects on complex diseases. They also have low resolution on account of the limited number of meioses from one generation to the next within families (Risch and Merikangas 1996).

Given that high-penetrance genes co-segregating in affected families have turned out to be relatively rare, association studies have become the far more common and more powerful tool to investigate genetic relationships (Claussnitzer et al. 2020). They rely on historical recombination events from millions of years of evolution and thus do not require pedigree information or controlled crosses to identify genetic variants associated with the phenotype. In addition, because most association studies leverage the phenomenon of linkage disequilibrium (LD) to localize such variants, they can detect causal loci within narrower regions and allow for genetic mapping at a finer scale than linkage studies (Xiong and Guo 1997). That is, association studies do not require the *direct* evaluation of postulated causal variants. Rather, they may utilize LD to *indirectly* evaluate genetic variants neighboring those assayed (see Chap. 2 on LD). Moreover, genome-wide association studies (GWAS) allow investigators to broadly search the genome for disease-causing variants in a manner that is relatively agnostic to previous biological knowledge.

The fundamental approach to any genetic association study is based on the following premise: compare the frequency of the genetic characteristic of interest across individuals with different values for the phenotype of interest. Consider, for example, a single-nucleotide polymorphism (SNP) with effect allele A under investigation in a standard analysis of a binary phenotype (Fig. 5.1). To determine whether or not the SNP is associated with the phenotype, one would calculate the frequency of the effect allele in cases and controls. When the frequency is greater in individuals with the phenotype than in those without it, then the effect allele is positively associated with the phenotype (as in the figure). When the opposite is true, then the effect allele is inversely associated. In GWAS, these associations are estimated for every SNP measured across the entire genome.

Genomic research traverses genetic sequence information, protein products, and the eventual expression of traits. It may also utilize a range of organisms; only one facet is the study of humans. Our focus in this chapter is on population-based genetic association studies in humans, in which data are derived from unrelated individuals. Relative to family-based association studies, population-based studies are the more common—and often more powerful—approach to the evaluation of genetic associations. In describing types of association studies (Sect. 5.2), considerations in their design (Sect. 5.3), measurement of genetic information (Sect. 5.4), and analytical techniques (Sect. 5.5), we aim to provide a basis on which readers can build their own efforts to characterize associations between genetic polymorphisms and measured phenotypes.

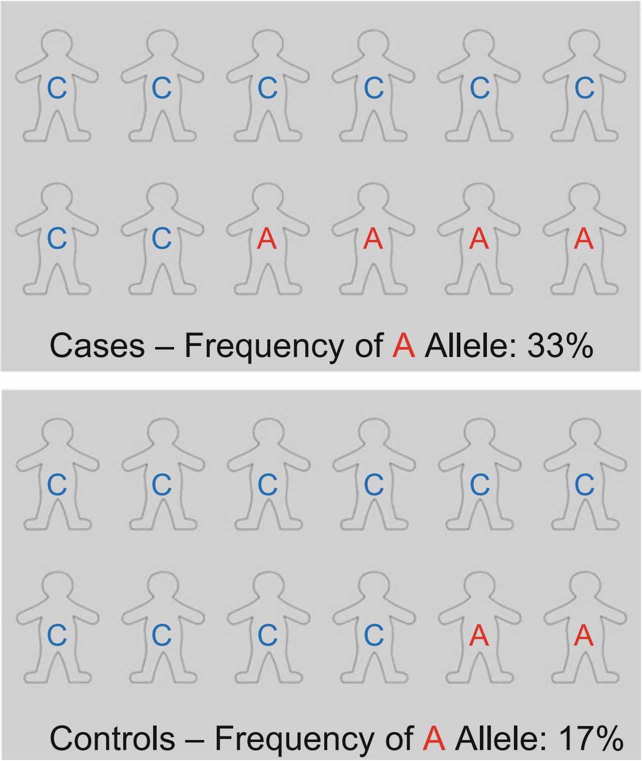

C
C
C
C
C
C
C
C
A
A
A
A
Cases – Frequency of A Allele: 33%
C
C
C
C
C
C
C
C
C
C
A
A
Controls – Frequency of A Allele: 17%

Fig. 5.1 Standard approach to a genetic association study

5.2 Types of Association Studies

Before recent technologies enabled larger-scale investigations, efforts to decipher the genetic basis of disease were predominantly supported by candidate gene studies (Claussnitzer et al. 2020). GWAS have since become an important avenue for undertaking agnostic evaluation of the association between common genetic variants and risk of disease (Claussnitzer et al. 2020). Here we describe these most common designs for genetic association studies, and Fig. 5.2 summarizes some of their differences with respect to the number of variants they can address and the sample sizes they require. In brief, as investigators shift from discovery to confirmation of associations, the number of markers investigated tends to decrease, while the number of samples should increase. Fine-mapping studies, however, do not require particularly large sample sizes as they evaluate a limited number of variants.

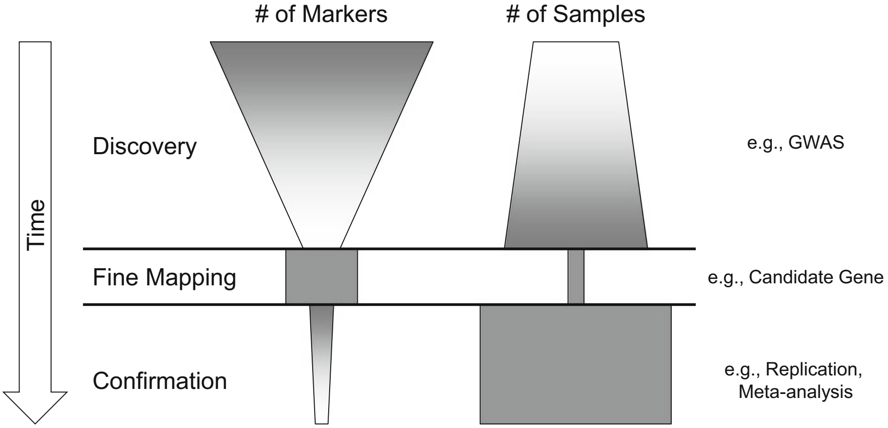


Fig. 5.2 Overview of genetic association study designs

5.2.1 Candidate Gene Studies

Candidate gene studies overcome some of the issues of linkage analysis by focusing on associations between disease and specific variants plausibly involved with the disease a priori. These studies became pervasive following the realization that genetic variants contributing to the risk of complex disease were likely to have individually weak effects (Claussnitzer et al. 2020).

Candidate gene studies generally evaluate several SNPs within a single gene under the assumption that the SNPs capture information about the underlying genetic variability of the gene (even if the SNPs are not the true causal variants). They may do so either *directly*, by evaluating postulated causal variants, or *indirectly*, by leveraging LD. Sufficiently large candidate gene studies are able to detect weak effects due to common variants, though it is important to note that they too become underpowered as variants become more rare (Risch and Merikangas 1996). In addition, their focus on particular genes means that they ignore much of the genome.

Many early candidate gene studies were underpowered, and results went largely unreplicated (Cordell and Clayton 2005). It was also unclear what should actually constitute a candidate gene. Traditionally, lists of candidate genes were compiled after an extensive manual biomedical literature review. The process to identify candidate genes then evolved to incorporate automated text-mining procedures, selection of genes belonging to specific biological pathways, and/or prioritization based on gene characteristics such as degree of conservation or proximity to known loci (Piro and Di Cunto 2012). Since GWAS have come into the picture, however, the role of candidate gene studies has become increasingly coherent as a fine-mapping approach. These studies can be targeted toward regions of the genome in which GWAS find strong hits in order to see which findings are replicable and thus more likely to be true associations.

5.2.2 Genome-Wide Association Studies

5.2.2.1 Background

Increased throughput, scalability, and speed have enabled investigators to undertake GWAS (Claussnitzer et al. 2020)—research that would have been far too complex to consider even 20 years ago. It has become possible to simultaneously measure hundreds of thousands of SNPs due to technological advances in array-based genotyping (Wang et al. 1998). The number of variants that may be assayed by these SNP arrays rapidly increased at the same time that array prices steadily decreased. At present, arrays can directly measure millions of SNPs while providing relatively high coverage of common genetic variation across the human genome (Jorgenson and Witte 2006; Lindquist et al. 2013; Nelson et al. 2013; Xing et al. 2016; Wojcik et al. 2018).

The genetic content of such arrays was facilitated by the development of technology that allows for large-scale sequencing efforts in combination with the sequencing of the human genome (Lander et al. 2001; Venter et al. 2001). Beginning in 2002, the International Haplotype Map (HapMap) Project undertook an effort to catalog the common genetic variants that occur in human beings. It was also determined that a substantial portion of this variation can be efficiently captured by a subset of "tag" SNPs via the phenomenon of LD among neighboring SNPs (Daly et al. 2001; Gabriel et al. 2002; International HapMap Consortium 2003; International HapMap Consortium 2005) and that this structure varies across ancestral populations (International HapMap Consortium 2003; International HapMap Consortium 2005; Frazer et al. 2007).

Unlike candidate gene studies, GWAS are not hypothesis-driven; they do not require a priori specification of the genes or polymorphisms that are conjectured to be associated with the phenotype of interest. Rather, they quantify DNA sequence variations from across the entire human genome in an attempt to pinpoint genetic risk factors for common diseases. In designing an array for genome-wide assessment, a primary objective should thus be to capture as much common variation in the human genome as possible.

5.2.2.2 Multistage Study Designs

When GWAS first became popular, the high cost of SNP arrays and necessity for large sample sizes (to achieve sufficient statistical power to detect the anticipated modest associations among hundreds of thousands of SNPs) (Witte et al. 2000) motivated the development and use of *multistage* GWAS designs (Thomas et al. 2005). Decreasing SNP array costs have made multistage designs for GWAS less essential, but we choose to briefly describe them for two reasons: (1) an overview is important for understanding historical studies, and (2) as we move into the post-GWAS era, next-generation sequencing of entire genomes may be sufficiently expensive to once again make multistage designs relevant.

In the initial discovery stage of a multistage design, a subset of the study sample is genotyped using genome-wide SNP arrays. Then the most strongly associated

SNPs are genotyped with a less expensive genotyping platform in the remaining samples. The procedure prioritizes the most promising SNPs for evaluation in additional stages and can pinpoint associated regions for fine mapping. The optimal division of samples across stages depends on a number of factors, but in general, the most efficient approach entails the inclusion of approximately one-third to one-half of the samples in the initial stage and the remaining samples in follow-up stages (Skol et al. 2006, 2007). The number of noteworthy SNPs that should be tested depends on the sample sizes in the respective stages, the number of false-negative results that one is willing to accept, and whether or not one wishes to incorporate SNP information (e.g., proximity to the nearest gene or likelihood of being functional) (Chen and Witte 2007; Roeder and Wasserman 2009; Roshan et al. 2011; Thomas et al. 2009). Ideally, at least 1% of the first stage SNPs should be typed in the second stage (Skol et al. 2006). One must also decide whether the early follow-up stages should be treated as part of a replication or joint analysis.

5.2.2.3 Limitations

Despite their numerous strengths, GWAS carry several notable limitations. First, it is important to note that most variants discovered via GWAS are only associated with, and not causal for, disease. Even when an association is real and statistically reproducible in other datasets, another untyped variant in LD with the associated SNP may still be the causal variant. Determining the factors underlying results can be extremely challenging and require separate fine-mapping and mechanistic studies. That many of the associations detected to date are not in gene regions can make the findings yet more complicated (Buniello et al. 2019). These issues limit our understanding of the biological basis of results and our ability to implement preventive or therapeutic measures.

Second, findings from GWAS thus far account for only a limited amount of disease heritability (Maher 2008; Nolte et al. 2017). Most SNPs detected by GWAS show a small magnitude of effect. That said, as sample sizes for GWAS are increasing, studies are detecting and replicating a larger number of trait-associated variants. That we are now also able to examine essentially the entirety of common variation across the genome (at least indirectly) allows us to explain an increasing proportion of heritability. So too does our ability to assess the contribution of rare variants. The polygenic model of heritability is becoming increasingly accepted; many risk variants with small effect sizes are thought to underlie disease risk.

Finally, GWAS have not yet sufficiently distinguished between individuals with low- and high-risk disease. In general, screening tests based on SNPs detected by GWAS to date may have low positive (and negative) predictive value for disease and thus limited utility in a diagnostic setting (Kraft et al. 2009; Ware 2006). As more SNPs are discovered, however, combining them into polygenic risk scores (PRS) efficiently summarizes individuals' genetic susceptibility profiles, thereby improving phenotypic prediction (Torkamani et al. 2018). PRS have the potential to personalize risk estimates and improve the discriminatory ability of screening tests (Mavaddat et al. 2019; Toland 2019). For example, a 2015 study created a risk score of 105 SNPs that was strongly associated with prostate cancer risk among non-Hispanic whites (*P* value: 1.0×10^{-211}) (Hoffmann et al. 2015). More recently, a PRS for breast cancer based on 313 variants demonstrated strong predictive performance (AUC = 0.630) and identified 19% of women who could be eligible for early screening at age 40 (Mavaddat et al. 2019). Still, few individuals will carry large numbers of risk alleles from GWAS, though essentially all individuals will carry some risk alleles. Screening for them in the general population is thus unlikely to be cost-effective, unless individuals receive genome-wide evaluations. In addition, predictive models may have worse performance in ancestral populations other than those in which the models were discovered, because effect size estimates will be diluted when SNPs in populations with one set of LD patterns (e.g., Europeans) are applied to populations with a different set of LD patterns (e.g., African Americans) (Carlson et al. 2013). Note also that justification for genetic testing additionally depends on the existence of effective interventions.

5.2.3 Mendelian Randomization

In some instances, genetic variation can be leveraged toward evaluating causal relationships between exposures and outcomes that may be challenging to investigate in traditional observational studies. By using a genetic predictor of exposure as an instrumental variable*, Mendelian randomization* circumvents issues of confounding and reverse causation that often afflict epidemiological studies. While the method has been around for several decades (Gray and Wheatley 1991; Katan 1986; Smith and Ebrahim 2003), its use has exploded with the ever-increasing discovery of trait-associated variants and modern statistical methods for high-dimensional genetic data. In general, its implementation requires the identification of a set of genetic variants that is predictive of the exposure of interest followed by the performance of instrumental variable analyses (Burgess et al. 2013; Pierce and Burgess 2013).

As with all instrumental variable approaches, Mendelian randomization is premised on three assumptions: (1) the genetic instrument is associated with the exposure, (2) the genetic instrument shares no common causes with the outcome, and (3) the genetic instrument only affects the outcome through exposure. The first assumption is easily satisfied by selecting genetic variants that are strongly associated with the exposure of interest, such as those reaching genome-wide significance. The second assumption can be at least partially verified by assessing associations between genetic instruments and known confounders. The third assumption, however, cannot be substantiated empirically. Nevertheless, sensitivity analyses can help evaluate the consistency and robustness of observed results (Bowden et al. 2017; Haycock et al. 2016).

5.2.4 Transcriptome-Wide Association Studies

Among the more recent methodological developments in genetic association studies is the *transcriptome-wide association study* (TWAS) (Gamazon et al. 2015; Gusev et al. 2016). Without relying on directly measured expression levels, TWAS aim to identify genes associated with complex traits. By using an external reference set of individuals with genetic and transcriptomic data, one can impute gene expression levels in the target study population and evaluate associations with the outcome. Extensions of this approach allow for implementation with summary statistics rather than individual-level data, making TWAS an increasingly popular study design (Barbeira et al. 2018). Furthermore, because associations at the gene expression level often have clearer functional interpretations than associations with individual risk variants, TWAS have the potential to offer insights distinct from those offered by GWAS. Testing for associations with genes rather than SNPs also reduces the multiple testing burden, thereby improving statistical power for discovery. TWAS are, however, limited by the comprehensiveness of gene expression reference panels both across different tissues and for populations of non-European ancestry. Furthermore, although the genetic architecture of gene expression allows for reasonable imputation accuracy, gene expression can also be influenced by non-genetic, external factors.

5.2.5 Replication and Meta-analysis

Findings from a single genetic association study are not generally sufficient to instill confidence in results. Rather, results should be validated in independent samples and combined with other studies to bolster sample size.

5.2.5.1 Replication

Early genetic association studies frequently yielded results that failed to reproduce in independent samples (Hirschhorn et al. 2002; Ioannidis 2006; Ioannidis et al. 2001; Lohmueller et al. 2003). Why the surfeit of false positives? Historically, studies of candidate markers or genes often had small sample sizes, inappropriate thresholds for statistical significance, and/or low prior probabilities of true associations (Chanock et al. 2007; Hirschhorn and Altshuler 2002; Ioannidis 2005; Manolio et al. 2008; Mutsuddi et al. 2006; Wacholder et al. 2004). Even now, investigators are conscious of "winner's curse," whereby the effect estimates from initial discovery studies are consistently biased upward (Lohmueller et al. 2003; Goring et al. 2001; Huang et al. 2018). They are generally more attentive to winner's curse for genetic association studies than for other epidemiological investigations because the former most often test a large number of exposures. The gold standard for substantiating results from genetic association studies has thus become *replication* in independent samples. Replication has become important (and essentially required for publication) to externally validate the credibility of genetic associations.

Studies designed for the purposes of replication should ensure that sample sizes are sufficiently large to detect associations of the hypothesized magnitudes. In fact, sample sizes should ideally be larger than those of the initial study so as to account for overestimation in the original sample (unless one only wishes to replicate a limited number of variants). The larger the sample, the better success replication studies will have in reproducing results from and identifying false positives generated by the initial study. Replication studies should also evaluate the same ancestral population as the discovery study and, ideally, the same genetic variant with respect to the same definition of phenotype. Successful replication then entails finding the same direction of association (for the same effect allele) at a predetermined threshold for statistical significance. What that threshold should be is somewhat controversial; some investigators expect that associations be replicated at a genome-wide significance level, whereas others apply a less conservative threshold based on evaluating a smaller number of variants in the replication sample. Still others are not as concerned with the statistical significance of the replication association as they are with the significance of the joint analysis of discovery and replication.

In some cases, studies are not designed exclusively for the purposes of replication. Rather, colleagues may help one another replicate their strongest results by looking them up in independent, existing "discovery" studies. Once results are confirmed in the original target populations, investigators may also choose to evaluate associations in populations of varying ancestries. Results that replicate from these studies are often said to *generalize*, meaning that the effect is relevant to multiple human populations. In contrast to replication, studies conducted for generalization should draw from an ancestral population different from the discovery population.

It should be noted that while replication has become standard practice to corroborate genetic associations, it may not be as necessary as it once was. As genetic association studies have become increasingly sizeable and larger numbers of markers have been genotyped in large replication samples, the statistical power to detect modest effects has substantially increased. As a result, the potential for winner's curse has decreased. Still, replication inspires confidence in findings and remains customary for genetic association studies.

5.2.5.2 Meta-analysis

Results from multiple studies or even multiple stages of the same study can be combined into a single result via *meta-analysis*. Meta-analytical methods synthesize results from analyses that examine the same hypothesis without accessing individual-level data (as mega-analytical studies would). In doing so, they considerably boost the sample size and power for examining the hypothesis and thus may achieve a more precise estimate of the association of interest. Several software packages are available for the implementation of meta-analysis for GWAS, among which are METASOFT (Han and Eskin 2011), METAL (Willer et al. 2010), GWAMA (Magi and Morris 2010), PLINK (Purcell et al. 2007), and GenABEL/MetABEL (Aulchenko et al. 2007). Available features in most of these packages were summarized in a side-by-side comparison (Evangelou and Ioannidis

2013). In addition, there exist tools to meta-analyze results from populations of varying ethnicities (Hong et al. 2016; Morris 2011).

Meta-analytical methods can also be used to discover novel genetic loci with pleiotropic effects and to explore associations across phenotypes or disease subtypes. Association analysis based on subsets (ASSET) is a flexible meta-analysis framework that can evaluate associations for a given SNP across phenotypes and identify the combination of associated traits that maximizes the overall test statistic (Bhattacharjee et al. 2012). In addition to boosting power in the presence of heterogeneity, attractive features of ASSET are its ability to account for sample overlap across contributing studies and its internal correction for the multiple tests required by the subset search. ASSET has been applied to a number of traits, among which are multiple cancers (Fehringer et al. 2016) and immune-related diseases (Marquez et al. 2018; Zhu et al. 2018).

To conduct a rigorous meta-analysis, all studies should be subject to a standard quality control procedure that determines which SNPs are included in each study. It is a fundamental assumption of meta-analysis that the studies provide independent information, so it is also critical to ensure that there not be any overlap in the samples included from each study. In addition, the design of each study incorporated into a meta-analysis should ideally be similar; the measurement of covariates and phenotypes should be analogous, analytic procedures should be comparable, and covariate adjustment should be standardized (Zeggini and Ioannidis 2009). It is also important that all studies report results using the same reference allele and mode of inheritance. Imputation is often required to ensure that all studies in a meta-analysis offer data about the same SNPs (discussed further below).

The most common method to estimate an average effect across studies is fixed-effects modeling that weights each study effect based on its inverse variance. Mixed-effects models may also be used when there is substantial heterogeneity of effects across studies; their random effect parameters can help account for the heterogeneity. Regardless of the model selected for meta-analysis, it is important to quantify the differences across studies, particularly given that it is rare that studies perfectly fulfill the stringent criteria for meta-analysis. The most commonly used measures to do so are the Q statistic and I^2 index (Evangelou and Ioannidis 2013; Huedo-Medina et al. 2006; Panagiotou et al. 2013).

5.3 Design of Association Studies

The first step toward obtaining meaningful results from any genetic association study is designing it effectively. Investigators must always define appropriate phenotypes, designate a valid study population, and ensure a sufficient sample size. In this section, we outline some of these key elements that should be contemplated in conceiving new studies.

5.3.1 Quantitative Versus Qualitative Traits

There are two primary classes of phenotypes that one might wish to evaluate with a genetic association study, namely *quantitative* and *qualitative* (most often binary case-control). Quantitative traits generally have higher statistical power to detect genetic effects, and the interpretation of effects is often more straightforward. For a genetic variant that influences a quantitative trait, each allele or genotype class may be interpreted as affecting a unit change in the level of the trait. Alternatively, one might opt to study subjects at the extremes of a quantitative trait distribution to maximize power per genotyped individual for detecting associations (Huang and Lin 2007; Guey et al. 2011).

Many diseases do not have meaningful or well-established quantitative measures. In such scenarios, individuals are commonly classified as either affected or unaffected, and studies most often implement a case-control design. Frequencies of genetic variants observed in cases are compared with those observed in controls in order to evaluate whether an association between genes and disease exists. It is important to note that for a complex phenotype (e.g., metabolic syndrome) or one that is diagnosed over a long period (e.g., Alzheimer's disease), there may be some measurement error in dichotomizing individuals as cases or controls. Still, many association studies of binary traits have been extremely successful in detecting genetic variants correlated with disease (see Chap. 7 on what we have learned from GWAS).

5.3.2 Subject Selection

The most important facet of subject selection is ensuring that subjects are representative of their source population (Wacholder et al. 1992). For a case-control study in which cases with a particular disease are compared to unaffected controls, this means that controls should be individuals who, if diseased, would be cases. Whenever controls are not selected to represent the source population of the cases, spurious associations may result. Consider, for example, a scenario in which controls are selected from a different ancestral population from cases. In such a circumstance, control subjects might have fundamentally different allele frequencies in the SNPs of interest relative to cases. As a result, one is likely to find associations between these SNPs and disease even in the absence of true associations. This particular bias is called population stratification and can, if unaccounted for, confound GWAS. We will discuss methods to control for population stratification later in this chapter.

Cases are commonly recruited from a specific population, hospital, or disease registry. Depending on the study design, controls may either be unrelated (population-, hospital-, or registry-based) or family members of the cases. Controls are also commonly matched to cases with respect to ancestry, age, and sex.

Even without rigorous control selection, many GWAS have been successful at detecting highly replicated variants. Due to the high cost of subject recruitment and genotyping, investigators sometimes use genotype information from controls who have been recruited into prior studies and that has been made publicly available to researchers (e.g., via the database of genotypes and phenotypes (dbGaP)) (Luca et al. 2008; Burton et al. 2007; Paltoo et al. 2014). The inclusion of controls from public databases can also increase statistical power without affecting costs (Ho and Lange 2010). The potential bias arising from the use of such "convenience" or "public" controls is mitigated by the low measurement error in SNP genotyping, the absence of recall bias when studying inherited variants, large sample sizes, stringent criteria for statistical significance, and rigorous replication of findings. Nevertheless, the use of convenience controls may result in the confounding of associations due to population stratification (discussed further below). One should thus address the bias analytically with genetic information (Devlin and Roeder 1999; Price et al. 2006; Pritchard and Rosenberg 1999; Mitchell et al. 2014). One must also consider the potential for batch effects due to differences in genotyping quality control procedures and phenotype misclassification in individuals not thoroughly screened for common diseases. These issues can be assessed in small subsets of the sample by comparing genotype concordance in re-genotyped individuals or by conducting sensitivity analyses of the phenotype (Mitchell et al. 2014). Restricting the use of controls to those genotyped on the same platform and from the same genetic ancestry as cases may prevent or reduce these biases (Sinnott and Kraft 2012).

5.3.3 Sample Size

As with any study, it is critical that genetic association studies include a sufficiently large sample size to ensure good statistical power. The power of a study depends on the unknown frequency and effect size of the causal genetic variant(s) for which one is searching. Whenever SNPs in LD with the true causal variant are genotyped rather than the causal variant itself, power is reduced; the sample size required will be inflated proportionally to the inverse of the correlation between the genotyped and causative markers. There undoubtedly exist genetic variants that have a small effect on disease but that have not been detected due to insufficient sample size. In general, it is rare for successful GWAS to include fewer than 1000 cases and 1000 controls, and many include substantially larger numbers of individuals. Among the largest GWAS conducted to date have investigated smoking initiation (n = 1,232,091) (Liu et al. 2019), educational attainment (n = 1,131,881), blood pressure traits (n = 1,006,863) (Evangelou et al. 2018), and risk tolerance (n = 975,353) (Karlsson Linner et al. 2019).

5.4 Measurement of Genetic Information

Accurate measurement of genetic information is yet another crucial component of a reliable genetic association study. The study design is likely to inform the appropriate category of measurement broadly, but each method requires nuanced decision-making to achieve the highest quality and most relevant results. Here we discuss some of the most common tools utilized to measure genetic information and some considerations to contemplate in deciding on methods.

5.4.1 Common Variants

SNP genotyping arrays are currently the most pervasive tool utilized for evaluating genetic information. GWAS in particular typically employ microarray-based tag SNP genotyping techniques that capture common variation in the human genome. The arrays for early GWAS generally contained between 100,000 and 500,000 variants identified in databases such as HapMap. More current chips include approximately one million or more variants. No matter the set of variants, the array is then typed in a specific set of individuals. The arrays have generally been designed to measure variants at or above a minor allele frequency of 5%, though they may even miss some common variation (Jorgenson and Witte 2006). More recently, however, microarrays have been designed to detect variants down to a minor allele frequency of 1% (Hoffmann et al. 2011a, b). The platforms most often come from one of two companies: Illumina (San Diego, CA) or Affymetrix (Santa Clara, CA; now owned by Thermo Fisher Scientific) (Hindorff et al. 2009).

One consideration in designing a chip for assessment is determining the set of SNPs required to capture common variation in the population of interest. Consider, for example, the number of SNPs required to capture variation across the genome for African versus European populations. When the latter emigrated from Africa, they experienced a bottleneck that reduced the population size and resulting genetic variation. They thus have more LD than the former (see Chap. 8 on human demographic history). As a result, the chip used for a study of an African population requires more SNPs to obtain the same overall genomic coverage.

5.4.2 Rare Variants

The recent considerable expansion of the human population and negative selection of deleterious alleles over time have resulted in low allele frequencies for many disease-causing variants. Consequently, rare variants with substantial effects may remain untyped by standard genotyping assays. In addition, whole-genome sequencing is not generally cost-effective for the evaluation of rare variants, because the sample sizes required for association studies are normally much too large. More effective methods to identify rare disease-causing variants involve utilizing *exome sequencing* or *exome genotyping arrays* to investigate coding variation, even though these approaches ignore potentially important parts of the genome.

Sequencing and capture technologies are now able to accurately determine the sequence of nearly all protein-coding variants in humans (Choi et al. 2009; Gnirke et al. 2009; Ng et al. 2009; Teer and Mullikin 2010). They allow researchers to detect and genotype variants found in particular individuals without requiring that the variants be previously ascertained and included on genome-wide genotyping arrays. Originally, such technologies were most often utilized for (and successful at) identifying genetic contributors to many Mendelian disorders. More recently, the technologies have been leveraged to assay the exome as a popular approach for evaluating associations between rare variants and complex phenotypes. Exome sequencing selects the entire set of human exons as the sequencing target (Gnirke et al. 2009; Hodges et al. 2007). In contrast, exome arrays concentrate on a fixed set of variants. Regardless of the platform used to evaluate the exome, the variants assessed have functional implications that are *relatively* easy to derive. Still, due to reduced penetrance, sample sizes required for detecting associations with complex traits are generally larger than those required for the evaluation of Mendelian disorders. As exome sequencing costs have come down, however, it has become increasingly feasible to conduct well-powered studies. It is just important that they increase sample sizes in proportion to the rarity of causal variants.

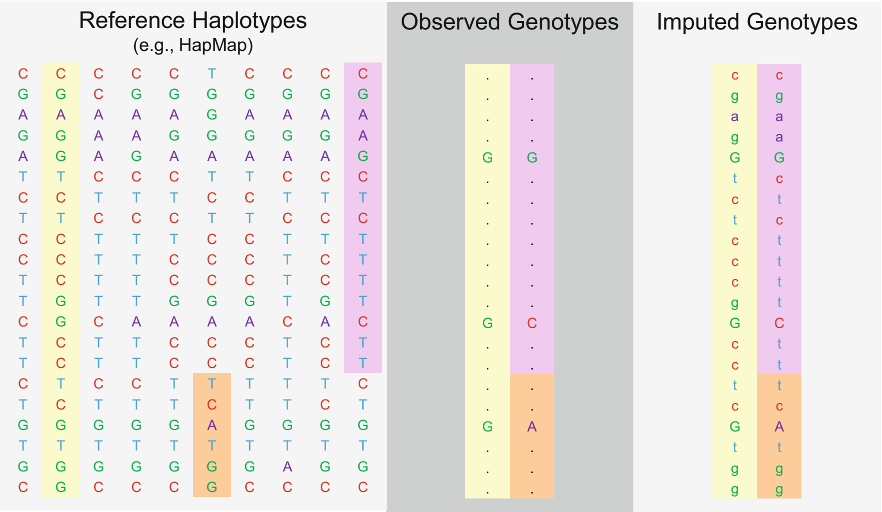
5.5 Data Analysis

Upon completing the measurement of both genotype and phenotype, one must consider the appropriate methods for the analysis of the data. In addition to thinking through the statistical methods that should be applied, one must also assess the data for their quality, consider covariates that should be accounted for, and potentially incorporate information from external sources. Below we identify some key considerations for analyzing genetic association studies.

5.5.1 Quality Control

Before analyzing the data from any study of genetic association, it is imperative that the genotyping be subject to a number of quality control checks. Samples that come from various sources may be processed in different ways or measured at different times, which can result in systematic differences across batches. One must also evaluate the proportion of samples that are successfully genotyped and test for Hardy–Weinberg equilibrium. Issues with these metrics could indicate genotyping problems that affect all of the SNPs in the sample. As such, one should remove SNPs that fail predefined quality standards from further consideration. Using SNP genotypes and external LD information on the underlying structure of a genetic region (e.g., from the TOPMed imputation reference panel), one can impute some of the untyped variants and variants that fail quality control. Doing so allows for a more thorough and powerful evaluation of potential associations across the genome (Huang et al. 2009; Marchini and Howie 2008; Marchini et al. 2007).

Note that sequencing (as opposed to genotyping) also requires appreciable quality control efforts, but their description is beyond the scope of this chapter.



Reference Haplotypes (e.g., HapMap)										Observed Genotypes		Imputed Genotypes	
C	C	C	C	C	T	C	C	C	C	.	.	c	c
G	G	C	G	G	G	G	G	G	G	.	.	g	g
A	A	A	A	A	G	A	A	A	A	.	.	a	a
G	G	A	A	G	G	G	G	G	A	.	.	g	a
A	G	A	G	A	A	A	A	A	G	G	G	G	G
T	T	C	C	C	T	T	C	C	C	.	.	t	c
C	C	T	T	T	C	C	T	T	T	.	.	c	t
T	T	C	C	C	T	T	C	C	C	.	.	t	c
C	C	T	T	T	C	C	T	T	T	.	.	c	t
C	C	T	T	C	C	C	T	C	T	.	.	c	t
T	C	T	T	C	C	C	T	C	T	.	.	c	t
T	G	T	T	G	G	G	T	G	T	.	.	g	t
C	G	C	A	A	A	A	C	A	C	G	C	G	C
T	C	T	T	C	C	C	T	C	T	.	.	c	t
T	C	T	T	C	C	C	T	C	T	.	.	c	t
C	T	C	C	T	T	T	T	T	C	.	.	t	t
T	C	T	T	T	C	T	T	C	T	.	.	c	c
G	G	G	G	G	A	G	G	G	G	G	A	G	A
T	T	T	T	T	T	T	T	T	T	.	.	t	t
G	G	G	G	G	G	G	A	G	G	.	.	g	g
C	G	C	C	C	G	C	C	C	C	.	.	g	g

Fig. 5.3 Schematic of genotype imputation modified from Li and colleagues (2009). Observed genotypes are compared to haplotypes in a reference panel to fill in unobserved genotypes

5.5.2 Data Imputation

Above, we discussed the exploitation of LD to capture common variation in the human genome without directly genotyping every SNP. Using LD patterns and haplotype frequencies (e.g., from the 1000 Genomes Project or TOPMed imputation reference panel), it is possible to *impute* data for SNPs that are not directly genotyped (Fig. 5.3) (Li et al. 2009). First, directly genotyped markers are compared to a variant-dense reference panel that contains haplotypes drawn from the same population as the study sample. A collection of shared haplotypes is then identified, and genotypes missing from the study panel can be inferred from the matching reference haplotypes. Because the study sample may match multiple reference haplotypes, one might opt to give a score or probability for an imputed marker rather than a definitive allele. In such scenarios, uncertainty can be incorporated into the analysis of imputed data, typically with Bayesian methods (Marchini et al. 2007). A less computationally intensive method involves pre-phasing, in which haplotypes are first estimated for every individual, followed by genotype imputation using the reference panel for each haplotype. This method also makes it faster to execute the imputation step with different reference panels as they become updated, since the genotypes need only be phased once and the estimated haplotypes are saved for future use (Howie et al. 2012). Note that imputation is especially useful for meta-analyzing results across studies that rely on different genotyping platforms.

5.5.3 Analysis of Common Variants

The conventional analysis plan for genome-wide data is a series of statistical tests that examine each locus independently for an association with the phenotype of interest. In the case of binary phenotypes, the simplest approach to these tests is *contingency tables* of counts of disease status by either genotype or allele count (Clarke et al. 2011). One can use a series of χ^2 or Cochran-Armitage trend tests to evaluate the independence of the rows and columns of each table. More commonly used than contingency tables is a regression approach, linear for quantitative traits and logistic for case-control traits, with categorical predictor variables for the genotypes. Regression models are generally favored because they allow adjustment for covariates, such as principal components (PC) of genetic ancestry.

For quantitative phenotypes, one can use a linear regression model framework, $\mathbf{y} = \alpha + \mathbf{x}\boldsymbol{\beta} + \mathbf{c}\boldsymbol{\gamma}$, to model association with phenotype, where $\mathbf{x}$ is a matrix (or vector) of genotypes, $\mathbf{c}$ is a matrix of covariates such as ancestral PCs, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the corresponding vectors of regression coefficients. For binary phenotypes, one would implement a logit link function. Regardless of the link function for the outcome, the $\boldsymbol{\beta}$s are the parameters of interest, and we can test the null hypothesis of no association between $\mathbf{x}$ and $\mathbf{y}$, H_0: $\boldsymbol{\beta} = \mathbf{0}$.

In addition to maximum likelihood-based regression models, linear mixed models (LMM) have become increasingly popular in GWAS, motivated by the computational challenges of analyzing datasets with large numbers of subjects and genetic variants. One of the most attractive features of LMM is their ability to control for confounding due to population structure by directly modeling relatedness among individuals, thereby improving power relative to standard GWAS with adjustment for PCs (Yang et al. 2014; Zhang et al. 2010; Zhou and Stephens 2012). The most recent addition to this class of methods, BOLT-LMM, adopts a Bayesian perspective by imposing a prior distribution on SNP effect sizes. It does not require computing or storing a genetic relationship matrix, which substantially reduces computational time compared to other methods (Loh et al. 2015). However, applying BOLT-LMM to case-control data can be problematic since genetic effects are estimated on the observed 0–1 scale rather than the odds ratio scale. As a result, transformations are required to make LMM-based results for binary traits comparable with logistic regression (Lloyd-Jones et al. 2018).

Regardless of the structure of the phenotypic data, there are several ways in which one might code the genotype data for association tests; the choice made should reflect the assumed mode of inheritance and genetic effect. In GWAS, the genotypes are usually coded as 0, 1, or 2 to reflect the number of effect alleles. This coding assumes that each additional copy of the variant allele increases the phenotype or log risk of disease by the same amount. The approach is fairly robust to incorrect assumptions about the mode of inheritance and has reasonable power to detect both additive and dominant genetic effects. It may, however, be underpowered if the true mode of inheritance is recessive (Lettre et al. 2007). If one believes the mode of inheritance to be recessive, then one may use an alternative genetic model that assumes that two copies of the risk allele are necessary to result in phenotype susceptibility. For such models, the heterozygote and wild-type homozygote are collapsed into a single category, and the genotypic exposure is treated as binary. The genotypic exposure is also treated as binary for models that assume a dominant mode of inheritance, but the heterozygote and mutant homozygote are collapsed separately from the wild-type homozygote. These dominant models assume that a single copy of the risk allele is expected to result in phenotype susceptibility. Both recessive and dominant models force heterozygotes to have the same risk or mean phenotype as one of the homozygotes. If investigators do not have an a priori hypothesis as to the mode of inheritance, they may choose to assess several different genetic models. Doing so, however, requires additional corrections for multiple testing. Another option when there is no a priori hypothesis would be to avoid any assumptions about how the risk for heterozygotes compares with both homozygotes. In such codominant models, maintaining the three distinct genotype classes requires two degrees of freedom, while the other models require only one, thereby making the latter more attractive if the genetic effect approximately follows one of their modes of inheritance.

To visualize the results from analyses of common variants, particularly from GWAS, investigators often generate *Manhattan plots* (e.g., Fig. 5.4). The x-axis of these scatter plots is a chromosomal position, and the y-axis shows the P value for association with the phenotype. Each point on the plot represents a single SNP, and the height of each point depicts the strength of association between the SNP and the phenotype. Manhattan plots with genome-wide significant results often exhibit clear peaks where SNPs in LD show comparable signals. Those with points seemingly scattered at random should be viewed with some skepticism.

Quantile-quantile (Q-Q) plots are another important visualization tool to evaluate potential bias or quality control problems in GWAS results (e.g., Fig. 5.5). These plots present the expected distribution of association test statistics for all SNPs on the x-axis against the observed values on the y-axis. Deviation from the $x = y$ line suggests a systematic difference between cases and controls across the whole genome (such as that which might occur in the presence of population stratification). One should rather hope to see the plotted points fall on the $x = y$ line until a curve at the very end representing any true associations.

5.5.4 Analysis of Rare Variants

While many common variants that contribute to complex diseases have been identified, the majority of variants contributing to disease susceptibility have yet to be described. Rare variants, which are unlikely to be captured by GWAS focusing on common SNPs, undoubtedly contribute to phenotype as well (Frazer et al. 2009; Gorlov et al. 2008). Unfortunately, detecting associations between individual rare variants and phenotypes can be difficult, even with large sample sizes; the low frequency of rare variants in the population results in low power (Gorlov et al. 2008; Altshuler et al. 2008; Li and Leal 2008). To increase power, researchers have

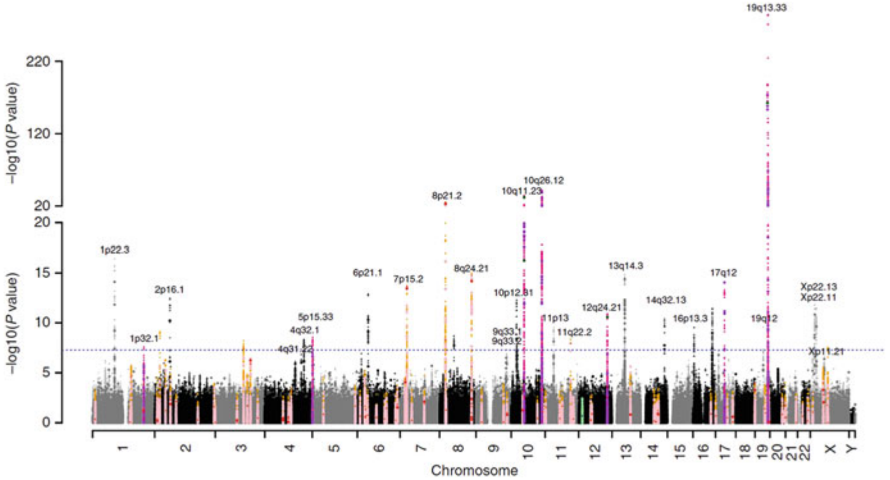


Fig. 5.4 Example Manhattan plot from a recent GWAS of prostate-specific antigen (PSA) levels (Hoffmann et al. 2017). *P* values are for variant associations with log-transformed PSA levels, adjusted for age and ancestry PCs using a linear regression model. Black and grey peaks indicate novel findings. Dark purple and magenta indicate previously reported PSA level-associated genotyped and imputed hits, respectively, and light purple and magenta indicate those within 0.5 Mb of previously reported hits that were replicated at genome-wide significance. Dark pink and red points denote previously reported prostate cancer SNPs genotyped and imputed, respectively, and pink and orange indicate those within 0.5 Mb of previously reported prostate cancer SNPs genotyped and imputed. Dark blue and green points denote the previously reported genotyped and imputed, respectively, SNPs associated with PSA levels only (and not prostate cancer), and light blue and green those within 0.5 Mb previously reported hits. Circles denote genotyped SNPs, and triangles represent imputed SNPs

developed a number of methods to evaluate the collective effect of multiple rare variants within and across genomic regions (Li and Leal 2008; Asimit and Zeggini 2010; Morgenthaler and Thilly 2007; Larson et al. 2017; Santorico and Hendricks 2016).

The two primary approaches to rare variant analyses are burden tests (Morgenthaler and Thilly 2007; Asimit et al. 2012; Li et al. 2012; Madsen and Browning 2009; Morris and Zeggini 2010; Zawistowski et al. 2010) and variance component tests (Neale et al. 2011; Pan 2009; Wu et al. 2010; Wu Michael et al. 2011). The simplest burden approach collapses rare variants into a single group by counting individuals who possess *at least* one rare variant in the genomic region under study and then tests for frequency differences across phenotypic groups. A limitation of burden tests is their assumption that all alleles have the same direction of effect; in the presence of both protective and deleterious variants, power can be substantially reduced. Burden tests also have reduced power in regions with a large number of non-causal variants. These limitations are addressed by variance component tests, the most common of which is the sequence kernel association test (SKAT) (Wu Michael et al. 2011). SKAT aggregates genetic information across variants using a

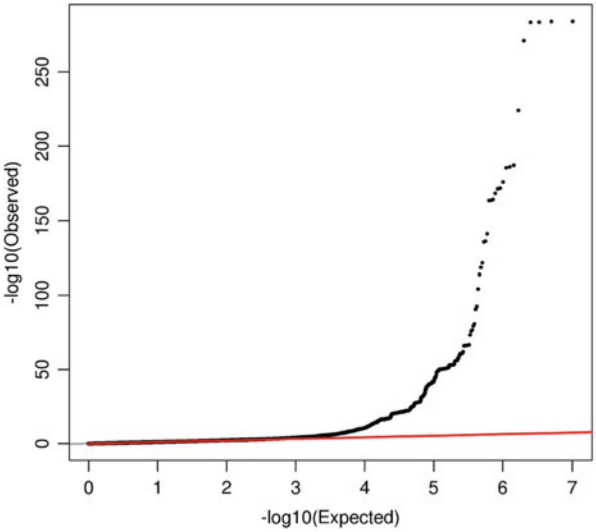


Fig. 5.5 Example Q-Q plot of results from a recent GWAS of PSA levels (Hoffmann et al. 2017). Because there were so many positive results, we see a substantial curve representing true associations at the end

kernel function. To test the null hypothesis that a set of rare variants does not impact the phenotype, one can compute the variance component score statistic Q, which is equal to $(\mathbf{y} - \overline{\mathbf{y}})'\mathrm{K}(\mathbf{y} - \overline{\mathbf{y}})$, where $\overline{\mathbf{y}}$ is the predicted mean of $\mathbf{y}$ under the null hypothesis of no association, adjusting for covariates $\mathbf{c}$, and the kernel K is an $n \times n$ matrix that defines the genetic similarity among individuals. The SKAT framework has expanded to create a family of tests accommodating a range of scenarios (Wu et al. 2013; Lee et al. 2012), including combination tests for common and rare variants (Ionita-Laza et al. 2013), time-to-event models (Chen et al. 2014), and multiple phenotypes (Dutta et al. 2019). Most recently, rare variant tests based on generalized linear mixed models have been proposed (Chen et al. 2019), as have flexible sliding-window approaches that account for LD structure (Li et al. 2019).

5.5.5 Incorporating External Information into Association Study Analyses

5.5.5.1 Gene Set Analysis

Analyses of data from GWAS can test multi-marker combinations of SNPs. Such *gene set analyses* can be used to determine whether groups of functionally related genes defined a priori are associated with a phenotype. Given that complex disease may result from a sum of changes across genes in a biological pathway, it makes sense to evaluate genes in a pathway as a set. These analyses aim to identify gene sets with coordinated expression changes that would not be detected by single variant methods (e.g., testing one SNP at a time).

Gene set analysis generally consists of four steps: (1) determining the gene sets to be tested, (2) selecting an appropriate set of hypotheses, (3) carrying out corresponding statistical tests, and (4) evaluating the statistical significance of said tests. Regarding step 2, there are three standard null hypotheses against which investigators most often test (Dinu et al. 2009; Nam and Kim 2008; Tian et al. 2005). The first is the competitive null hypothesis, which states that the genes in a set have the same association with phenotype as genes in the rest of the genome. The second is the self-contained null hypothesis, which asserts that the genes in a set are not associated with the disease phenotype. The third option, the mixed null hypothesis, declares that none of the sets under consideration is associated with the disease.

The set of hypotheses selected largely informs the tests that should be used for analysis. To obtain a test statistic for the competitive null, a measure of association should first be computed for each gene and the phenotype of interest. For genes in a given set, the association measures should then be combined. To evaluate the statistical significance of the combined test statistic, it should be compared against the distribution under the null hypothesis, obtained by permuting the association measures (Tian et al. 2005). The procedure is similar to obtaining a test statistic for the self-contained null hypothesis, but the null distribution should be generated by permuting the phenotypes across samples (Tian et al. 2005). Regardless of the test statistic, larger magnitudes indicate increasing significance, and the sign indicates the direction of change in phenotype.

The gene sets that are deemed significant are likely to depend on the choice of methods implemented to analyze gene set associations (Elbers et al. 2009a, b). Oftentimes, gene set analyses lack sufficient statistical power to detect gene sets consisting of markers only weakly associated with disease, and they are prone to several sources of bias, among which are gene set size, LD patterns, and overlapping genes (Elbers et al. 2009b; Cantor et al. 2010; Hong et al. 2009; Wang et al. 2011; Sun et al. 2019). It is important to consider and address all of these limitations when interpreting results from gene set analyses.

5.5.5.2 Hierarchical Modeling

Hierarchical modeling leverages the abundance of bioinformatic data characterizing the structural and functional roles of common variants analyzed for GWAS (Cantor et al. 2010; Wang et al. 2010). It aims to incorporate a priori biological knowledge via Bayesian methods (Cardin et al. 2012), stabilize effect and variance estimation of SNP associations (Aragaki et al. 1997; Evangelou et al. 2014), and improve the selection of SNPs for further evaluation (Witte 1997; Witte and Greenland 1996). It also addresses issues of multiple comparisons in analyses of GWAS. Rather than perform traditional single-locus analyses, hierarchical models output “knowledge-based” estimates of SNP effects, thereby improving the ranking of results from GWAS.

The hierarchical modeling approach uses higher-level “priors” to model the parameters of interest as random variables with a joint distribution that is a function of hyperparameters (Witte 1997). In addition to information about **x** and **y** (as defined above), one also utilizes information about similarities among the components of β. For example, one might assume that associations corresponding to markers that are located near one another on a particular chromosome might be similar. Conditional on this additional information, one may fit a second-stage generalized model for the expectation of β: $f_2(\beta \mid \mathbf{Z}) = \delta + \mathbf{Z}\pi$. According to this model, f_2 is a strictly increasing link function, and **Z** is a second-stage design matrix expressing the similarities among the β. Hierarchical (i.e., posterior) estimates are obtained by combining results from the different level models (Witte 1997).

5.5.6 Interactions

GWAS present an opportunity to go beyond single-locus analyses and into the realm of *gene-gene interactions* throughout the genome. Given the number of SNPs generally evaluated in GWAS, it would prove intractable to evaluate all pairwise combinations. Instead, one can reduce the set of SNPs to further investigate via one of several methods (McAllister et al. 2017). The first is to select an arbitrary significance threshold for the set of single-locus analyses. One can then evaluate all pairwise interactions between SNPs falling below the threshold, or between such SNPs and all other SNPs. Implementing this method, however, will preclude the discovery of combinations of markers that affect a significant change in disease risk even when the individual markers’ marginal effects are statistically undetectable. An alternative approach is to restrict the analysis of interactions to SNPs with an established biological function or within a particular protein family. A general comment regarding all analyses of interaction is that the scale (additive or multiplicative) on which they are evaluated will impact the results.

The evaluation of gene-gene interactions is not limited to model-based methods. Multifactor dimensionality reduction (MDR) was developed to reduce the dimensionality of multilocus data so as to improve the ability to detect gene-gene interactions. MDR pools genotypes into high-risk and low-risk groups, thereby reducing data to a single dimension. The method is nonparametric and model-free—one need not make hypotheses regarding the values of any parameters or assume any particular mode of inheritance (Motsinger and Ritchie 2006). The details of MDR analyses have been well described (Hahn et al. 2003; Ritchie et al. 2001, 2003).

The study of *gene-environment interactions* is another critical component of understanding the biological mechanisms of complex disease, heterogeneity across studies, and susceptible subpopulations (Dick et al. 2015). Until recently, gene-environment interaction studies have been largely carried out using candidate approaches. Such studies require the identification of genes with related biological functionality as well as knowledge of the mode of action through which environmental factors affect the genes of interest (Rava et al. 2013).

With the advent of high-throughput technology, investigators are exploring gene-environment interactions at the genome-wide level. They are also realizing some of the fundamental challenges of doing so. The genome-wide approach does not make use of prior knowledge of biological processes and pathways. In addition, the stringent significance threshold required due to the number of statistical tests may preclude the identification of significant interactions.

Analysis approaches can focus on environmental interactions with single genes, multiple genes, and/or biological pathways (Thomas 2010). Alternatively, one can utilize available biomarker data that may reflect intermediate phenotypes to establish informative priors in a hierarchical model framework (Li et al. 2012). Regardless, it bears recognition that statistical interaction does not conclusively indicate causality. Variants showing interaction may not be causal, and interactions may be significant for reasons other than true association (as is true for any other type of association). Still, the identification of gene-environment interactions with respect to disease risk is of fundamental public health relevance.

5.5.7 Incorporating Covariates

5.5.7.1 Population Stratification

Population-based association studies are susceptible to a form of confounding known as *population stratification*. It occurs whenever the gene of interest shows pronounced variation in allele frequency across ancestral subgroups of the population, and these subgroups differ in their baseline risk of disease. In extreme scenarios, population stratification can result from *cryptic relatedness*, wherein some individuals in an ostensibly unrelated population are actually related.

In a sample with population stratification in any form, SNPs with large allele frequency differences across groups are likely to be associated with the trait under study. The first step toward dealing with the bias is to ensure that cases and controls are well-matched in the study design phase. One can then evaluate the extent of residual population stratification via *Q*–*Q* plots and their associated inflation factor, lambda (λ). The latter is defined as the ratio of the median of the observed test statistics relative to the expected median and reflects the excess false-positive rate. When the value of lambda is inflated, one can adjust the test statistics by dividing them by lambda, thereby reducing them, and then recalculating the associated *P* values (Devlin and Roeder 1999).

In recent years, the more common approach to the management of population stratification has been the measurement of the ancestry of each sample in the dataset using PC methods (Price et al. 2006; Falush et al. 2003). These methods cluster individuals together based on their ancestral populations, often by comparing them with an external reference population such as the 1000 Genomes Project or TOPMed imputation reference panel. With the results, one can then exclude samples that are extremely different from the main clusters of individuals and then include the top 10 or so PCs as covariates in association analyses. A criticism of PCs is that they are unable to differentiate between true signal due to polygenicity and confounding due to population stratification. An alternative adjustment method can be used that requires LD score regression, which quantifies the association between LD and test statistics within a GWAS using a reference panel, in order to calculate a correction factor for genomic control in GWAS analysis (Bulik-Sullivan et al. 2015).

Methods that account for cryptic relatedness specifically tend to be more complex and are largely beyond the scope of this chapter. Software packages such as KING (Manichaikul et al. 2010) can be implemented to identify closely related individuals, who can subsequently be removed from the analytical population. There also exist approaches to deal with more distant relatedness as part of data analysis (Price et al. 2010).

5.5.7.2 Addressing Other Confounding

Genetic association studies are distinct from traditional epidemiological studies in that behavioral and environmental factors are unlikely to confound the associations of interest. Despite this, to draw clinically relevant conclusions, it is critical to properly account for patient-level covariates that may confound the associations under investigation. For example, associations may be confounded by sex whenever allele frequencies differ between the sexes (i.e., for sex-linked traits) (Clayton 2009). Other associations may be confounded by age whenever tag SNPs are in LD with both longevity SNPs and causal SNPs for a phenotype that most commonly occurs in late-life. If one does not adjust for the necessary covariates, one might find spurious associations due to sampling artifacts or bias in the study design. It is important to note that covariate adjustment may reduce statistical power because it requires additional degrees of freedom.

5.5.7.3 Improving Precision

Covariates may also be included in tests of association in an attempt to improve the precision of estimates. If a behavioral or environmental factor is associated with a quantitative phenotype under study independently of the genes of interest, then its inclusion is often beneficial. The covariate can explain some of the variability in the outcome, thereby reducing noise and increasing power (Mefford and Witte 2012). For binary traits, the story is more complicated. Inclusion of a covariate associated only with the outcome may actually reduce power for case-control association studies (Pirinen et al. 2012). There do exist methods, however, that leverage information about covariates to increase power in association studies of binary traits (Zaitlen et al. 2012).

5.5.8 Multiple Testing

Genetic association studies generally test hundreds of thousands of associations and may also examine multiple phenotypes and/or the results from various genetic models and covariate adjustments. The enormous number of resulting hypothesis tests must be adjusted for multiple comparisons, lest a large number of false-positive associations be detected. One approach to management of this issue of *multiple* *testing* is to limit the number of association tests executed. For example, in a study of multiple SNPs and a single phenotype, one might consider running only a single test per SNP. If the number of SNPs is sufficiently large, however, one will have to further adjust for the number of tests performed. Doing so is helpful to address issues of multiple comparisons but it is also important to note that the significance cutoffs described below are somewhat arbitrary and do not reflect the potential clinical or biological importance of an association (Witte et al. 1996).

5.5.8.1 Bonferroni and Number of Effective Independent Tests

The most straightforward and the most stringent means by which to correct for multiple testing is the *Bonferroni adjustment*. It adjusts the conventional type I error rate of 0.05 to 0.05/k, where k is the total number of tests performed. The adjustment assumes that the hypothesis tests are independent, thereby making this approach quite conservative in the context of correlated tests.

To make the strategy more applicable to the scenario of SNPs in LD, one can estimate the effective number of independent SNPs in lieu of the total number of SNPs in the Bonferroni adjustment (Nyholt 2004). Because the number of effective independent SNPs will always be less than or equal to the total number of genotyped SNPs, this approach is less conservative than the standard Bonferroni correction. For GWAS, the generally accepted alpha-level for statistical significance based on the effective number of genome-wide tests is 5×10^{-8}. This concept of *genome-wide significance* should be used only when hypotheses are tested on the genomic scale. It is not appropriate for candidate gene studies or replication studies, for which the number of effective independent tests is substantially lower.

5.5.8.2 Permutation Testing

Permutation testing is a more computationally intensive method with which to adjust for multiple testing. It is less conservative than the Bonferroni adjustment because it incorporates the correlation between genotypes and/or phenotypes. It does so by randomly shuffling phenotypes in the dataset, effectively removing any association between phenotypes and genotypes, while maintaining the correlation among genotypes resulting from LD within an individual, and then testing for association again. Random reassignment of the data and association testing is repeated some prespecified number of times (generally in the thousands), and all of the test statistics for the associations of interest are computed for each permuted dataset. A permuted P value can then be obtained by comparing the original test statistic to the distribution of test statistics from the permuted datasets. Several statistical packages implement permutation testing, though the most commonly used is PLINK (Purcell et al. 2007).

5.5.8.3 False Discovery Rate

Another method to account for multiple testing is the *false discovery rate (*Benjamini and Hochberg 1995*;* Brzyski et al. 2017*)*. Rather than control the family-wise error rate as does the Bonferroni adjustment, the false discovery rate controls the expected proportion of false discoveries among significant results.

Under a null hypothesis of no true associations in a GWAS dataset, *P* values would conform to a uniform distribution between zero and one. The false discovery rate essentially corrects for the expected number of false discoveries under this null distribution. While it typically allows researchers to reject more null hypotheses than would a Bonferroni adjustment, it may still be overly rigorous in the context of GWAS or large-scale candidate gene association studies. In such scenarios, one may implement weighted or stratified false discovery rates to achieve greater power to detect true associations in subsets of SNPs with a higher proportion of true positives than in the full set of SNPs (Genovese et al. 2006; Greenwood et al. 2007; Roeder et al. 2006).

5.5.8.4 Bayesian Approach

Rather than correct for multiple comparisons via traditional frequentist methods, one might choose to implement a Bayesian approach to the false discovery rate. Under a Bayesian framework, the Bayes factor quantifies the strength of evidence for an association between a SNP and phenotype. It is weighed against the prior probability of an association to arrive at the posterior probability (Stephens and Balding 2009). The calculation does not reference the number of SNPs tested. While the expected number of false-positive associations will increase as more tests are performed, so too will the number of true positive associations under a reasonable set of assumptions. As such, the ratio of true to false positives will remain roughly constant. Several software packages (e.g., SNPTEST (Marchini et al. 2007) and BIMBAM (Servin and Stephens 2007)) accommodate genome-wide Bayesian analyses.

5.6 Concluding Remarks

Well-executed genetic association studies can contribute immensely to our understanding of the underpinnings of disease. For meaningful conclusions to be drawn, it is critical that they be designed with an appropriate population consisting of a sufficient number of subjects. This number will depend upon the design selected—candidate gene, GWAS, or otherwise—and should take into account the methodological nuances thereof. Accurate measurement of genetic information is also integral to the success of an association study, as are the quality control checks that validate it. Then, statistical analyses must consider the specific research question at hand, so as to make decisions that will best answer it.

There remain many gaps in our understanding and many association studies that have the potential to fill them in moving forward. Since the proposal of an exposome in 2005 (Wild 2005), investigators have striven to conceive of methods that incorporate all of the exposures that individuals experience in a lifetime into the study of their genetics. They have also been busy considering the question of pleiotropy so as to identify genes that affect multiple, sometimes seemingly unrelated, phenotypes. Many are developing methods that will better address rare variants and interactions. The collection of these efforts will further improve our understanding of the disease process, risks, and response to therapy in this era of genomic discovery.

Acknowledgments This work was supported by National Institutes of Health grants R25CA112355, R01CA088164, and R01CA201358.

References

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881–888

Aragaki CC, Greenland S, Probst-Hensch N, Haile RW (1997) Hierarchical modeling of gene-environment interactions: estimating NAT2 genotype-specific dietary effects on adenomatous polyps. Cancer Epidemiol Biomark Prev 6:307–314

Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annu Rev Genet 44:293–308

Asimit JL, Day-Williams AG, Morris AP, Zeggini E (2012) ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. Hum Hered 73:84–94

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. Bioinformatics 23:1294–1296

Barbeira AN, Dickinson SP, Bonazzola R et al (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun 9:1825

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B 57:289–300

Bhattacharjee S, Rajaraman P, Jacobs KB et al (2012) A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. Am J Hum Genet 90:821–835

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Bowden J, Del Greco MF, Minelli C, Davey Smith G, Sheehan N, Thompson J (2017) A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Stat Med 36:1783–1802

Brzyski D, Peterson CB, Sobczyk P, Candes EJ, Bogdan M, Sabatti C (2017) Controlling the rate of GWAS false discoveries. Genetics 205:61–75

Bulik-Sullivan B, Loh P-R, Finucane H et al (2015) LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 47:291–295

Buniello A, MacArthur JAL, Cerezo M et al (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47:D1005–D1012

Burgess S, Butterworth A, Thompson SG (2013) Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol 37:658–665

Burton PR, Clayton DG, Cardon LR et al (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678

Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet 86:6–22

Cardin NJ, Mefford JA, Witte JS (2012) Joint association testing of common and rare genetic variants using hierarchical modeling. Genet Epidemiol 36:642–651

Carlson CS, Matise TC, North KE et al (2013) Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. PLoS Biol 11:e1001661

Chanock SJ, Manolio T, Boehnke M et al (2007) Replicating genotype-phenotype associations. Nature 447:655–660

Chen GK, Witte JS (2007) Enriching the analysis of genomewide association studies with hierarchical modeling. Am J Hum Genet 81:397–404

Chen H, Lumley T, Brody J et al (2014) Sequence kernel association test for survival traits. Genet Epidemiol 38:191–197

Chen H, Huffman JE, Brody JA et al (2019) Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. Am J Hum Genet 104:260–274

Choi M, Scholl UI, Ji W et al (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 106:19096–19101

Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT (2011) Basic statistical analysis in genetic case-control studies. Nat Protoc 6:121–133

Claussnitzer M, Cho JH, Collins R et al (2020) A brief history of human disease genetics. Nature 577:179–189

Clayton DG (2009) Sex chromosomes and genetic association studies. Genome Med 1:110

Cordell HJ, Clayton DG (2005) Genetic association studies. Lancet 366:1121–1131

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Dick DM, Agrawal A, Keller MC et al (2015) Candidate gene-environment interaction research: reflections and recommendations. Perspect Psychol Sci 10:37–59

Dinu I, Potter JD, Mueller T et al (2009) Gene-set analysis and reduction. Brief Bioinform 10:24–34

Dutta D, Scott L, Boehnke M, Lee S (2019) Multi-SKAT: general framework to test for rare-variant association with multiple phenotypes. Genet Epidemiol 43:4–23

Elbers CC, van der Schouw YT, Wijmenga C, Onland-Moret NC (2009a) Comment on: Perry et al. (2009) interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. Diabetes 58:1463–1467. e9; author reply e10

Elbers CC, van Eijk KR, Franke L et al (2009b) Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol 33:419–431

Evangelou E, Ioannidis JPA (2013) Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 14:379–389

Evangelou M, Dudbridge F, Wernisch L (2014) Two novel pathway analysis methods based on a hierarchical model. Bioinformatics 30:690–697

Evangelou E, Warren HR, Mosen-Ansorena D et al (2018) Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat Genet 50:1412–1425

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Fehringer G, Kraft P, Pharoah PD et al (2016) Cross-cancer genome-wide analysis of lung, ovary, breast, prostate, and colorectal cancer reveals novel pleiotropic associations. Cancer Res 76:5103–5114

Frazer KA, Ballinger DG, Cox DR et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251

Gabriel SB, Schaffner SF, Nguyen H et al (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229

Gamazon ER, Wheeler HE, Shah KP et al (2015) A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47:1091–1098

Genovese CR, Roeder K, Wasserman L (2006) False discovery control with p-value weighting. Biometrika 93:509–524

Gilbert W, Maxam A (1973) The nucleotide sequence of the lac operator. Proc Natl Acad Sci USA 70:3581–3584

Gnirke A, Melnikov A, Maguire J et al (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27:182–189

Goring HH, Terwilliger JD, Blangero J (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 69:1357–1369

Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82:100–112

Gray R, Wheatley K (1991) How to avoid bias when comparing bone marrow transplantation with chemotherapy. Bone Marrow Transplant 7(Suppl 3):9–12

Greenwood CM, Rangrej J, Sun L (2007) Optimal selection of markers for validation or replication from genome-wide association studies. Genet Epidemiol 31:396–407

Guey LT, Kravic J, Melander O et al (2011) Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. Genet Epidemiol 35:236–246

Gusev A, Ko A, Shi H et al (2016) Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 48:245–252

Hahn LW, Ritchie MD, Moore JH (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19:376–382

Han B, Eskin E (2011) Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am J Hum Genet 88:586–598

Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey SG (2016) Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. Am J Clin Nutr 103:965–978

Hindorff LA, Sethupathy P, Junkins HA et al (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367

Hirschhorn JN, Altshuler D (2002) Once and again-issues surrounding replication in genetic association studies. J Clin Endocrinol Metab 87:4438–4441

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4:45–61

Ho LA, Lange EM (2010) Using public control genotype data to increase power and decrease cost of case–control genetic association studies. Hum Genet 128:597–608

Hodges E, Xuan Z, Balija V et al (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet 39:1522–1527

Hoffmann TJ, Kvale MN, Hesselson SE et al (2011a) Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics 98:79–89

Hoffmann TJ, Zhan Y, Kvale MN et al (2011b) Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics 98:422–430

Hoffmann TJ, Van Den Eeden SK, Sakoda LC et al (2015) A large multiethnic genome-wide association study of prostate cancer identifies novel risk variants and substantial ethnic differences. Cancer Discov 5:878–891

Hoffmann TJ, Passarelli MN, Graff RE et al (2017) Genome-wide association study of prostate-specific antigen levels identifies novel loci independent of prostate cancer. Nat Commun 8:14248

Hong MG, Pawitan Y, Magnusson PK, Prince JA (2009) Strategies and issues in the detection of pathway enrichment in genome-wide association studies. Hum Genet 126:289–301

Hong J, Lunetta KL, Cupples LA, Dupuis J, Liu CT (2016) Evaluation of a two-stage approach in trans-ethnic meta-analysis in genome-wide association studies. Genet Epidemiol 40:284–292

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44:955–959

Huang BE, Lin DY (2007) Efficient association mapping of quantitative trait loci with selective genotyping. Am J Hum Genet 80:567–576

Huang L, Li Y, Singleton AB et al (2009) Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet 84:235–250

Huang QQ, Ritchie SC, Brozynska M, Inouye M (2018) Power, false discovery rate and Winner's curse in eQTL studies. Nucleic Acids Res 46:e133

Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, Botella J (2006) Assessing heterogeneity in meta-analysis: Q statistic or I2 index? Psychol Methods 11:193–206

International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2:e124

Ioannidis JP (2006) Common genetic variants for breast cancer: 32 largely refuted candidates and larger prospects. J Natl Cancer Inst 98:1350–1353

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29:306–309

Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X (2013) Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet 92:841–853

Jorgenson E, Witte JS (2006) Coverage and power in genomewide association studies. Am J Hum Genet 78:884–888

Karlsson Linner R, Biroli P, Kong E et al (2019) Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. Nat Genet 51:245–257

Katan MB (1986) Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet 1:507–508

Kraft P, Wacholder S, Cornelis MC et al (2009) Beyond odds ratios—communicating disease risk based on genetic profiles. Nat Rev Genet 10:264–269

Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Larson NB, McDonnell S, Cannon Albright L et al (2017) gsSKAT: rapid gene set analysis and multiple testing correction for rare-variant association studies using weighted linear kernels. Genet Epidemiol 41:297–308

Lee S, Emond MJ, Bamshad MJ et al (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 91:224–237

Lettre G, Lange C, Hirschhorn JN (2007) Genetic model testing and statistical power in population-based association studies of quantitative traits. Genet Epidemiol 31:358–362

Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83:311–321

Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10:387–406

Li R, Conti DV, Diaz-Sanchez D, Gilliland F, Thomas DC (2012) Joint analysis for integrating two related studies of different data types and different study designs using hierarchical modeling approaches. Hum Hered 74:83–96

Li Z, Li X, Liu Y et al (2019) Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. Am J Hum Genet 104:802–814

Lindquist KJ, Jorgenson E, Hoffmann TJ, Witte JS (2013) The impact of improved microarray coverage and larger sample sizes on future genome-wide association studies. Genet Epidemiol 37:383–392

Liu M, Jiang Y, Wedow R et al (2019) Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet 51:237–244

Lloyd-Jones LR, Robinson MR, Yang J, Visscher PM (2018) Transformation of summary statistics from linear mixed model association on all-or-none traits to odds ratio. Genetics 208:1397–1408

Loh PR, Tucker G, Bulik-Sullivan BK et al (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet 47:284–290

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177–182

Luca D, Ringquist S, Klei L et al (2008) On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82:453–463

Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5:e1000384

Magi R, Morris AP (2010) GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 11:288

Maher B (2008) Personal genomes: the case of the missing heritability. Nature 456:18–21

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26:2867–2873

Manolio TA, Brooks LD, Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118:1590–1605

Marchini J, Howie B (2008) Comparing algorithms for genotype imputation. Am J Hum Genet 83:535–539. author reply 539-540

Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913

Marquez A, Kerick M, Zhernakova A et al (2018) Meta-analysis of immunochip data of four autoimmune diseases reveals novel single-disease and cross-phenotype associations. Genome Med 10:97

Mavaddat N, Michailidou K, Dennis J et al (2019) Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. Am J Hum Genet 104:21–34

McAllister K, Mechanic LE, Amos C et al (2017) Current challenges and new opportunities for gene-environment interaction studies of complex diseases. Am J Epidemiol 186:753–761

Mefford J, Witte JS (2012) The covariate's dilemma. PLoS Genet 8:e1003096

Mitchell BD, Fornage M, McArdle PF et al (2014) Using previously genotyped controls in genome-wide association studies (GWAS): application to the stroke genetics network (SiGN). Front Genet 5

Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res 615:28–56

Morris AP (2011) Transethnic meta-analysis of genomewide association studies. Genet Epidemiol 35:809–822

Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34:188–193

Motsinger AA, Ritchie MD (2006) Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies. Hum Genomics 2:318–328

Mutsuddi M, Morris DW, Waggoner SG, Daly MJ, Scolnick EM, Sklar P (2006) Analysis of high-resolution HapMap of DTNBP1 (Dysbindin) suggests no consistency between reported common variant associations and schizophrenia. Am J Hum Genet 79:903–909

Nam D, Kim SY (2008) Gene-set approach for expression pattern analysis. Brief Bioinform 9:189–197

Neale BM, Rivas MA, Voight BF et al (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7:e1001322

Nelson SC, Doheny KF, Pugh EW et al (2013) Imputation-based genomic coverage assessments of current human genotyping arrays. G3 (Bethesda) 3:1795–1807

Ng SB, Turner EH, Robertson PD et al (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276

Nolte IM, van der Most PJ, Alizadeh BZ et al (2017) Missing heritability: is the gap closing? An analysis of 32 complex traits in the lifelines cohort study. Eur J Hum Genet 25:877–885

Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74:765–769

Paltoo DN, Rodriguez LL, Feolo M et al (2014) Data use under the NIH GWAS data sharing policy and future directions. Nat Genet 46:934–938

Pan W (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol 33:497–507

Panagiotou OA, Willer CJ, Hirschhorn JN, Ioannidis JPA (2013) The power of meta-analysis in genome-wide association studies. Annu Rev Genomics Hum Genet 14:441–465

Pierce BL, Burgess S (2013) Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. Am J Epidemiol 178:1177–1184

Pirinen M, Donnelly P, Spencer CC (2012) Including known covariates can reduce power to detect genetic effects in case-control studies. Nat Genet 44:848–851

Piro RM, Di Cunto F (2012) Computational approaches to disease-gene prediction: rationale, classification and successes: computational disease-gene prediction. FEBS J 279:678–696

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11:459–463

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228

Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

Rava M, Ahmed I, Demenais F, Sanchez M, Tubert-Bitter P, Nadif R (2013) Selection of genes for gene-environment interaction studies: a candidate pathway-based strategy using asthma as an example. Environ Health 12:56

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Ritchie MD, Hahn LW, Roodi N et al (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–147

Ritchie MD, Hahn LW, Moore JH (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24:150–157

Roeder K, Wasserman L (2009) Genome-wide significance levels and weighted hypothesis testing. Stat Sci 24:398–413

Roeder K, Bacanu SA, Wasserman L, Devlin B (2006) Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet 78:243–252

Roshan U, Chikkagoudar S, Wei Z, Wang K, Hakonarson H (2011) Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. Nucleic Acids Res 39:e62–e62

Santorico SA, Hendricks AE (2016) Progress in methods for rare variant association. BMC Genet 17(Suppl 2):6

Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3:e114

Sinnott JA, Kraft P (2012) Artifact due to differential error when cases and controls are imputed from different platforms. Hum Genet 131:111–119

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2007) Optimal designs for two-stage genome-wide association studies. Genet Epidemiol 31:776–788

Smith GD, Ebrahim S (2003) ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 32:1–22

Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet 10:681–690

Sun R, Hui S, Bader GD, Lin X, Kraft P (2019) Powerful gene set analysis in GWAS with the generalized Berk-Jones statistic. PLoS Genet 15:e1007530

Teer JK, Mullikin JC (2010) Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet 19:R145–R151

Thomas D (2010) Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu Rev Public Health 31:21–36

Thomas DC, Haile RW, Duggan D (2005) Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet 77:337–345

Thomas DC, Casey G, Conti DV, Haile RW, Lewinger JP, Stram DO (2009) Methodological issues in multistage genome-wide association studies. Stat Sci 24:414

Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ (2005) Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA 102:13544–13549

Toland AE (2019) Polygenic risk scores for prostate cancer: testing considerations. Can J Urol 26:17–18

Torkamani A, Wineinger NE, Topol EJ (2018) The personal and clinical utility of polygenic risk scores. Nat Rev Genet 19:581–590

Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. Science 291:1304–1351

Wacholder S, McLaughlin JK, Silverman DT, Mandel JS (1992) Selection of controls in case-control studies. I Principles. Am J Epidemiol 135:1019–1028

Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 96:434–442

Wang DG, Fan JB, Siao CJ et al (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11:843–854

Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z (2011) Gene set analysis of genome-wide association studies: methodological issues and perspectives. Genomics 98:1–8

Ware JH (2006) The limitations of risk factors as prognostic tools. N Engl J Med 355:2615–2617

Wild CP (2005) Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomark Prev 14:1847–1850

Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26:2190–2191

Witte JS (1997) Genetic analysis with hierarchical models. Genet Epidemiol 14:1137–1142

Witte JS, Greenland S (1996) Simulation study of hierarchical regression. Stat Med 15:1161–1170

Witte JS, Elston RC, Schork NJ (1996) Genetic dissection of complex traits. Nat Genet 12:355–356. author reply 357–358

Witte JS, Elston RC, Cardon LR (2000) On the relative sample size required for multiple comparisons. Stat Med 19:369–372

Wojcik GL, Fuchsberger C, Taliun D et al (2018) Imputation-aware tag SNP selection to improve power for large-scale, multi-ethnic association studies. G3 (Bethesda) 8:3255–3267

Wu R, Kaiser AD (1968) Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. J Mol Biol 35:523–537

Wu Michael C, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89:82–93

Wu R, Taylor E (1971) Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. J Mol Biol 57:491–511

Wu MC, Kraft P, Epstein MP et al (2010) Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet 86:929–942

Wu MC, Maity A, Lee S et al (2013) Kernel machine SNP-set testing under multiple candidate kernels. Genet Epidemiol 37:267–275

Xing C, Huang J, Hsu YH et al (2016) Evaluation of power of the Illumina HumanOmni5M-4v1 BeadChip to detect risk variants for human complex diseases. Eur J Hum Genet 24:1029–1034

Xiong M, Guo SW (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. Am J Hum Genet 60:1513–1531

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014) Advantages and pitfalls in the application of mixed-model association methods. Nat Genet 46:100–106

Zaitlen N, Lindstrom S, Pasaniuc B et al (2012) Informed conditioning on clinical covariates increases power in case-control association studies. PLoS Genet 8:e1003032

Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S (2010) Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet 87:604–617

Zeggini E, Ioannidis JP (2009) Meta-analysis in genome-wide association studies. Pharmacogenomics 10:191–201

Zhang Z, Ersoz E, Lai CQ et al (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355–360

Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44:821–824

Zhu Z, Lee PH, Chaffin MD et al (2018) A genome-wide cross-trait analysis from UK biobank highlights the shared genetic architecture of asthma and allergic diseases. Nat Genet 50:857–864

Identity by Descent in the Mapping of Genetic Traits

6

Elizabeth A. Thompson

Abstract

This chapter shows how the descent of genome from an ancestor to currently observed descendants results in *identity by descent* (IBD) in current individuals and hence similarities in their DNA at genetic marker loci. Conversely, data on the marker genotypes of individuals provides inferences of shared descent of genome in current individuals, not just genome-wide but in specific genome regions. Regions where shared genome accords with phenotypic similarities for a trait provide evidence of causal DNA at some location in the region. The chapter considers both data observed on defined pedigree structures, and data on population members whose pedigree relationships may be remote and are unknown. We take a model-based approach, deriving probabilities of IBD and likelihoods of mapping parameters, given observed genetic data. We first consider probabilities of gene IBD among individuals and across a chromosome, using either a known pedigree or a population-based model. We then consider probabilities of genotypic and phenotypic data on individuals, conditional on latent IBD. Thence IBD may be inferred from marker genotypes, combining information from multiple SNP markers. Finally, we show how location-specific realizations of IBD can be used to address questions of gene mapping. By focusing on IBD, we unify pedigree and population-based approaches.

E. A. Thompson (✉)
Department of Statistics, University of Washington, Seattle, WA, USA
e-mail: eathomp@uw.edu

K. E. Lohmueller, R. Nielsen (eds.), *Human Population Genomics*,
https://doi.org/10.1007/978-3-030-61646-5_6

6.1 Introduction

6.1.1 Identity by Descent

Genetic similarities, whether at the population level or between close relatives, result from coancestry. Copies of DNA that descend to current individuals from a single copy of DNA in a common ancestor are, with high probability, of the same allelic type, implying greater phenotypic similarity. Such DNA is said to be *identical by descent* or IBD, and this concept is of key importance in analysis of phenotypic variation, across species of plants and animals (including humans), and across traits including disease traits, selected traits in agriculture, and normal variation.

Defining IBD is not straightforward. At every point in the genome, among any group of organisms, the coalescent ancestry will at some point converge in the *most recent common ancestor* (MRCA), and relative to this point all the organisms are IBD. Although models for this time of the MRCA of a pair of haploid genomes across genetic loci can be used for inference (Li and Durbin, 2011), for genetic mapping it will be important to have models for the changing patterns of IBD among individuals across the genome. We therefore define IBD relative to an ancestral population or time point of interest. In studies of data on defined pedigrees, IBD has often been measured relative to the pedigree founders, but these founders are members of a population and may be related or inbred relative to an earlier point in the population's history.

The choice of the time-depth of interest for analyses in IBD will depend on the scientific question. New variants arising in the distant past descend to current individuals and remain in linkage disequilibrium (LD) with the genetic background on which they arose. This ancient coancestry (IBD) gives rise to the patterns of LD we see in populations today and provides information on population structure and demographic history. At the other end of the scale, IBD among close relatives may be important in analysis of family data, but generally close pedigree relationships are known. In this chapter we focus on IBD relative to ancestors at a time-depth of 10 to 40 generations. For human populations, this is beyond the depth for which pedigree information is available or, even if available, provides useful information on coancestry of genome. On the other hand, the last 1000 years of human history encompasses a large part of the huge expansion of the human species and has established the current patterns of genetic variation within and among local populations.

DNA is copied from parents to offspring in accordance with the process of meiosis. Throughout this chapter we restrict attention to nuclear autosomal chromosomes; that is, we consider neither the sex (X and Y) chromosomes nor mitochondrial DNA. At any given locus, the process is as specified by Mendel's first law (Mendel, 1866). A randomly chosen one of the two homologous copies of the parental DNA is copied to an offspring. A fundamental feature of the process is that this random choice is made independently in distinct meioses. Figure 6.1 shows a schematic view of DNA descent at a locus in a small family. At this locus, brothers

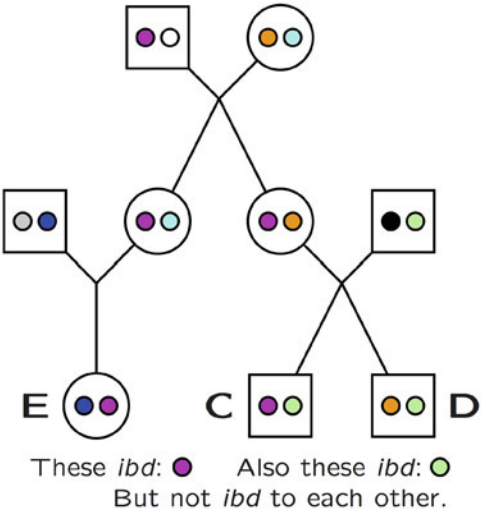


Fig. 6.1 Descent of DNA in a pedigree. Males are designated by squares; females by circles. Individuals *C* and *D* are brothers, and *E* is their cousin, since the mothers of these three individuals are sisters

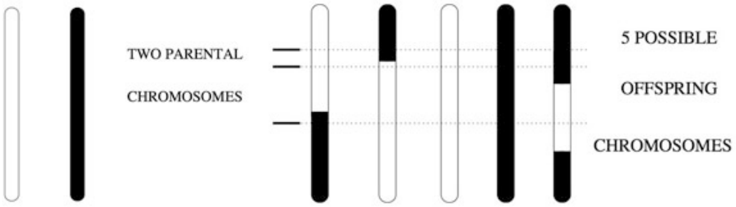


Fig. 6.2 Descent of a chromosome from parent to gamete. Shown is a pair of parental homologous chromosomes together with five examples of potential gamete chromosomes. One gamete from the parent will be transmitted to an offspring. At closer locations, the DNA in an offspring chromosome has higher probability of deriving from the same parental chromosome

C and *D* share their paternal DNA IBD, and cousins *E* and *C* share their maternal DNA IBD from their grandfather.

The process of meiosis also determines the copying of parental DNA across a chromosome. During the formation of gametes, in the first meiotic division, homologous chromosomes can exchange DNA leading to alternating segments of DNA from each of the two parental chromosomes (Fig. 6.2). The points at which the offspring chromosome switches between the parent's maternal and paternal chromosomes are *crossovers*. Typically, for short chromosomes, close to 50% will have no crossover, so that an intact parental chromosome is transmitted. Most of the remaining 50% will have one crossover, with a small probability of more crossovers. In larger chromosomes, there is a high probability of multiple crossovers. On average, the distance between crossover points is of order 10^8 base pairs (bp).

From parents to offspring, DNA is inherited in long segments, but over multiple generations, repeated meioses break the IBD DNA in current individuals into smaller and smaller segments.

6.1.2 From Descent to Gene Mapping

The fundamental framework of genetic epidemiology was formalized by Elston and Stewart (1971), who defined the three components of a genetic model: population, transmission, and penetrance. DNA variation in a population results from evolutionary processes and demographic history, arising via mutation, and modified by selection and random genetic drift. The parameters of the population model are allele, genotype, and haplotype frequencies, which are often assumed known in genetic epidemiological studies. The model for transmission of DNA from parents to offspring is provided by the process of meiosis, as summarized in Mendel's first law and the recombination probabilities across the genome (Sect. 6.1.1). The parameters of the population process normally relate only to the genetic map that provides recombination probabilities between any two loci, although it could also include models of segregation distortion or genetic interference.

Finally, the penetrance model defines the probability relationship between an individual's genotype at the relevant locus or loci and the observable data. For genetic markers, this relationship is straightforward. Normally unphased genotypes at each marker locus are directly observed, although a model for typing error may be included. For genetic epidemiological traits of interest, the penetrance model is often the least certain and most complex component of the model, and successful inference will often require careful analysis of a range of possible models.

The goal of genetic mapping is to determine the genome locations of DNA that affects a phenotypic trait of interest. This mapping relies on co-inheritance of DNA at marker loci of known location and of DNA inferred to affect the trait. At the population level, co-inheritance of DNA leads to linkage disequilibrium (LD), which is the basis of association mapping. In a defined pedigree (Fig. 6.1), the dependence in inheritance between a trait and a genetic marker provides evidence that the DNA affecting the trait is in proximity to the genetic marker locus. In between these extremes, even in the absence of known pedigree relationships, evidence that individuals of similar trait phenotype share DNA IBD in particular regions of the genome provides evidence that these regions harbor causal loci. This is the basis of IBD-based genetic mapping.

6.1.3 Outline of the Chapter

The remainder of the chapter is divided into three main sections. The focus is on related individuals, who may therefore share genome IBD, but the relationships among the individuals may be known or unknown. Within each section we consider both pedigree data (known relationships) and population-level data (unknown relationships). We present many of the ideas through numerical examples, but the reader should not (unless they wish to) be concerned with the details of computations and derivations. Focus instead on the qualitative message in the numbers provided: do the results make sense? and why does a given table or result provide insight into the approach to and goals of genetic mapping?

In Sect. 6.2 we consider probabilities of the underlying IBD in related individuals. At any locus, even a small number of gametes can share IBD in many different ways. The changes in the IBD pattern across a chromosome that result from recombination events in ancestral meioses add additional complexity. Additionally, these ancestral processes have high variance. Against these complexities are the facts that, on a bp scale, IBD changes slowly across the chromosome and that, in populations, relatively simple prior models for IBD can provide a basis for inference. Sections 6.2.3 and 6.2.4 show the importance of IBD. Given a specification of the IBD and a penetrance model, probabilities of genotypes and phenotypes, jointly across sets of observed individuals can be computed, without further reference to the descent structure that gave rise to the IBD.

While Sect. 6.2.3 provides probabilities of marker genotype data given a pattern of IBD, in Sect. 6.3 we consider the reverse problem—the inference of IBD from genetic marker data. As dense genetic marker data become increasingly available, and traits of interest become increasingly complex, there has been a shift in the paradigm of joint analysis of trait and marker data for purposes of gene mapping. Whereas models for complex traits may involve several genetic loci, each genetic marker corresponds to a single locus and simple models apply. By first analyzing the marker data to obtain patterns of IBD across a chromosome among observed individuals, direct joint consideration of marker and trait data may be avoided. Instead the patterns of IBD inferred from marker data may be used to investigate multiple trait models and hypotheses or even multiple traits observed on subsets of the same individuals (see, e.g., Chapman et al. 2015, Peter et al. 2016 and Saad et al. 2016). Efficient methods for realizing, estimating, and storing complex IBD summaries based on genetic marker data are key to success of this approach. This applies both in the presence of defined pedigree structures and also in populations: Sect. 6.3 considers both cases.

Finally, in Sect. 6.4 we consider approaches to genetic mapping of loci underlying phenotypes of interest, using the IBD inferred from genetic marker data. Both classical and modern approaches can be phrased in terms of IBD, and placing analyses in this framework shows there is no fundamental difference between pedigree-based and population-based approaches. Indeed, framing the problem in terms of IBD allows the combination of pedigree and population data. For close relatives, where relationships can be well-validated, the assumed pedigree is useful, not least in providing phase information on individual haplotypes. However, the location-specific IBD resulting from more remote relationships is often better inferred without reference to an assumed pedigree, and this IBD may be used in exactly the same way as pedigree-based IBD in genetic mapping algorithms. A final conclusion is thus that, with modern genetic marker data, it is not a choice between pedigree and population data: gene mapping relies on having related individuals, but the relationships do not need to be known.

The original version of this chapter was written in 2013–2016. HSG4, the 4th edition of the *Handbook of Statistical Genomics* (Balding et al., 2019), was developed in 2017–2018. Thus, although the focus is slightly different, there is significant overlap between the current chapter and Chapter 20 of HSG4 (Thompson, 2019).

6.2 Probabilities of IBD

6.2.1 IBD in Defined Relatives

In a defined pedigree relationship, the probabilities of IBD are determined by the pedigree. As an example, we consider again the pair of cousins E and C in the pedigree of Fig. 6.1. At a single locus, the probability of IBD of their maternal gametes from any one of the four grandparental genes is $(1/2)^4 = 1/16$, since that copy of the DNA must be chosen in each of four independent meioses. Of course, E and C cannot share other than their maternal genomes IBD. Thus, in total, the probability that E and C share DNA IBD at a locus is $4 \times (1/16)$ or 1/4.

Suppose that, at a particular locus, DNA in one of the four genomes of the shared grandparent couple does descend both to E and to C, then E and C do share genome IBD at this locus. Moving along the chromosome, we assume, for simplicity, that recombination in any meiosis occurs at random at an average rate of 1 per 10^8bp. Since a recombination event in any one of the four connecting independent meioses will break this IBD, IBD is broken at a rate 4×10^{-8} per bp, and the expected distance until IBD is broken is 25×10^6 bp (25 Mbp), which is of order 25 centiMorgans (cM). That is, an IBD segment in first cousins has expected length 25 cM. Note however that this is different from the expected length of segment surrounding a known IBD point, since the segment will extend in both directions along the chromosome, for a total expected length of 50 cM. This apparent anomaly is the phenomenon of size-biased sampling. It is well known in statistics (Cox, 1962) but has caused some confusion in the recent IBD literature. It is important to distinguish between a randomly chosen IBD segment and the segment surrounding a randomly chosen point of IBD.

Consider now the brothers C and D. At a locus they receive their maternal DNA IBD with probability 1/2 and independently receive their paternal DNA IBD with probability 1/2. Thus they share both homologues IBD with probability 1/4, neither with probability 1/4. With the remaining probability 1/2, they share one of their two homologues IBD. Since recombination in either of two meioses breaks a maternal IBD segment, the expected length of such a segment is 50 cM and likewise of a paternal IBD segment. The four meioses to the sibs from their parents are independent, and every recombination in any one of the four parental meioses results in a state change. Thus, along the chromosome, the IBD state remains constant for an average of 25 cM. Switches in state occur from 1 to 0 or 2 IBD or from 0 or 2 to 1 IBD (Fig. 6.3).

Fig. 6.3 The states and state changes in a pair of full sibs

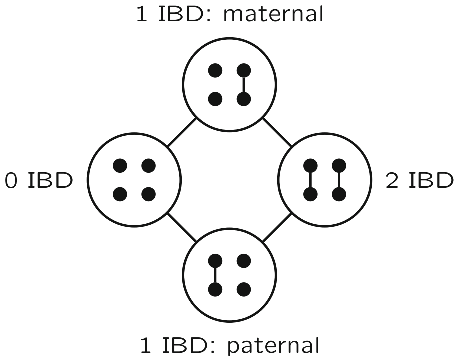


Table 6.1 Probabilities of IBD states among E, C, and D at any point in the genome, given the pedigree relationship of E with her sibling cousins C and D. Here, $\equiv$ denotes IBD among the specified gametes, and $\not\equiv$ denotes non-IBD

	$C_p \equiv D_p$	$C_p \not\equiv D_p$	Total
$E_m \equiv C_m \equiv D_m$	1/16	1/16	1/8
$E_m \equiv C_m \not\equiv D_m$	1/16	1/16	1/8
$E_m \equiv D_m \not\equiv C_m$	1/16	1/16	1/8
$E_m \not\equiv C_m \equiv D_m$	3/16	3/16	3/8
$E_m,\ C_m\ D_m$ all $\not\equiv$	1/8	1/8	1/4
Total	1/2	1/2	1

We now introduce the concept of more general states of IBD at a locus, using this same example. The probabilities of the ten possible states are shown in Table 6.1 and may be derived as follows. There is probability 1/2 that C and D share their paternal DNA IBD at a locus ($C_p \equiv D_p$) and probability 1/2 that they do not ($C_p \not\equiv D_p$). Also, this IBD is independent of any IBD among the maternal genomes of C, D, and E. Now, for E's maternal gamete E_m to be IBD to either of the maternal gametes C_m or D_m, the same one of the four grandparental genes that descends to E must also descend to the mother of C and D: probability $4 \times (1/8) = 1/2$. The probability this same DNA is copied to both D_m and C_m, to D_m but not C_m, and to C_m but not D_m is then each $(1/2) \times (1/2) = 1/4$, giving the first three rows of Table 6.1. Now also, there is total probability 1/2 of IBD between the maternal gametes C_m and D_m of C and D, so that

$$\begin{aligned}\Pr(E_m \not\equiv C_m \equiv D_m) &= \Pr(C_m \equiv D_m) - \Pr(E_m \equiv C_m \equiv D_m)\\ &= 1/2 - 1/8 = 3/8,\end{aligned}$$

and the fourth row of the table follows. The final row then follows from the known column totals. We see that even in this small example, the complexity of IBD patterns increases rapidly as more gametes are considered. Here there are just five relevant gametes, and a simple pedigree relationship, but there are already ten possible IBD combinations.

Considering changes among the ten states of Table 6.1 across the genome is also more complex, despite the independence of the two meioses that determine IBD between C_m and D_m and those that relate the mother of C and D to E. Additionally the rates of moving out of a given IBD state are no longer the same. C and D will share their maternal genomes for an average length of 50 cM, but IBD with E will be more rapidly broken, because of the greater number of intervening meioses. Also, it is no longer sufficient to consider only the rate of breaking an IBD chain. For example, E_m can switch directly from being IBD with C_m to being IBD with D_m.

Despite the rapidly increasing complexity of IBD states as more individuals are considered, the specification of inheritance in a defined pedigree is straightforward. Suppose all the meioses of a pedigree are indexed by m, $m = 1, \ldots, M$; M is twice the number of non-founders in the pedigree, since each non-founder has a maternal and a paternal meiosis giving rise to their maternal and paternal gametes. Suppose L locations of interest across a chromosome are indexed by loci j, $j = 1, \ldots, L$. We define, for each meiosis m and location j

$$\begin{aligned} S_{mj} &= 1 \quad \text{if the parent's paternal DNA is transmitted} \\ S_{mj} &= 0 \quad \text{if the parent's maternal DNA is transmitted} \end{aligned} \tag{6.1}$$

Then Mendel's first law states that meioses m are independent and that

$$\Pr(S_{mj} = 1) = \Pr(S_{mj} = 0) \;\; = \;\; 1/2. \tag{6.2}$$

Secondly, under the assumption of no genetic interference (Haldane, 1919), the crossover points in the gametes transmitted to offspring (Fig. 6.2) occur independently and at random at rate 0.01 per cM. Then the $\{S_{mj}; j = 1, \ldots, L\}$ have a Markov dependence over j. This can be expressed as

$$\Pr(S_{mj} = 1 \mid S_{m',j'}, (m', j') \neq (m.j)) = \Pr(S_{mj} = 1 \mid S_{m,(j-1)}, S_{m,(j+1)}) \tag{6.3}$$

That is, given all the other $S_{m'j'}$, S_{mj} depends only on the values $S_{m,(j-1)}$ and $S_{m,(j+1)}$ for the same meiosis m and the two neighboring loci. The vector of components S_{mj} over the values of m for any given locus j is known as the *inheritance vector* at locus j (Lander and Green, 1987).

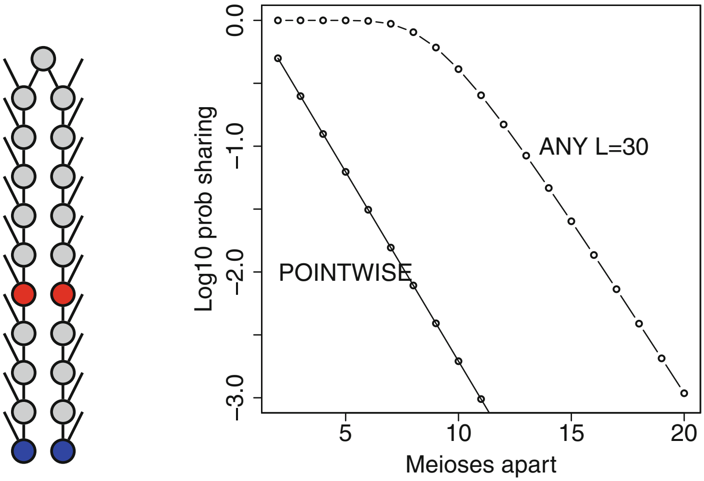
Equations (6.2) and (6.3) provide easy methods for simulation of the descent of genome in a defined pedigree. At independently inherited loci, it is simply an application of Mendel's first law, with the parents maternal or paternal DNA each being transmitted independently in every meiosis. Across the genome, the copying switches between copying from the parent's maternal and paternal DNA at a rate determined by the genetic map. Under the model of no genetic interference (Haldane, 1919), the distance to the next switch point on each chromosome can be generated as an exponential random variable with mean 100 cM. The values of S_{mj} are then determined at the specified discrete locations j. For each j, or indeed jointly over j, the S_{mj} determine which founder genome descends to each haploid genome of each current individual and hence the IBD state among current individuals. Thus Monte Carlo estimates of the probabilities of IBD patterns, both at a locus and across loci, can be very efficiently obtained.

6.2.2 IBD in Populations

In this section we consider IBD at the population level, when no pedigree relationship is specified. However, to motivate the discussion, we consider first the case of two individuals who have a single common ancestor, such that they are separated by a total of m meioses. For example, half-kth-cousins have a single common ancestor $(k + 1)$ generations ago and are separated by $m = 2(k + 1)$ meioses.

The probability of sharing genome IBD decreases by a factor of (1/2) with each additional meiosis. The formula for sharing any of an autosomal genome length L Morgans is more complicated (Donnelly, 1983) but also decays exponentially for larger numbers of meioses. Figure 6.4 shows these probabilities on a log scale, as a function of the number of separating meioses m. At a point, the probability of IBD decays rapidly, from 0.1 at $m \approx 4$ meioses, to 0.01 at $m \approx 8$, to 0.001 at $m \approx 11$. For a genome of length $L = 30$ Morgans, the probability of some IBD remains high for $m \leq 8$ but then starts to decrease to 0.148 at $m = 12$ and to 0.001 at $m = 20$. While 15% of pairs separated by 12 meioses will share some genome IBD from their common ancestor, this reduces to 1 in 1000 pairs for a separation of 20 meioses.

A very different picture results from considering the lengths of an IBD segment, given that it exists. IBD resulting from a chain of m meioses is broken by recombination at rate proportional to m so that length of IBD segments are of order m^{-1} Morgans. Even for $m = 20$, given there is a segment of IBD, this segment is expected to be several Mbp long.



	$m = 12$	$m = 20$
Probability of IBD at point	0.0005	2×10^{-6}
Probability of any IBD ($L = 30$)	0.148	0.001
Expected length of IBD segment	8.5 Mbp	5 Mbp

Fig. 6.4 IBD in remote half-cousins

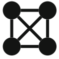
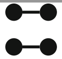
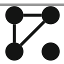
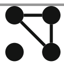
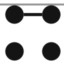
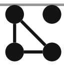
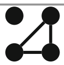
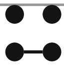
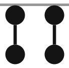
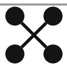
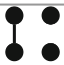
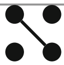
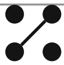
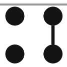
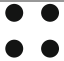
Even for a pair of individuals, the IBD of 0, 1, or 2 gametes at a location (Fig. 6.3) are not the only possibilities. If the parents of an individual are related, then the individual is inbred, and the two gametes within an individual may be IBD at some locations. For four gametes, there are 15 possible IBD states; the left-hand columns of Table 6.2 show the states in pictogram form. An alternative way to specify the IBD is by specifying the partition of the four gametes into the IBD subsets, and this specification is also given in Table 6.2. For example, in state 7 of Table 6.2 the maternal gamete A_m is IBD to both gametes B_p and B_m of B. The two subsets of this IBD partition are thus $\{A_p\}$ alone and the three IBD gametes $\{A_m, B_p, B_m\}$. One advantage of specification as a partition is that it extends to any number of individuals or set of haploid genomes.

For model-based inference of IBD from genetic data, a prior model for IBD is required; see Sect. 6.3.4. Given a pedigree, the probabilities of Mendelian segregation and the process of meiosis provide probabilities of each IBD partition or *state*. However, in the absence of a known pedigree, an alternative approach is necessary. One natural model is that of the *Ewens Sampling Formula* (ESF: Ewens 1972), which provides a one-parameter model for partitions of an exchangeable set of gametes. The ESF probabilities are in general written in terms of a_i, the number of subsets of size i; this specification is given in the next column of Table 6.2. For example, in state 7, with partition $\{\{A_p\}, \{A_m, B_p, B_m\}\}$, there is one subset of size 1 and one of size 3: $a_1 = a_3 = 1$. Since, under the ESF model, the gametes are exchangeable, all IBD partitions with the same set of values of a_i must have the same probability. For example, in Table 6.2, states 2, 9, and 10 have the same probability since each has $a_2 = 2$, even though in state 2 the IBD is between the two gametes within each individual and in states 9 and 10 there are two pairs of inter-individual IBD.

For our purposes the probabilities are most easily parameterized in terms of β, which is the pairwise probability of IBD between any two gametes. In terms of Ewens' classical parameter θ of genetic variation, $\beta = 1/(1 + \theta)$. For the case of four gametes the relative probability of each state is also given in Table 6.2. Each term is normalized by the column sum $(1+\beta)(1+2\beta)$ to give the probability. Again, in the example of state 7, there is one non-IBD factor $(1-\beta)$ and two IBD factors β to link the other three gametes. The general formula for multiple gametes is beyond the scope of this chapter, but the interested reader may consult Tavaré and Ewens (1997). It can also be checked that every pair of gametes has IBD probability β. For example, $A_m \equiv B_p$ in states 1, 3, 7, 10, and 13 for a total probability

$$\frac{6\beta^3 + 2 \times 2\beta(1-\beta) + \beta^2(1-\beta) + \beta(1-\beta)^2}{(1+\beta)(1+2\beta)}$$
$$= \frac{\beta(6\beta^2 + 5\beta(1-\beta) + (1-\beta)^2)}{(1+\beta)(1+2\beta)} = \beta \qquad (6.4)$$

Table 6.2 The 15 IBD partitions at a locus, among the four gametes of two individuals. For two individuals A and B, the paternal (p) and maternal m gametes are denoted and depicted as in Fig. 6.3 with individual A above and B below

State		Partition	Ewens' $\{a_i\}$	Probability	Kinship
1		$\{A_p, A_m, B_p, B_m\}$	$a_4 = 1$	$6\beta^3$	1
2		$\{A_p, A_m\}, \{B_p, B_m\}$	$a_2 = 2$	$\beta^2(1-\beta)$	0
3		$\{A_p, A_m, B_p\}, \{B_m\}$	$a_1 = a_3 = 1$	$2\beta^2(1-\beta)$	1/2
4		$\{A_p, A_m, B_m\}, \{B_p\}$	$a_1 = a_3 = 1$	$2\beta^2(1-\beta)$	1/2
5		$\{A_p, A_m\}, \{B_p\}, \{B_m\}$	$a_2 = 1,\ a_1 = 2$	$\beta(1-\beta)^2$	0
6		$\{A_p, B_p, B_m\}, \{A_m\}$	$a_1 = a_3 = 1$	$2\beta^2(1-\beta)$	1/2
7		$\{A_p\}, \{A_m, B_p, B_m\}$	$a_1 = a_3 = 1$	$2\beta^2(1-\beta)$	1/2
8		$\{A_p\}, \{A_m\}, \{B_p, B_m\}$	$a_2 = 1,\ a_1 = 2$	$\beta(1-\beta)^2$	0
9		$\{A_p, B_p\}, \{A_m, B_m\}$	$a_2 = 2$	$\beta^2(1-\beta)$	1/2
10		$\{A_p, B_m\}, \{A_m, B_p\}$	$a_2 = 2$	$\beta^2(1-\beta)$	1/2
11		$\{A_p, B_p\}, \{A_m\}, \{B_m\}$	$a_2 = 1,\ a_1 = 2$	$\beta(1-\beta)^2$	1/4
12		$\{A_p, B_m\}, \{A_m\}, \{B_p\}$	$a_2 = 1,\ a_1 = 2$	$\beta(1-\beta)^2$	1/4
13		$\{A_p\}, \{A_m, B_p\}, \{B_m\}$	$a_2 = 1,\ a_1 = 2$	$\beta(1-\beta)^2$	1/4
14		$\{A_p\}, \{A_m, B_m\}, \{B_p\}$	$a_2 = 1,\ a_1 = 2$	$\beta(1-\beta)^2$	1/4
15		$\{A_p\}, \{A_m\}, \{B_p\}, \{B_m\}$	$a_1 = 4$	$(1-\beta)^3$	0

The classical *coefficient of kinship* between two individuals A and B is the probability that gametes segregating from each of A and B are IBD at any point in the genome. The final column of Table 6.2 gives this probability conditional on the IBD state among the four gametes of A and B. Note there are only four possibilities. If there is no IBD between the individuals (states 2, 5, 8, and 15), the value is 0.

For one between-individual IBD link (states 11, 12, 13, and 14), the value is 1/4. If all four gametes are IBD, the value is 1, and the remaining 6 states each gives value 1/2. From Equation (6.4) and its analogues for other gamete pairs, it is seen that, under the ESF model, β is both the kinship between A and B and the inbreeding coefficient of each individual (the probability of IBD between the individual's two gametes).

The IBD state among gametes changes along a chromosome due to ancestral recombination events. For a pair of gametes, Leutenegger et al. (2003) proposed a simple model in which potential changes occur at rate α and at a potential change point the new (possibly unchanged) state is IBD or non-IBD with probability β and $(1 - \beta)$, respectively. This gives rise to an equilibrium pairwise IBD probability β and to alternating segments of IBD and non-IBD. The lengths of these segments are exponentially distributed with expectations $1/\alpha(1-\beta)$ and $1/\alpha\beta$, respectively. This model is based on the the consideration of a single chain of ancestry (Fig. 6.4) and does not reflect a more complex situation where there are multiple paths of ancestry of varying numbers of meioses between the two gametes. However, it is flexible enough to provide a useful prior distribution for IBD.

Modeling the changes in IBD among multiple gametes along a chromosome is a challenging problem. The full *ancestral recombination graph* is too complex a model for genome-wide use. Simple approximations cannot accommodate the range of changes that can occur or fail to mimic the types of changes that do occur. For example, an extension of the model of Leutenegger et al. (2003), which samples from the ESF at each potential change point, would allow immediate changes from state 1 to state 15, in Table 6.2, or from state 9 to state 10, whereas no single ancestral recombination event could accomplish these changes.

One model that has proved useful is that proposed by Brown et al. (2012) which applies to any number of gametes and has the ESF as its equilibrium distribution. This model allows for the move of any one gamete into, out of, or between any two IBD subsets at each potential transition point. Potential transitions occurs at rate α, which is a surrogate for recombination rate. The two parameters β and α together control the overall level of IBD and the lengths of chromosome over which a subset of gametes will remain IBD. This model also does not accommodate all possible transitions. For example, an ancestral recombination that is ancestral to two current gametes may move them together into another IBD subset. However, provided other changes are allowed for with some small probability, this model also provides a useful prior (Zheng et al., 2014).

6.2.3 Probabilities of Genotypic Data Given IBD

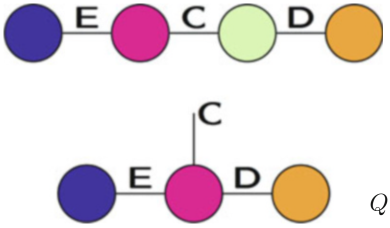
We now consider the relationship between latent IBD and the probabilities of marker genotypes. At a specific locus, in a specific gamete, the probability of the allelic type of the DNA is simply the population allele frequency. We will label SNP alleles as u and v, with population frequencies q and $(1 - q)$. At any locus the observed genotype $G(A)$ of any individual A is thus uu, uv, or vv. More generally, for a marker with k alleles ($k > 2$), we will denote the alleles by v_i and the population frequencies by q_i ($\sum_{i=1}^{k} q_i = 1$). It is assumed that appropriate allele frequencies are known from population databases or other sources.

The basic premise is that IBD DNA is of the same allelic type and that non-IBD DNA copies are of independent allelic type. Mutations may cause IBD DNA to differ in allelic type, but the probability of mutation is generally much less than of typing error. We consider typing error in Sect. 6.2.4. The independence of non-IBD and the appropriate population allele frequencies are harder issues, since both depend on the frame of reference. If IBD is measured relative to a particular time-point or ancestral population, then it is the allele frequencies in that population that govern the probability that a set of IBD gametes has each the given allelic type. However, these allele frequencies are generally unknown, and instead current population estimates are used. IBD is more reliably inferred using only common genetic variants, for which population allele frequencies are more easily estimated and more stable over time.

As a simple example, we consider again the small family of Fig. 6.1 and assume the descent of IBD that is shown in that figure. That is, at the locus of interest, cousins E and C share their maternal gametes IBD from their grandfather, and sibs C and D share their paternal DNA IBD. This is shown graphically in the upper left graph of Fig. 6.5. In this *IBD graph*, the observed individuals E, C, and D are depicted as edges, and the notes denote the DNA. Where two or edges impinge on a node, those individuals shared that DNA IBD. Each edge joins the two DNA nodes that represent the two gene copies carried by that diploid individual.

Consider the probability that, at the marker locus of interest, E and C each has the homozygous genotype uu, while D is a heterozygote, uv. It is immediately clear that the three nodes (blue, magenta, and green) in E and C are of type u, while the remaining (orange) node of D is of type v (see right part of Fig. 6.5). The probability that any node is type u is q, the population frequency of allele u, and likewise $(1-q)$ for v. In the total of four nodes, there are three of type u and one of type v for a total probability of $q^3(1-q)$.

Now suppose it is inferred that, owing to some previously unspecified relationship between the grandfather and the father of the sibs C and D, the DNA shown as magenta and as green are in fact IBD. The IBD graph is simply modified by



●	●	●	●	Q
u	u	u	v	$q^3(1-q)$
u	u	$-$	v	$q^2(1-q)$

$$Q = \Pr(G(E) = uu,\ G(C) = uu,\ G(D) = uv)$$

Fig. 6.5 Example of probabilities of marker genotypes given IBD

merging these two nodes (Fig. 6.5, lower left). Individual C now carries two copies of the sane (magenta) DNA node, which is shared also with E and D. There are three nodes in total, two of type u and one of type v, for a total probability $Q = q^2(1-q)$.

While these examples are very simple, they make important points. First, once the IBD graph and allele frequencies are given, the joint genotype probabilities are easily determined. The pedigree structure or population history that gave rise to the IBD graph is no longer relevant once the IBD graph is known. Second, at any given marker, only observed individuals are included in the IBD graph. Unlike classical pedigree computations, there is no need to sum out over possible genotypes of unobserved individuals. Third, the assignment of allelic types to the DNA nodes is trivial. If any individual is homozygous, then the only possible assignment (assuming there is one) is immediately determined by extending from that initial constraint. If all individuals are heterozygous, there may be two alternate assignments whose probabilities must be summed. For example, a single heterozygous uv genotype not sharing DNA nodes with other individuals has genotype probability $q(1-q) + (1-q)q = 2q(1-q)$, since either allele may be assigned to either node. Regardless of the number of possible alleles at a locus and regardless of the complexity of the IBD graph, there are always 0, 1, or 2 possible allelic assignments for each connected component of the graph. Of course, on separate components of the total graph we simply take the product of the probabilities on each component.

6.2.4 Probabilities of Phenotypic Data Given IBD

In this section, we extend the above ideas to phenotypic data. Where a phenotype allows for several underlying genotypes, it is necessary to sum over the possible assignments of allelic types to DNA nodes. This computation may be accomplished sequentially across the graph, using methods that are standard in the area of graphical models (Lauritzen, 1992). Suppose, for example, that individuals E, C, and D have (discrete or continuous) phenotypes Y_E, Y_C, and Y_D and that a genetic model for the trait provides the probabilities of each individual's phenotype given the unobserved allelic types of his/her DNA at the trait locus. Additionally, these penetrance probabilities may depend on other observed covariates such as the age and gender of each individual. Them we may compute the probability of the observed phenotypes as

$$\Pr(Y_E, Y_C, Y_D) = \sum_{\textcolor{blue}{\bullet}} \sum_{\textcolor{magenta}{\bullet}} (\Pr(Y_E | \textcolor{blue}{\bullet}, \textcolor{magenta}{\bullet}) q(\textcolor{blue}{\bullet}) q(\textcolor{magenta}{\bullet}) \\ \sum_{\textcolor{green}{\bullet}} (\Pr(Y_C | \textcolor{magenta}{\bullet}, \textcolor{green}{\bullet}) q(\textcolor{green}{\bullet}) \sum_{\textcolor{orange}{\bullet}} (\Pr(Y_D | \textcolor{green}{\bullet}, \textcolor{orange}{\bullet}) q(\textcolor{orange}{\bullet})))) \tag{6.5}$$

where $q(\bullet)$ denotes the population allele frequency of the allelic type of DNA node $\bullet$. That is, proceeding from right to left, we may first use the information on individuals D and sum over the possible allelic types of node $\textcolor{orange}{\bullet}$ for each value of the type of node 🟢. Then we can incorporate the data on individual C and for each value of 🔴 sum over the possible values of 🟢. Finally we may incorporate the data on E and sum over the possible allelic types of the two remaining DNA nodes. By processing the summation in this way, we can break the overall sum into smaller feasible computations.

Although our example is of a small pedigree, this is for ease of presentation only. Even for more much larger and more complex IBD graphs, these computations are generally feasible. Where IBD graph components are small, probabilities can be computed even under models for which the trait is controlled by genotypes at several loci. Moreover, once the IBD graph is given, the source of that IBD information is irrelevant. The IBD graph contains all the relevant information on the impact of shared ancestry on joint phenotype probabilities.

A special case of phenotypic data arises with marker data where an allowance is made for typing error. That is, the true marker genotype provides probabilities for the marker phenotype; the observed marker "genotype" may not be the true one. In practice, it is important to use a model that allows for the possibility of error, so that IBD nodes are not of necessity of the same allelic type. It is not necessary to have a model that precisely reflects biological or technological genotyping processes. One simple error model for single genotypes is due to Leutenegger et al. (2003). In the case of IBD, with probability $(1 - \varepsilon)$ the alleles are the same, and of type v_i with probability q_i, but with probability ε the Hardy-Weinberg frequencies are used (Table 6.3). This allows for heterozygous genotypes even in segments where the individual's two gametes are IBD, whether due to typing error, mutation, or other causes.

For larger numbers of individuals, one error model that makes probability computations on an IBD graph straightforward is a generalization of the model of Leutenegger et al. (2003). With an error parameter ε, the probability of genotypes $\mathbf{g}$ is modeled as

$$\Pr(\mathbf{g}|\varepsilon, \text{IBD graph}) = (1 - \varepsilon)\Pr(\mathbf{g}|\varepsilon = 0, \text{IBD graph}) + \varepsilon\Pr(\mathbf{g}| \text{ no IBD}) \quad (6.6)$$

That is, on any connected component, with probability $(1 - \varepsilon)$ there is no error, while with probability ε there is some error, and then the probability is computed as if there is no IBD among the individuals represented in that IBD graph component.

For small numbers of individuals, or low levels of IBD, the model of Equation (6.6) works well, but in some cases more complex models are needed. For

Table 6.3 The probabilities of an individual's genotype at a k-allele locus. Any two distinct possible alleles at the locus are denoted v_i and v_{i*} $(1 \le i < i^* \le k)$. The population frequencies of alleles v_i and v_{i*} are q_i and q_{i^*}, respectively

	Non-IBD	IBD
$v_i\ v_i$	q_i^2	$(1 - \varepsilon)q_i + \varepsilon q_i^2$
$v_i\ v_{i^*} (i < i^*)$	$2q_i q_{i^*}$	$\varepsilon 2 q_i q_{i^*}$

SNP genotypes, another model is that each allele is independently toggled to its alternative with probability ε (Zheng et al., 2014). Since some loci are more error-prone, ε may be made locus-dependent. With this or a more general error model, genetic marker become in effect discrete trait phenotypes. Then probability computations require the general summation method exemplified in Equation (6.5).

6.3 Inferring in Pedigrees and Populations

6.3.1 IBD Given Marker Data on Relatives

In this section we will consider the inference of IBD from genetic marker data. We consider first the case where the pedigree structure of the observed individuals is known and assumed correct. Marker genotypes for some individuals for some subsets of loci may be missing, but we assume that, if observed, the marker genotypes are without error. As an example of the principles involved, we consider the example of a pair of full sibs. At any locus, sibs share their maternal/paternal genome IBD with probability 1/2. In the absence of genetic marker data, there are prior probabilities 1/4, 1/2, and 1/4 (respectively) that they share 0, 1, or 2 gene copies IBD (Sect. 6.2.1).

More generally, we will denote genetic marker (usually SNP) data by $\mathbf{X}$ and a specification of the IBD to be inferred by $\mathbf{Z}$. Section 6.2.3 showed how probabilities $\Pr(\mathbf{X} \mid \mathbf{Z})$ of genetic marker data on relatives could be easily computed given the a pattern of IBD at the marker locus. Conversely, given a prior probability $\Pr(\mathbf{Z})$ of IBD and genetic marker data, conditional probabilities of IBD can be obtained. By Bayes theorem:

$$\Pr(\mathbf{Z} \mid \mathbf{X}) \quad \propto \quad \Pr(\mathbf{X} \mid \mathbf{Z})\,\Pr(\mathbf{Z}). \tag{6.7}$$

For pairs of individuals in a known pedigree relationship, there are well-established methods for computing these prior probabilities (Karigl, 1981).

Suppose the SNP genotypes of the pair of sibs at three linked loci are as shown in Table 6.4. As in Sect. 6.2.3 we will denote the SNP alleles as u and v, while q_j and $(1 - q_j)$ will now denote the frequency of u and of v at locus j. The reader should not struggle with details of the computation but consider only whether the results make qualitative sense. Given the allele frequencies shown, then the probabilities that the sibs share 0, 1, or 2 DNA copies IBD are computed using Equation (6.7) and are given in Table 6.4. Note the effect of the u allele frequency, q_j. The two sibs have the same genotypes at locus-1 and locus-3, but at locus-3 the u allele is the rare allele, giving much stronger weight to IBD sharing between the two sibs. Only at locus-2 do the two sibs have the same genotype and so can share 2 copies IBD ($Z = 2$). The genotypic data raises the probability that they do so to 0.4, which is higher than above the prior probability 0.25.

However, if the loci are linked, this computation does not take into account all the information available; there is dependence in the IBD state Z at linked loci. Suppose

Table 6.4 Single-locus probabilities of IBD in a sib pair for the example data

			$Z = 0$	$Z = 1$	$Z = 2$	$Z = 0$	$Z = 1$	Z=2
		Pr(Z)				0.25	0.5	0.25
loc-j	q_j	Data X_j	Pr($X_j \mid Z_j$)			Pr($Z_j \mid X_j$)		
loc-1	0.9	$uu,\ uv$	0.146	0.081	0.000	0.474	0.526	0.000
loc-2	0.5	$uv,\ uv$	0.250	0.250	0.500	0.200	0.400	0.400
loc-3	0.1	$uu,\ uv$	0.002	0.009	0.000	0.091	0.910	0.000

Table 6.5 Transition probabilities in sib IBD states at two loci at recombination probability $\rho =$ 0.05

$Z =$	0	1	2	0	1	2
0	c^2	$2c(1-c)$	$(1-c)^2$	0.819025	0.17195	0.009025
1	$c(1-c)$	$1-2c(1-c)$	$c(1-c)$	0.085975	0.82805	0.085975
2	$(1-c)^2$	$2c(1-c)$	c^2	0.009025	0.17195	0.819025

the recombination fraction between adjacent loci 1 and 2, and 2 and 3, are each $\rho = 0.05$. Then there is no change in the maternal [paternal] DNA sharing between adjacent loci if either both or neither of the meioses from the mother [father] to the two sibs is recombinant. The probability of this is $c = (1-\rho)^2 + \rho^2 = 0.905$. This leads to the probabilities of transition between the IBD states $Z = 0, 1, 2$ from one locus to another at recombination distance ρ. The matrix of these transition probabilities is shown in Table 6.5. It can be checked that, for any value of c, this matrix has the required equilibrium single-locus probabilities 1/4, 1/2, 1/4 for $Z = 0,\ 1,\ 2$.

We can now compute the probability of the IBD state at locus-2, given the data at all three loci. Let Z_j denote the IBD state at locus j and X_j the SNP genotypes at locus j. Note that the data at a locus depend only on the IBD state at that locus. Thus, for example, X_1, X_2, and X_3 are conditionally independent given the IBD state Z_2 at the middle locus. Then

$$\begin{aligned}\Pr(X_1, X_2, Z_2) &= \sum_{Z_1} \Big(\Pr(X_1|Z_1)\Pr(Z_2|Z_1)\Pr(Z_1)\Big)\Pr(X_2|Z_2)\\ &= 0.0083,\ \ 0.0099,\ \ 0.0019 \text{ for } Z_2 = 0, 1, 2.\\ \Pr(X_3|Z_2) &= \sum_{Z_3} \Big(\Pr(X_3|Z_3)\Pr(Z_3|Z_2)\Big)\\ &= 0.0030,\ \ 0.0076,\ \ 0.0016 \text{ for } Z_2 = 0, 1, 2.\end{aligned}$$

Combining these we have

$$\Pr(X_1, X_2, X_3, Z_2) \propto (2.490,\ 7.524,\ 0.304) \times 10^{-5} \text{ for } Z_2 = 0, 1, 2.$$

and normalizing these gives the probability of Z_2 given X_1, X_2, X_3 as approximately $(0.24, 0.73, 0.03)$ for $Z_2 = 0, 1, 2$. Note that whereas the data at locus-2 increased the relative probability that $Z_2 = 2$, incorporating the data at loci 1 and 3 greatly decreases the probability since the state of 2 IBD is impossible at each of these flanking loci. For $Z_2 = 2$, a recombination would be required both between locus-1 and locus-2 and between locus-2 and locus-3.

6.3.2 Monte Carlo Realization of IBD in Defined Pedigrees

The principles that underlie the computations of Sect. 6.3.1 are that the data at each marker locus depends only on the latent IBD state at that locus and that the transitions in latent IBD state follow a Markov process across the chromosome. This is the classic framework for a *hidden Markov model* (HMM) which enables many computations to be made. In fact, IBD is in general not Markovian, since many different patterns of inheritance may give rise to the same IBD state among observed individuals. However, in the absence of genetic interference, the indicators of maternal or paternal transmission in a meiosis are indeed Markov (see Equation (6.1)). This Markov nature of *inheritance vectors* (Sect. 6.2.1) has been used by many to enable computations on pedigrees. For example, as implemented in the MERLIN software (Abecasis et al., 2002), probabilities of IBD pairwise among the members of a small pedigree may be computed at each marker locus, conditional on the joint marker data on all pedigree members and on marker genotypes at all loci. More generally, as seen in Sect. 6.2.1, the inheritance vector at a locus determines the IBD state at that locus, among all members of the pedigree.

While location-specific probabilities of IBD are useful, there are good reasons to prefer Monte Carlo realizations of latent inheritance vectors or IBD. First, it is not feasible to consider probabilities of IBD jointly across multiple loci, but a set of realizations directly display the segmental nature of inheritance and can identify locations of recombination events that change the IBD among observed individuals. Second, the variation in a set of realizations provides a measure of uncertainty in the IBD information which cannot be captured in a single probability. In using IBD inferred from marker data in subsequent genetic analyses, it is important to have some measure of this uncertainty.

The same HMM methods that allow computation of IBD probabilities also allow Monte Carlo realization of IBD conditional on genetic marker data (Thompson, 2000). On small pedigrees, where exact computation is feasible, independent realizations may be generated. On larger pedigrees, where the space of inheritance vectors at each locus is too large for exact computation, Markov chain Monte Carlo (MCMC) methods must instead be used (Sobel and Lange, 1996; Thompson, 2000). However, the same principles apply. The Markov dependence of inheritance vectors and the dependence of marker genotypes only on the inheritance at that marker locus enable realizations of inheritance jointly across loci and among individuals to be made efficiently. The only requirement is that, at each marker location j, $\Pr(\mathbf{X}_j \mid \mathbf{Z}_j)$ can be easily computed (see Sect. 6.2.3). However, this requirement does generally impose the restriction that marker genotypes are assumed to be observed without error.

If multiple realizations of inheritance across a chromosome are to be realized and used in subsequent genetic analyses, it is necessary to store them compactly. Note that in any meiosis, crossovers (switches between transmission of maternal and paternal DNA) occur on average only every 10^8 bp. Thus, rather than storing inheritance vectors at each location, it is more efficient to store only the initial value and the bp locations of successive switches. The inheritance vector at any location may then be efficiently determined and consequently the IBD graph among individuals observed for a trait of interest. Only the IBD graph is relevant to subsequent analyses.

6.3.3 Inference of Realized Kinship or Relatedness

A pedigree provides a very strong prior on probabilities of IBD at a locus (Sect. 6.2.1), but as genetic marker data become more and more informative, this prior is increasingly unnecessary. Moreover, for more remote relatives, IBD is highly variable. In the example of Fig. 6.4, only 1 in 1000 pairs of individuals separated by 20 meioses will share any autosomal IBD, but if they do they will share (on average) 5 Mbp. Other examples are considered by Donnelly (1983), while Hill and Weir (2011) give a more extensive review of the variation in realized proportions of genome shared given different patterns and degrees of pedigree relatedness.

Therefore, with the current availability of dense SNP marker data, there has been an explosion of interest in the recent literature in the estimation of realized kinship from genotypic data. More often this is phrased in terms of *realized relatedness*, or of the proportion of genome shared IBD by pairs of individuals, but this is simply twice the realized kinship, which is, in turn, a function of the realized 4-gamete IBD states across the genome (Table 6.2).

The most widely used measure of realized relatedness based on genotypes is the *genetic relatedness matrix* of GRM; see, e.g., Hayes et al. (2009). The GRM is estimated as follows. As previously, at any SNP locus j, we have alleles u and v with frequencies q_j and $(1 - q_j)$. The genotype x_{ij} of an individual i can be specified by the number of u alleles he carries: $x_{ij} = 2, 1, 0$ for genotypes uu, uv, and vv, respectively. Under a model of sampling alleles from the current population, x_{ij} has expectation $2q_j$ and variance $2q_j(1 - q_j)$. For two individuals i and k, the (i, k) entry of the GRM is the empirical correlation between the allele counts x of i and k:

$$A_{ik} = \frac{1}{L}\sum_{j=1}^{L} \frac{(x_{ij} - 2q_j)(x_{kj} - 2q_j)}{2q_j(1 - q_j)} \tag{6.8}$$

where L is the total number of loci genotyped. A more general form is

$$A_{ik} = \sum_{j=1}^{L} w_j \, \frac{(x_{ij} - 2q_j)(x_{kj} - 2q_j)}{2q_j(1 - q_j)} \tag{6.9}$$

With $w_j = 1/L$ we obtain the previous estimator, while $w_j = p_j(1 - p_j)/\sum_{l=1}^{L} p_l(1 - p_l)$ provides a form that is more robust to small allele frequencies; see, for example, VanRaden (2008).

One major deficiency of estimators where the weights depend only on allele frequencies is that they do not account for allelic associations among loci (LD). One form of weighting to accommodate LD was developed by Speed et al. (2012), while S. Sverdlov developed an alternative approach that is used in Wang et al. (2017). The top two panels of Fig. 6.6 show the increased precision of estimation of realized kinship (relatedness/2) by accounting for LD. The left-hand panel is for the GRM estimate of realized kinship, $A_{ik}/2$ where A_{ik} is as in Equation (6.8). The right-hand panel shows the results for Equation (6.9) with weights w_j computed according to the LD-weighting developed in Wang et al. (2017). Shown are simulation results for 1000 pairs of second cousins, and the histograms are of the difference between the estimated realized kinship and the actual realized kinship in each pair. Note

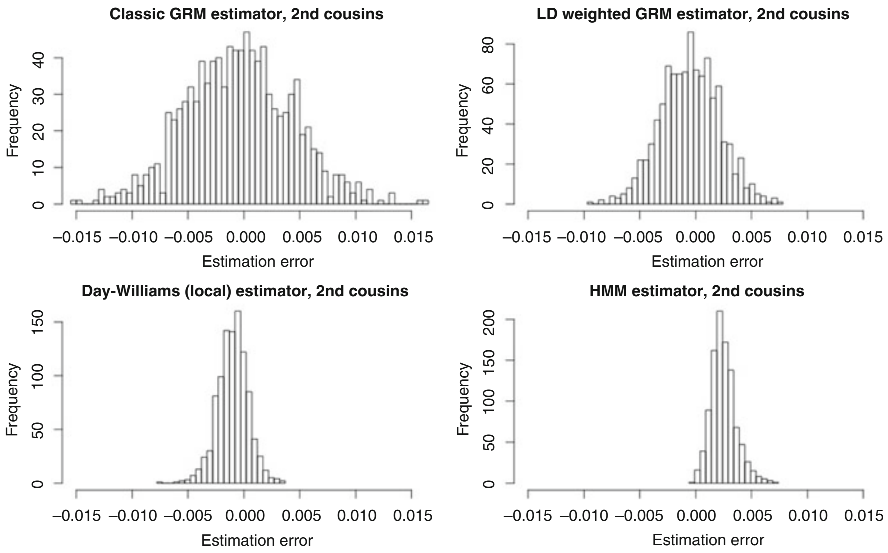


Fig. 6.6 Histograms of estimation errors of 1 estimators on 1000 simulated second cousin pairs. The values are the difference between the estimated global realized kinship and the actual (simulated) global realized kinship, computed at 169,751 SNP marker positions. The estimators are the classic GRM (6.8), an LD-weighted version of the form (6.9), and estimates based on the local DW approach (DW), and on the HMM method (Sect. 6.3.4). (The figure is due to Bowen Wang, based on the study by Wang et al. 2017)

that we are not here attempting to estimate the pedigree kinship. The histogram of differences between the pedigree value and actual realized values has a larger spread than even the upper left histogram based on Equation (6.8).

There is however a more serious deficiency in estimators of the form (6.9); this is that they do not take the physical locations of SNPs into account. We have already seen that IBD occurs even in remote relatives as a few long segments. Additionally, SNPs are individually very uninformative; information about IBD should be combined across local SNPs to provide more accurate estimates of the probability of IBD at each point in the genome. Such estimates can then be averaged across the genome to estimate the realized proportion of genome shared IBD. One such method was proposed by Day-Williams et al. (2011) and is denoted DW. They use the four between-individual comparisons of allelic sharing at loci in windows across the genome, to obtain estimates of local kinship. These are then smoothed across the genome, subject to constraints that at each point the value is 0, 1/4, 1/2, or 1 (Table 6.2). An alternative is to estimate the IBD state for the four gametes using an HMM approach and the model of Brown et al. (2012) for changes in the 15 states across the genome (Sect. 6.2.2): details are given in Sect. 6.3.4 below. This method provides estimates of realized kinship at points across the genome, and these may then be combined into a genome-wide estimate.

The two lower histograms of Fig. 6.6 show the results using the local IBD estimation methods of Day-Williams et al. (2011), denoted DW, and of Brown et al. (2012) denoted HMM. It is seen that incorporating the segmental nature of DNA into the inference process greatly improves the precision of estimation of realized kinship. However, these methods are more computationally intensive and also show bias. The DW method tends to underestimate IBD, especially in the presence of inbreeding. The HMM method tends to overestimate IBD in the presence of LD. For this reason, an LD-weighted version of the HMM estimator provides further improvement. These and other estimators are further discussed by Wang et al. (2017).

6.3.4 IBD Given Marker Data in Populations

In the previous section, the focus was on estimating the genome-wide proportions shared IBD between two individuals. However, for gene mapping, we may be interested in the joint pattern of IBD among several observed individuals, not only pairwise measures. Second, for mapping we are interested in IBD at specific locations across the genome. Third, we may wish to consider segments of IBD and changes in IBD across genome locations. Estimates of relatedness such as Equation (6.9) do not take into account the physical linkage among loci, treating them as an exchangeable collection of SNPs; any permutation of the SNPs will provide the same result. By contrast, estimates of location-specific IBD rely on the genetic marker map and depend jointly on the SNPs in the genome region. Each SNP alone provides little evidence, but segments of IBD typically encompass many SNPs.

A defined pedigree provides prior probabilities of IBD among individuals at a locus and across a chromosome (Sect. 6.2.1). However, population genetics also provides probabilities of coancestry and IBD for individuals sampled from a population. In the context of modern highly informative SNP data, the prior distribution has relatively less weight, and genetic marker data can provide strong evidence of segments of IBD among individuals not known to be related. Formerly, where marker data were sparse both in the genome and among individuals, the highly informative pedigree prior was a necessity for successful inference. With modern data, it is often unnecessarily constraining. Further, ancestral pedigrees may be inaccurate and cannot be validated from current genetic data. Except among the current generations of sampled individuals, where the genetic data may be used to validate the pedigree, the use of a pedigree prior is often best avoided.

Model-based inference of IBD requires allele frequencies, and the relevant allele frequencies are those in the reference population relative to which IBD is measured. Sharing of a rare variant allele among individuals provides strong evidence of IBD, but on average common allelic variation provides more information. Moreover, for a rare variant, it is difficult to either quantify the strength of the evidence or to assess its uncertainty; even the concept of a population allele frequency may be problematic. In contrast, common SNP variation is ancient and relatively stable. While each SNP alone provides little information, segments of IBD generally encompass large numbers of SNPs. Whether or not data are phased, and whether or not local haplotype frequencies are used in estimation, it is the combination of data from multiple contiguous SNPs that provides the evidence of IBD.

As described in Sect. 6.2.2, a flexible two-parameter prior model for IBD between the pair of gametes of an individual was introduced by Leutenegger et al. (2003). In this case there are just two possible IBD states at each locus; the two gametes are IBD ($Z = 1$) or they are not ($Z = 0$). The assumption of a Markov process of transitions between the two states again gives an HMM structure that allows inference of IBD segments. The model is shown schematically in Fig. 6.7. The latent IBD state consists of alternating segments of IBD ($Z = 1$) and non-IBD ($Z = 0$) between the two gametes. In an IBD segment, the allelic types at marker loci are, with high probability, of the same allelic type. In non-IBD segments they are of independent allelic types. More precisely, the data model was given in Table 6.3 in Sect. 6.2.4. At each marker locus, given non-IBD, we have Hardy-Weinberg genotype probabilities. In the case of IBD, a small "error" probability ε allows for the possibility that IBD DNA may be, or be recorded, as of

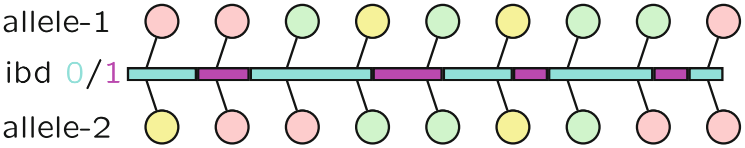


Fig. 6.7 Model for inferring IBD between two gametes

different allelic types. While the exact form of the error model is not important, it is important to allow this flexibility: generally, zero probabilities of latent states should be avoided in modeling data. Since the model has an HMM structure, standard algorithms give probabilities $\Pr(Z_j \mid \mathbf{Y})$, of the IBD state Z_j at each locus j, given the data $\mathbf{X}$ of the allele types on the gametes over all loci. Further, as in Sect. 6.3.2, we may instead obtain realizations of IBD $\{Z_j;\, j = 1, \ldots, \ell\}$ given $\mathbf{X}$, jointly over j.

There have been a number of extensions of the basic model, including to analyses of genotypes of pairs of individuals as implemented in the well-known PLINK software (Price et al., 2006). A more general prior model for IBD across the genome and among IBD among any number of gametes was given in Sect. 6.2.2. This model was used by Brown et al. (2012) on sets of four gametes, using either haplotypic or genotypic data and by Zheng et al. (2014) for multiple gametes but assuming haplotypic data. Moltke et al. (2011) also proposed a model for multiple gametes and used it to study IBD in a set of five individuals. All these approaches have the same basic framework and objective. That is, patterns of IBD across a chromosome among sets of gametes are to be inferred from genetic marker data. The IBD process is approximated by a Markov process, and the allelic or genotypic data at each locus depends only on the underlying IBD state, giving rise to an HMM.

There is one significant approximation in these models in that linkage disequilibrium (LD) is ignored. That is, there is no direct dependence of allelic types between loci. While it is the allelic similarity of gametes across multiple loci that results in inference of IBD, haplotypic similarities are not modeled directly. Allele frequencies are incorporated into the model, and for common SNP variation, these are normally adequately accurately known, but haplotype frequencies are often less well established. While ignoring LD is a model mis-specification that can result in false inferences of IBD (Fig. 6.6), over-compensation for LD can lead to failure to detect IBD (Brown et al., 2012). A model that does include LD in the inference of IBD is that due to Browning and Browning (2010), implemented in the BEAGLE package but at the expense of a simplified IBD model. This approach works very well in large samples from large populations, where IBD levels are low and haplotype frequencies can be well-estimated.

These methods also all face another issue: as the number of gametes n increases, the number of possible IBD states at each locus increases very rapidly, being the number of partitions of n items. For the 12 gametes of 6 individuals, there are more than 4 million possible states. The example considered by Moltke et al. (2011) was for just five individuals and a limited gene region. Zheng et al. (2014) considered 860 SNP markers over a region of 10 Mbp and succeeded in realizing joint IBD among 40 gametes but assumed the availability of haplotypic marker data. Neither of these approaches is scalable to chromosome-wide inferences of joint IBD among multiple gametes from genotypic marker data.

6.4 IBD-Based Genetic Mapping

Any genetic mapping procedure aims to detect the genomic locations of DNA variation underlying a trait of interest, by reference to a genetic map of markers that have known locations in the genome. An association test directly considers the dependence between marker genotypes (**X**) and trait phenotypes (**Y**). However, these allelic associations arise from the descent of DNA from common ancestors to different individuals within a population or to different populations. It is therefore useful to consider the associations between **X** and **Y** through the lens of descent **Z** and specifically patterns of IBD among individuals observed for the trait inferred at locations across the genome. Throughout this section we assume that **Z** contains all the information needed for analysis of association between **X** and **Y**: that is, we assume that **X** and **Y** are conditionally independent given **Z**.

6.4.1 Mapping from IBD in Pedigrees

We first consider the pedigree context, in which prior probabilities of IBD are provided by the specified pedigree relationships among individuals. At locations across a chromosome, the genetic marker data **X** provide probabilities of IBD or realizations of the *inheritance vectors* which determine the location-specific descent of DNA through a pedigree (Sect. 6.3.2).

In genetic mapping, the use of location-specific IBD in affected individuals of known pedigree relationship has long been established as a powerful approach. This may be IBD in affected sib pairs (Suarez et al., 1978) or more general relative pairs (Weeks and Lange, 1988) or even the two parental gametes of individuals affected by a rare recessive trait (Lander and Botstein, 1987). In each case, individuals sharing the trait have increased probability of sharing genome IBD at causal loci. Observation of the same genome regions showing IBD across multiple pairs of affected relatives provides the linkage signal. Significance of the observations can be readily assessed, since the known pedigree relationship provides the null distribution of any IBD-based test statistic.

IBD-based tests for linkage have been extended to larger groups of relatives, and a wide variety of test statistics have been developed (McPeek, 1999). Joint sharing of genome IBD by multiple affected relatives generally provides stronger evidence, since the pedigree-based prior probability of this multiple sharing event is generally smaller. An advantage of an IBD-based approach is that it is relatively robust to allelic heterogeneity: within a pedigree the affected individuals are likely to carry the same causal mutation IBD. If there are large pedigrees with multiple affected individuals available, even locus heterogeneity is a lesser concern, since even a single pedigree can provide sufficient evidence of linkage.

In the genetic mapping of loci underlying quantitative traits (QTL), IBD-based approaches also have a long history. Haseman and Elston (1972) developed an approach to mapping QTL by considering the (negative) correlation between the squared difference in trait values in sibs and the marker-based IBD probability. More generally, the variance component approaches to QTL mapping in the SOLAR software (Almasy and Blangero, 1998; Blangero et al., 2000) use pairwise location-specific IBD probabilities computed conditionally on the pedigree structure and on observed marker data to model covariances among relatives and map QTL.

A simple version of the model for QTL detection is as follows. The vector of trait observations $\mathbf{Y}$ over the individuals is modeled as

$$\mathbf{Y} = \mu\mathbf{1} + \sigma_a\mathbf{g} + \tau_j\mathbf{w}_j + \sigma_e\mathbf{e} \tag{6.10}$$

Here μ is the overall mean, which may more generally include other fixed effects and covariates, $\mathbf{g}$ is a vector of genome-wide genetic (polygenic) effects, $\mathbf{w}$ is a vector of location-specific effects, and $\mathbf{e}$ is a vector of independent individual residuals. Thus Var($\mathbf{e}$) is the identity matrix $\mathbf{I}$, and $\mathrm{Var}(\mathbf{w}_j) = 2\Phi_j$ where Φ_j is the matrix of between-individual kinships at location j (Table 6.2). The variance of the genome-wide effect $\mathbf{g}$ is $\mathrm{Var}(\mathbf{g}) = 2\Psi$ where Ψ is the matrix of genome-wide kinships. For any pair of individuals, the term in the matrix Φ_j may be obtained for any location j from realizations of descent conditional on marker data $\mathbf{X}$ (Sect. 6.3.2). For each pair of individuals, these values may be averaged across the genome to obtain an estimate of Ψ, although in the past a pedigree-based expectation was often used for Ψ (Sect. 6.2.1).

A log-likelihood ratio can be used to test whether there is an effect specific to any location j in the genome. The purpose of the genome-wide term ($\sigma_a > 0$) is to absorb effects of genes other than at the test location j, in order to provide greater power and precision in detecting the effect at locus j. The general model of Equation (6.10) can be compared to the model in which there is no effect at location j ($\tau_j^2 = 0$):

$$\ell_j = \log\left(\frac{\max_{\sigma_a^2,\tau_j^2,\sigma_e^2} L(\sigma_a^2, \tau_j^2, \sigma_e^2; \Phi_j, \Psi)}{\max_{\sigma_a^2,\sigma_e^2} L(\sigma_a^2, \tau_j^2 = 0, \sigma_e^2; \Psi)}\right) \tag{6.11}$$

For pedigrees with not too many observed individuals, maximization over parameters, and hence computation of the test statistic (6.11), is not computationally intensive.

6.4.2 IBD in Pedigree-Based Likelihoods

In model-based testing for an association between genetic marker data $\mathbf{X}$ and trait data $\mathbf{Y}$, one may consider either the probability $\Pr(\mathbf{Y}|\mathbf{X})$ or the probability $\Pr(\mathbf{X}|\mathbf{Y})$. The latter is the basis of association studies, in which individuals are selected on the basis of their trait data, $\mathbf{Y}$ (e.g., cases and controls). Genotypes $\mathbf{X}$ are then compared between these two groups. For a quantitative trait, or when a more general trait model is desired, it is more natural to consider **Y** given **X**. The same applies to IBD-based genetic mapping, where the marker data **X** are used to provide information about IBD, **Z**. In the analogue of population-based association tests, we consider differences in inferred IBD in pairs conditional on their trait status (see Sect. 6.4.3 below). In QTL mapping we model **Y** given the inferred IBD as in Equation (6.10) above. In this section, we consider more generally the probability $\Pr(\mathbf{Y}|\mathbf{X})$ in the case where the pedigree structure is known.

As in classical linkage analysis (Smith, 1953; Morton, 1955), the goal is to test for dependence between inheritance of DNA at specific genome locations and the inheritance of DNA underlying a trait. The pedigree relationships among individuals are assumed known, and genetic marker data **X** are available for some individuals, for markers with known locations in the genome. Trait data **Y** are also available, and a model is assumed for the relationship between the allelic type of latent causal DNA and the trait of interest. In the following, Γ_X will denote the probability model for the marker data **X**, which involves marker allele frequencies and locations which are assumed known, and Γ_Y will denote the model for the trait, which specifies frequencies of trait alleles, and the probabilities of phenotypes given latent trait genotypes. The parameter λ is a set of locations at which, with any model, there is hypothesized to be causal DNA. The full model is $\Gamma = (\Gamma_X, \Gamma_Y, \lambda)$. The full set of all locations at which causal DNA is hypothesized in any model to be considered will be denoted Λ; each λ is a subset of Λ. For models with a single trait locus, $\lambda = \{j\}$ and Λ is the set of j at which likelihoods are to be computed.

As before, we assume that **Y** and **X** are conditionally independent given **Z** and denote by $Z(\lambda)$ the IBD jointly at locations specified by λ. For single-locus trait models $\lambda = \{j\}$, we write $Z(\lambda) = Z_j$. Then

$$\Pr(\mathbf{Y} \mid \mathbf{X}; \Gamma) \;=\; \sum_{\mathbf{Z}} \Pr(\mathbf{Y} \mid Z(\lambda); \Gamma_Y, \lambda) \Pr(Z(\lambda) \mid \mathbf{X}; \Gamma_X) \qquad (6.12)$$

There are several issues inherent in the use of Equation (6.12). First, even though **Z** is required only at locations specified in λ and only among individuals observed for the trait, direct computation is infeasible, except in cases of small pedigrees and simple trait models. If the trait model involves more than a single trait locus, so λ is not a single point, computation of the joint probabilities of $Z(\lambda)$ given the marker data **X** across the chromosome is not possible. Next, even for a single hypothesized location $\lambda = \{j\}$ for the causal DNA, if there are more than three related individuals observed for the trait, the number of possible IBD states among them is too large for practical computation of $\Pr(Z_j|\mathbf{X})$.

However, a Monte Carlo approach is feasible. The conditional probability of **Y** given **X** may be rewritten as

$$\Pr(\mathbf{Y} \mid \mathbf{X}; \Gamma) = \mathrm{E}(\Pr(\mathbf{Y} \mid Z(\lambda); \Gamma_Y) \mid \mathbf{X})$$

where the expectation is over the values of $Z(\lambda)$ given $\mathbf{X}$. The methods of Sect. 6.3.2 allow a large number, N, of patterns of IBD $\mathbf{Z}^{(k)}$ ($k = 1, \ldots, N$) across the chromosome to be sampled jointly for all relevant locations j in any collection of models $\{\lambda : \lambda \subset \Lambda\}$ conditional on the joint marker data on individuals and across the chromosome. For any specific hypothesis λ, the required probability may be estimated as

$$\widehat{\Pr}(\mathbf{Y} \mid \mathbf{X}; \Gamma) \;=\; \frac{1}{N}\sum_{k=1}^{N} \Pr(\mathbf{Y} \mid Z(\lambda)^{(k)}; \Gamma_Y), \quad Z(\lambda)^{(k)} \sim \Pr(\cdot|\mathbf{X}; \Gamma_X). \tag{6.13}$$

Given the realizations of $Z(\lambda)$, the estimate requires only $\Pr(\mathbf{Y} \mid Z(\lambda))$ for each hypothesized λ. If the model for trait phenotypes involves only a single locus $\lambda = \{j\}$, and writing $Z_j = Z(\{j\})$, the probability $\Pr(\mathbf{Y} \mid Z_j)$ may be computed by the methods of Sect. 6.2.4. The IBD graph approach outlined there can be extended to two-locus trait models (Su and Thompson, 2012).

For a single-locus trait model, Equation (6.13) is analogous to that first proposed by Lange and Sobel (1991) but is here phrased in terms of IBD rather than latent genotypes of individuals. Sampling and efficient storage of a collection of IBD realizations across the genome greatly facilitates analysis. Since Equation (6.12) separates the marker model Γ_X from the trait data $\mathbf{Y}$ and model Γ_Y, the analysis of the marker data may be performed once only. Computation of likelihoods directly from the stored IBD graphs allows these IBD graphs realized from marker data to be used not only for different hypothesized trait locations as in Lange and Sobel (1991) but also for different trait models and even for different traits observed on subsets of the same set of individuals.

On a given pedigree component, many different realizations, across many loci, and of different inheritance vectors, may give rise to the same IBD graph. The probability $\Pr(\mathbf{Y} \mid Z_j^{(k)}; \Gamma_Y)$ should be computed only once for each equivalent IBD graph. Given a sample of IBD graphs, each across a chromosome, there are algorithms to determine when IBD graphs are genetically equivalent (Koepke and Thompson, 2013). This can greatly increase efficiency of the trait data portion of the analysis, especially when the same collection of marker-based IBD graphs are to be used in analyses of multiple trait models or for data on multiple traits.

There are several other issues in the use of Equation (6.12). While the values of $\Pr(\mathbf{Y}|\mathbf{X}; \Gamma) = \Pr(\mathbf{Y}|\mathbf{X}; \Gamma_X, \Gamma_Y, \lambda)$ can be compared for different hypothesized values of λ, there is no baseline as to what should be expected for given sets of marker data $\mathbf{X}$. The classical human genetics approach has been to compare the value of (6.12) with the probability $\Pr(\mathbf{Y}; \Gamma_Y)$ under the same trait segregation model but in the absence of marker data. Note that this baseline "marker-free" null model is different from the null model of QTL mapping (Equation (6.11)), which is widely used in the plant and animal literature (Lander and Botstein, 1987). In that case the null model is of a zero effect of the DNA at a specific genome location j ($\tau_j^2 = 0$).

In the days before the existence of genome-wide genetic marker maps, a comparison of the marker-based $\Pr(\mathbf{Y}|\mathbf{X};\,\Gamma)$ with $\Pr(\mathbf{Y};\,\Gamma_Y)$ had a sound foundation (Smith, 1953; Morton, 1955). However, this is less meaningful when values of λ or locations j are spread across the genome, and genetic markers are likewise distributed genome-wide. Further, the unconditional trait probability $\Pr(\mathbf{Y};\,\Gamma_Y)$ may not be computable for data observed on very large complex pedigrees or for complex trait models. Second, even when easily computed, there remains the choice of Γ_Y. Trait models have a number of parameters, for each latent trait locus and genotype. Maximization over these parameters is often impractical, and the likelihood (6.12) is often sensitive to model choice. We return to these issues in Sect. 6.4.4 below.

6.4.3 Mapping from IBD in Populations

Just as when the pedigree relationships among individuals are known (Sect. 6.4.1), population-based IBD mapping relies on excess IBD among individuals of similar phenotype, relative to some null model or comparison group. As an example, we consider IBD-based mapping in a case-control study (Browning and Thompson, 2012). To avoid the issues of inferring IBD from marker data, we will assume that the local IBD between pairs of individuals is known with certainty.

Recall that in a simple association test, the frequency of a SNP allele in N_1 cases is compared with that in N_2 controls:

$$\left(\frac{1}{2N_1}\sum_{\text{cases}} X_i - \frac{1}{2N_2}\sum_{\text{controls}} X_i\right) \tag{6.14}$$

where $X_i = 0, 1, 2$ is number of alleles of specified type in i. By analogy, in an IBD-based test, we compare the frequency of IBD between M_1 case-case pairs and M_2 other pairs (case-(non-case) or (non-case)–(non-case)):

$$\left(\frac{1}{M_1}\sum_{\text{case-case}} Z_i - \frac{1}{M_2}\sum_{\text{other}} Z_i\right) \tag{6.15}$$

where $Z_i = 1$ or 0 as the pair does/does not share genome by descent at test location.

Just as in an association test, we must allow for population heterogeneity or structure. In an association test, there may be similarities among cases and/or among controls that are unrelated to the trait. Likewise in an IBD-based test, there may be different degrees of relatedness among cases from among controls, due to the methods of sampling or ascertainment. The average IBD scores within each group in Equation (6.15) may be adjusted for the genome-wide average within in each group.

To assess significance, a null distribution is required. Whereas in a known pedigree, Mendelian segregation provides an appropriate null distribution, in a population there is no such framework. However, just as in an association test, permutation of case-control labels provides a null distribution of the test statistic (6.15) under which there is no association between IBD at the test location and the case-control status of individuals. Since IBD is on a scale of Mbp, at most 3,000 tests can cover the genome. This results in a multiple testing burden that is several orders of magnitude less than that for SNP based GWAS.

To show that population-based IBD mapping can work, we present the details of part of the study undertaken by Browning and Thompson (2012). A coalescent simulation including selection and mutation provided a base population with effective size $N_e = 10^4$, over a 200 kbp region of chromosome, representing a functional gene region. The population was then run forward, and, at some later time point, IBD relative to the base population was scored in descendant individuals. In the example summarized here, the effective size of the recent population was $N_e = 10^5$ and the time-depth of IBD was $G = 25$ generations.

For the purposes of association testing, the best SNP in alternating 1 kb blocks was retained, for a total of 100 SNPs (Fig. 6.8). The five blocks of the central 10kb (schematically representing the exons of the gene) also contained causal variants that arose in the population simulation. Individuals with ≥ 1 causal variant alleles in the five central 1kb blocks are designated as cases with probability 0.1, providing sufficient information for a mapping signal, while still modeling a trait of low penetrance.

Tables 6.6 and 6.7 summarize the relevant results from Browning and Thompson (2012). The values in Table 6.6 show the range of properties of causal variants with different amounts of selection over 100 independent simulations. When selection is weak ($s = 0.0005$), there are somewhat more causal variants, at frequencies up to about 0.5%, but haplotypes carrying causal variants are not rare. These haplotypes have frequency from 4.5% to 13%, and normally there is a high association between at least one of the causal variants and one of the common SNPs. However, when selection is stronger ($s \geq 0.002$), the frequencies of causal variants are much lower, and the total frequency of haplotypes carrying causal alleles is of the order of 1%.

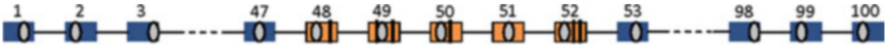


Fig. 6.8 The alternating 1 kbp blocks over a 200 kbp region with the five central blocks containing causal variants. (Figure from Browning and Thompson 2012)

Table 6.6 Properties of the simulated causal variants at different levels of selection. Selection is measured as the deficiency in fitness of allele carriers relative to non-carriers

Selection	# var.	var.freq.	total freq. of var-hap.	max assoc R^2 w/marker SNP
0.0005	11–16	0.00015–0.0060	0.045–0.13	0.91–1.00
0.001	9–14	0.00010–0.0031	0.019–0.050	0.28–1.00
0.002	8–13	0.00010–0.0020	0.0097–0.031	0.06–0.52
0.005	7–10	0.000088–0.001	0.0045–0.011	0.03–0.16

Table 6.7 Comparison of the power of IBD-based and association tests

Selection	# cases= # controls	power assoc.	power IBD	association vs. IBD
0.0005	500	0.87	0.57	assoc.
0.001	500	0.65	0.53	Not-Sig
0.002	1000	0.53	0.87	IBD
0.005	3000	0.47	0.90	IBD

In this case there is rarely a detectable association between any causal variant and any of the 100 common marker SNPs.

The results in Table 6.7 follow naturally. Here the size of the study is chosen to provide intermediate power values for easier comparison. The association test uses Equation (6.14) at each of the 100 common marker SNPs, and the IBD-based test uses the statistic (6.15) evaluated at the locations of these common SNPs. When selection is weak ($s = 0.0005$), the association test has higher power. However, when selection is stronger, so that each causal variant has lower frequency, an IBD-based test performs better than an association test. Allelic heterogeneity is a major problem for association testing, unless there is at least one variant with sufficiently high frequency to show association. In contrast, an IBD-based test is less affected by allelic heterogeneity, since each case-case pair has a higher chance of carrying the same causal allele, even though this allele may differ among pairs.

6.4.4 Model-Based Mapping Likelihoods in Populations

In Sect. 6.3.2 we saw how IBD $\mathbf{Z}$ could be sampled conditional on marker data $\mathbf{X}$ and a known pedigree structure. In Sect. 6.4.2 we saw how these realizations of $\mathbf{Z}$ could be used to compute a likelihood function (6.13) for use in inferring the locations of DNA underlying trait phenotypes $\mathbf{Y}$. In Sect. 6.3.4, we saw how IBD can be realized conditional on marker data in the absence of pedigree information. Finally we now show how these population-based realizations can also be used in genetic mapping. In fact, once the marker data $\mathbf{X}$ have been used to provide realizations of IBD, $\mathbf{Z}$, it is largely irrelevant whether or not they were made conditionally on a known pedigree structure.

For the general trait models considered in Sect. 6.4.2, it is usually insufficient to have only pairwise measures of IBD. Even for a single-locus trait model, with hypothesized causal DNA at location j, the probability $\Pr(\mathbf{Y} \mid Z_J; \Gamma_Y)$ (Equation (6.13)) will depend on the joint IBD state among the individuals observed for the trait. While extension of the methods of Sect. 6.2.4 allows efficient computation of this probability for any specified Z_j , the number of possible IBD states at a locus is huge (Sect. 6.3.4), and effective realization of $\mathbf{Z}$ given $\mathbf{X}$ is a difficult problem. The MCMC methods developed by Moltke et al. (2011) and by Zheng et al. (2014) are not scalable.

Glazner and Thompson (2015) proposed an alternate approach, in which a joint IBD state among multiple individuals is built up successively from pairwise inferences. The 15-state HMM is run on pairs of individuals as in Brown et al. (2012). However, in adding individuals to a joint configuration, the IBD trajectories across the chromosome are constrained by previously realized IBD. Using this method in a simulated example, Glazner and Thompson (2015) showed that joint IBD realized in the absence of an assumed pedigree structure can be used to recover a likelihood across genome locations j using Equation (6.13). The "gold standards" are likelihoods that would be obtained if IBD were perfectly imputed from marker data $\mathbf{X}$: $\Pr(\mathbf{Y} \mid Z_j; \Gamma_Y)$ at multiple locations j across a chromosome. For IBD realized using haplotypic marker data the approximation is very good, but for genotypic marker data, $\mathbf{X}$, it is less so. Additionally, the method becomes computationally intensive, and performance degrades, as larger sets of related individuals have observed phenotypes, $\mathbf{Y}$. There are additional problems also in the use of Equation (6.13) for likelihood-based mapping in the absence of a pedigree. First, the usual baseline probability $\Pr(\mathbf{Y}; \Gamma_Y)$ is not available; without a pedigree there is no basis to compute it. Second, no constant baseline works well: the likelihood (6.13) is affected by the inferred level of IBD across the chromosome, and for IBD estimated without the constraints of a pedigree structure, this can vary widely.

Because of unsolved problems in computationally feasible and effective ways to realize joint IBD in the absence of a pedigree structure, we return to the pairwise model of Equation (6.10) for our final example of IBD-based mapping in the absence of an assumed pedigree. In this model, the local kinships Φ_j are now estimated using the methods of Sect. 6.3.4, and the genome-wide kinship Ψ is estimated by averaging the local kinships Φ_j across the genome. This approach was first taken by Day-Williams et al. (2011). They used the estimator of local pairwise kinship Φ_j outlined in Sect. 6.3.3 and denoted DW. The resulting estimators, smoothed across the chromosome, may additionally be constrained so that each Φ_j for each pair of individuals takes the values 0, 1/4, 1/2, or 1 (Table 6.2). By contrast, the joint realizations of IBD among individuals produced by Glazner and Thompson (2015) can be immediately reduced to a set of pairwise realizations of the 15 states of Table 6.2 and hence to local kinships Φ_j; these estimates are denoted GT. At each j, and for each pair of individuals, the averages across realizations can also be constrained to the values 0, 1/4, 1/2, and 1.

The two estimation methods GT and DW were compared on a simulated example using the variance component log-likelihood ratio (6.11) computed at locations j across the chromosome. Since there was no evidence of a genome-wide genetic effect, for simplicity it was assumed that $\sigma_a^2 = 0$. The two sets of local estimates were each used constrained and unconstrained. Each of the four log-likelihood curves were compared with the curve that would be obtained if the actual $\mathbf{Z}$ and hence the realized Φ_j were known at each location j. Details may be found in Glazner and Thompson (2015), but generally the GT estimators performed better than the DW estimators. Also, whereas for GT there was little difference in the results between the constrained and unconstrained versions, for DW the constraint had negative impact especially in regions of high IBD. As noted in Sect. 6.3.3, the DW method underestimates IBD particularly in regions of within-individual IBD and multi-gamete IBD.

If a genetic model requires only local pairwise estimates of IBD, the HMM method of Brown et al. (2012) provides an alternative method. In this case the 15-state HMM is simply run separately on all pairs of observed individuals, to provide estimates of Φ_j across all j. The result of applying this approach to the simulated example of Glazner and Thompson (2015) is shown in Fig. 6.9, together with the results from the the DW and GT approaches. It is seen that the HMM method almost perfectly recovers the actual realized pairwise IBD in this example. Given the difficulties of estimation of IBD among multiple individuals, the use of models and methods that require only pairwise estimates provide an attractive alternative.

In Sect. 6.4.2 we showed how, on a defined pedigree, realizations of $\mathbf{Z}$ given genetic marker data $\mathbf{X}$ could be used to provide estimates of linkage likelihoods $\Pr(\mathbf{Y} \mid \mathbf{X}, \Gamma)$ for locations $\lambda = \{j\}$ across a chromosome. In this section we have shown how the same may be accomplished in populations, using a population prior model for IBD (Sect. 6.2.2). Once realizations of $\mathbf{Z}$, jointly among individuals and across a chromosome, are obtained conditional on genetic marker data $\mathbf{X}$, there is no essential difference whether these were made with or without the assumption of a pedigree structure. The pedigree structure provides a more informative, and sometimes overly constraining, prior, but modern SNP data at multiple markers can compensate for the lack of pedigree information.

This raises the attractive possibility of combination of pedigree and population-based IBD. The pedigrees of any family study exist within a population, and founders within and between pedigrees may be related. Methods to combine IBD inferred within pedigrees with that inferred among founder members have been implemented (Saad et al., 2016), but there are several issues. First founders of pedigree structures are often unobserved: populations-based inference of IBD is

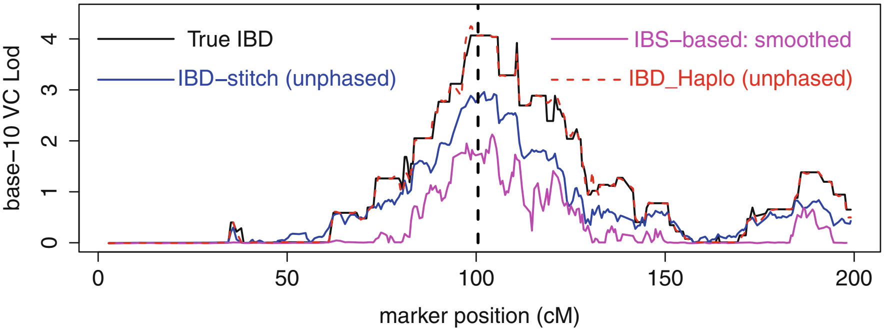


Fig. 6.9 Comparison of DW (magenta), GT (blue), and HMM (red-dashed) estimators of the log-likelihood ratio (6.11) on the example of Glazner and Thompson (2015). The black line shows the value that would be obtained if the true realized pairwise IBD were known as each point in the genome

practical only for individuals with fully observed genotypic data. Second, there are multiple individuals who may be related outside the defined pedigree structures. Population-based realizations of IBD jointly among multiple individuals are intractable.

One case where analysis is tractable is that of a set of parent-offspring trios for whom there are unphased genotype data. Population-based modeling of the IBD between the two parents can be combined with the modeling of the two meioses from parents to offspring. Joint analysis of the six genomes of the trio is important: pairwise analyses are much less informative and can give conflicting results. With joint analysis, one may, for example, compare the location-specific IBD probability between the two genomes of the offspring (i.e., *autozygosity*) with that expected given the IBD probabilities between the parental genomes. Hence one may detect regions of the genome in which autozygosity of surviving offspring is significantly lower than expected given the inferred segments of IBD in the parents, indicating possible selection against autozygosity in these regions.

In more general pedigrees, there is an additional complication in combining population- and pedigree-based IBD estimates. Within a pedigree, the maternal/paternal origins of haplotypes can be realized where there are informative data. In other cases, for example, for the two haplotypes of founders, there is no information (even with data) on which haplotype is maternal and which paternal, but this is irrelevant to within-pedigree IBD. In the population-based context, even if the IBD inference implies fully correct phasing, it is likewise arbitrary which haplotype is designated the maternal/paternal one of the individual. Combining population and pedigree IBD faces the intractable challenge of resolving the multiple pairings of each individual's two haploid genomes with the labeling in each population-based realization of IBD. Thus while IBD provides a natural unifying framework in which to combine pedigree- and population-based inferences, there remain challenges for successful implementation of methods.

6.5 Summary

We have shown how IBD $\mathbf{Z}$ can be inferred from genetic marker data $\mathbf{X}$ and then used to provide evidence for genome locations at which the DNA variants may be causal for trait phenotypes $\mathbf{Y}$. Rather than considering directly the association between $\mathbf{X}$ and $\mathbf{Y}$, we consider this association through the lens of descent $\mathbf{Z}$. In fact, a basic assumption of our models is that $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$. In the three main sections of the text, we have considered: first, probability models for $\mathbf{Z}$; second, inference of $\mathbf{Z}$ from$\mathbf{X}$; and third, use of this inferred or realized $\mathbf{Z}$ to map DNA underlying $\mathbf{Y}$

In Sect. 6.2 the focus was on the models for IBD or $\mathbf{Z}$. We considered probability models for $\mathbf{Z}$, both in defined pedigrees and among members of a population. It is important to consider not only IBD at separate locations but also how it changes across a chromosome. Because DNA descends generation to generation is large segments, even remote relatives will (if they share any genome IBD) share segments that are of order of millions of base pairs long and will contain many SNP markers. Also in Sect. 6.2 we considered probabilities of marker data, **X**, and trait data, **Y**, conditionally on **Z**. The key point here is that, for these probabilities, it is irrelevant whether or not the pedigree structure is known.

In Sect. 6.3 we turn to the inference of **Z** given marker data **X**. We first consider the case of defined relatives, where the pedigree structure provides a strong, but sometimes overly constraining, prior. Usually not all members of a pedigrees are typed, but pedigrees can only be validated among current individuals for whom marker genotypes are available. In some studies also there may be issues of genotypic error; for computational reasons, pedigree-based methods assume that marker data are observed without error. We then turn to the inference of IBD in population, the purpose of the IBD model being to provide a flexible and tractable prior for inference. One important aspect of this flexibility is that is allows for error in the observation of marker genotypes. SNP typing is quite accurate, but there are very large numbers of SNPs. We consider first genome-wide measures of IBD. We point out issues with methods that treat all SNPs equally and take no account of their dependence due either to allelic association (LD) or to their physical locations. We describe one method of adjusting for LD, but our major focus again is on the segmental natural of DNA descent. Since individual SNPs are very uninformative, combining multiple SNPs in detecting IBD segments is of key importance. Using a model for the dependence of descent across SNP markers has two important consequences. First, estimates of genome-wide IBD proportions are greatly improved. Second, and essential for gene mapping purposes, **Z** is realized at locations across the genome: the actual locations of segments of IBD are detected.

Finally, in Sect. 6.4 we show how **Z** inferred from marker data **X** can be used to map DNA that is causal to trait data **Y** against the genetic marker map. Again we consider first the case of pairs or groups of individuals whose pedigree relationships are known. This includes approaches such as that of affected relative pairs for binary traits and variance component models for mapping quantitative trait loci (QTL). We then extend to more general models for phenotypes **Y** and show how realizations of **Z** conditional on **X** can be used to obtain Monte Carlo estimates of linkage likelihoods $\Pr(\mathbf{Y}|\mathbf{X})$ for a specified trait model and specified hypotheses of the location(s) of causal DNA. Next we return to populations and consider an IBD-based analogue of case-control studies, showing that where different rare variants in a functional gene can cause the trait, the IBD-based approach outperforms a GWAS test. IBD-based tests can address allelic heterogeneity both in pedigrees and in populations. Finally, we return to linkage likelihoods, on the basis of IBD inferred in populations where pedigree relationships are unknown. We consider the complexities of multi-individual IBD and suggest that often a variance component model that requires only pairwise IBD may be more useful. However, the more fundamental message is that it is largely irrelevant to subsequent analysis whether IBD is inferred under a population model or on a defined pedigree. All pedigrees exist within a broader population framework; the IBD framework permits the combination of population and pedigree information.

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101

Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1211

Balding DJ, Moltke I, Marioni J (eds) (2019) Handbook of statistical genomics, 4th edn. Wiley, Oxford, UK

Blangero J, Williams JT, Almasy L (2000) Robust LOD scores for variance component-based linkage analysis. Genet Epidemiol 19(Suppl. 1):S8–S14

Brown MD, Glazner CG, Zheng C, Thompson EA (2012) Inferring coancestry in population samples in the presence of linkage disequilibrium. Genetics 190:1447–1460

Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet 86:526–539

Browning SR, Thompson EA (2012) Detecting rare variant associations by identity by descent mapping in case-control studies. Genetics 190:1521–1531

Chapman NH, Nato AQ Jr, Bernier R, Ankeman K, Sohi H, Munson J, Patowary A, Archer M, Blue EM, Webb SJ, Coon H, Raskind WH, Brkanac Z, Wijsman EM (2015) Whole exome sequencing in extended families with autism spectrum disorder implicates four candidate genes. Hum Genet 134:1055–1068

Cox DR (1962) Renewal theory. Methuen and Co., London, UK

Day-Williams, AG, Blangero J, Dyer TD, Lange K, Sobel EM (2011) Linkage analysis without defined pedigrees. Genet Epidemiol 35:360–370

Donnelly KP (1983) The probability that related individuals share some section of genome identical by descent. Theor Popul Biol 23:34–63

Elston RC, Stewart J (1971) A general model for the analysis of pedigree data. Hum Hered 21:523–542

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3:87–112

Glazner CG, Thompson EA (2015) Pedigree-free descent-based gene mapping from population samples. Hum Hered 80:21–35

Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8:229–309

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19

Hayes BJ, Visscher PM, Goddard ME (2009) Increased accuracy of artificial selection by using the realized relationship matrix. Genome Res 91:47–60

Hill WG, Weir BS (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res Camb 93:47–64

Karigl G (1981) A recursive algorithm for the calculation of gene identity coefficients. Ann Hum Genet 45:299–305

Koepke HA, Thompson EA (2013) Efficient testing operations on dynamic graph structures using strong hash functions. J Comput Biol 20:551–570

Lander ES, Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236:1567–1570

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci (USA) 84(8):2363–2367

Lange K, Sobel E (1991) A random walk method for computing genetic location scores. Am J Hum Genet 49:1320–1334

Lauritzen SJ (1992) Propagation of probabilities, means and variances in mixed graphical association models. J Am Stat Assoc 87:1098–1108

Leutenegger A, Prum B, Genin E, Verny C, Clerget-Darpoux F, Thompson EA (2003) Estimation of the inbreeding coefficient through use of genomic data. Am J Hum Genet 73:516–523

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493–496

McPeek MS (1999) Optimal allele-sharing statistics for genetic mapping using affected relatives. Genet Epidemiol 16:225–249

Mendel G (1866) Experiments in plant hybridisation. In: Bennett JH (ed) English translation and commentary by R. A. Fisher. Oliver and Boyd, Edinburgh, 1965

Moltke I, Albrechtsen A, Hansen T, Nielsen FC, Nielsen R (2011) A method for detecting IBD regions simultaneously in multiple individuals – with applications to disease genetics. Genome Res 21:1168–1180

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318

Peter B, Wijsman EM, Nato AQ Jr, Matsushita M, Chapman KL, Stanaway IB, Wolff J, Oda K, Gabo VB, Raskind WH (2016) Genetic candidate variants in two multigenerational families with childhood apraxia of speech. PLOS One 11(4):e0153864

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Saad M, Nato AQ, Grimson FL, Leweis SM, Brown L, Blue EM, Thornton TA, Thompson EA, Wijsman EM (2016) Identity-by-descent estimation with population- and pedigree-based imputation in admixed family data. BMC Proc 10(Suppl 7):295–301

Smith CAB (1953) Detection of linkage in human genetics. J Roy Stat Soc B 15:153–192

Sobel E, Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 58:1323–1337

Speed D, Hemani G, Johnson MR, Balding DJ (2012) Improved heritability estimation from genome-wide SNPs. Am J Hum Genet 91:1011–1021

Su M, Thompson EA (2012) Computationally efficient multipoint linkage analysis on extended pedigrees for trait models with two contributing major loci. Genet Epidemiol 38:602–611

Suarez BK, Rice J, Reich T (1978) The generalized sib pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 42:87–94

Tavaré S, Ewens WJ (1997) The multivariate Ewens distribution. In: Discrete multivariate distributions. Wiley, New York, pp 232–246

Thompson EA (2000) Statistical inferences from genetic data on pedigrees. In: Volume 6 of NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics, Beachwood

Thompson EA (2019) Descent-based gene mapping in pedigrees and populations, chapter 20. In: Balding DJ, Moltke I, Marioni J (eds) Handbook of statistical genomics, 4th edn. Wiley, Oxford, UK, pp 573–596

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Wang B, Sverdlov S, Thompson EA (2017) Efficient estimation of realized kinship. Genetics 205:1063–1078

Weeks DE, Lange K (1988) The affected pedigree member method of linkage analysis. Am J Hum Genet 42:315–326

Zheng C, Kuhner MK, Thompson EA (2014) Joint inference of identity by descent along multiple chromosomes from population samples. J Comput Biol 21:185–200

What Have We Learned from GWAS?

7

Benjamin F. Voight

Abstract

The last 15 years have witnessed the development and application of well-powered genome-wide association studies (GWAS), an approach in which a large number of genetic markers across the entire genome are tested for association with complex phenotypes in large, unrelated cohorts. This approach has led to hundreds of bona fide and reproducible associations between genotype and phenotype in human populations, with additional studies poised to dramatically increase this number in the near term. In the midst of this discovery process, a retrospective pause is warranted to consider how the field has evolved and what practically has been learned over this short time period. The success of GWAS as an approach to uncover the biological basis for disease required a number of key innovations and developments, both methodologically and technologically. Once those hurdles had been overcome, a number of basic insights have been revealed about complex traits directly from GWAS, most notably regarding the polygenic architecture of complex disease, shared genetic susceptibility across ethnicities, and success discovering previously unknown biology underlying disease progression. The following chapter describes these and other insights learned along the way, with examples from trait analyses and empirical observation. I conclude with considerations for future expected experiments as GWAS leaves the phase of locus discovery to systematic epiphany over the biological underpinnings of complex traits in humans.

B. F. Voight (✉)
Department of Systems Pharmacology and Translational Therapeutics and Department of Genetics, The University of Pennsylvania – Perelman School of Medicine, Philadelphia, PA, USA
e-mail: bvoight@upenn.edu

K. E. Lohmueller, R. Nielsen (eds.), *Human Population Genomics*,
https://doi.org/10.1007/978-3-030-61646-5_7

7.1 Introduction

The last 15 years has witnessed a paradigm shift in studies of complex phenotypes in humans. The basic idea—that by surveying a large number of genetic markers across the entire genome in a large number of individuals, markers linked to genes that associate with complex traits or disease can be identified—was envisioned almost a hundred years ago by the scientific giants Hermann Muller, Thomas Morgan, and Alfred Sturtevant, among others. Despite the transformative features the approach has had on scientific activity and productivity, the adoption of genome-wide association studies (GWAS) has not been completely uncontroversial within scientific and public spheres (Wade 2009; Goldstein 2009; Hirschhorn 2009; Visscher et al. 2012). While answers to any scientific question often spawn still deeper levels of inquiry, the debate within the scientific community about the magnitude of discovery which has been emitted from GWAS certainly should *not* be misconstrued by a broader audience to mean that the activity was without meaning or scientific value. In the midst of this discussion, a retrospective pause is warranted to consider how the field has evolved and what has been learned. Such insight will fully advantage the next set of experiments—in addition to and beyond GWAS—designed to advance the biological and genetic understanding of these conditions.

Several critical advances and insights have been made in the process of performing these studies. First, the process of deploying genome-wide studies required a number of critical insights about how to design such studies, statistical approaches and uniform tools to analyze the data generated from them, and best practices on how to integrate analysis across multiple data sets into clear queries of association, variant by variant. After these technical best practices were worked out in detail, the application of the experimental process has generated a number of key insights about the architecture of complex traits, the underlying biology, and the potential mechanisms of action. A fundamental second message is that the application of these studies has led unambiguously to the conclusion that common disease can be studied, understood, and characterized through systematic studies of common genetic variants. A third message is that these common variants are linked to causal ones and that they explain a fraction of heritability (though not completely); an observation that underscores the heavily polygenic nature of these phenotypes. And finally, as specific examples of genes and mechanisms are elucidated (Musunuru et al. 2010), there is increasing evidence that networks underlying complex traits are emerging (Cotsapas et al. 2011), demonstrating that mechanisms of disease etiology and the pathways contributing to them can be identified from genome-wide association signals.

In the following, I describe the progression of the genome-wide approach toward the dissection of complex traits and the insights obtained over the past and present. My intention in these sections is not to build a protective bulwark defending the genome-wide association approach as somehow the “best value” for the dollar spent on the approach. Rather, my goal is to highlight specific findings where the genome-wide approach has clearly added value in the analysis of genetic data and the understanding of complex traits.

Based on the studies that have been initiated, there are a number of clear directions to continue the process of uncovering the architectural organization, biological insight, and evolutionary history of the genes and pathways that contribute to these traits. Given some of the clues suggested by GWAS that have yet to be fully and systematically enumerated, I conclude with a description of several future research threads that might be anticipated in the coming years.

7.2 Lessons over Eight Years of Association Studies (2006–2013)

Looking back to the discussion leading up to the first published genome-wide associations for traits in 2006/2007 (The Diabetes Genetics Initiative 2007; Wellcome Trust Case Control Consortium 2007), one can be reminded of the uncertainty and subsequent debate waged around the scientific merit and technical feasibility of these studies. (Interested readers can consult a prescient review with the rationale and issues of the subject written just before this time by Hirschhorn and Daly (2005) or a historical perspective from Bodmer and Bonilla (2008).)

While the design of such studies appeared tractable, concerns surrounding the details of how such studies would be technically implemented were raised. A first intellectual hurdle was overcome in 1999, with a prediction based on empirical observation and simulations that the selection of as many as ~500,000 well-chosen markers would be required to survey common variation throughout the entire human genome (Kruglyak 1999); a number of SNPs were within the costs of the technologies used to ascertain them. However, the scope of such experiments raised concerns about the technical capability of genotyping technologies to provide accurate and unbiased genotypes in that number of markers in thousands of people. Further still, concerns about the appropriate statistical thresholds by which significance could be declared, how observations would be replicated and validated, and if studies that perform statistical analysis accounting for effects known to induce confounding (Pritchard and Rosenberg 1999) would work in practice. Today, the application of these analyses and technical quality control of data produced by these technologies are routine, owing to the careful and systematic groundwork laid down by researchers that contributed to the first studies published on the subject.

A further point of the discussion focused on the question if the architecture of common complex traits and diseases could even be dissected by such a design. Proponents who argued in favor hypothesized that common genetic variants likely contribute to common disease, motivating the international Hapmap Project (International HapMap Consortium 2005, 2007; International HapMap 3 Consortium 2010) that formed the basis of the design of array technologies that facilitate high-throughput genetics underpinning GWAS. One alternative model to this view suggested that common disease could indeed be influenced by rare variants rather than common ones (Pritchard 2001; Schork et al. 2009; Gibson 2012), thus motivating technologies designed to characterize this spectrum of alleles carefully from selected patient populations (Cirulli and Goldstein 2010). Empirical observation supports the model that interindividual risk to complex traits is influenced by a spectrum of alleles that are rare and common and weak and highly penetrant and that the relative magnitude and effect in this spectrum depends on many factors, including the intrinsic complexity of the trait (Iyengar and Elston 2007) as well the historical fitness consequences of the phenotype studied (Eyre-Walker 2010; Simons et al. 2014; Lohmueller 2014).

7.2.1 Study Design, Quality Control, and the Search for Technical Biases

One of the major contributions of the genome-wide association approach was the development of rigorous approaches to the control of the quality of samples, genotypes, and the fundamental praxis of careful evaluation of data for technical or other biases that can induce excess false-positive associations. It is obvious to state that the best study is one where every sample and all polymorphisms are accurately genotyped at high fidelity. However, even in the best-designed study, high-throughput genomics often trades off some accuracy (ideally with errors distributed randomly) for enhanced information collection. Before the first studies were performed, it was difficult to predict all sources of errors but more importantly which type of error would routinely induce false positives if not controlled. For example, some of the first genotype-calling methods (e.g., the dynamic modeling algorithm (Di et al. 2005)) were, for some polymorphic sites, biased against heterozygous genotype calls (Rabbee and Speed 2005). Miscalling heterozygotes—either incorrectly or by not calling—at best can result in reduced power for association or, at worst, could result in false-positive association if cases and controls were unbalanced in genotype calling (Zeggini and Morris 2011). Today, the field enjoys a number of highly accurate genotype-calling algorithms paired with workflows that instantiate a number of quality checks for samples and genotyping data, for example, checks for the effects of “batches” of samples, gender misclassification, various tests for missing genotypes, examination of quantile-quantile plots, etc. Interested readers can consult a number of areas where checks have been enumerated in detail elsewhere (Neale and Purcell 2008; Weale 2010) and codified into association testing software to facilitate quality evaluations for these purposes (Purcell et al. 2007).

After many studies have been performed, one broad lesson that has been understood now is that errors largely can be traced back to the study design, the source of sample DNA, how samples are processed through the lab, the technology used to generate data, and the algorithm used to perform genotyping. This basic insight has then fed back into the study design and the protocol by which genotyping is performed. In many cases, this has resulted in a closer working relationship between the statisticians and epidemiologists who helped plan the study, the technicians in the lab generating the data, and the analysts who work with the output of these labors. A corollary from this principle is that the task of robustly checking data quality—ensuring that patterns match expectation—is the first, foremost, and most labor-intensive activity for any genome-wide study, especially for new data types on cutting-edge technologies. For example, the first GWAS testing of common copy-number polymorphism data with disease followed the same trajectory as for single-nucleotide polymorphisms but required additional checks and filters for quality (Myocardial Infarction Genetics Consortium 2009). As the field moves forward into new data types and technologies each with their own biases, the establishment of a procedure of stringent quality control will be essential to ensure the pace of discovery and the avoidance of false-positive associations (see retraction from Sebastiani et al. 2011).

7.2.2 Addressing Confounding from Population Stratification

After the technical control and application of high-quality data had been achieved, an essential biological phenomenon to control in the GWAS design is population ancestry. It had been known for some time that spurious, false-positive associations between marker and phenotype could occur when the prevalence of the phenotype of interest differed across sampled populations and if one does not take this fact into account when sampling and performing association testing (Knowler et al. 1988; Campbell et al. 2005). Though well-matched association designs were far more statistically powerful and cheaper than the previous linkage or family-based studies (Risch and Merikangas 1996), there was a brief time where it was not entirely clear how one could achieve ancestry matching in practice. As the need arose, several statistical developments preceding the GWAS era arose, either to correct the distribution of generated test statistics (Devlin and Roeder 1999), proposing either formal tests for cases and controls matching (Pritchard and Rosenberg 1999), using the genetic data to infer ancestry directly (Pritchard et al. 2000), or summarize the genetic similarities of samples using principal components analysis and control for those vectors in association testing (Price et al. 2006). (Readers interested further in the subject should refer to previous chapters that further detail methods.) Sufficed to say, the application of these approaches in the context of genome-wide association studies was highly successful, as overtime, discoveries across a range of phenotypes continue to be consistently supported by the influx of additional data sets and across a range of ethnic groups.

7.2.3 Threshold for Declaring Significant Genome-Wide Results

Another point that required resolution was the determination of an appropriate genome-wide statistical threshold that maintained the appropriate error rates, given the number of markers tested. If Kruglyak's estimate of markers was correct (that ~500,000 markers would be needed), what would such a threshold actually turn out to be, especially given the number of not only markers directly assayed and tested but also implicit (but partially correlated) tests of additional markers not assayed but indirectly tested due to linkage disequilibrium? What emerged was a consensus of observations from multiple lines of inquiry designed to estimate the threshold empirically from data—permutations, principal components methods, haplotype-block counting, and more—all of which triangulated on a threshold around 5×10^{-8}. This specific value maintained equal to or slightly lower than the required error rates given the number and haplotype structure around common variants tested (Johnson et al. 2010). While discussion about the appropriateness of this specific choice continues today, the emergence of an overall threshold allowing specific findings to stand out in the context of the entire genome (rather than candidate genes) was critical to identify bona fide observations which could be statistically comparable across studies and across phenotypes, observations that would stand up to scrutiny and replication efforts. The fact that now thousands of associations exceed this threshold and have been independently replicated, across multiple ethnicities (Saxena et al. 2012), validates the central claim that GWAS can identify common genetic variants that contribute to complex traits.

7.2.4 Best Practices for In Silico Statistical Imputation

While the direct testing of markers on genotyping arrays was a major advance to understanding complex disease, two central challenges emerged as multiple technologies, and data sets emerged to perform the task. First, after the issue of statistical thresholds had been addressed, direct testing as many of the ~2.5 million markers found in population databases (International HapMap Consortium 2005, 2007) would likely be the best shot at detecting association with any SNP. Second, as new studies increasingly used different genotyping technologies often with different and only partially overlapping SNP panels, it was clear that a way to summarize evidence of association across a common SNP panel would be desirable. These two facts and an emergently clear picture about the landscape of haplotypes in genetic data motivated the development of statistical methods that use panels of reference haplotypes from population studies to statistically infer the genotypes of untyped markers. The intuition for how this process works comes fundamental to the idea of linkage disequilibrium, in a multi-locus context. If one can predict with high accuracy ($r^2 > 0.95$, for example) the genotype of one SNP given another, one does not need to know the precise SNP genotypes at one site to test the other. Now, consider that information on linkage extends beyond simply pairs of SNPs, but to multiple SNPs that exist on specific haplotypes. If one can match the haplotype in question closely to combinations of those that have been previously observed, then one can probabilistically infer the genotypes at all untyped markers that also reside on those haplotypes with relatively high accuracy. This process (called *imputation*) now has many robust and specific implementations in software (Browning and Browning 2009; Howie et al. 2009; Li et al. 2010). The principles of how best to incorporate genotypes at SNPs that are not completely certain in the association with phenotype are relatively well understood, at least for common variation (Marchini and Howie 2010). Both the utilization of imputation to test markers for association in GWAS and the fact that consistent replication and validation using further direct genotyping of imputed variants show empirically that imputation as a practice for common variation does work and helps to improve power to associate variants with phenotype.

7.2.5 Collecting Evidence Across Studies via Meta-Analysis

Once directly genotyped and imputed SNPs have been collected across multiple studies for a specific trait, it is natural and statistically desirable to pool evidence across studies into a single assessment of association. The most direct way to perform this analysis is direct with genotypes, in a stratified analysis (e.g., the Cochran-Mantel-Haenszel procedure). However, this analysis approach is not directly feasible when informed consent and patient protections for genotype data do not allow the transmission of clinical data to external actors. In this case, it is desirable to summarize association statistics study to study, by estimates of the effect of the SNP alleles to phenotype and the associated error as well as by *p*-values, and meta-analyze those data. The meta-analysis of association data in this way raised a number of details that needed to be worked out to obtain faithful and unconfounded estimates of association across studies. Some of those issues were determining the most powerful and appropriate statistical models used for meta-analysis; controlling for the “forward and reverse-strand ambiguous” nature of A/T and C/G polymorphisms and ensuring that the same allele for each SNP is tested for each contributing study; thresholds to filter out SNPs passing into meta-analysis based on frequency, quality of imputation, and sample sizes; or how stratification and genomic control ought to be applied to the subsequent data. While the technical details did take time to work out (de Bakker et al. 2008), ultimately, these and other issues surrounding the analysis were resolved effectively. The combination of methodology and workflow to appropriately combine data to maintain power and discover new loci has had a tremendous impact, not least of all in terms of the number of discoveries as a meta-analytical consortium for glycemic, cardiovascular, autoimmune-mediated, and psychiatric disease grow larger in sample sizes as additional studies are added. The details and issues surrounding how to amass information across multiple studies in a systematic and comparable way are a critical step for complex trait studies to collectively analyze a dozen of studies together, and those details will be all the more important as new technologies and studies amass ever-increasing detail of genetic variation in patient populations.

7.3 Uncovering the Biology and Architecture of Complex Traits

Today, the human genetics field has identified hundreds of reproducible associations across a range of genetic and pharmacogenetic traits (Hindorff et al. 2009). While these data represent only a fraction of the underlying genetic contributions, advancements across numerous traits have led to clear insight into the architecture, as well as clues about the pathophysiology and mechanisms contributing to complex traits. One can expect additional observations, features, and partial resolution of some of these questions to emerge as studies continue to amass and specific observations are taken into functional and experimental systems for further mechanistic proof and understanding.

7.3.1 Complex Traits Are Differentially Complicated

One clear message is that complex traits broadly share architectural features in common, but not all traits are uniformly easy to dissect. One can see such features by comparing the trends in the sample sizes that have led to discoveries across a handful of traits (Fig. 7.1). Not plotted here are examples in the pharmacogenetics context, where a few point mutations explain a large fraction of patients with adverse drug response (SEARCH Collaborative Group et al. 2008; Ge et al. 2009), but we should be reminded of these cases for their relatively simple architecture. Overall, these trends indicate a roughly linear relationship between sample size and discovery at the current stage of analysis for these traits. This implies that, eventually, many traits will enjoy a similar degree of success in locus discovery efforts, as a function of samples contributing. Also, these trends indicate different slopes across traits that imply that each trait has an intrinsic but variable mapping difficulty. The differences in difficulty could be explained by many factors (e.g., the underlying genetic model, effect sizes, frequencies of risk variants, the extent of genetic heterogeneity, etc.), and further dissection of the architecture of traits is a source of active research.

Also contributing to the trait complexity is the degree to which associations are shared across phenotypes. For example, great success has been met mapping genetic association for a range of autoimmune disease (AID, e.g., type-1 diabetes (T1D), inflammatory bowel disease, rheumatoid arthritis). As associations have poured in, one leading observation is that SNPs with proven association to one trait are often (and nonrandomly) associated with multiples AIDs (Cotsapas et al. 2011). This observation is strongly suggestive of shared pathways and biological networks across diseases, as well as sets of networks that are disease-specific (e.g., the INS locus for T1D). Metabolic, cardiovascular, and anthropometric networks also appear to share genetic associations in common, though not nearly the extent as AID: commonality is more rarely shared among individual SNPs but is instead localized to discrete genomic locations. As traits share an overlap in genetic causality, one

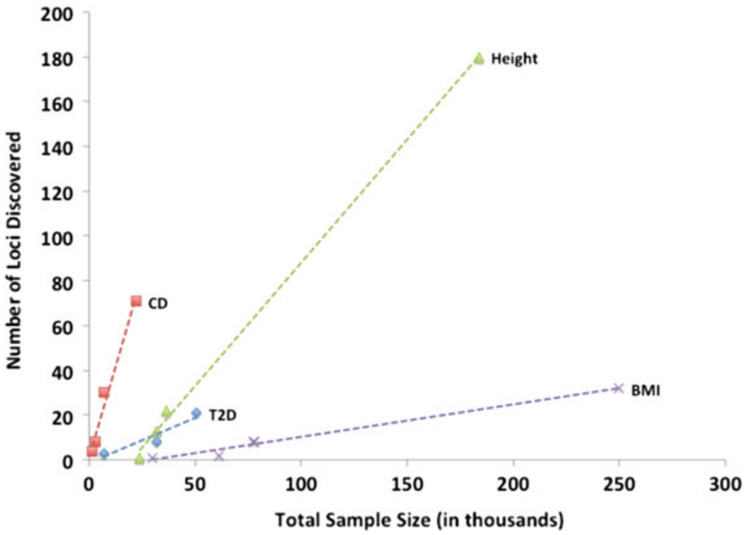


Fig. 7.1 Total number of genetic associations discovered as a function of total sample size collected for Crohn's disease (CD), height, type 2 diabetes (T2D), and body mass index (BMI). Approximated data was taken from published meta-analysis (plus replication) cohort sizes for each trait. Here, one can note that (**a**) traits all seem linearly related (increases in sample size linearly increase the number of loci discovered) but that (**b**) the rates differ across traits

can expect biological insight for one trait to also translate into the others, thereby increasing the pace of understanding the underlying etiology.

7.3.2 The (Un)Explained Heritability for Complex Traits

A laudable goal for complex trait studies is to systematically identify all genetic factors responsible, thus explaining as much of the variability in the trait, at a population level, as possible. Under some simple assumptions, the total phenotypic variability for a trait in a population can be mathematically described as a linear combination of genetic (additive, dominance, interactive) and environmental (common as well as random) terms (see Tenesa and Haley 2013 for a detailed review). The proportion of genetic relative to total phenotypic variance is referred to as H^2, or broad-sense heritability. A typical assumption is that dominance and interaction term nearly zero, and thus, we consider additive genetic variance relative to total narrow-sense heritability, or h^2. Calculating the amount of narrow-sense heritability that genetic loci contribute is straightforward: under this additive model, the variance is given by $2p(1\text{-}p)a^2$, where p is the allele frequency and a the additive effect of the allele on the trait, typically in units of standard deviation. Thus, based on a "current" set of discoveries of genetic associations, along with estimates of heritability from twin or family studies, one can determine how much heritability remains to be explained by genetics. One intellectual quandary that has emerged from complex trait mapping studies is that, taken collectively, the current set of genetic associations does not entirely explain the entire interindividual predisposition to these traits.

For example, the current estimates for human height suggest that while trait is upward of 80% heritable, established genetic associations account for only 5–15% of that heritability (Lango Allen et al. 2010). Similar reports have been shown for a range of anthropometric, cardiovascular, glycemic, and autoimmune traits (Manolio et al. 2009). Thus, where does the remaining genetic explanation to this “missing” heritability reside? A wide variety of potential culprits could explain this difference: additional risk could reside within genetic variation which is lower frequency, rare, or structural (i.e., deletions or duplications) that has historically been difficult to assay, excess polygenicity from hundreds of weakly penetrant common variants which have yet to reach statistical thresholds for significance, epigenetic phenomenon, parent of origin effects, or within interactions between genes (Maher 2008; Eichler et al. 2010). Another possibility could be that population calculations might actually overestimate the total heritability, due to simplifying assumptions about how the traits are modeled with respect to interactions (Zuk et al. 2012). Given numerous potential sources to explain the unexplained, it is clear that no single factor will completely equalize the population and genetic estimates of heritability.

7.3.3 The Data Are Consistent with a Strong Polygenic Component

One quite clear source of missing heritability can be explained, in part, by the large polygenic component underlying complex traits. The evidence for this claim has been delivered by the development of statistical methods that infer the number of loci that are left to be found, given the power to discover loci which already have been seen (Park et al. 2010), predictive models which use large numbers of common variants in a primary stage to predict disease in a second stage (International Schizophrenia Consortium et al. 2009), or approaches which directly seek to estimate the heritability contributed by common variants in large genome-wide studies (Yang et al. 2011a). These approaches have resulted in many studies that quantify the heritability for a range of complex traits. In the example of human height, while established associations at 180 loci can explain 10% of the variation in height, common variation overall could explain as much as 45% of the variance (Yang et al. 2011b; Lango Allen et al. 2010). The results from the application of all these methods are the summary conclusion that, for each trait, several thousand genetic loci are linked with common variation tested in GWAS that remain to be discovered.

7.3.4 Complex Traits Share a Genetic Basis in Common Across Populations

One observation from GWAS is that the genetic basis of this disease is indeed shared across populations descended from different ancestries, though the details can vary across associated sites. Taking T2D as an example trait, this has been primarily observed, where sufficient statistical power exists, at genetic associations identified in one population that also associate (individually or in aggregate) in other populations (Waters et al. 2010; Saxena et al. 2012). In several cases, the same physical regions and even the same SNPs have been implicated by association studies across ethnic groups (e.g., *TCF7L2* or *KCNQ1*), which imply the existence of a common underlying etiology and mechanism. However, in other cases, different or additional genetic factors at the locus are also associated given the best available data (e.g., *HMGA2* or *CDKN2A/2B*), and in some cases, the estimated effect of genotype to phenotype is heterogeneous across populations (e.g., *CDKAL1*). These observations could be compatible with shared biology, but variable mechanisms underlie disease susceptibilities across populations. Alternatively, the data could be compatible with simply a lack of statistical power in testing due to small sample sizes coupled with a lack of a comprehensive variant map across ethnicities to resolve the true, underlying causal variant(s) associated across groups. While a comprehensive answer to this question will require extensive fine mapping and genotyping in large samples across ethnicities, it is overwhelmingly clear that further genetic investigations along this track are essential to identify novel biological underpinnings of disease.

7.3.5 The Identification of Unknown Mechanisms for Disease Pathogenesis

The preoccupation with reconciling the question of missing heritability is certainly important for many applications, but should not detract from the goal of translating genetic findings into biological knowledge, hypotheses for a mechanism, and actionable intelligence for treatment. The most promising aspect of GWAS is that the design has provided a rigorous procedure by which biology and mechanisms related to disease can be discovered. This can clearly be seen as, for a range of cardiovascular, anthropometric, and glycemic traits, previously known monogenic causes of these diseases are preferentially found by common variant association (Voight et al. 2010; Teslovich et al. 2010). But further still, it is also increasingly clear that previously unknown causes and mechanisms of disease have also been clearly identified by GWAS. Those examples include, but are not limited to, the complement pathways for age-related macular degeneration, autophagy, Th17, and interleukin-23 pathways in Crohn's disease, multiple sclerosis, and other autoimmune-mediated diseases, hedgehog and TGF-β signaling, chromatin remodeling, histones for human stature (Lango Allen et al. 2010), and gene regulation by microRNAs MIR-137 for schizophrenia (Williams et al. 2011) as well as Lin28/let-7 for glucose metabolism and risk to diabetes (Zhu et al. 2011). In addition to recapitulating known and discovery of unknown biology, these studies have also recapitulated association at established pharmacological interventions for disease (statins, thiazolidinediones, sulfonylureas, and several others), despite the relatively weak effect sizes associated at each locus. Also, genetic studies on adverse drug response (SEARCH Collaborative Group et al. 2008) or treatment response (Ge et al. 2009) have identified factors strongly predictive and relevant in clinical outcomes. Interestingly, genes nearest to established associations may not always be causal: a recent report for an association for T2D indicated strong evidence for long-range enhancer activity regulating a gene further away (*IRX3/5*) than the most proximal one (*FTO*). This interaction and subsequently follow-up effort implicated brown fat and the biology of non-shivering thermogenesis as a result (Smemo et al. 2014). As further results from GWAS are taken forward into model systems (Musunuru et al. 2010) or in vitro studies to further elucidate the mechanism, one can expect the pace of understanding to proceed at a steady pace given the labor and time these experiments require.

7.3.6 The Importance of Phenotypic Collection in Study Designs

Another clear but obvious message is that phenotypic ascertainment can greatly impact the genetic study and associations discovered. An initial example of the importance was first observed in the first published GWAS studies for type 2 diabetes (T2D) (The Diabetes Genetics Initiative 2007). In that study, the design explicitly matched the body mass index (BMI) of cases with that of controls, with the goal of identification of variants for T2D not mediated through obesity. As a result, the study directly controlled for the effect of a strong genetic effect for obesity, variation near the *FTO/IRX3/5* locus. Subsequently, other studies without phenotypic ascertainment in this way discovered and replicated variation at that site, both with obesity and T2D (Frayling et al. 2007). As the field progresses toward next-generation genetic studies, with more sophisticated phenotypic measurements or extreme sampling (Guey et al. 2011), careful attention to the collection process (and under what conditions those collections occur or are associated with genotype) will be increasingly important. In addition, collection of a battery of additional trait measurements (endophenotypes) will be valuable for diseases where clinical boundaries are hard to describe or assign with precision (e.g., psychiatric conditions, Crohn's disease vs. ulcerative colitis)

7.3.7 Challenges with Interpreting the Distribution of Effects and Frequencies

The increasing number of polymorphisms associated with complex traits facilitates the first glimpses of the distribution of effect sizes and the frequency spectrum of alleles that underlie them. For many traits with association that are known, this distribution across a range of phenotypes is, with occasional outliers, trending toward "U-shaped," where the effect size increases as risk alleles become infrequent (e.g., Speliotes et al. (2010), Fig. 1c). However, this distribution is largely biased due to the power of the study to discover effects, given sample sizes collected. A more formal way to present this discussion is to formally test hypotheses about frequencies and effect sizes, building confidence sets of models of both parameters that fail to be rejected by the observed distribution of association scores in meta-analysis. As an example, power analysis focused on type 2 diabetes (and the sample sizes collected to date), assuming a simple additive genetic model, and across the range of "surveyable" genome, we can easily reject genetic models which postulate that additional common variants (~20–80% frequency) with modest effect or greater (odds ratio of 1.2 or greater) remain to be discovered (Park et al. 2010).

While it might be tempting to model the distribution of effect sizes and allele frequencies underlying complex disease, it is important to note that, except in a few rare circumstances, the actual causal alleles are not known with precision. More than likely, lead associations at common variation are simply strongly linked with variation that is indeed causal. An alternative hypothesis is that common variation is actually linked with rare, casual, coding mutations (potentially at very long distances), and thus the observed association at common sites with traits is merely "synthetic" (Dickson et al. 2010). While this hypothesis has some theoretical hurdles to overcome in explaining the majority of common variant associations (Wray et al. 2011; Anderson et al. 2011), empirical genetic studies are slowly indicating that low-frequency and rare variation cannot explain all common variant association signals. Coupled with examples of mechanistic explanations for complex traits (Musunuru et al. 2010), noncoding variation increasingly appears implicated in the biology of complex diseases. Systematic answers along all of these lines will require a comprehensive panel of variation across the frequency spectrum surveyed in a large number of samples.

7.3.8 Observational Epidemiology and Genetics Need Not Always Agree

Prior to initiating genetic studies, epidemiological and heritability studies are first initiated to quantify (and demonstrate) that specific clinical or physiological phenotypes segregate within families and are amenable to genetic dissection. As such, numerous epidemiological studies on a range of traits have been performing and continue to be performed. One feature that emerges from such studies, beyond the estimates of heritability for traits, is that many of these traits are correlated with one another. This can be due to not only how the traits are measured but also the intrinsic biology of the traits. This observation has been exploited with important practical effect, for example, in using plasma-lipid levels as biomarkers to predict the incidence of heart attack (Emerging Risk Factors Collaboration 2009). Other correlations across traits are also well-known: between systolic and diastolic blood pressure measures, triglyceride levels with high-density lipoprotein (HDL) cholesterol, body mass index and waist-to-hip or waist circumference measures, and many other examples.

Observed epidemiological correlation between traits can be intuitively thought of as the aggregated effect of all genetic (and environmental) perturbations across traits. However, even if the aggregated effect results in a significant correlation between traits, individual genetic factors are not required to contribute to both and could even have opposing effects. An outstanding example of this phenomenon comes again from one of the first published well-powered GWAS (The Diabetes Genetics Initiative 2007). That study observed a coding polymorphism at the glucokinase regulatory protein, *GCKR*, initially associated with triglyceride levels but subsequently with fasting glucose levels (Orho-Melander et al. 2008) and type 2 diabetes susceptibility (Dupuis et al. 2010). However, in contrast to observational epidemiology predicting a positive correlation among triglycerides, fasting glucose levels, and susceptibility to T2D, the allele associated with increased triglycerides was associated with decreased fasting glucose levels and lower risk to T2D at genome-wide levels of significance. A resolution to this paradox was proposed after careful mutational and mechanistic study indicating that the effect is due to a reduction in regulatory effect by fructose-6 phosphate-mediated inhibition of *GCKR*, resulting in increased glucokinase activity in the liver (Beer et al. 2009). This effect is predicted to enhance glycolytic flux, promoting hepatic glucose metabolism and elevated concentrations of malonyl-CoA, a substrate for de novo lipogenesis (Beer et al. 2009). While the story around *GCKR* is an unusual one, other examples are emerging (Kilpeläinen et al. 2011). It is this spectrum of unique variation discovered by GWAS, along with the notion that such variation empirically exists and can be found, that will contribute to the understanding of the mechanisms underlying disease susceptibility and trait biology.

7.4 What Lies Just Beyond the Horizon for GWAS?

In the immediate term, there are several genetic experiments that follow from what has already been learned from GWAS and should be anticipated in the coming years. The central theme that underlies each of them is the strategy to collect (and genetically test) large genetic data sets within previously uncharacterized human populations, join them together with other genetic data sets across populations and phenotypes, and finally integrate them with new, high-throughput genomics technologies. Systematic application of each of these threads is aimed to (1) help identify the polygenic contribution of diseases by systematic genetic study, (2) pinpoint causal genetic variation and hypotheses of the mechanism by fine mapping established loci for disease, and (3) generate hypotheses for networks and pathways by integrating with functional data sets.

7.4.1 Custom Genotyping Arrays Technologies for Genetic Studies

The central findings from GWAS, where hundreds of associations to multiple (and related) traits have been identified, are two central questions, among several others. First, can further genetic investigation identify additional associated regions with these traits? And second, can detailed genetic investigation localize causal variants where an established hit has been identified? Addressing these two questions, however, requires assaying hundreds of thousands of genetic variants from thousands of individuals, which is an expensive, labor-intensive, and time-consuming process. To make this process cost-efficient and streamlined in terms of producing an analysis, 2010 saw an effort to develop custom-array genotyping technologies, which are built directly from the information from leading association studies efforts and have been initiated. These include, among potentially others in development, the IBC-chip, the Metabochip for cardiovascular, metabolic, and anthropometric traits, and the ImmunoChip for AIDS (Keating et al. 2008; Cortes and Brown 2011; Voight et al. 2012). In addition, a custom array developed to comprehensively genotype low-frequency and rare variation based on discoveries from exome-sequencing projects for the coding genome, the Exomechip, has also been developed. These technologies offer data in a well-powered, second stage, to facilitate replication and fine-mapping genetic studies, as well as machine-learning or pathway/module-based methods to identify networks related to disease and to characterize their architecture. The first set of studies using these technologies are starting now to be published (Trynka et al. 2011, Morris et al. 2012; CARDIoGRAM Consortium 2013; Global Lipids Genetics Consortium 2013), and one should expect a surge of new locus discoveries and statistical methods that utilize this approach to data collection and technology.

7.4.2 Analysis of Multiple Phenotypic Measurements and Outcomes

The past 5 years of study has focused a large fraction of effort toward the analysis of individual phenotypes one at a time. An observation emerging from such studies is that multiple traits overlap in association, either at SNPs or at discrete physical locations of the genome. This is perhaps the most notable for autoimmune disease (Cotsapas et al. 2011), though several instances for metabolic (as previously discussed) and psychiatric diseases have also been observed. This set of observations clearly justify more expansive genetic studies which analyze multiple phenotypes simultaneously either using summary data or jointly in multivariate regression models. These approaches can be expected to uncover genetic loci with compelling associations that have not yet been captured by existing studies due to a lack of power.

One hypothesis along this line of inquiry posits that regions with associations to multiple traits are uniquely positioned to help identify causal genes, tissues, and mechanisms of action contributing to them. This hypothesis relies on the assumption that pathway modules and mechanisms that contribute genetic risk to multiple traits are the same across traits (or at least, for related traits). It is still a theoretical possibility that, locus-to-locus, different associations point to different (but proximally located) genes in an interval. Multiple SNPs could still potentially involve a single gene actor, but implicate different mechanisms important for different traits. For example, the ~1 Mb region upstream of the well-known proto-oncogene, c-MYC, has multiple associations to different cancers across tissues (Ghoussaini et al. 2008), with a leading hypothesis that this is due to *cis*-expression modules for various regulatory elements which are tissue-specific. A primary analysis will be to evaluate the pathways around regions where genome-wide significance for one trait has been established and multiple associations with additional related traits also reside.

An additional line of inquiry will involve comparisons across trait groups where prior evidence of biological overlap across traits will certainly be of interest—for example, autoimmune associations with metabolic and psychiatric disease or cancer with metabolic and cardiovascular traits. There are at least some examples in data where potentially coincident associations have been found (e.g., the regions around *BCL11A* and the *CDKAL1* with Crohn's disease and T2D; the MHC region with multiple autoimmune disease and psychiatric disease). While the exact gene candidates and mechanisms remain to be elucidated, exploratory analysis at this stage can begin to rule out (or hone in on) regions of the genome that could contribute risk, collectively. By combining data from multiple genetic scans for psychiatric disease (schizophrenia, bipolar disorder, attention deficit disorder, autism), new genetic loci directly contributing to the shared burden of disease have been reported (Williams et al. 2011), with additional support for a shared, common genetic comorbidity among several pairs of traits (Cross-Disorder Group of the Psychiatric Genomics Consortium 2013).

7.4.3 Genetic Fine Mapping for Complex Trait Loci

One activity that can be expected uniformly across complex traits will be to apply fine-mapping procedures to characterize the allele spectrum, identify new variants, and generate hypotheses of causal variants and genes at established loci. Early and highly illustrative examples of the strategy have already been performed (Graham et al. 2007), and the first large-scale fine-mapping studies amassing data across multiple ethnicities are starting to be reported (DIAGRAM Consortium 2014). The strategy begins by first sequencing a large panel of individuals at association regions to discover all polymorphic sites that could be related to disease, and second genotype all variants in a large number of samples and perform systematic conditional analyses to identify signals which explain the lead association signal, as well as additional associations independent of the leading signal. This specific design has been prohibitive until very recently, because of two key developments: first, availability of custom array genotyping technologies to perform the genotyping in a cost-efficient way and, second, approaches which allow the systematic conditional analysis on specific genetic variants using summary data, without the need to share individualized genotypes, prohibited due to patient privacy protections (Yang et al. 2012). Systematic application of these studies to hundreds of genetic loci identified by GWAS promises to discover additional independent alleles contributing to disease (explaining additional missing heritability) and a graded structure of high- to low-risk haplotypes contributing to risk. These haplotypes will then allow specific candidates for causality to emerge, which can then be tested in functional or experimental systems.

7.4.4 Genetic Studies of Complex Traits across Ethnicities

Despite the higher prevalence and resulting disproportionate public health burden, genetic studies of complex disease in non-European populations have not taken off nearly at the rate of their European counterparts (Bustamante et al. 2011). Fortunately, and recently, this trend appears to be shifting, for important scientific reasons. Well-powered GWAS for a range of disease traits across a range of ethnic groups (e.g., Asian, Indian, African-American, Latino) have recently been published (Saxena et al. 2012; SIGMA Type 2 Diabetes Consortium 2013; DIAGRAM Consortium 2014) with many more underway. Many of these studies are also actively working to pool data together and perform meta-analysis and fine-mapping efforts (DIAGRAM Consortium 2014), forming collaborations with existing groups who have focused on studies primarily in European populations.

These efforts are very important, as studies across populations worldwide can identify new genetic risk factors that are more common or operate at stronger effects across populations (Rosenberg et al. 2010; Pulit et al. 2010), and can help improve the quantification of genetic risk prediction across ethnicities. A potentially useful application of data across ethnicities is to exploit different patterns of LD across groups to aid in fine-mapping efforts that seek to identify causal alleles for disease. While this approach does have challenges, the combination of the well-powered primary stages, custom-array genotyping technologies, and a catalog of genetic variation down to low frequencies identified by the 1000 Genomes Project (1000 Genomes Project Consortium 2012) will test this hypothesis very directly in the immediate future. Furthermore, at genomic sites where an association has been established beyond a reasonable doubt, it is not unreasonable to imagine that additional variants statistically independent of the established association might exist and may segregate differently across ethnic groups, particularly if they are low-frequency or private to specific ethnicities. Thus, these data may help not only to triangulate on the gene causal for disease but also to help uncover the genetic mechanisms underlying disease predisposition across populations.

7.4.5 Integration of System-Based and "Omics" Approaches with Genetics

Given the number of established associations of human traits that have been discovered, along with the many polygenic contributions from numerous locations in the genome which have not yet met genome-wide significance, a key line of research is to utilize this collection of genetic data to uncover biological modules, pathways, and processes which contribute to disease susceptibility. Computational methods utilizing databases of text, gene expression, or protein-protein interaction networks have been applied to some success in evaluating the hypothesis that genes localizing nearby GWAS associations are more likely connected than expected by chance (Zhu et al. 2008; Raychaudhuri et al. 2009; Rossin et al. 2011). The results of applying these methods certainly suggest biological modules underlying disease, but do not always return a clear picture of the nature of those pathways (and causal candidates for them). Statistical methods such as gene-set enrichment analysis offer a complementary approach to tests on prespecified networks for the enrichment of statistical association (Segrè et al. 2010). This strategy offers a principled approach to hypothesis testing but assumes that the biological networks are known and can be specified ahead of time. Further research which incorporates functional genomics data (expression from RNA sequencing, transcription factor binding, histone occupancy, hypersensitivity assays) to further identify genes of action, molecular mechanism, tissues of importance, and pathophysiology will be increasingly available as large-scale experiments are underway (e.g., the genotype-tissue expression project http://www.genome.gov/gtex/), the ENCODE project, etc.). Given emerging evidence that suggests variants identified by GWAS are enriched for expression quantitative trait loci (Nicolae et al. 2010), integrating this information is almost certain to be of value. Recent success stories for fetal hemoglobin levels (Bauer et al. 2013) and type 2 diabetes (Pasquali et al. 2014) offer some promising early examples, whereby human genetic association data was identified within annotated elements from high-throughput assays (in both cases, enhancer elements), which were followed up and supported by additional functional studies.

7.4.6 Uniting Findings from GWAS with Sequencing Studies

While GWAS studies have and will continue, they will gradually be met with high-powered and detailed high-throughput sequencing efforts for disease traits. The primary aim of these studies is to comprehensively catalog variation at low and rare allele frequencies and to test that spectrum of variation for its relationship to disease. We should expect high power to *discover* variation implicated in diseases but dramatically low power to demonstrate compelling *association* with complex traits beyond a reasonable doubt (Guey et al. 2011). As a result, it will take some time before sequencing studies come into their own. However, the discovery of some rare or low-frequency variation identified by sequencing could certainly be integrated with findings from GWAS of complex traits in several ways; probably the most obvious is directed (rather than genome-wide) hypothesis testing for specific genes that segregate common variation implicated in the traits. Thus, methods, which expand the inference process to include both common and rare genetic variation from human studies, along with functional or other genomics information, should certainly be anticipated in the coming years.

7.5 Closing Thoughts

The field of human genetics is now engaged in systematically validating theoretical models worked out 100 years ago by R.A. Fisher, A.H. Sturtevant, T.H. Morgan, and many others, whose work formulated an expectation for the architecture of complex traits. In large part, the exciting time we live is due to the successful application of genome-wide studies that has characterized genetic variation in large numbers of individuals and tested for association in relatively unbiased (with respect to underlying biology) ways. Beyond the scientific merits of the approach, the systematic application of this central idea across a range of traits and diseases continues to revolutionize not only the business of how such science can be accomplished efficiently (in terms of technology and partnerships with industry) but also the social dynamics of how science is accomplished (e.g., the increasing emphasis on large, collaborative scientific activities and the public dissemination of results from large-scale studies).

In the process of embarking on genome-wide studies of common variation, development by the community of clear standards, pipelines, and rigor for how to relate genetic variation to complex traits has clear implications for the next wave of study designs and technologies. It is the case that next-generation sequencing studies have a number of very specific technical challenges that did not challenge genome-wide studies of common variants. As a result, these studies have an extra layer of complexity and challenge associated with them. However, we enter the development of new pipelines and best practices; we can take comfort in the knowledge that there are genetic discoveries to be made. The lessons learned along with best practices learned from GWAS give a basal understanding of what to watch out for, an expectation for the distribution of alleles (and regions of the genome implicated in disease), and open questions which the next wave of experimental practices are extraordinarily well-suited to answer.

At this stage, it is possible now to envision a world in the near future where a large number of practically actionable insights have been made across a range of traits, leading to new and improved interventions, risk prediction, and modalities for prevention. However, it is equally clear that much scientific ground remains to be traversed before that vision can be fully realized. There is the unique opportunity to translate established genetic findings into credible insight into the structure and mechanism underlying human traits and pathophysiology of disease. If successfully applied, these insights have the potential to radically and positively impact human health. If the next 10 years of biological discoveries met with only a fraction of the success as has been achieved from genetic discovery, a tremendous benefit on the health of human society is very likely to be achieved.

Acknowledgements The author is indebted to Mark J. Daly for the motivating idea and most of the data contributing to Fig. 7.1.

References

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65

Anderson CA, Soranzo N, Zeggini E, Barrett JC (2011) Synthetic associations are unlikely to account for many common disease genome-wide association signals. PLoS Biol 9:e1000580

Bauer DE, Kamran SC, Lessard S, Xu J, Fujiwara Y, Lin C, Shao Z, Canver MC, Smith EC, Pinello L, Sabo PJ, Vierstra J, Voit RA, Yuan GC, Porteus MH, Stamatoyannopoulos JA, Lettre G, Orkin SH (2013) An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science 342:253–257

Beer NL, Tribble ND, McCulloch LJ, Roos C, Johnson PR, Orho-Melander M, Gloyn AL (2009) The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. Hum Mol Genet 18:4081–4088

Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40:695–701

Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210–223

Bustamante CD, Burchard EG, De la Vega FM (2011) Genomics for the world. Nature 475:163–165

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN (2005) Demonstrating stratification in a European American population. Nat Genet 37:868–872

CARDIoGRAMplusC4D Consortium (2013) Large-scale association analysis identifies new risk loci for coronary artery disease. Nat Genet 45:25–33

Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11:415–425

Cortes A, Brown MA (2011) Promise and pitfalls of the immunochip. Arthritis Res Ther 13(1):101

Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, Abecasis GR, Barrett JC, Behrens T, Cho J et al (2011) Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet 7:e1002254

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013) Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet 45:984–994

de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 17:122–128

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, Shen MM, Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S (2005) Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. Bioinformatics 21:1958–1963

DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet 46:234–244

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8:e1000294

Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, Wheeler E, Glazer NL, Bouatia-Naji N, Gloyn AL, Lindgren CM et al (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 42:105–116

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450

Emerging Risk Factors Collaboration (2009) Major lipids, apolipoproteins, and risk of vascular disease. JAMA 302:1993–2000

Eyre-Walker A (2010) Evolution in health and medicine Sackler colloquium: genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc Natl Acad Sci USA 107(Suppl 1):1752–1756

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B et al (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316:889–894

Ge D, Fellay J, Thompson AJ, Simon JS, Shianna KV, Urban TJ, Heinzen EL, Qiu P, Bertelsen AH, Muir AJ, Sulkowski M, McHutchison JG, Goldstein DB (2009) Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature 461:399–401

Ghoussaini M, Song H, Koessler T, Al Olama AA, Kote-Jarai Z, Driver KE, Pooley KA, Ramus SJ, Kjaer SK, Hogdall E et al (2008) Multiple loci with different cancer specificities within the 8q24 gene desert. J Natl Cancer Inst 100:962–966

Gibson G (2012) Rare and common variants: twenty arguments. Nat Rev Genet 13:135–145

Global Lipids Genetics Consortium (2013) Discovery and refinement of loci associated with lipid levels. Nat Genet 45(11):1274–1283

Goldstein DB (2009) Common genetic variation and human traits. N Engl J Med 360:1696–1698

Graham RR, Kyogoku C, Sigurdsson S, Vlasova IA, Davies LR, Baechler EC, Plenge RM, Koeuth T, Ortmann WA, Hom G et al (2007) Three functional variants of IFN regulatory factor 5 (*IRF5*) define risk and protective haplotypes for human lupus. Proc Natl Acad Sci USA 104:6758–6763

Guey LT, Kravic J, Melander O, Burtt NP, Laramie JM, Lyssenko V, Jonsson A, Lindholm E, Tuomi T, Isomaa B et al (2011) Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. Genet Epidemiol 35:236–246

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367

Hirschhorn JN (2009) Genomewide association studies–illuminating biologic pathways. N Engl J Med 360:1699–1701

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5:e1000529

International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460:748–752

Iyengar SK, Elston RC (2007) The genetic basis of complex traits: rare variants or “common gene, common disease”? Methods Mol Biol. 376:71–84

Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, O’Brien SJ (2010) Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics 11:724–724

Keating BJ, Tischfield S, Murray SS, Bhangale T, Price TS, Glessner JT, Galver L, Barrett JC, Grant SF, Farlow DN, Chandrupatla HR, Hansen M, Ajmal S, Papanicolaou GJ, Guo Y, Li M, Derohannessian S, de Bakker PI, Bailey SD, Montpetit A, Edmondson AC, Taylor K, Gai X, Wang SS, Fornage M, Shaikh T, Groop L, Boehnke M, Hall AS, Hattersley AT, Frackelton E, Patterson N, Chiang CW, Kim CE, Fabsitz RR, Ouwehand W, Price AL, Munroe P, Caulfield M, Drake T, Boerwinkle E, Reich D, Whitehead AS, Cappola TP, Samani NJ, Lusis AJ, Schadt E, Wilson JG, Koenig W, McCarthy MI, Kathiresan S, Gabriel SB, Hakonarson H, Anand SS, Reilly M, Engert JC, Nickerson DA, Rader DJ, Hirschhorn JN, Fitzgerald GA (2008) Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. PLoS One 3(10):e3583

Kilpeläinen TO, Zillikens MC, Stanákova A, Finucane FM, Ried JS, Langenberg C, Zhang W, Beckmann JS, Luan J, Vandenput L, Styrkarsdottir U et al (2011) Genetic variation near IRS1 associates with reduced adiposity and an impaired metabolic profile. Nat Genet 43:753–760

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 43:520–526

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T et al (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832–838

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34:816–834

Lohmueller KE (2014) The impact of population demography and selection on the genetic architecture of complex traits. PLoS Genet 10(5):e1004379

Maher B (2008) Personal genomes: the case of the missing heritability. Nature 456:18–21

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A et al (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499–511

Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, Steinthorsdottir V, Strawbridge RJ, Khan H, Grallert H, Mahajan A, Prokopenko I, Kang HM, Dina C, Esko T, Fraser RM, Kanoni S, Kumar A, Lagou V, Langenberg C, Luan J, Lindgren CM, Müller-Nurasyid M, Pechlivanis S, Rayner NW, Scott LJ, Wiltshire S, Yengo L, Kinnunen L, Rossin EJ, Raychaudhuri S, Johnson AD, Dimas AS, Loos RJ, Vedantam S, Chen H, Florez JC, Fox C, Liu CT, Rybin D, Couper DJ, Kao WH, Li M, Cornelis MC, Kraft P, Sun Q, van Dam RM, Stringham HM, Chines PS, Fischer K, Fontanillas P, Holmen OL, Hunt SE, Jackson AU, Kong A, Lawrence R, Meyer J, Perry JR, Platou CG, Potter S, Rehnberg E, Robertson N, Sivapalaratnam S, Stančáková A, Stirrups K, Thorleifsson G, Tikkanen E, Wood AR, Almgren P, Atalay M, Benediktsson R, Bonnycastle LL, Burtt N, Carey J, Charpentier G, Crenshaw AT, Doney AS, Dorkhan M, Edkins S, Emilsson V, Eury E, Forsen T, Gertow K, Gigante B, Grant GB, Groves CJ, Guiducci C, Herder C, Hreidarsson AB, Hui J, James A, Jonsson A, Rathmann W, Klopp N, Kravic J, Krjutškov K, Langford C, Leander K, Lindholm E, Lobbens S, Männistö S, Mirza G, Mühleisen TW, Musk B, Parkin M, Rallidis L, Saramies J, Sennblad B, Shah S, Sigurðsson G, Silveira A, Steinbach G, Thorand B, Trakalo J, Veglia F, Wennauer R, Winckler W, Zabaneh D, Campbell H, van Duijn C, Uitterlinden AG, Hofman A, Sijbrands E, Abecasis GR, Owen KR, Zeggini E, Trip MD, Forouhi NG, Syvänen AC, Eriksson JG, Peltonen L, Nöthen MM, Balkau B, Palmer CN, Lyssenko

V, Tuomi T, Isomaa B, Hunter DJ, Qi L, Wellcome Trust Case Control Consortium; Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium, Shuldiner AR, Roden M, Barroso I, Wilsgaard T, Beilby J, Hovingh K, Price JF, Wilson JF, Rauramaa R, Lakka TA, Lind L, Dedoussis G, Njølstad I, Pedersen NL, Khaw KT, Wareham NJ, Keinanen-Kiukaanniemi SM, Saaristo TE, Korpi-Hyövälti E, Saltevo J, Laakso M, Kuusisto J, Metspalu A, Collins FS, Mohlke KL, Bergman RN, Tuomilehto J, Boehm BO, Gieger C, Hveem K, Cauchi S, Froguel P, Baldassarre D, Tremoli E, Humphries SE, Saleheen D, Danesh J, Ingelsson E, Ripatti S, Salomaa V, Erbel R, Jöckel KH, Moebus S, Peters A, Illig T, de Faire U, Hamsten A, Morris AD, Donnelly PJ, Frayling TM, Hattersley AT, Boerwinkle E, Melander O, Kathiresan S, Nilsson PM, Deloukas P, Thorsteinsdottir U, Groop LC, Stefansson K, Hu F, Pankow JS, Dupuis J, Meigs JB, Altshuler D, Boehnke M, MI MC, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44(9):981–990

Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM et al (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466:714–719

Myocardial Infarction Genetics Consortium (2009) Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet 41:334–341

Neale BM, Purcell S (2008) The positives, protocols, and perils of genome-wide association. Am J Med Genet B Neuropsychiatr Genet 147B:1288–1294

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6:e1000888

Orho-Melander M, Melander O, Guiducci C, Perez-Martinez P, Corella D, Roos C, Tewhey R, Rieder MJ, Hall J, Abecasis G et al (2008) Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations. Diabetes 57:3112–3121

Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42:570–575

Pasquali L, Gaulton KJ, Rodríguez-Seguí SA, Mularoni L, Miguel-Escalada I, Akerman I, Tena JJ, Morán I, Gómez-Marín C, van de Bunt M, Ponsa-Cobas J, Castro N, Nammo T, Cebola I, García-Hurtado J, Maestro MA, Pattou F, Piemonti L, Berney T, Gloyn AL, Ravassard P, Skarmeta JL, Müller F, McCarthy MI, Ferrer J (2014) Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat Genet 46:136–143

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Pulit SL, Voight BF, de Bakker PI (2010) Multiethnic genetic association studies improve power for locus discovery. PLoS One 5:e12600

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

Rabbee N, Speed TP (2005) A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 22:7–12

Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC, International Schizophrenia Consortium, Purcell SM, Sklar P, Scolnick EM, Xavier RJ, Altshuler D, Daly MJ (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet 5:e1000534

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M (2010) Genome-wide association studies in diverse populations. Nat Rev Genet 11:356–366

Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, International Inflammatory Bowel Disease Genetics Consortium, Cotsapas C, Daly MJ (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet 7:e1001273

Saxena R, Elbers CC, Guo Y, Peter I, Gaunt TR, Mega JL, Lanktree MB, Tare A, Castillo BA, Li YR et al (2012) Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci. Am J Hum Genet 90:410–425

Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19:212–219

SEARCH Collaborative Group, Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, Gut I, Lathrop M, Collins R (2008) SLCO1B1 variants and statin-induced myopathy–a genomewide study. N Engl J Med 359:789–799

Sebastiani P, Solovieff N, Puca A, Hartley SW, Melista E, Andersen S, Dworkis DA, Wilk JB, Myers RH, Steinberg MH, Montano M, Baldwin CT, Perls TT (2011) Retraction. Science 333:404

Segrè AV, DIAGRAM Consortium, MAGIC investigators, Groop L, Mootha VK, Daly MJ, Altshuler D (2010) Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet 6:e1001058

SIGMA Type 2 Diabetes Consortium (2013) Sequence variants in *SLC16A11* are a common risk factor for type 2 diabetes in Mexico. Nature 506:97–101

Simons YB, Turchin MC, Pritchard JK, Sella G (2014) The deleterious mutation load is insensitive to recent population history. Nat Genet 46(3):220–224

Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gómez-Marín C, Aneas I, Credidio FL, Sobreira DR, Wasserman NF, Lee JH, Puviindran V, Tam D, Shen M, Son JE, Vakili NA, Sung HK, Naranjo S, Acemel RD, Manzanares M, Nagy A, Cox NJ, Hui CC, Gomez-Skarmeta JL, Nóbrega MA (2014) Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*. Nature 507(7492):371–375

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J, Magi R, Randall JC et al (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 42:937–948

Tenesa A, Haley CS (2013) The heritability of human disease: estimation, uses and abuses. Nat Rev Genet 14:139–149

Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ, Johansen CT et al (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466:707–713

The Diabetes Genetics Initiative (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316:1331–1336

Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, Bakker SF, Bardella MT, Bhaw-Rosun L, Castillejo G, de la Concha, et al. (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet 43:1193–1201

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90:7–24

Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ et al (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42:579–589

Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, Burtt NP, Fuchsberger C, Li Y, Erdmann J, Frayling TM, Heid IM, Jackson AU, Johnson T, Kilpeläinen TO, Lindgren CM, Morris AP, Prokopenko I, Randall JC, Saxena R, Soranzo N, Speliotes EK, Teslovich TM, Wheeler E, Maguire J, Parkin M, Potter S, Rayner NW, Robertson N, Stirrups K, Winckler W, Sanna S, Mulas A, Nagaraja R, Cucca F, Barroso I, Deloukas P, Loos RJ, Kathiresan S, Munroe PB, Newton-Cheh C, Pfeufer A, Samani NJ, Schunkert H, Hirschhorn JN, Altshuler D, McCarthy MI, Abecasis GR, Boehnke M (2012) The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet 8(8):e1002793

Wade N (2009) Genes show limited value in predicting disease. The New York Times. http://www.nytimes.com/2009/04/16/health/research/16gene.html

Waters KM, Stram DO, Hassanein MT, Le Marchand L, Wilkens LR, Maskarinec G, Monroe KR, Kolonel LN, Altshuler D, Henderson BE, Haiman CA (2010) Consistent association of type 2 diabetes risk variants found in Europeans in diverse racial and ethnic groups. PLoS Genet 6:e1001078

Weale ME (2010) Quality control for genome-wide association studies. Methods Mol Biol 628:341–372

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678

Williams AL, Jacobs SB, Moreno-Macías H, Huerta-Chagoya A, Churchhouse C, Márquez-Luna C, García-Ortíz H, Gómez-Vázquez MJ, Burtt NP, Aguilar-Salinas CA, González-Villalpando C, Florez JC, Orozco L, Haiman CA, Tusié-Luna T, Altshuler D, Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011) Genome-wide association study identifies five new schizophrenia loci. Nat Genet 43:969–976

Wray NR, Purcell SM, Visscher PM (2011) Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol 9:e1000579

Yang J, Lee SH, Goddard ME, Visscher PM (2011a) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88:76–82

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG et al (2011b) Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet 43:519–525

Yang J, Ferreira T, Morris A, Medland SE, The GIANT Consortium, The DIAGRAM Consortium, Madden PA, Heath AC, Marin NG, Montgomery GW et al (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 44:369–375

Zeggini E, Morris A (eds) (2011) Analysis of complex disease association studies: a practical guide. Elsevier, London

Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet 40:854–861

Zhu H, Shyh-Chang N, Segrè AV, Shinoda G, Shah SP, Einhorn WS, Takeuchi A, Engreitz JM, Hagan JP, Kharas MG et al (2011) The Lin28/let-7 axis regulates glucose metabolism. Cell 147:81–94

Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci USA 109:1193–1198

Part III

Human Evolutionary Population Genetics

Inferring Human Demographic History from Genetic Data

8

Jeffrey D. Wall

Abstract

Genetics and archeology are two fields that provide complementary sources of information on our history as a species. We review what has been learned about human population history from recent genetic studies, including studies of geographic structure, population bottlenecks, and recent population growth. We also describe recent studies that highlight the importance of admixture in human history, from the mixing of different continental populations in the Americas over the past several centuries to ancient admixture between humans and Neanderthals tens of thousands of years ago.

8.1 Introduction

Making inferences about human demographic history has been one of the primary goals of human evolutionary studies for decades. Here demography refers to the history of population relationships, migrations, admixture, and changes in population size if we trace our ancestors back in time over the past 1–2 million years. Archeology and physical anthropology have contributed tremendously to our understanding of past demographic events (e.g., Jones et al. 1994), and the basic outlines of our history are well established. The genus *Homo* first arose in Africa roughly 2–2.5 million years ago, representing ancestors that were bipedal, used tools, and had brain sizes substantially larger than those of chimpanzees. Some of these early hominins then migrated to parts of tropical and temperate Eurasia 1.5–2.0 million years ago, followed by a later migration into southern Europe. More

J. D. Wall (✉)
Institute for Human Genetics, University of California, San Francisco, CA, USA
e-mail: jeff.wall@ucsf.edu

K. E. Lohmueller, R. Nielsen (eds.), *Human Population Genomics*,
https://doi.org/10.1007/978-3-030-61646-5_8

recently, anatomically modern humans first appear in the fossil record 150–200 thousand years ago (Kya) in Eastern and Southern Africa. Modern humans then spread across the rest of the inhabitable world, mostly replacing (but with low levels of hybridization and admixture) the indigenous hominin groups that they encountered. Within this broad outline though, there are many specific details that have yet to be worked out.

In this review, we will focus on what we have learned from genetic data over the past 50 years. Roughly speaking, genetic similarity can be used as a proxy for relatedness—the more similar two sequences are, the more likely they are to share a recent common ancestor. By examining DNA sequences from many individuals at once and by developing computational and statistical models for what these sequences should look like under different scenarios, we can begin to piece together aspects of our species' history in ways that other approaches cannot. This review highlights some of the insights that have come from the study of human genetic data. It is not meant to be comprehensive, but rather a personal opinion on the most important discoveries over the past several decades.

8.2 Human Population Structure

The first large-scale studies of human genetic variation involved classical markers such as blood groups and allozyme polymorphisms (e.g., Brues 1954; Edwards and Cavalli-Sforza 1964). These studies showed that the vast majority of genetic variation (~85–95%) is shared among different populations, while a much smaller fraction differentiates populations (reviewed in Cavalli-Sforza et al. 1994). Subsequent studies of microsatellite, genotype, and resequencing data (e.g., Rosenberg et al. 2002; International HapMap Consortium 2005, 2007, 2010; Wall et al. 2008; 1000 Genomes Project Consortium 2010, 2012, 2015; GenomeAsia 100K Consortium 2019) have found similar levels of population structure between human populations (as measured by F_{ST}). These results have been interpreted as human populations being much more (genetically) similar to each other than different from each other, and we now know that the level of population structure in humans is similar to what is found in many other species. At the time these results formed an important counterweight to the "candelabra model" of Coon (1962), which posited that different continental groups had been mostly isolated for hundreds of thousands of years and which was used to justify segregationist policies in the United States (Jackson 2001).

Despite the overall similarity of different human groups at specific genetic markers, the small differences at many markers in aggregate can be used to accurately distinguish between different human populations (e.g., Pritchard et al. 2000; Patterson et al. 2006). For example, Rosenberg et al. (2002) showed that 377 microsatellite markers were sufficient to distinguish individuals from the major continental groups, such as sub-Saharan Africans, Europeans, Asians, Melanesians, and Native Americans. In addition, for recently admixed individuals, these methods can estimate the proportion of ancestry that comes from multiple ancestral groups

(Pritchard et al. 2000; Falush et al. 2003). When more markers are available, such as from standard SNP genotyping chips or whole-genome sequence data, individuals from closely related populations (e.g., French vs. Spanish, Chinese vs. Japanese) can be distinguished from one another (e.g., Patterson et al. 2006; Jakobsson et al. 2008; Novembre et al. 2008; Tishkoff et al. 2009; Lu and Xu 2013; GenomeAsia 100K Consortium 2019).

The methods and studies described above made no a priori assumptions about the makeup of different human populations. For recently admixed individuals (e.g., African–Americans), more powerful methods can identify the precise portions of their genome that were inherited from different ancestral populations, provided that individuals from each of the ancestral populations have been genotyped (e.g., Sankararaman et al. 2008; Price et al. 2009; Bryc et al. 2010; Wall et al. 2011; Baran et al. 2012). These studies have confirmed the heterogenous nature of Latinos and African–Americans—for example, the estimated proportion of European ancestry in a sample of 181 Mexican controls varied from ~0% to ~100% (Fig. 8.1, Choudhry et al. 2006). In addition, detailed admixture studies have uncovered surprising examples of admixture over the past several thousand years, including low (<5%) levels of West African ancestry in Southern Europeans (Moorjani et al. 2011),

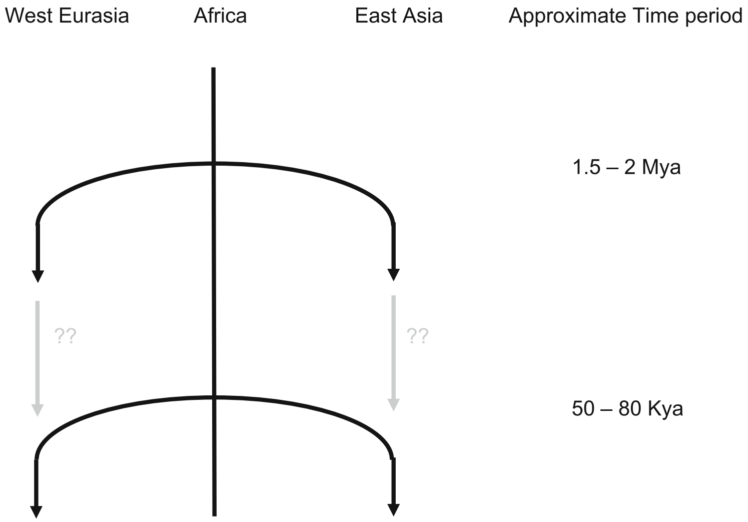


Fig. 8.1 Simplified schematic describing the qualitative differences between historical models of modern human origins. Hominins first evolve in Africa and later colonize temperate Eurasia roughly 1.5 – 2 million years ago. Much later, modern humans evolved in Africa 150–200 thousand years ago and then spread to the rest of the world starting 50–80 thousand years ago. Models differ in the amount of genetic contribution extinct hominins in Eurasia made to the contemporary gene pool. The recent African origin and replacement model predicts that this contribution was negligible, and the recent African origin and hybridization model predicts a small contribution, while the multiregional model predicts that the archaic human contribution was large

substantial Khoesan ancestry in South African Bantu speakers (Henn et al. 2011), and substantial non-African ancestry in the East African Maasai (Wall et al. 2013).

8.3 Effective Population Size

The current human population size is in excess of seven billion and is thought to have been in the millions for all of recorded history (Tellier 2009). However, genetic sequencing studies have found that extant humans are quite similar to each other, with the average sequence divergence between two individuals around 0.1% (Li and Sadler 1991; Wall et al. 2008; 1000 Genomes Project Consortium 2010, 2012, 2015). This corresponds to an effective population size (i.e., time-averaged number of breeding adults) of around 20,000–25,000, which is vastly less than our estimates of the recent census size. While there are reasons why the census size might be more than the effective population size (e.g., nonrandom mating, variance in reproductive success across individuals, unequal sex ratio, natural selection, etc., cf. Caballero 1994), the magnitude of the difference is still surprising. In contrast, great ape species with geographically restricted ranges generally have effective population sizes larger than humans (Fischer et al. 2006; Prado-Martinez et al. 2013). This suggests that despite being able to colonize virtually the whole world, our species must have had a much smaller population size (and likely a more restricted geographic range) for much of its evolutionary history. Researchers have generally considered models of recent exponential population growth from a small initial population to explain the discrepancy between census and effective population size (e.g., Slatkin and Hudson 1991; Marjoram and Donnelly 1994; Gutenkunst et al. 2009; Gravel et al. 2011).

The initial analyses of genetic variation in human mitochondrial DNA (mtDNA) estimated a relatively recent time to the most recent common ancestor (TMRCA) and observed an excess of rare variants over equilibrium expectations (Cann et al. 1987; Vigilant et al. 1991). These data were then used to estimate a time of onset of recent explosive population growth of 60–120 thousand years ago (Kya; Rogers and Harpending 1992). However, mtDNA does not experience recombination so operates as a single genetic locus. As such, the observed patterns of genetic variation are sensitive to the effects of natural selection, and inferences of demographic parameters from mtDNA data are inherently untrustworthy. When comparable sequence polymorphism data was obtained from multiple nuclear regions, the skew toward rare variants was much smaller, suggesting that substantial mid-Pleistocene population growth was unlikely (Wall and Przeworski 2000; Voight et al. 2005; Gutenkunst et al. 2009). Current estimates based on resequencing data from thousands of individuals suggest that population growth started in the late Pleistocene (20–25 Kya) and that the growth rate has accelerated (i.e., super-exponential growth) in the past 5 thousand years (e.g., Coventry et al. 2010; Nelson et al. 2012; Tennessen et al. 2012; Gazave et al. 2014; Chen et al. 2015; Gao and Keinan 2016). This is consistent with the spread of agriculture (and a more steady food supply) being the primary innovation that enabled our recent explosive population growth. We note in passing though that there are several potential confounding factors that have not yet been adequately accounted for in studies of recent population growth, including the effects of purifying and background selection on the site frequency spectrum and the effects of aggregating different populations into larger groups such as "Europeans" or "African-Americans."

8.4 Population Bottlenecks

The early analyses of human mtDNA variation assumed a founder model, whereby the TMRCA corresponded to the "founding" of a population by a few genetically similar individuals (e.g., Cann et al. 1987). This in part led to the idea that our species had undergone a drastic population bottleneck (i.e., a temporary reduction in effective population size) during the mid- to late-Pleistocene (e.g., Cann et al. 1987; Gibbons 1993). Population genetics theory, though, suggests that the specific TMRCA has very little correlation with the effective population size at that time. In particular, the recent TMRCA of mtDNA might be due to random chance or the action of natural selection. Since demographic events such as bottlenecks are expected to affect genetic variation across the whole genome, it is straightforward to analyze nuclear sequence polymorphism data to assess the strength of evidence for a species-wide bottleneck. Analyses of the HLA region, which has extremely high levels of diversity due to diversifying selection, show that the human effective population size has not dropped much below 10,000 for all of our species' history (Takahata 1993; Ayala 1995), and analyses of unlinked, putatively neutral, autosomal regions came to similar conclusions (Sjödin et al. 2012).

In contrast, there is strong evidence that at least some human populations have experienced a recent population bottleneck. Simulations suggest that population bottlenecks lead to a reduction in levels of nucleotide and haplotype diversity, an increase in levels of linkage disequilibrium (LD), and a skew in the distribution of allele frequencies (Fay and Wu 1999; Reich et al. 2001; Wall et al. 2002). Studies of microsatellite (Tishkoff et al. 1996; Rosenberg et al. 2002), single-nucleotide polymorphism (SNP, e.g., Conrad et al. 2006; Jakobsson et al. 2008), and sequence (Frisse et al. 2001; Livingston et al. 2004; Voight et al. 2005; 1000 Genomes Project 2010, 2012, 2015; Mallick et al. 2016) data consistently show that all non-African populations have less variation and more LD than all sub-Saharan African populations. The simplest explanation for this pattern is that all non-African populations have experienced at least one population bottleneck in their recent history. This is consistent with the recent African origin and replacement model of human evolution (Stringer and Andrews 1988), which posits that modern humans first evolved in sub-Saharan Africa 150–200 Kya and that modern humans later expanded and replaced (without admixture) the indigenous "archaic" humans they encountered in the rest of the world. It is also consistent with the recent African origin and hybridization model (Brauer 1989), which is identical to the previous model but allows for a limited amount of hybridization between modern and archaic humans, but not consistent with some models of modern human evolution such as the multiregional model (Wolpoff et al. 1984), which claims that modern humans evolved from archaic humans simultaneously in Africa, Europe, and Asia (see Fig. 8.1 for a schematic of these models). Under the multiregional model, there would not be strong systematic differences in levels of LD across continents, and we would expect greater degrees of population differentiation (as measured by F_{ST}) between different human continental groups.

One potential model for explaining global patterns of human genetic variation is the serial bottleneck model (Ramachandran et al. 2005; DeGiorgio et al. 2009), which posits that non-African populations experienced a series of founder bottlenecks as they left Africa and spread across the rest of the world. Under this model, populations that are further away from the putative origin of modern humans (in Eastern or Southern Africa) have experienced more bottlenecks and are expected to have decreasing levels of variation and increasing levels of LD. Genomic data are mostly consistent with these expectations (DeGiorgio et al. 2009; 1000 Genomes Project Consortium 2010, 2012, 2015; Luca et al. 2011; Mallick et al. 2016), but a denser sampling of the world's populations is needed to more fully test this theory. If the serial bottleneck model were mostly correct, it might help explain why inferences of population history based on mtDNA are qualitatively similar to ones based on autosomal data (see also Sect. 8.6).

8.5 Sex Ratio

Human effective population sizes vary not only across time but also between males and females. In principle, this gender-specific difference can be estimated by looking at patterns of genetic diversity on the X chromosome relative to the autosomes. In a randomly mating population with equal numbers of males and females, there are 3 X chromosomes for every four autosomes, and we expect to observe a 3:4 ratio in levels of diversity in X vs. autosome comparisons. However, if the numbers of breeding females and breeding males are unequal, there will be a resulting skew in this 3:4 ratio (Caballero 1994). For example, in polygynous societies, there is greater variation in male reproductive success, since many males have no wives, while some have many. This in turn reduces the male effective population size and leads to an increased ratio of X to autosome diversity.

However, there are many other factors that can affect relative levels of diversity on sex chromosomes versus autosomes. Past changes in population size (e.g., population bottlenecks) can lead to temporary variation in the ratio of sex chromosome to autosome diversity levels (Fay and Wu 1999), while recent positive selection, which decreases levels of genetic variation due to hitchhiking effects, is expected to be more effective on the X chromosome than the autosomes. (This is because recessive advantageous mutations cannot easily spread on the autosomes but can on the X chromosome since their advantageous effect is unmasked in males.) Analyses of human polymorphism data show the effects of both of these evolutionary processes (Hammer et al. 2008, 2010; Keinan et al. 2009; Arbiza et al. 2014). Overall, the ratio of X to autosome diversity increases with increasing distance from genes, consistent with greater hitchhiking effects on the X chromosome. This result is also apparent from linkage-disequilibrium-based estimates of X and autosome population sizes (Lohmueller et al. 2010). In addition, systematic differences between continental groups are likely the result of shared demographic processes such as a population bottleneck associated with the exodus of modern humans out of Africa (Arbiza et al. 2014). In summary, while it is tempting to speculate on the relative prevalence of polygyny vs. polyandry across human history, it is extremely difficult to separate out the effects of other population processes that differentially affect autosomal and sex-linked levels of diversity.

8.6 Estimating Demographic Parameters

Recently, researchers have started to develop computational and statistical methods for jointly estimating multiple demographic parameters (e.g., split times and migration rates) from DNA sequence data. Generally, these methods must navigate a tradeoff between statistical rigor and biological realism, since the statistically "optimal" approach of full maximum likelihood on autosomal sequence data from multiple individuals is computationally infeasible for the foreseeable future. Below, some of the basic approaches researchers have used are outlined, as well as some recent applications to human data.

The major computational burden of inference methods comes from the modeling of intragenic recombination. One way of reducing the computational burden is to assume a model with no recombination (e.g., Nielsen and Wakeley 2001; Drummond and Rambaut 2007; Gronau et al. 2011) and to apply this model to mtDNA data, data from short autosomal regions without visible evidence of recombination, or single diploid genome sequences. While the no-recombination assumption reduces the usefulness of these methods, this is partially counteracted by the ability to employ full-likelihood techniques (generally using Bayesian Markov chain Monte Carlo algorithms) to efficiently utilize the information contained in the data. For example, Gignoux and colleagues used a large mtDNA sequence data set to infer the rates of recent population growth in different human populations (Gignoux et al. 2011). Their results suggest that most of the population growth happened within the past 8000 years, consistent with recent analyses of autosomal sequence data (described above in Sect. 8.3).

Researchers have also tried the opposite tactic of assuming free recombination between all sites (Nielsen 2000; Marth et al. 2004; Garrigan 2009; Gutenkunst et al. 2009; Nielsen et al. 2009; Kamm et al. 2017, 2019). Under this assumption, sequence data can be summarized by the site frequency spectrum (SFS, the distribution of the number of SNPs with different allele frequencies), and the expected relative values for the SFS can be calculated computationally or analytically for complex demographic models (e.g., Garrigan 2009; Gutenkunst et al. 2009; Bhaskar et al. 2015; Kamm et al. 2017, 2019). These methods have the benefit of being able to handle genome-wide polymorphism data in a computationally efficient manner but at the cost of making a biologically unrealistic assumption and ignoring an

Table 8.1 Comparison of demographic estimates (with confidence intervals in parentheses) from Gutenkunst et al. (2009), Gravel et al. (2011), Malaspinas et al. (2016), and Steinrücken et al. (2019)

Parameter	Gutenkunst	Gravel	Malaspinas	Steinrücken
$T_{\text{EU-AS}}$	21.2 (17.2–26.5)	23 (21–27)	42 (29–55)	54 (52–55)
$T_{\text{EU-AF}}$	140 (40–270)	51 (45–69)	127 (83–171)	NA
$m_{\text{EU-AS}}$	9.6 (2.3–17.4)	3.1 (1.8–3.9)	2.2–3.6 (1.4–6.5)	NA
$T_{\text{PAP-AS}}$	NA	NA	58 (51–72)	113 (110–115)

Here $T_{\text{EU-AS}}$ refers to the split time between European and East Asian populations, $T_{\text{EU-AF}}$ refers to the split time between European and West African populations, $m_{\text{EU-AS}}$ refers to the migration rate ($\times 10^{-5}$) between European and East Asian populations, and $T_{\text{PAP-AS}}$ refers to the split time between Papuan and East Asian populations. NA refers to parameters that were not calculated. Note that Steinrücken have a different parameterization for gene flow between European and East Asian populations which is not directly comparable to the migration rates of the other studies

important component of the data (LD). Application of these methods to human data have generally focused on the demographic history of continental European, East Asian, and West African populations (e.g., Gutenkunst et al. 2009; Gravel et al. 2011; Malaspinas et al. 2016) and occasionally between European, East Asian, and Melanesian populations (Malaspinas et al. 2016; Steinrücken et al. 2019). The parameter estimates obtained by these methods can vary substantially (Table 8.1), but are not directly comparable to each other due to varying demographic and model assumptions. It is also unclear whether dates from ancient DNA studies (e.g., Fu et al. 2013) are consistent with some of the more recent estimates.

Other approaches, which can make more realistic assumptions about intragenic recombination, replace the data with one or more summary statistics in order to achieve computational tractability. These methods then use approximate Bayesian computation (e.g., Patin et al. 2009) or composite likelihood (Voight et al. 2005; Wall et al. 2009) methodology to estimate demographic parameters. Approximate likelihood methods are appealing because, unlike the previously described approaches, they have the potential to exploit the information contained in LD to help estimate demographic parameters. However, there are two main drawbacks that might limit their wider use. First, it is not easy to choose summaries of the data that are both easy to calculate and informative about demographic history. Second, these methods are all extremely computationally intensive, and they are currently unable to handle the analyses of large-scale (e.g., genome-wide) data sets without access to powerful (e.g., hundreds of nodes) computer clusters.

One final promising approach, the pairwise sequentially Markovian coalescent (PSMC) developed by Li and Durbin (2011), uses single diploid genome sequences to estimate the trajectory of past population sizes over time. Population genetics theory predicts that the distribution of coalescent times (i.e., the distribution of times until the two copies of a diploid sequence share a common ancestor) depends directly on the effective population size (both past and present) from which the sample was drawn. While coalescent times cannot be directly observed, they can be estimated using the sequence divergence between two haploid sequences—older coalescent times generally lead to greater divergence between sequences. Li and Durbin (2011) use a hidden Markov model to estimate the coalescent times between two haploid sequences sequentially across the genome (the times vary across the genome because of recombination). The distribution of times is then used to estimate the trajectory of past population sizes in the history of the population containing the sample. While currently limited to analyses of a single genome, the PSMC provides a novel way of utilizing genome-wide data to make inferences about human history, and the results so far are consistent with previous findings of a population bottleneck in non-Africans and recent growth in all populations (Li and Durbin 2011). More recent work has used the sequentially Markovian assumption in other theoretical frameworks, allowing for the analyses of larger sample sizes and population splits (Sheehan et al. 2013; Schiffels and Durbin 2014; Terhorst et al. 2017; Steinrücken et al. 2019).

8.7 Ancient Admixture

After modern humans evolved in Africa 150–200 Kya, they quickly expanded to colonize the rest of the inhabitable world. As they did, they encountered other hominin groups that already occupied the rest of Africa and Eurasia. These other groups, often called "archaic" humans, included Neanderthals, Denisovans and *Homo erectus* in Eurasia, *H. floresiensis* in island Southeast Asia, and several unnamed groups within sub-Saharan Africa (Klein 2000; Trinkaus 2005; Rightmire 2009). The extent to which the expanding modern humans interacted and interbred with the various archaic human groups is still unclear, though some interbreeding must have occurred (see below). We first discuss how ancient DNA from archaic hominins has changed our perspective on this issue. Then we describe other indirect methods for inferring the existence of ancient admixture.

Direct Evidence for Ancient Admixture The isolation and sequencing of a portion of the mtDNA hypervariable region from the Neanderthal-type specimen opened up a new avenue of research for human evolutionary studies (Krings et al. 1997). Subsequent work has generated whole mtDNA sequences from several Neanderthals (Briggs et al. 2009) and a putative *H. heidelbergensis* individual (Meyer et al. 2014), low-coverage draft genomes from a Neanderthal (Green et al. 2010) and a Denisovan (Reich et al. 2010), and high-coverage genomes from a Denisovan (Meyer et al. 2012) and two Neanderthals (Prüfer et al. 2014, 2017). Denisovans were a group of archaic humans whose only remains have been found in a single cave in Southern Siberia. They are known almost exclusively from their DNA, with very little morphological information available from the limited fossil remains that have been found (Bennett et al. 2019; Viola et al. 2019). Analyses of the Denisovan genome have shown that they are distant cousins of Neanderthals (Reich et al. 2010).

Studies of Neanderthal mtDNA found that the Neanderthal sequence was outside of the range of normal human variation, suggesting that any Neanderthal contribution to the modern human gene pool was limited (Krings et al. 1997; Serre et al. 2004). However, an analysis of the initial draft Neanderthal genome showed that Neanderthals were more closely related to all non-African populations than to sub-Saharan African populations (Green et al. 2010). The most likely explanation for this is that Neanderthals and the ancestors of non-African populations interbred and exchanged genes, perhaps in the Middle East 60–90 Kya (Sankararaman et al. 2012), leading to greater genetic similarity. This result was initially surprising, since researchers initially assumed that any interbreeding between Neanderthals and modern humans would have occurred in Europe 30–40 Kya. Instead, more detailed analyses show that East Asian individuals tend to have greater Neanderthal ancestry than do European ones (e.g., Wall et al. 2013). While this could be due to weaker purifying selection in East Asian populations (Sankararaman et al. 2014), it more likely reflects a separate Neanderthal admixture event after the initial divergence of European and East Asian populations or a recent dilution of Neanderthal ancestry in the ancestors of European populations (Kim and Lohmueller 2015; Vernot and Akey 2015). Thus far, a separate Neanderthal admixture event seems more likely (see, e.g., Villanea and Schraiber 2019).

Similar studies using the draft Denisovan genome also turned up a surprising result—aboriginal Australians, Melanesians, and Philippine Negrito groups derived 4–5% of their genome from Denisovans, whereas all other human populations tested showed much smaller levels of Denisovan ancestry (e.g., Reich et al. 2010, 2011; Sankararaman et al. 2016; Vernot et al. 2016; GenomeAsia 100K Consortium 2019). This strongly suggests that the colonization of Melanesia involved a separate migration out of Africa than did the colonization of mainland Eurasia, perhaps along the Southern coast of Asia (the "Southern route" hypothesis, reviewed by Oppenheimer 2009). If so, the historical range of the Denisovans must have stretched far south of the Siberian cave where the fossil remains were found. There is also growing evidence for multiple separate Denisovan admixture events in East and Southeast Asia (Browning et al. 2018; GenomeAsia 100K Consortium 2019; Jacobs et al. 2019).

Indirect Evidence for Ancient Admixture While the direct comparison between modern human DNA and archaic human DNA described above has been extremely informative, it has been limited by the scarcity of archaic human fossils with sufficient amounts of ancient DNA. Other, indirect methods are needed to detect potential ancient admixture between modern humans and other archaic human groups in East Asia and Africa. These methods rely on the observation that ancient admixture, if it occurred, will leave large, discrete chunks of introgressed sequence that are substantially different from orthologous modern human sequences. These chunks can then be detected by searching for unusual patterns of LD (e.g., Wall 2000; Plagnol and Wall 2006; Wall et al. 2009), in an analogous way to how local ancestry is estimated across the genomes of recently admixed individuals (e.g., Price et al. 2009; Baran et al. 2012). Analyses of human polymorphism data have found evidence for ancient admixture in sub-Saharan African populations (Garrigan et al. 2005; Hayakawa et al. 2006; Wall et al. 2009; Hammer et al. 2011; Lachance et al. 2012; Hsieh et al. 2016; Durvasula and Sankararaman 2020; Wall et al. 2019), though quantifying the amount or the timing of admixture depends on specific assumptions about modern human population structure. Taken at face value, these studies suggest that admixture between modern and archaic humans may have been a relatively common occurrence, involving Neanderthals, Denisovans, and one or more archaic human groups (in sub-Saharan Africa) for which no ancient DNA samples are currently available.

8.8 Other Ancient DNA Studies

The previous section describes what insight has been gained from studies of admixture between modern and archaic human groups. Ancient DNA studies involving only modern human samples have also been very informative about human demographic history. In particular, a comparison between an ancient DNA sample and modern human sequences can provide information on genetic affinities between the population that the ancient sample comes from and extant human groups. For example, Rasmussen et al. (2010) generated a high-coverage whole-genome sequence from a 4000-year-old hair sample from Greenland and found that the genome sequence obtained was most similar to sequences from extant Siberian populations. They concluded that the population from which the hair sample came represented a separate migration than the one that gave rise to modern Native Americans and Inuit. In another example, Rasmussen et al. (2011) sequenced a 100-year-old hair sample from an aboriginal Australian; their analyses provided additional support for the early (>50 Kya) colonization of Melanesia and the Southern route hypothesis. Other studies have also elucidated the recent movements and population affinities of groups across Eurasia (e.g., Lipson et al. 2018; McColl et al. 2018; Harney et al. 2019; Narasimhan et al. 2019; Shinde et al. 2019).

Finally, ancient DNA studies have helped to understand the relationship between Neolithic farmers in Europe and contemporary European Paleolithic hunter-gatherers roughly 6–10 Kya. Specifically, a long-standing debate has centered on the extent to which the Neolithic transition was cultural, with populations adopting farming technologies from their neighbors, versus demic, with farming populations replacing nonfarmers. So far, most of the early ancient DNA studies have suggested that demic diffusion was predominant (e.g., Bramanti et al. 2009; Malmstrom et al. 2009; Haak et al. 2010; Skoglund et al. 2012; but see Sampietro et al. 2007). Later studies have been able to quantify the relative contributions of various ancestral European groups, including ones that were previously unknown, and have highlighted the enormous complexity in European demographic history over the past ~40,000 years (e.g., Lazaridis et al. 2014; Seguin-Orlando et al. 2014; Jones et al. 2015; Lipson et al. 2017).

8.9 Conclusion

With more than 100,000 modern human genome sequences publicly available and several hundred thousand more genome sequences already generated but not yet public, we are entering a time when the amount of (modern human) genetic data available for demographic inference is essentially unlimited. The computational, statistical, and theoretical methods for dealing with these vast amounts of data are still limited though, and we anticipate that advances on the methodological front will lead to new insights into human demographic history in the coming decade. Since human fossils are rare and appreciable amounts of ancient DNA from fossils rarer still, it is difficult to predict what influence future ancient DNA studies will have on human evolutionary studies. However, based on the results of the past few years, it is safe to assume that occasional future fossil finds (with sufficient amounts of endogenous DNA) will have a major influence on our understanding of human history.

References

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65

1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–74

Arbiza L, Gottipati S, Siepel A, Keinan A (2014) Contrasting X-linked and autosomal diversity across 14 human populations. Am J Hum Genet 94:827–844

Ayala FJ (1995) The myth of Eve: molecular biology and human origins. Science 270:1930–1936

Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, Rodriguez-Cintron W, Chapela R, Ford JG, Avila PC et al (2012) Fast and accurate inference of local ancestry in Latino populations. Bioinformatics 28:1359–1367

Bennett EA, Crevecoeur I, Viola B, Derevianko AP, Shunkov MV, Grange T, Maureille B, Geigl EM (2019) Morphology of the Denisovan phalanx closer to modern humans than to Neanderthals. Sci Adv 5:eaaw3950

Bhaskar A, Wang YXR, Song YS (2015) Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Res 25:268–279

Bramanti B, Thomas MG, Haak W, Unterlaender M, Jores P, Tambets K, Antanaitis-Jacobs I, Haidle MN, Jankauskas R, Kind CJ et al (2009) Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. Science 326:137–140

Brauer G (1989) The evolution of modern humans: a comparison of the African and non-African evidence. In: Mellars P, Stringer C (eds) The human revolution: behavioural and biological perspectives on the origins of modern humans. Edinburgh University Press, Edinburgh, pp 123–154

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajkovic D, Kucan Z et al (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325(5938):318–321

Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM (2018) Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. Cell 173:53–61

Brues AM (1954) Selection and polymorphism in the A-B-O blood groups. Am J Phys Anthropol 12:559–597

Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo JM, Wambebe C, Tishkoff SA et al (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci USA 107:786–791

Caballero A (1994) Developments in the prediction of effective population size. Heredity 73:657–679

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ

Chen H, Hey J, Chen K (2015) Inferring very recent population growth rate from population-scale sequencing data: using a large-sample coalescent estimator. Mol Biol Evol 32:2996–3011

Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, Tsai HJ, Naqvi M, Phong A, Ung N et al (2006) Population stratification confounds genetic association studies among Latinos. Hum Genet 118:652–664

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 38:1251–1260

Coon CS (1962) The origins of races. Alfred A. Knopf, New York

Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, Hixson JE, Rea TJ, Muzny DM, Lewis LR et al (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun 1:131

DeGiorgio M, Jakobsson M, Rosenberg NA (2009) Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. Proc Natl Acad Sci USA 106:16057–16062

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214

Durvasula A, Sankararaman S (2020) Recovering signals of ghost archaic introgression in African populations. Sci Adv 6(7):eaax5097. https://doi.org/10.1126/sciadv.aax5097

Edwards AWF, Cavalli-Sforza LL (1964) Reconstruction of evolutionary trees. In: Heywood VH, McNeill J (eds) Phenetic and phylogenetic classification (systematics association pub. #6). London, pp 67–76

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Fay JC, Wu CI (1999) A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol Biol Evol 16:1003–1005

Fischer A, Pollack J, Thalmann O, Nickel B, Paabo S (2006) Demographic history and genetic differentiation in apes. Curr Biol 16:1133–1138

Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A (2001) Gene conversion and different population histories may explain the contrast between polymorphsim and linkage disequilibrium levels. Am J Hum Genet 69:831–843

Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, Paabo S (2013) DNA analysis of an early modern human from Tianyuan cave, China. Proc Natl Acad Sci USA 110:2223–2227

Gao F, Keinan A (2016) Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. Genetics 202:235–245

Garrigan D (2009) Composite likelihood estimation of demographic parameters. BMC Genet 10:72

Garrigan D, Mobasher Z, Severson T, Wilder JA, Hammer MF (2005) Evidence for archaic Asian ancestry on the human X chromosome. Mol Biol Evol 22:189–192

Gazave E, Ma L, Chang D, Coventry A, Gao F, Muzny D, Boerwinkle E, Gibbs RA, Sing CF, Clark AG et al (2014) Neutral genomic regions refine models of recent rapid human population growth. Proc Natl Acad Sci USA 111:757–762

GenomeAsia100K Consortium (2019) The GenomeAsia 100K project enables genetic discoveries across Asia. Nature 576:106–111

Gibbons A (1993) Pleistocene population explosions. Science 262:27–28

Gignoux CR, Henn BM, Mountain JL (2011) Rapid, global demographic expansions after the origins of agriculture. Proc Natl Acad Sci USA 108:6044–6049

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, 1000 Genomes Project, Bustamante CD (2011) Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci USA 108:11983–11988

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY et al (2010) A draft sequence of the Neandertal genome. Science 328:710–722

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. Nat Genet 43:1031–1034

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695

Haak W, Balanovsky O, Sanchez JJ, Koshel S, Zaporozhchenko V, Adler CJ, Der Sarkissian CS, Brandt G, Schwarz C, Nicklisch N et al (2010) Ancient DNA from European early neolithic farmers reveals their near eastern affinities. PLoS Biol 8:e1000536

Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD (2008) Sex-biased evolutionary forces shape genomic patterns of human diversity. PLoS Genet 4:e1000202

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Cox MP, Wall JD (2010) The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. Nat Genet 42:830–831

Hammer MF, Woerner AE, Mendez FL, Watkins JD, Wall JD (2011) Genetic evidence for archaic admixture in Africa. Proc Natl Acad Sci USA 108:15123–15128

Harney E, Nayak A, Patterson N, Joglekar P, Mushrif-Tripathy V et al (2019) Ancient DNA from the skeletons of Roopkund Lake reveals Mediterranean migrants in India. Nat Commun 10:3670

Hayakawa T, Aki I, Varki A, Satta Y, Takahata N (2006) Fixation of the human-specific CMP-N-Acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution. Genetics 172:1139–1146

Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodriguez-Botigue L, Ramachandran S, Hon L, Brisbin A et al (2011) Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci USA 108:5154–5162

Hsieh P, Woerner AE, Wall JD, Lachance J, Tishkoff SA et al (2016) Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in central African pygmies. Genome Res 26:291–300

International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

Jackson JP (2001) In ways unacademical: the reception of Carleton S. Coon's the origin of races. J History Biol 34:247–285

Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, Mondal M, Pagani L et al (2019) Multiple deeply divergent Denisovan ancestries in Papuans. Cell 177:1010–1021.e32

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R et al (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998–1003

Jones S, Martin RD, Pilbeam DR (eds) (1994) The Cambridge encylopedia of human evolution. Cambridge University Press, Cambridge

Jones ER, Gonzalez-Fortes G, Connell S, Siska V, Eriksson A et al (2015) Upper Palaeolithic genomes reveal deep roots of modern Eurasians. Nat Commun 6:8912

Kamm JA, Terhorst J, Song YS (2017) Efficient computation of the joint sample frequency spectra for multiple populations. J Comp Graph Stat 26:182–194

Kamm J, Terhorst J, Durbin R, Song YS (2019) Efficiently inferring the demographic history of many populations with allele count data. J Am Stat Assoc 115:1472–1487

Keinan A, Mullikin JC, Patterson N, Reich D (2009) Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat Genet 41:66–70

Kim BY, Lohmueller KE (2015) Selection and reduced population size cannot explain higher amounts of Neandertal ancestry in East Asian than in European human populations. Am J Hum Genet 96:454–461

Klein RG (2000) The earlier Stone age of southern Africa. S Afr Archaeol Bull 55:107–122

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Paabo S (1997) Neandertal DNA sequences and the origin of modern humans. Cell 90:19–30

Lachance J, Vernot B, Elbers CC, Ferwerda B, Froment A, Bodo JM, Lema G, Fu W, Nyambo TB, Rebbeck TR et al (2012) Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell 150:457–469

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S et al (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409–413

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493–496

Li WH, Sadler LA (1991) Low nucleotide diversity in man. Genetics 129:513–523

Lipson M, Szécsényi-Nagy A, Mallick S, Pósa A, Stégmár B et al (2017) Parallel palaeogenomic transects reveal complex genetic history of early European farmers. Nature 551:368–372

Lipson M, Cheronet O, Mallick S, Rohland N, Oxenham M et al (2018) Ancient genomes document multiple waves of migration in Southeast Asian prehistory. Science 361:92–95

Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA (2004) Pattern of sequence variation across 213 environmental response genes. Genome Res 14:1821–1831

Lohmueller KE, Degenhardt JD, Keinan A (2010) Sex-averaged recombination and mutation rates on the X chromosome: a comment on Labuda et al. Am J Hum Genet 86:978–980

Lu D, Xu S (2013) Principal component analysis reveals the 1000 genomes project does not sufficiently cover the human genetic diversity in Asia. Front Genet 4:127

Luca F, Hudson RR, Witonsky DB, Di Rienzo A (2011) A reduced representation approach to population genetic analyses and applications to human evolution. Genome Res 21:1087–1098

Malaspinas AS, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, Bergström A, Athanasiadis G, Cheng JY, Crawford JE et al (2016) A genomic history of Aboriginal Australia. Nature 538:207–214

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, Skoglund P et al (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538:201–206

Malmstrom H, Gilbert MTP, Thomas MG, Brandstrom M, Stora J, Molnar P, Andersen PK, Bendixen C, Holmlund G, Gotherstrom A et al (2009) Ancient DNA reveals lack of continuity between Neolithic hunter-gatherers and contemporary Scandinavians. Curr Biol 19:1758–1762

Marjoram P, Donnelly P (1994) Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics 136:673–683

Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166:351–372

McColl H, Racimo F, Vinner L, Demeter F, Gakuhari T et al (2018) The prehistoric peopling of Southeast Asia. Science 361:88–92

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C et al (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222–226

Meyer M, Fu Q, Aximu-Petri A, Glocke I, Nickel B, Arsuaga JL, Martinez I, Gracia A, Bermudez de Castro JM, Carbonell E, Pääbo S (2014) A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature 505:403–406

Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D (2011) The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet 7:e1001373

Narasimhan V, Patterson N, Moorjani P, Rohland N, Bernardos R et al (2019) The formation of human populations in South and Central Asia. Science 365:eaat7487

Nelson MR, Wegman D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D et al (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337:100–104

Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154:931–942

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158:885–896

Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, Indap A, Bustamante CD, Clark AG (2009) Darwinian and demographic forces affecting human protein coding genes. Genome Res 19:838–849

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR et al (2008) Genes mirror geography within Europe. Nature 456:98–101

Oppenheimer S (2009) The great arc of dispersal of modern humans: Africa to Australia. Quat Int 202:2–13

Patin E, Laval G, Barreiro LB, Salas A, Semino O, Santachiara-Benerecetti S, Kidd KK, Kidd JR, Van der Veen L, Hombert JM et al (2009) Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. PLoS Genet 5:e1000448

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

Plagnol V, Wall JD (2006) Possible ancestral structure in human populations. PLoS Genet 2:e105

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G et al (2013) Great ape genetic diversity and population history. Nature 499:471–475

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet 5:e1000519

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C et al (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43–49

Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlevic P, Hajdinjak M, Vernot B, Skov L, Hsieh P, Payregne S et al (2017) A high-coverage Neandertal genome from Vindija cave in Croatia. Science 358:655–658

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA 102:15942–15947

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, Bertalan M et al (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463:757–762

Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrectsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T et al (2011) An aboriginal Australian genome reveals separate human dispersals into Asia. Science 334:94–98

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R et al (2001) Linkage disequilibrium in the human genome. Nature 411:199–204

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL et al (2010) Genetic history of an archaic hominin group from Denisova cave in Siberia. Nature 468:1053–1060

Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AMS, Ko YC, Jinam TA, Phipps ME et al (2011) Denisovan admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet 89:516–528

Rightmire GP (2009) Out of Africa: modern human origins special feature: middle and later Pleistocene hominins in Africa and Southwest Asia. Proc Natl Acad Sci USA 106:16046–16050

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9:552–569

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385

Sampietro ML, Lao O, Caramelli D, Lari M, Pou R, Marti M, Bertranpetit J, Lalueza-Fox C (2007) Palaeogenetic evidence supports a dual model of Neolithic spreading into Europe. Proc Biol Sci 274:2161–2167

Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in admixed populations. Am J Hum Genet 82:290–303

Sankararaman S, Patterson N, Li H, Pääbo S, Reich D (2012) The date of interbreeding between Neandertals and modern humans. PLoS Genet 8:e1002947

Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J et al (2014) The genomic landscape of Neanderthal ancestry in present-da humans. Nature 507:354–357

Sankararaman S, Mallick S, Patterson N, Reich D (2016) The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. Curr Biol 26:1241–1247

Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. Nat Genet 46:919–925

Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas AS, Manica A et al (2014) Genomic structure in Europeans dating back at least 36,200 years. Science 346:1113–1118

Serre D, Langaney A, Chech M, Teschler-Nicola M, Paunovic M, Mennecier P, Hofreiter M, Possnert G, Paabo S (2004) No evidence of Neandertal mtDNA contribution to early modern humans. PLoS Biol 2:E57

Sheehan S, Harris K, Song YS (2013) Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics 194:647–662

Shinde V, Narasimhan V, Rohland N, Mallick S, Mah M et al (2019) An ancient Harappan genome lacks ancestry from steppe pastoralists or Iranian farmers. Cell 179:729–735

Sjödin P, Sjöstrand AE, Jakobsson M, Blum MGB (2012) Resequencing data provide no evidence for a human bottleneck in Africa during the penultimate glacial period. Mol Biol Evol 29:1851–1860

Skoglund P, Malmstrom H, Raghavan M, Stora J, Hall P, Willerslev E, Gilbert MT, Gotherstrom A, Jakobsson M (2012) Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science 336:466–469

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562

Steinrücken M, Kamm J, Spence JP, Song YS (2019) Inference of complex population histories using whole-genome sequences from multiple populations. Proc Natl Acad Sci USA 116:17115–17120

Stringer CB, Andrews P (1988) Genetic and fossil evidence for the origin of modern humans. Science 239:1263–1268

Takahata N (1993) Allelic genealogy and human evolution. Mol Biol Evol 10:2–22

Tellier LN (2009) Urban world history: an economic and geographical perspective. Presses de l'Universite du Quebec, Montreal

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G et al (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337:64–69

Terhorst J, Kamm JA, Song YS (2017) Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet 49:303–309

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M (1996) Global patterns of linkage disequilibrium at the *CD4* locus and modern human origins. Science 271:1380–1387

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O et al (2009) The genetic structure and history of Africans and African Americans. Science 324:1035–1044

Trinkaus E (2005) Early modern humans. Annu Rev Anthropol 34:207–230

Vernot B, Akey JM (2015) Complex history of admixture between modern humans and Neandertals. Am J Hum Genet 96:448–453

Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB et al (2016) Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science 352:235–239

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507

Villanea FA, Schraiber JG (2019) Multiple episodes of interbreeding between Neanderthal and modern humans. Nat Ecol Evol 3:39–44

Viola BT, Gunz P, Neubauer S, Slon V, Kozlikin MB, Shunkov MV et al (2019) A parietal fragment from Denisova cave. Am J Phys Anthropol 168(S68):258

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci USA 102:18508–18513

Wall JD (2000) Detecting ancient admixture in humans using sequence polymorphism data. Genetics 154:1271–1279

Wall JD, Przeworski M (2000) When did the human population size start increasing? Genetics 155:1865–1874

Wall JD, Andolfatto P, Przeworski M (2002) Testing models of selection and demography in *Drosophila simulans*. Genetics 162:203–216

Wall JD, Cox MP, Mendez FL, Woerner A, Severson T, Hammer MF (2008) A novel DNA sequence database for analyzing human demographic history. Genome Res 18:1354–1361

Wall JD, Lohmueller KE, Plagnol V (2009) Detecting ancient admixture and estimating demographic parameters in multiple human populations. Mol Biol Evol 26:1823–1827

Wall JD, Jiang R, Gignoux C, Chen GK, Eng C, Huntsman S, Marjoram P (2011) Genetic variation in native Americans, inferred from Latino SNP and resequencing data. Mol Biol Evol 28:2231–2237

Wall JD, Yang MA, Jay F, Kim SK, Durand EY, Stevison LS, Gignoux C, Woerner A, Hammer MF, Slatkin M (2013) Higher levels of Neanderthal ancestry in east Asians than in Europeans. Genetics 194:199–209

Wall JD, Ratan A, Stawiski E, GenomeAsia 100K Consortium (2019) Identification of African-specific admixture between modern and archaic humans. Am J Hum Genet 105:1254–1261

Wolpoff MH, Wu X, Thorne AG (1984) Modern *Homo sapiens* origins: a general theory of hominid evolution involving the fossil evidence from East Asia. In: Smith FH, Spencer F (eds) The origins of modern humans: a world survey of the fossil evidence. Liss, New York, NY, pp 411–483

Natural Selection, Genetic Variation, and Human Diversity

9

Leslie S. Emery and Joshua M. Akey

Abstract

Patterns of human genetic diversity observed in present-day individuals provide insight into our species evolutionary history and regions of the genome that have been influenced by natural selection. In this chapter, we discuss methods and models that have been used to better understand how selection has impacted human variation and review studies of natural selection in humans and the stories they reveal about human history.

9.1 Introduction

A comprehensive understanding of how natural selection has shaped extant patterns of human genomic variation remains an important, yet elusive, goal. Although it is becoming increasingly clear that a significant proportion of the human genome has been impacted by selection (Hahn 2008; Akey 2009), many outstanding questions remain. These questions include identifying the precise causal alleles that have been targets of selection, the timing, strength, and form of selection exerted on causal variants, the mechanistic and molecular pathways selection acts upon, and ultimately what factors led to differences in fitness among alleles at a given locus. Answering these questions not only provides a window into human history and the historical forces that have shaped our past but also provides fundamental insights into hominin evolution, mechanisms of evolutionary change, and the heritable basis of disease.

L. S. Emery · J. M. Akey (✉)
Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA
e-mail: jakey@princeton.edu

K. E. Lohmueller, R. Nielsen (eds.), *Human Population Genomics*,
https://doi.org/10.1007/978-3-030-61646-5_9

In this chapter, we discuss the prevalence of selection in the human genome and how natural selection interacts with demographic forces to shape extant patterns of variation. Focusing particularly on the action of positive selection, we will review current knowledge of where and when selection has acted upon the human genome, based on both detailed candidate gene studies and genomic scans for selection. Additionally, we will discuss advances in models of selection and statistical methods that are necessary for a comprehensive description of the action of selection on the human genome. Finally, we will highlight how an understanding of selection's effects on variation can inform studies of human population differentiation.

9.2 Amount of Selection in the Human Genome

We will begin by discussing the amount of balancing, negative, and positive selection that has been identified in genomic studies. Detailed descriptions of these types of selection and how they affect genetic variation can be found in Chap. 4.

9.2.1 Balancing Selection

The two subtypes of balancing selection—overdominance and frequency-dependent selection have been described in Chap. 4. Overdominance and frequency-dependent selection are rare for most human traits, but both are particularly important for immune-related genes (Charlesworth 2006). Several notable examples of overdominance have been described including the following: the cystic fibrosis ΔF508 allele of the *CFTR* gene may also prevent childhood asthma (Schroeder et al. 1995); phenylketonuria-causing alleles of the *PAH* gene show molecular evidence for overdominance, though the selective pressure is as yet unknown (Krawczak and Zschocke 2003); and sickle cell trait alleles and thalassemia alleles at the *HBB* gene provide resistance against malaria (Allison 1954; Quintana-Murci and Barreiro 2010).

The X-linked gene *G6PD* also shows signatures of overdominance due to selective pressure from the *Plasmodium* malaria parasites. *G6PD* encodes the enzyme glucose-6-phosphate dehydrogenase, which is responsible for replenishing supplies of NADPH from NADP$^+$. Alleles that drastically reduce the activity of the G6PD enzyme result in oxidative damage to the red blood cells, because G6PD is the only enzyme replacing NADPH in these cells (Verrelli et al. 2002). These deficiency alleles are found at high frequencies in areas with a high prevalence of malaria, including sub-Saharan Africa and the Mediterranean. Experimental studies show that deficiency alleles in both heterozygous females and hemizygous males reduce the risk of malaria infection by about half (Verrelli et al. 2002). The region exhibits unusually high levels of polymorphism consistent with balancing selection. Although this molecular signature is not completely conclusive, the confluence of molecular, phenotypic, and population genetic data make *G6PD* one of the most convincing cases of balancing selection in the human genome and one of the first to be identified by molecular evidence (Tishkoff et al. 2001; Sabeti et al. 2002; Verrelli et al. 2002).

Frequency-dependent selection is common in host-pathogen interactions, and it is known to be an important force acting on alleles at the human leukocyte antigen (HLA) genes, which will be discussed in detail below. A recent study examined SNP data in European and African American samples and found evidence for long-term, frequency-dependent selection at 60 genes, including immunity genes, keratins, membrane channels, and cellular structure genes (Hahn 2008; Andrés et al. 2009). Although balancing selection undoubtedly plays an important role in shaping variation at perhaps dozens of genes, it is likely the rarest type of selection acting on the human genome (Akey 2009; Andrés et al. 2009).

9.2.2 Negative Selection

The effects of negative selection are difficult to quantify for a single locus, so studies have focused on describing its impact genome-wide (Reed et al. 2005; Charlesworth 2006; McVicker et al. 2009; Lohmueller et al. 2011a). Unexpectedly low levels of diversity near functional elements would provide evidence for ongoing negative selection in the human genome, and this pattern is borne out by studies showing reduced variation near evolutionarily conserved elements (Schroeder et al. 1995; McVicker et al. 2009; Lohmueller et al. 2011a). Variation is even further reduced near human exons, and estimates show that selection has reduced variation by 19–26% on the autosomes and by 12–40% on the X chromosome (Krawczak and Zschocke 2003; McVicker et al. 2009). Although hitchhiking caused by positive selection also makes a contribution to this reduced variation, widespread weak purifying selection combined with positive selection in a small fraction of the genome can explain the observed patterns (Allison 1954; Reed et al. 2005; McVicker et al. 2009; Quintana-Murci and Barreiro 2010; Lohmueller et al. 2011a). We can assume, therefore, that negative selection is acting on a large proportion of the functional elements in the human genome.

9.2.3 Positive Selection

Much debate has centered on the relative contribution of purifying and positive selection affecting genetic variation (Verrelli et al. 2002; Stephan 2010). Current evidence suggests that weak negative selection is widespread, whereas positive selection has affected a smaller portion of the genome more strongly. As a test of this hypothesis, Hernandez et al. searched for a reduction of neutral genetic diversity surrounding amino acid human substitutions. This is another signature of selective sweeps (Fig. 9.1a). They found that the reduction in neutral diversity surrounding amino acid substitutions was similar to that surrounding putatively neutral substitutions. Thus, they concluded that fewer than 10% of amino acid substitutions in the human lineage were fixed by selective sweeps. A more recent

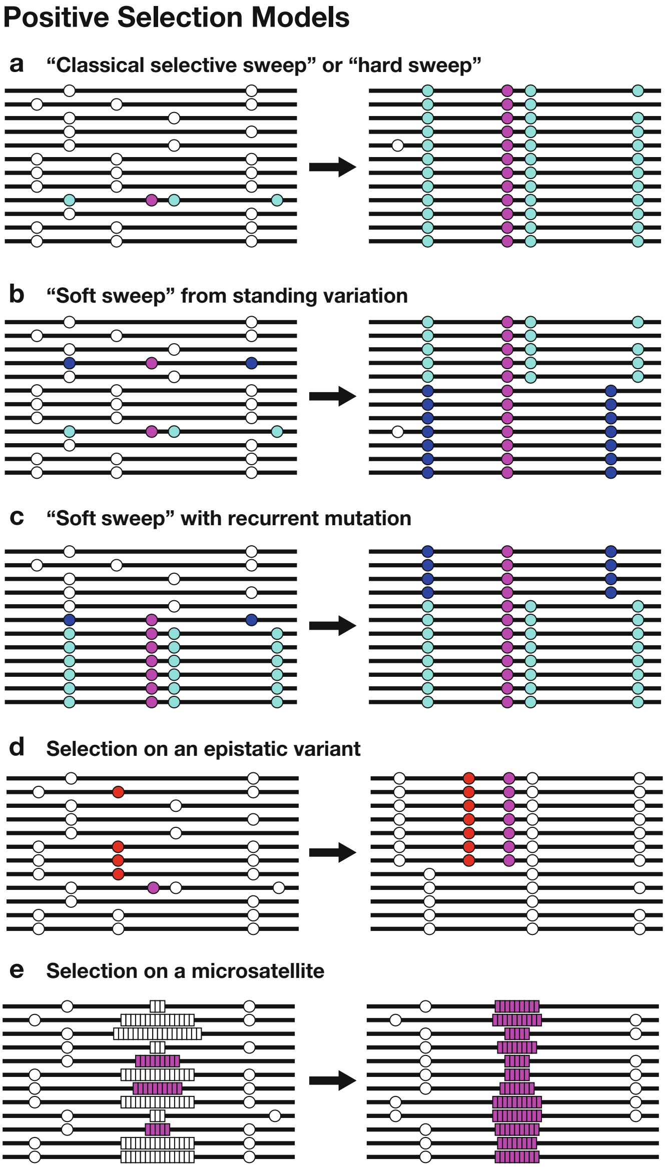

Positive Selection Models
a "Classical selective sweep" or "hard sweep"
b "Soft sweep" from standing variation
c "Soft sweep" with recurrent mutation
d Selection on an epistatic variant
e Selection on a microsatellite

Fig. 9.1 Classical and alternative models of positive selection. Pink: causal variant under positive selection. Cyan: hitchhiking variants. Blue: hitchhiking variants on a second haplotype background. Red: epistatic variant that affects the fitness of the selected variant. Left and right panels denote patterns of variation before and after selection, respectively. (**a**) A new mutation arises and is immediately advantageous. (**b**) An existing variant with multiple haplotype backgrounds is newly advantageous due to a novel selective pressure. (**c**) During a selective sweep, the advantageous variant arises again due to a mutation on a second haplotype background. (**d**) A new mutation arises that is advantageous only in the presence of another variant due to epistatic interactions. (**e**) A microsatellite locus is advantageous if the number of repeats is within a specific range; all other repeat lengths have a lower fitness

study by Enard et al. found that after controlling for genomic features, a greater fraction of amino acid substitutions may have been fixed by positive selection (Enard et al. 2014). However, Enard et al. concluded background selection is still the predominant type of selection affecting neutral diversity across the genome. Thus, while negative selection has affected a larger proportion of the genome, positive selection has had a stronger effect at the modest number of regions it has influenced (Verrelli et al. 2002; Lohmueller et al. 2011a).

Because advantageous alleles provide information about specific human adaptations, positive selection has been widely studied. Both interspecific and intraspecific adaptations are of interest for studying positive selection in humans. First, positive selection specific to the human lineage can provide information about unique human characteristics. These human-specific adaptations can be identified by looking for alleles that have become fixed in *Homo sapiens* but are absent in the other great apes (Tishkoff et al. 2001; Sabeti et al. 2002; Verrelli et al. 2002; Nielsen et al. 2007). Adaptive alleles identified by such interspecific comparisons are usually the products of completed selective sweeps, and therefore, the surrounding linked neutral variation is in the process of recovering. Recent positive selection detectable by intraspecific comparisons is more pertinent to the characterization of current human genetic variation. Adaptation of specific human populations to their respective environments and subsistence strategies is responsible for most of these ongoing or recently completed selective sweeps. Most recent work to characterize the action of selection in the human genome has focused on adaptations shaped by recent positive selection, and we will therefore focus on this topic for the remainder of the chapter.

9.3 Demography and Selection

Extant patterns of human genomic variation are jointly influenced by natural selection and human demographic history. Thus, it is necessary to understand and account for the effects of demographic history when making inferences about natural selection. To this end, here, we briefly summarize aspects of human demographic history that are relevant for interpreting signals of selection in humans. In particular, as described below, demographic history influences the amount of genetic variation available for selection to act upon and impacts the signatures imparted on standing levels of genetic variation that often complicate inferences of selection. A more detailed discussion of inference of human demography can be found in Chap. 8 by Wall.

Anatomically modern humans arose in Africa over 200,000 years ago, and a relatively small group of individuals began what is now recognized as the out-of-Africa dispersal between 50,000–100,000 years ago (Goldstein and Chikhi 2002; Cavalli-Sforza 2007). The out-of-Africa dispersal ultimately led to the peopling of the Middle East, Europe, Asia, and the Americas. Notably, the recent African origin of modern humans and subsequent out-of-Africa dispersal likely explain a number of features of human genetic variation such as lower levels of genetic diversity compared to other great apes (Kaessmann et al. 2001) and higher levels of genetic diversity in African compared to non-African populations (Vigilant et al. 1991; Jorde et al. 2000; Tishkoff et al. 2009). Genetic data is consistent with several models of dispersal, including a serial founder effect (Ramachandran et al. 2005; Deshpande et al. 2009; Barbujani and Colonna 2010) or an isolation-by-distance model (Handley et al. 2007). Regardless of the appropriate model, the essential consequence of the peopling of different geographic regions is that each subsequent dispersal was associated with a reduction in population size, often referred to as "population bottlenecks" (Cavalli-Sforza 2007). Moreover, as humans dispersed into new environments, selective pressure to adapt to emergent climates, diets, and pathogens occurred, resulting in geographically restricted adaptation. In more recent human history, particularly with the advent of agriculture, population sizes expanded dramatically, which also likely created new selective pressures (Cavalli-Sforza 2007). Indeed, it has been suggested that recent population growth has led to rapid adaptation and an accelerated rate of evolution (Hawks et al. 2007).

The evolutionarily recent divergence of human populations and relatively high rates of migration and admixture between them have the important consequence that most of the variation observed in humans is found among all populations, and not between them (Barbujani et al. 1997; Jorde et al. 2000), resulting in clinal patterns of genetic diversity (Handley et al. 2007). Population history determines which alleles are found where, thus limiting the available substrates for selection to act upon. Indeed, the geographic distribution of SNPs that are highly differentiated between populations (and possibly subject to selection) is virtually the same as the distribution of randomly chosen SNPs used to determine population structure (Coop et al. 2009). Such highly differentiated SNPs are important possible examples of local adaptation, but other local adaptations are likely the result of subtle allele frequency shifts and parallel adaptation in similar environments (Hancock et al. 2010a, b; Tennessen and Akey 2011). Accounting for the shared history between populations is very important for identifying and interpreting the evidence for local adaptions in specific populations (Tennessen and Akey 2011).

The population expansions and contractions experienced at various times in human history also complicate the detection of advantageous alleles because it is well-documented that such demographic events can produce patterns of variation similar to those left by the hitchhiking effect and background selection (Tajima 1989; Przeworski 2002; Stajich and Hahn 2005; Li et al. 2012). Population expansions can also mimic the genomic signatures of selection in certain conditions (Excoffier et al. 2009). Selection events may also be associated with demographic changes, such as population reduction caused by selective pressure, making the two forces more difficult to distinguish (Li et al. 2012). It may seem that demography and selection cannot be distinguished, but a reasonable way to address this problem is to employ an outlier approach, which assumes that demographic events will affect variation throughout the genome, while selective events act at individual loci (Cavalli-Sforza 1966). Thus, advantageous alleles can be detected by identifying loci that exhibit anomalous patterns of variation compared to the rest of the genome and are therefore outliers in the distribution of a statistic of interest (Akey et al. 2002).

Finally, human demographic changes have a significant impact on our ability to correctly identify the signatures of selection in the first place. The statistical power to detect selection depends on the interaction between local recombination rate, the dominance and selection coefficients of the selected allele, and whether the selected allele is new or selected from standing variation (Teshima et al. 2006). Recent population bottlenecks reduce the power of haplotype homozygosity tests to detect signatures of selection (Pickrell et al. 2009). Recent admixture, which is an important consideration in modern human populations, decreases the power of most neutrality test statistics but actually increases the power of Fay and Wu's *H* (Lohmueller et al. 2011b). Cryptic admixture can also obscure other signals of selection (Lohmueller et al. 2011b). The unique demographic histories of different populations also impart differing statistical power, meaning that we are better able to detect selection in some populations than in others (Lohmueller et al. 2011b). In addition to affecting statistical power, demography affects the false discovery rate of tests for selection. Recessive selected alleles, alleles selected from standing variation, and recent population bottlenecks all contribute to higher false discovery rates (Teshima et al. 2006). New approaches are being developed to address these issues, such as the composite of multiple signals (Grossman et al. 2010, 2013). Likelihood approaches, machine learning, and approximate Bayesian computation methods may also provide powerful ways to disentangle the effects of demography and selection in the near future (Li et al. 2012). However, the severe tradeoff between false-positive and false-negative rates in human populations means that some selection events will always remain beyond our detection abilities (Teshima et al. 2006; Li et al. 2012).

9.4 Empirical Studies of Positive Selection

9.4.1 Single Gene Studies

The earliest inferences of positive selection in humans focused on candidate gene studies, where there was an a priori hypothesis that a particular gene may have fitness consequences. Candidate gene studies were also the only practical way to test hypotheses about selection given the limitations in DNA sequencing technology at the time. Perhaps the most well-studied and understood example of positive selection arising from candidate gene studies is that of *LCT*, which encodes for the enzyme lactase and digests the sugar lactose (Fig. 9.2). In mammals, *LCT* is expressed, and thus lactase is produced, during infancy and early childhood when an individual is dependent on a milk diet, but the gene is turned off following weaning (Harris and Meyer 2006). This pattern of *LCT* expression is also the ancestral state in humans, but adaptive mutations have arisen in several pastoral populations that allow lactase production to continue into and throughout adulthood. Patterns of genetic variation around *LCT* show evidence for an incomplete or

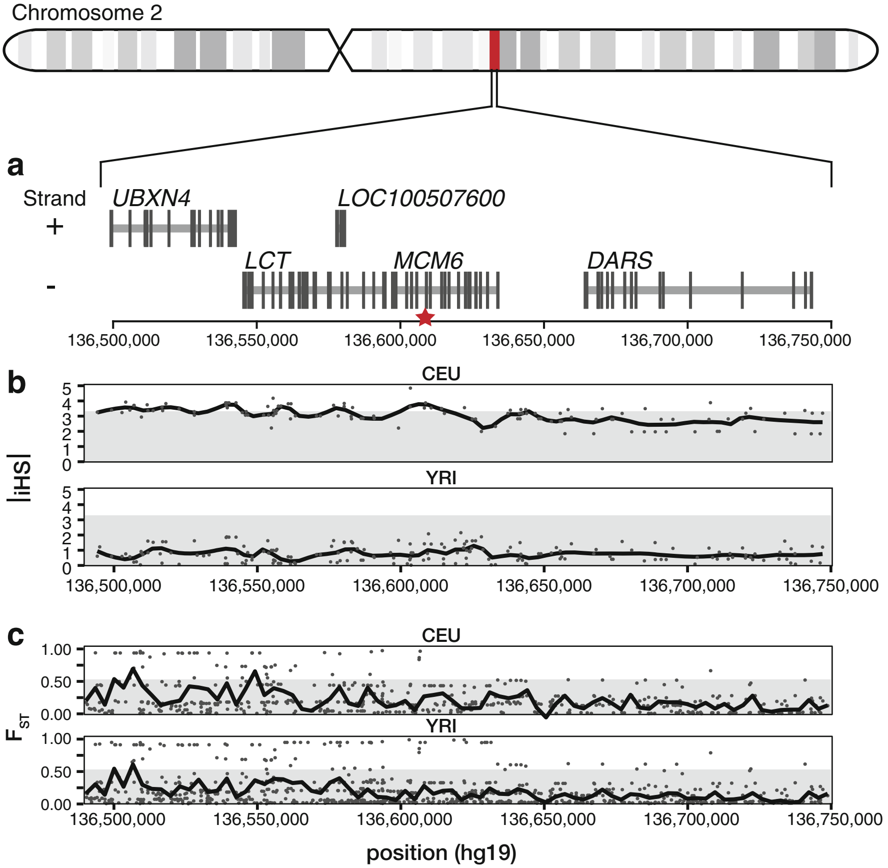


Fig. 9.2 Signal of selection surrounding the *LCT* locus. (**a**) *LCT* and nearby RefSeq genes on chromosome 2. Genes on the top line are on the + strand. and genes on the bottom line are on the—strand. The red star marks the location of rs4988235, one of the polymorphisms in an enhancer region that is associated with lactase persistence (Bersaglieri et al. 2004). (**b**) The absolute value of the integrated haplotype score (iHS) for SNPs from the HapMap phase 2 data in three populations, from Voight et al. (2006) Populations are (CEU) Utah residents with Northern and Western European ancestry and (YRI) Yoruba in Ibadan, Nigeria. The black line is a LOESS-smoothed curve of the plotted points. The gray box denotes values that exceed the 95th percentile of genome-wide iHS scores, indicating outlier scores that may show evidence for positive selection. (**c**) F_{ST}, a measure of population differentiation, for variants from the 1000 Genomes project phase 1 data. Populations are the same as above. The black line is a LOESS-smoothed curve of the plotted points. The gray box denotes values that exceed the 90th percentile of genome-wide F_{ST} values, indicating outlier values that may show evidence of extreme population differentiation

ongoing selective sweep, including long, high-frequency haplotypes (Fig. 9.2b) (Bersaglieri et al. 2004; Tishkoff et al. 2007; Enattah et al. 2008) and high levels of population differentiation (Fig. 9.2c). Haplotypes at the *LCT* gene that do not carry the advantageous mutation exhibit normal levels of variation (Tishkoff et al. 2007).

Strikingly, studies of different pastoral populations have identified several distinct causative lactase-persistence alleles in an upstream intron of the gene *MCM6* (Enattah et al. 2002, 2008; Tishkoff et al. 2007; Ingram et al. 2009; Gallego Romero et al. 2012; Jones et al. 2013), which shows enhancer activity (Olds and Sibley 2003); thus, there has been convergent adaptive evolution influencing transcriptional regulation of the *LCT* gene.

Another well-documented example of positive selection is found at the gene *DARC*, which encodes the Duffy antigen receptor for chemokines. The malaria parasite *Plasmodium vivax* requires the DARC protein to infect red blood cells, but the FY*O allele prevents expression of *DARC* exclusively in the red blood cells and therefore provides malaria resistance (Harris and Meyer 2006). FY*O is essentially fixed in sub-Saharan Africa, where *P. vivax* was likely a strong selective pressure in the past, but it is found at very low frequency in non-Africans. Phenotype and population frequency information support the hypothesis that positive selection drove FY*O to fixation in Africans, while non-Africans experienced no such selection (Hamblin et al. 2002). In concordance with this hypothesis, FY*O has a very high level of differentiation from non-Africans as measured by F_{ST}, when compared to the FY*A and FY*B alleles (Hamblin et al. 2002). The *DARC* region also shows a skew toward rare variants (positive Tajima's D) and an excess of high-frequency derived variants (Hamblin et al. 2002). Notably, *DARC* shows significantly lower variation within Africans than in non-Africans, which is opposite the pattern expected given our knowledge of human migration history (Hamblin et al. 2002).

Although other success stories arising from candidate gene studies exist, such approaches of analyzing loci individually for signatures of selection are fraught with difficulty. Most importantly, interpreting patterns of genetic variation for a single locus at a time is difficult because of the confounding influence that demographic history imparts on patterns of DNA sequence variation (Akey et al. 2004; Stajich and Hahn 2005; Akey 2009; Li et al. 2012). As technologies became available to more comprehensively survey patterns of human genetic variation in large sample sizes, candidate gene studies of selection gave way to genome-wide studies, which we discuss below.

9.4.2 Genome-Wide Scans for Selection

9.4.2.1 Insights from Genome-Wide Scans

Genome-wide scans provide a more comprehensive and unbiased approach to systematically search the genome for substrates of adaptive evolution and also enable more sophisticated approaches to disentangle the confounding effects of selection and demographic history. The first genome-wide scans for selection in humans were made possible by the development of dense genome-wide SNP genotype data and showed that selection footprint could be detected in such data (Akey et al. 2002). Genome-wide scans have proliferated, using a wide variety of statistical approaches and methods designed to identify genomic regions with unusual patterns of variation that may be indicative of selection (Table 9.1). These test statistics examine features including levels of population structure (Akey et al. 2002; Chen et al. 2010); amount and patterns of linkage disequilibrium (Sabeti et al. 2002, 2007; Voight et al. 2006); measures of the site frequency spectrum such as an excess of high frequency derived alleles or an excess of rare variants (Tajima 1989; Fay and Wu 2000; Grossman et al. 2010; Zhong et al. 2010, 2011); and differences between population- and pedigree-based recombination rate estimates (O'Reilly et al. 2008). It is important to note that no single test statistic is best suited for detecting all types of selection and their sensitivity and specificity vary widely (Ronald and Akey 2005; Biswas and Akey 2006; Sabeti et al. 2007). Additionally, some of the methods that have been developed are appropriate for identifying selective sweeps that have gone to completion (fixation of the causal adaptive allele), whereas others are most appropriate for ongoing or incomplete sweeps.

Although there are still many significant advances to be made, the dozens of published genome-wide scans for selection have provided a comprehensive first-pass picture of how, when, and where positive selection has affected the human genome. Even the earliest of studies were able to show that the signatures of natural selection are fairly evenly distributed throughout the genome rather than being found in discrete clusters (Akey et al. 2002; International HapMap Consortium 2005; Nielsen et al. 2005). Additionally, scans have thus far identified about 10% of the genome as being under recent positive selection within humans (Akey 2009). Furthermore, the number of regions identified as the targets of selection is substantial, whether testing for the signatures of complete or incomplete sweeps (Akey 2009).

In studies that have truly examined the whole genome, intergenic regions exhibit a surprising number of selection signatures (Akey 2009). Although some of these signals can be attributed to regulatory variants associated with nearby genes, some are so far from the nearest gene that other interesting explanations are possible. The functional substrate of selection in these regions could be regulatory elements, such as distant enhancers. They could also be attributed to functional non-coding RNAs, which we still know relatively little about. Another intriguing possibility is that some intergenic targets of selection are structural elements responsible for genome organization. For example, Williamson and colleagues observed many signals indicating positive selection acting on centromeres. These centromeric regions could be selected due to meiotic drive, since any variant that makes a chromosome more likely to end up in the oocyte than in a polar body during female meiosis will have a very strong selective advantage (Williamson et al. 2007). While it is evident that the targets of recent positive selection are not restricted to genic or gene-associated regions, these intergenic selected regions have been largely overlooked thus far and warrant further study.

A majority of genome-wide scans have either limited their analyses to regions around genes or focused on identified targets within or near genes. From these results, we can confidently say that a wide variety of genes in the human genome have been subject to recent positive selection. Some of these genes are also identified as the targets of older selective events, such as those detected by interspecific

Table 9.1 Studies of identifying genomic regions under selection using genome-wide scans

Study	Data	Statistic	Population(s)	# Candidates found	Candidate regions of note
1000 Genomes Project Consortium (2010)	Low coverage whole genomes	Population differentiation (F_{ST})	3 HapMap II	152 genes	*SLC24A5* *DARC* *EDAR* *SLC45A2* *PCDH15* *HERC1* *ALMS1*
Akey et al. (2002)	26.5 K SNPs	Population differentiation (F_{ST})	European-American African-American Chinese-American (The SNP Consortium)	174 regions	*CFTR* *APOB*
Amato et al. (2009)	4 M SNPs (HapMap III)	Population differentiation (F_{ST})	3 HapMap II		
Bhatia et al. (2011)	250 K or 900 K SNPs (imputed to 900 K for all)	Population differentiation (F_{ST})	Gambian African-American Nigerian		*LCT* *HBB* *HLA*
Cai et al. (2011)	4 M SNPs (HapMap II)	Shared genomic segment (SGS) (homozygosity tracts)	3 HapMap II	20 regions	
Chen et al. (2010)	3.6 M SNPs	XP-CLR (differentiation)	3 HapMap II N. and S. Europeans from POPRES	40 regions	*LCT* *HERC2*
Grossman et al. (2010)	3.1 M SNPs	CMS based on FST xp-EHH, delta-iHH, delta-DAF, and iHS	3 HapMap II	Tested only 185 candidate regions from past scans	*MATP* *LCT* *EDAR* *HERC2* *SLC24A5* *PCDH15*

(continued)

Table 9.1 (continued)

Study	Data	Statistic	Population(s)	# Candidates found	Candidate regions of note
Hancock et al. (2010a, b)	0.64M SNPs	Correlation of SNP frequency with environmental variable	61 pops: HGDP; Luhya, Maasai, Tuscan, Gujarati from HapMap III; !Kung, Amhara, Yup'ik, Chukchee, and Aborigine	?	*MTRR* *PLRP2* *KCNQ1* *ME3*
International HapMap Consortium (2005)	1 M SNPs	LD (rEHH), population differentiation (heterozygosity, low MAF, p_{excess})	3 HapMap I	926 SNPs 14 strong candidate regions	*ALMS1* *LCT*
International HapMap Consortium (2007)	3.1 M SNPs	LD (LRH and iHS)	3 HapMap II	200 regions	*HBB* *LCT* *HLA* *EDAR* *TRPV5* *TRPV6* *PCDH15*
International HapMap 3 Consortium (2010)	3.1 M SNPs	CMS based on FST, xp-EHH, delta-iHH, delta-DAF, and iHS	3 non-admixed HapMap III (TSI, LWH, MKK)	54 regions	*KITLG* *MLPH* *ANKH* *DPP7*
Johansson and Gyllensten (2008)	1.6 M SNPs	Joint analysis of population differentiation and LD (haplotype block length and *FST*)	Perlegen (Eur. Amer., Afr. Amer., Han Chinese)	23 regions	*EDAR* *LCT*
Kelley et al. (2006)	1.6 M SNPs assigned to 14,589 gene regions	Site frequency spectrum (Tajima's D)	Perlegen	385 genes	*TRPV6*

Kimura et al. (2007)	1 M SNPs	LD (rMHH)	3 HapMap I	17 regions	*DARC* *SLC24A5* *MATP* *EDAR*
Kimura et al. (2008)	500 K SNPs	LD (AREHH, rHH)	Melanesian, Polynesian		*IGF1R*
Lappalainen et al. (2010)	250 K or 500 K SNPs	LD (iHS, rEHH) and population differentiation (FST)	Finnish, Swedish, German, British—100–350 each; HapMap reference pops	60 regions	*APOE*
López Herráez et al. (2009)	900 K SNPs	LD (ln[Rsb])	HGDP (5 ea. Of 51 pops)	100 per world region considered 632 from all pops	*SLC24A5* *LCT* *TRPV5* *TRPV6* *EDAR*
Nielsen et al. (2009)	13,400 genes (sequence)	Site Frequency Spectrum (G2D, MWu-high, MWu-low) Pop diff (*F*ST)	European Americans African Americans	15 regions	*SLC45A2*
Nielsen et al. (2005)	HapMap II Chr. 2 and 148 genes from Seattle SNPs resequencing	Site frequency Spectrum (CLR, MWu)	SeattleSNPs – Europeans and Afr. Amer. 3 HapMap II		*LCT*
O'Reilly et al. (2008)	1.6 M or 1 M SNPs	LD (Ped/pop)	HapMap, Perlegen	4 regions	*CCR5*
Oleksyk et al. (2008)	200 K SNPs	Population differentiation (S^2F_{ST}) SFS (heterozygosity)	European-American, African-American	180 regions	*FOXP2* *CCR5*

(continued)

Table 1 (continued)

Study	Data	Statistic	Population(s)	# Candidates found	Candidate regions of note
Pickrell et al. (2009)	650 K SNPs	LD (iHS) Population differentiation (XP-EHH, F_{ST})	53 HGDP populations	60 regions (only looked at 10 most extreme scores in each world region)	*SLC24A5* *EDAR*
Sabeti et al. (2007)	3.1 M SNPs	LD (rEHH, iHS, XP-EHH)	3 HapMap II	22 regions	*EDAR* *SLC24A5* *HERC1* *PCDH15*
Tang et al. (2007)	1.5 M Perlegen SNPs 3.6 M HapMap SNPs	LD (ln[Rsb])	3 HapMap II, Perlegen	298 regions	*SLC24A5* *ALMS1* *EDAR* *LCT* *TRPV6*
Tennessen and Akey (2011)	100 K SNPs	Parallel divergence (F_{ST} in independent pop comparisons)	19 of HGDP pops	~5000 unique SNPs ~1300 unique genes	*ALMS1* *PCDH15* *HERC2* *APOB*
Tennessen et al. (2010)	25,769 kb exome sequence (~56,000 SNPs)		4 African, 6 European	281 previous candidate regions	*TRPV6* *SLC24A5*
Voight et al. (2006)	1 M SNPs	LD (iHS)	3 HapMap I	431 regions	*SLC24A5* *LCT*

Wang et al. (2006)	1.6 M SNPs	LD (ALnLH)	3 HapMap I, Perlegen	25,600 SNPs 112 genes	
Williamson et al. (2007)	100 K SNPs	Site frequency spectrum (CLR statistic)	Perlegen	101 gene regions	*EDAR* *HERC1* *PCDH15* *LCT*
Zhang et al. (2006)	100 K SNPs	LD (WGLRH)	3 Perlegen + YRI from HapMap	126 regions	
Zhong et al. (2011)	3.1 M SNPs	Population differentiation (xp-EHHST)	3 HapMap II	15 regions	*SLC24A5* *EDAR* *LCT* *HERC1*

comparisons looking for adaptively evolving loci during hominid evolution (Amato et al. 2009). Many of these genes have been implicated in selection by multiple studies; however, due to differences in the statistics and significance cutoffs used by each study, the overlap between sets of candidate regions is far from perfect. For instance, as of 2009, only ~14% of regions (722 regions) implicated in selection were identified by more than one genome-wide scan (Akey 2009). Encouragingly, many of the genes previously known to be targets of selection have also been detected by genome-wide scanning methods. In particular, *HBB*, *DARC*, and *LCT* have been detected by several different studies (Table 9.1). These regions serve as a "positive control," showing that genome-wide scans are identifying selected regions with independent supporting evidence for selection. Additionally, scans have identified numerous new possible targets of selection, some of which are identified in multiple studies. For example, *PCDH15* has been identified in at least six different genome-wide scans. This gene is important for retinal and cochlear function and, when mutated, can cause Usher syndrome type 1F (Ahmed et al. 2001) and autosomal recessive deafness 23 (Ahmed et al. 2003). *PCDH15* has also been associated with familial combined hyperlipidemia (Huertas-Vazquez et al. 2010). *EDAR* encodes the ectodysplasin A receptor and has been identified in 11 different genome-wide scans for selection. The EDAR protein mediates ectoderm-mesoderm interactions during development, and mutations in *EDAR* can prevent the normal formation of hair, sweat glands, and teeth in a disorder called hypohidrotic ectodermal dysplasia (Bryk et al. 2008; Mou et al. 2008). Perhaps the most compelling new signature of selection, identified by ten different studies, is *SLC24A5*, which encodes sodium/potassium/calcium exchanger 5, a protein required for proper melanogenesis. Although it was already shown that *SLC24A5* variants contribute directly to human variation in skin pigmentation (Lamason et al. 2005), previous evidence for selection was minimal.

By examining the candidate selection genes identified by studies so far, we can search for patterns in gene function. For example, gene ontology analyses of results from individual scans have suggested that the targets of recent positive selection are enriched for genes related to immunity, olfactory reception, metabolism, reproduction, pigmentation, and muscle development (Wang et al. 2006; Voight et al. 2006; Williamson et al. 2007; López Herráez et al. 2009). Furthermore, when results from multiple scans have been compiled, the best-supported candidate selection regions have proved to be enriched for various metabolic processes, signaling, development, and the cell cycle, among others (Akey 2009). Over and over again, the signatures of selection have been found in genes related to metabolic processes and pathways (López Herráez et al. 2009; Hancock et al. 2010b), which is consistent with previous hypotheses regarding the recent evolution of human metabolism (Neel 1962; Babbitt et al. 2011). One possibility is that humans have adopted a drastically different diet from that of our hominin ancestors and have therefore adopted extensively. Such recent metabolic evolution could be attributed to the development of agriculture and corresponding new food sources, as well as the availability of novel food sources in the environments encountered by humans during the peopling of the world (Hancock and Rienzo 2008; Hancock et al. 2010b; Babbitt et al. 2011). Although the signals of selection in metabolic genes are pervasive and striking, it is worth noting that many other unrelated pathways and processes have been subject to selection as well.

One important question that genome-wide scans have answered is whether the targets of the recent selection show evidence of selection in one population or all populations. Although the number of populations studied so far is disappointingly small, they usually represent a worldwide, rather than a regional, sample, so any sharing of candidate selection regions between populations is likely to represent a worldwide pattern. Several studies have so far observed a significant amount of overlap between candidate selection regions identified in disparate populations. One study found evidence of substantial levels of parallel divergence, i.e., sharing of signals of selection at a gene in phylogenetically independent populations (Tennessen and Akey 2011). Other studies that do not account for the shared ancestry between populations see even more evidence of shared targets of selection; for example, around 30% of targets identified in one study were shared between at least two worldwide regions (where the world's populations are clustered into six world regions) (Pickrell et al. 2009). Although these observations show that a non-negligible number of candidate selection regions are shared between populations to some extent, it is also clear that most detectable signatures of positive selection are geographically restricted. However, the geographic distribution of selected alleles is not markedly different from randomly selected variants (Coop et al. 2009). Furthermore, in most cases, the specific allele under selection cannot be identified; so, although a signal for selection may be seen in multiple populations, it remains to be seen whether the signal is produced by the same causative allele or not. The uniqueness of selective forces on different human populations is unsurprising in the context of human population history but has significant implications for translating medical applications between populations.

Another issue that genome-wide scans for selection hope to address is whether the signatures of recent selection we see in the human genome are largely attributable to alleles of small effect (conferring a small selective advantage) or large effect (conferring a large selective advantage). For the most part, this question remains unanswered. The strongest signals of selection are, of course, the easiest to detect; as a result, candidate selection regions identified so far are biased toward alleles of large effect (Akey et al. 2002). Since the signature of selection left in the genome is proportional to the selection coefficient for the advantageous allele, alleles of small effect are likely to remain outside of our detection ability even as methods improve. Some studies, however, have attempted to address this issue by looking for signatures of selection across many genes at once (Hancock et al. 2010b). In these studies, it is possible to detect a very subtle shift in allele frequency that is pervasive across many genes and, in some cases, correlated with variables related to environmental selective pressures (Hancock et al. 2010a).

9.4.2.2 Improving Genome-Wide Scans

Most genome-wide scans for selection performed so far have several major weaknesses that limit the applicability of their results. First, with existing methods, only the strongest signals of recent selection are detectable. While it is likely that signals of very weak selection will always remain beyond the limits of detection, the use of the outlier approach and other methods without formal significance thresholds makes it particularly difficult to sort out false positives and false negatives (Kelley et al. 2006). Furthermore, our ability to detect a signature of selection remains largely dependent upon levels of background linkage disequilibrium in the regions (O'Reilly et al. 2008; Stephan 2010). The result is that targets of recent selection are easier to identify when they are found in regions of unusually low-background LD (O'Reilly et al. 2008; Akey 2009). This issue has been addressed by some of the LD-based statistics, such as iHS and EHH, but more comprehensive methods should also be developed (O'Reilly et al. 2008). The problems of detection are exacerbated by conflicting results from various different neutrality test statistics. Since different statistics measure different kinds of signatures of selection, it is not entirely clear what combination of signals is most indicative of selection. Composite likelihood methods have been developed to address this problem, and their further development is warranted (Grossman et al. 2010).

Perhaps the most glaring drawback of existing genome-wide scans is the repeated use of the same few population panel SNP data sets, including SeattleSNPs (http://pga.gs.washington.edu), the SNP Consortium (Altshuler et al. 2000), the Perlegen SNPs (Hinds et al. 2005), and the HapMap SNPs (phase 1 (International HapMap Consortium 2005) or 2 (International HapMap Consortium et al. 2007)). Of the 31 studies presented in Table 9.1, only 8 (26%) examined other population data sets. While the use of the same data sets makes comparisons between studies easier, it also significantly reduces the applicability of those results to other data sets. Studies examining new populations are likely to provide much more novel information than yet another study on a panel of one African, one Asian, and one European population. Another major drawback to the existing data sets is the pervasive problem of ascertainment bias in SNP identification. SNP genotyping platforms were developed using nonuniform SNP discovery techniques that were biased toward identifying polymorphisms within European ancestry populations (Clark et al. 2005). Therefore, the information these SNP chips provide about variation in non-European populations is very difficult to interpret accurately. Furthermore, SNP data is too sparse to allow the detection of the causative variant underlying a signature of selection in most cases (Grossman et al. 2010). The use of whole-genome sequences for detecting selection will solve both problems of ascertainment bias and resolution, as shown in recent studies using whole-exome sequences to detect causative variants (Tennessen et al. 2010, 2012).

Finally, the usefulness of current genome-wide scans for recent selection is severely limited by the simplicity of selection models they employ. The vast majority of current genome-wide scans have been predicated on the model of the classic selective sweep, in which a new mutation is immediately beneficial upon its introduction into the population and quickly reaches fixation, sweeping along linked neutral variants (Fig. 9.1a). While some attention has focused on detecting signatures of incomplete or in-progress sweeps (Voight et al. 2006), that is the only variation on the classic sweep that has been accounted for in genome-wide scans. There are many ways in which selection has likely differed from this simplified model (Fig. 9.3), and it is unclear how robust current neutrality test statistics are to deviations from it (Przeworski et al. 2005; Pritchard et al. 2010; Hernandez et al. 2011). In particular, models of selection from standing variation are intuitively well-suited for recent human evolution, which has been characterized by adaptation to novel environments (Hermisson and Pennings 2005; Pritchard et al. 2010; Hernandez et al. 2011). In this model, a neutral allele is drifting in the population when a change in selective pressure makes it beneficial. The newly beneficial allele also reaches fixation quickly and drags along linked neutral variants, but because the allele had time to accumulate different haplotype backgrounds, the signature of the hitchhiking effect is more difficult to detect (Fig. 9.1b–c) (Hermisson and Pennings 2005; Przeworski et al. 2005). All selective sweep models are also simplified by considering only single-locus effects, despite the fact that epistatic effects can have a drastic impact on selection (Fig. 9.1d). Detection methods that employ gene networks or other multi-locus models have already begun to address this issue (Hancock et al. 2010a, b). While a few genome-wide scans have examined microsatellite variation for signatures of selection (Payseur et al. 2002; Kayser et al. 2003; Storz et al. 2004), no study to date has used a model of selection that explicitly accounts for their unique properties (e.g., Fig. 9.1e). Finally, although candidate gene analyses have revealed tantalizing signatures of selection at structural variants affecting amylase (Perry et al. 2007) and *APOBEC3* (Kidd et al. 2007), methods to identify structural variants under selection remain elusive. As selection detection methods are further developed, it will remain imperative that copy number variants, structural variants, indels, and microsatellites are not overlooked.

9.5 Selection and Population Differentiation

An outstanding question that has emerged from studies of natural selection in humans to date is how selection has influenced population differentiation (Barreiro et al. 2008). Previous interpretations of allele frequency differences within and between human populations have to a large extent been influenced by the imprecise and cultural constructions of race, which is of questionable biological meaning (Barbujani et al. 1997; Bamshad et al. 2004; Vitti et al. 2012). A more nuanced and precise understanding of worldwide patterns of genetic variation will provide insight into human evolutionary history and be an important framework for interpreting geographic patterns in the prevalence and burden of disease. Levels of genetic divergence between two populations is a complicated mosaic of many factors including time of population splitting, population sizes, rates of migration and admixture, and population-specific selective pressures. Therefore, a given level of genetic divergence between two populations could arise from myriad different evolutionary histories.

Substrates of recent natural selection are likely to show unusually high or low levels of differentiation relative to neutrally evolving genomic loci (Lewontin and

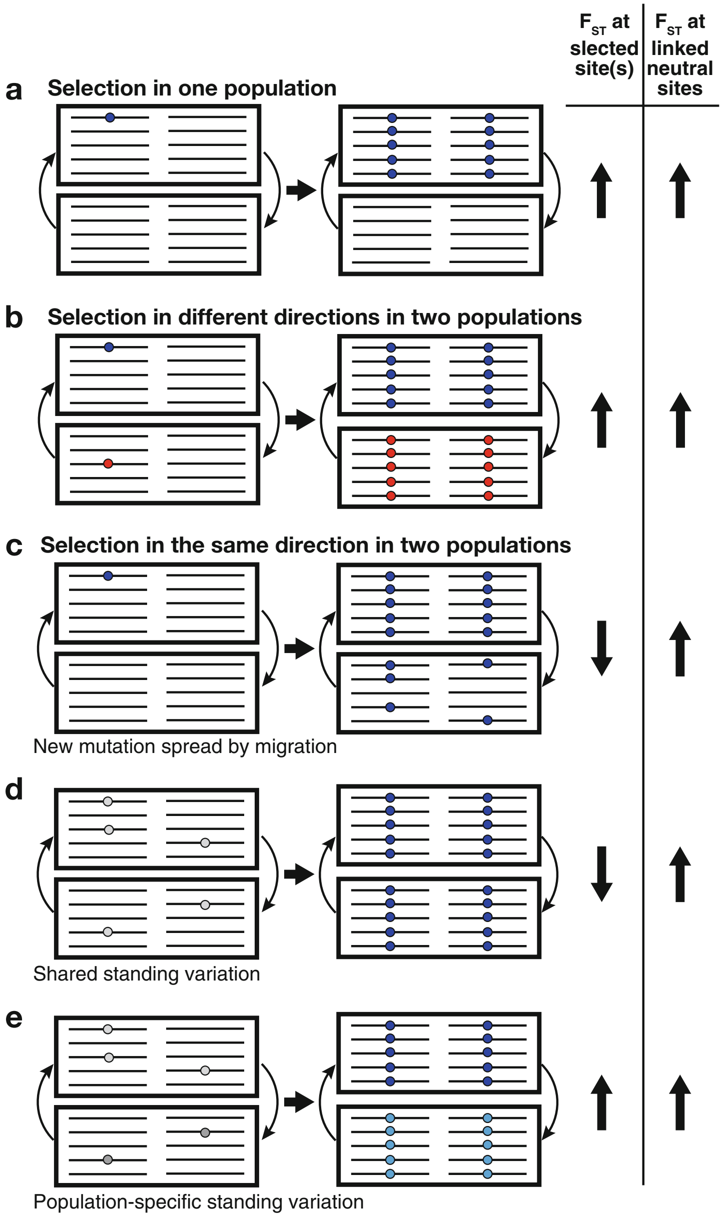

F_ST at slected site(s)
F_ST at linked neutral sites
a Selection in one population
b Selection in different directions in two populations
c Selection in the same direction in two populations
New mutation spread by migration
d
Shared standing variation
e
Population-specific standing variation

Fig. 9.3 The effects of positive selection on population differentiation between two populations, as measured by F_{ST}. Left and right panels denote patterns of variation before and after selection, respectively. Thick horizontal arrows denote selection, and thin curved arrows indicate migration. Each line is a chromosome in the population, and each circle is a variant segregating in the population. Thick vertical arrows denote increase or decrease in F_{ST} due to selection. Blue: positively selected causal variant for a high extreme trait. Red: positively selected causal variant for a low extreme trait. Cyan: a different positively selected causal variant for a high extreme trait. Light and dark gray: segregating variant that will become advantageous

Krakauer 1973; Nielsen 2005), and hitchhiking and background selection will also affect differentiation at linked neutral sites. The nature of these unusual patterns of differentiation depends largely on the type of selection acting there, as well as whether the selective pressure is acting in the same direction in each population. To understand the relationship between selection and population differentiation, we must first identify extremely differentiated (or undifferentiated) regions and then determine whether this pattern is in fact due to selection.

As discussed above, most targets of selection that have been identified to date are under selection in a subset of populations. Intuitively, population-specific balancing selection can make two populations appear to be more differentiated than they actually are at the locus of interest. The population affected by selection will maintain several intermediate frequency alleles, while the neutral variation in the other population will be randomly fixed or lost due to drift. If a locus is under positive selection in one population but neutral in another, this will also make the two populations appear to be more differentiated than they actually are (Fig. 9.3a), assuming migration rates are low (Slatkin and Wiehe 1998). The population undergoing selection will have a different major allele at a very high frequency. Taken to an extreme, the beneficial allele could be private to the selected population. In contrast, negative selection in one population, with neutrality in the other, will make it appear to be less differentiated from another population (Charlesworth et al. 1997). The population affected by negative selection will be purged of linked neutral variation, while the other population's diversity in the region will continue to be ruled by genetic drift.

If selection is acting in both populations of interest, the situation is more complicated and will depend on the origin of the selected allele, migration rates between populations, and geographic factors (Kim and Maruki 2011). In this case, long-term balancing selection can make the two populations appear to be less differentiated than they actually are at linked neutral sites (Charlesworth et al. 1997; Schierup et al. 2000). Both populations will be under selection to maintain ancient alleles, as well as to acquire and maintain many novel alleles. However, the theoretical models used to make this prediction lack much of the complexity found in empirical examples of both frequency-dependent and overdominant selection. In fact, the HLA loci, which comprise the best-known example of balancing selection in the human genome, exhibit extremely high population differentiation and an extremely old TMRCA that predates human population divergence (Meyer and Thomson 2001). This ancient TMRCA indicates very deep splits in the gene tree, which can appear to be deep population divergence. There are many HLA alleles that are private to a single population, and most alleles are found at drastically different frequencies in different populations. Measures of genetic distance between populations are extremely high at HLA compared to other genomic regions, even between very closely related populations. The only groups that show reduced distance measures are extremely isolated populations, which have experienced drastic founder effects (Vina et al. 2012). The discrepancy between the theoretical predictions and empirical observations is probably due to the use of simple biallelic models that lack the dynamic changes in selection strength that characterize balancing selection.

On the other hand, negative selection acting on both populations will result in lower differentiation between them at the site of selection. Both populations will experience selective pressure to maintain the ancestral allele and have any arising deleterious variants purged. Although specific examples of this phenomenon have not been described, genes that show modest levels of silent human polymorphism, but no divergence from chimpanzee at the protein level, are more likely to cause Mendelian diseases (Barreiro et al. 2008). We can infer from this that sites under pervasive purifying selection have limited capacity for population differentiation, mainly because allele frequencies will be low at selected variants in both population; however, linked neutral sites affected by background selection will actually show increased levels of differentiation due to lower within-population diversity levels (Charlesworth et al. 1997).

Positive selection acting on both populations can have several possible effects on levels of differentiation at a locus (Fig. 9.3b–e). Although a non-negligible proportion of identified targets of positive selection are shared among populations at a local, regional, or global scale (Pickrell et al. 2009), selection may be acting on vastly different causative alleles in different populations. Selection in two given populations could be acting on two different functions of the same gene, which would cause extremely high differentiation. Additionally, selection may be driving two populations to different extremes of the same phenotype (Fig. 9.3b), which would cause extremely high differentiation at the causative allele and linked neutral sites (Charlesworth et al. 1997). Differentiation between two populations under positive selection also depends on the variants available for selection to act upon. Even if positive selection is moving two populations toward the same trait, if different variants are available in each population, there will be high levels of resulting differentiation (Fig. 9.3e). Several notable examples of this kind of convergent evolution have been identified thus far. For instance, lactase persistence alleles of *LCT* have arisen separately in European (Bersaglieri et al. 2004), Middle Eastern (Enattah et al. 2008), and East African (Tishkoff et al. 2007) populations. Each population possesses a different lactase persistence allele—sometimes multiple alleles—with different haplotype backgrounds. So, although positive selection is driving these populations toward the same phenotypic trait, the result is still unusually high differentiation compared to the rest of the genome. A recent genome-wide scan identified numerous SNPs exhibiting unusually high levels of parallel divergence in two or more populations, suggesting convergent evolution (Tennessen and Akey 2011). Migration rates between the populations will also interact with positive selection to affect differentiation. If the beneficial allele arises in one population and then spreads to the other through migration, there will actually be a decrease in differentiation near the selected site, and the hitchhiking region will be narrower (Fig. 9.3c) (Santiago and Caballero 2005). On the other hand, if the beneficial allele is present in both populations before selection starts, as in the case of selection from standing variation, differentiation will be higher at hitchhiking neutral sites as long as migration rates are not too high (Fig. 9.3d) (Slatkin and Wiehe 1998).

The connection between phenotypic similarities and population relatedness becomes even more tenuous when discussing physical, rather than molecular, phenotypes. One notable example is adaptation to high-altitude living (Scheinfeldt and Tishkoff 2010). Although the concentration of atmospheric oxygen is the same at high altitudes, lower atmospheric pressure means that the partial pressure of oxygen is too low to induce gas exchange in the lungs. Without adaptation, long-term high-altitude living would be impossible. Populations living in the Andes, the Tibetan Plateau, and the Ethiopian Highlands all exhibit a lack of hypoxia at high altitudes. This shared phenotype is the result of convergent evolution and could be expected to cause genetic similarities between these populations that are otherwise distantly related. When the phenotype is examined more closely on a molecular level, however, Andeans, Ethiopian Highlanders, and Tibetans may have different hemoglobin and oxygen saturation levels (Beall et al. 2002; Huerta-Sanchez et al. 2013). Closer examination shows that even populations exhibiting similar molecular phenotypes have adaptations at different genes involved in the same pathway (the hypoxia-inducible factor 1, or HIF-1 pathway) (Bigham et al. 2009; Xu et al. 2011; Scheinfeldt et al. 2012; Huerta-Sanchez et al. 2013). Clearly, even detailed physiological phenotypes are not always a reliable indicator of population differentiation due to positive selection.

While much attention has been focused on visible phenotypic differentiation between populations (e.g., hair and skin pigmentation, hair texture, stature, and facial features), this focus is almost entirely unwarranted. Genome-wide scans for the targets of positive selection have detected the signatures of selection in genes associated with visible phenotypes—notably skin pigmentation via *SLC24A5* and hair texture at *EDAR*—but these signals are far outnumbered by those of other pathways, such as metabolic processes, immune responses, and olfaction (Akey 2009). While it is not entirely clear how much of human genetic variation is represented in visible phenotypic differences, we can speculate that the majority of molecular differences have no immediately visible physical manifestation. Therefore, the majority of population differentiation, whether due to selection or not, is likely to be found in specific molecular phenotypes that require detailed measurements to detect. Furthermore, the amount of population differentiation seen at selected loci is extreme compared to the rest of the genome and should not be used to draw sweeping conclusions about population differences. The temptation to conjecture about the meaning and significance of these selected population differences is high, but there is potential for such conjectures to be misused for harmful and discriminatory purposes (Vitti et al. 2012). Researchers should therefore interpret population adaptations with restraint and the use of appropriate supporting data.

9.6 Conclusions

The study of human evolution sits poised on the cusp of a new and comprehensive understanding of how natural selection and demographic history have influenced human genetic variation. The past century has seen great strides in this endeavor, including the description of the hitchhiking effect, background selection, and the development of models to describe positive, negative, and balancing selection. These developments have provided new and surprising insights into population differentiation, the persistence of deleterious and disease-causing alleles, and the peopling of the world. Studies of convergent evolution in disparate human populations on such traits as lactase persistence and living at high altitudes have exhibited the stochastic nature of adaptive evolution.

Empirical studies have provided an enormous body of knowledge on the signatures of selection across the human genome. These inferences show, for example, that negative selection has been weak but pervasive across much of the genome. Balancing selection has proven to be an important force at a handful of loci, with its complex interaction between hitchhiking and background selection leaving unique and striking patterns in patterns of local genetic variation. The influence of positive selection has significantly influenced patterns of human genetic variation, both directly and indirectly through genetic hitchhiking. Targets of adaptive evolution are scattered across the human genome, affecting many gene processes, pathways, and networks. Thus, human populations have adapted to their environments in ways numerous and diverse (Hahn 2008).

Even in the face of these vast new areas of scientific knowledge, the bulk of our insight into these topics is yet to come. Advances in next-generation sequencing technology have enabled the sequencing of hundreds of thousands of human genomes from increasingly diverse worldwide populations. Advances in statistical methods for detecting the signatures of selection will improve sensitivity and specificity of "candidate selected loci." Additionally, the development of more realistic models of selection will account for epistatic effects between genes, selection from standing variation, recurrent mutation, and selection at CNVs, indels, microsatellites, and other structural variants. Finally, armed with a detailed map of the signatures of selection across the human genome, we will be able to delineate the phenotypic consequences of adaptive evolution, make more accurate predictions of genetic risk for individuals, and gain insights into the connection between human populations and their evolutionary environments.

References

1000 Genomes Project Consortium (2010) A map of human genome variation from population - 422 scale sequencing. Nature 467:1061–1073

Ahmed Z, Riazuddin S, Bernstein SL et al (2001) Mutations of the protocadherin gene *PCDH15* cause Usher syndrome type 1F. Am J Hum Genet 69:25–34. https://doi.org/10.1086/321277

Ahmed ZM, Riazuddin S, Ahmad J et al (2003) *PCDH15* is expressed in the neurosensory epithelium of the eye and ear and mutant alleles are responsible for both *USH1F* and *DFNB23*. Hum Mol Genet 12:3215–3223. https://doi.org/10.1093/hmg/ddg358

Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19:711–722. https://doi.org/10.1101/gr.086652.108

Akey JM, Zhang G, Zhang K et al (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805–1814. https://doi.org/10.1101/gr.631202

Akey JM, Eberle MA, Rieder MJ et al (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2:e286. https://doi.org/10.1371/journal.pbio.0020286

Allison AC (1954) Protection afforded by sickle-cell trait against subtertian malareal infection. Br Med J 1:290–294

Altshuler D, Pollara VJ, Cowles CR et al (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407:513–516. https://doi.org/10.1038/35035083

Amato R, Pinelli M, Monticelli A et al (2009) Genome-wide scan for signatures of human population differentiation and their relationship with natural selection, functional pathways and diseases. PLoS One 4:e7927. https://doi.org/10.1371/journal.pone.0007927

Andrés AM, Hubisz MJ, Indap A et al (2009) Targets of balancing selection in the human genome. Mol Biol Evol 26:2755–2764. https://doi.org/10.1093/molbev/msp190

Babbitt CC, Warner LR, Fedrigo O et al (2011) Genomic signatures of diet-related shifts during human origins. Proc R Soc B Biol Sci 278:961–969

Bhatia G, Patterson N, Pasaniuc B, Zaitlen N, Genovese G, Pollack S, Mallick S, Myers S, Tandon A, Spencer C, Palmer CD, Adeyemo AA, Akylbekova EL, Cupples LA, Divers J, Fornage M, Kao WH, Lange L, Li M, Musani S, Mychaleckyj JC, Ogunniyi A, Papanicolaou G, Rotimi CN, Rotter JI, Ruczinski I, Salako B, Siscovick DS, Tayo BO, Yang Q, McCarroll S, Sabeti P, Lettre G, De Jager P, Hirschhorn J, Zhu X, Cooper R, Reich D, Wilson JG, Price AL (2011 Sep 9) Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. Am J Hum Genet 89(3):368–381. https://doi.org/10.1016/j.ajhg.2011.07.025.

Bamshad M, Wooding S, Salisbury BA, Stephens JC (2004) Deconstructing the relationship between genetics and race. Nat Rev Genet 5:598–609. https://doi.org/10.1038/nrg1401

Barbujani G, Colonna V (2010) Human genome diversity: frequently asked questions. Trends Genet 26:285–295. https://doi.org/10.1016/j.tig.2010.04.002

Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL (1997) An apportionment of human DNA diversity. Proc Natl Acad Sci 94:4516–4519

Barreiro LB, Laval G, Quach H et al (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40:340–345. https://doi.org/10.1038/ng.78

Beall CM, Decker MJ, Brittenham GM et al (2002) An Ethiopian pattern of human adaptation to high-altitude hypoxia. Proc Natl Acad Sci 99:17215–17218. https://doi.org/10.1073/pnas.252649199

Bersaglieri T, Sabeti PC, Patterson N et al (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74:1111–1120. https://doi.org/10.1086/421051

Bigham AW, Mao X, Mei R et al (2009) Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. Hum Genomics 4:79–90

Biswas S, Akey JM (2006) Genomic insights into positive selection. Trends Genet 22:437–446. https://doi.org/10.1016/j.tig.2006.06.005

Bryk J, Hardouin E, Pugach I et al (2008) Positive selection in east Asians for an *EDAR* allele that enhances NF-kappaB activation. PLoS One 3:e2209. https://doi.org/10.1371/journal.pone.0002209

Cai Z, Camp NJ, Cannon-Albright L, Thomas A (2011 Jun) Identification of regions of positive selection using Shared Genomic Segment analysis. Eur J Hum Genet 19(6):667–671. https://doi.org/10.1038/ejhg.2010.257

Cavalli-Sforza LL (1966) Population structure and human evolution. Proc R Soc B Biol Sci 164:362–379

Cavalli-Sforza LL (2007) Human evolution and its relevance for genetic epidemiology. Annu Rev Genom Human Genet 8:1–15

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2:e64. https://doi.org/10.1371/journal.pgen.0020064

Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res 70:155–174

Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Res 20:393–402. https://doi.org/10.1101/gr.100545.109

Clark AG, Hubisz MJ, Bustamante CD et al (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15:1496–1502. https://doi.org/10.1101/gr.4107905

Coop G, Pickrell JK, Novembre J et al (2009) The role of geography in human adaptation. PLoS Genet 5:e1000500. https://doi.org/10.1371/journal.pgen.1000500

Deshpande O, Batzoglou S, Feldman M (2009) A serial founder effect model for human settlement out of Africa. Proc R Soc B Biol Sci 276:291–300

Enard D, Messer PW, Petrov DA (2014) Genome-wide signals of positive selection in human evolution. Genome Res 24:885. https://doi.org/10.1101/gr.164822.113

Enattah NS, Sahi T, Savilahti E et al (2002) Identification of a variant associated with adult-type hypolactasia. Nat Genet 30:233–237. https://doi.org/10.1038/ng826

Enattah NS, Jensen TGK, Nielsen M et al (2008) Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. Am J Hum Genet 82:57–72. https://doi.org/10.1016/j.ajhg.2007.09.012

Excoffier L, Foll M, Petit R (2009) Genetic consequences of range expansions. Annu Rev Ecol Evol Syst 40:481–501

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Gallego Romero I, Basu Mallick C, Liebert A et al (2012) Herders of Indian and European cattle share their predominant allele for lactase persistence. Mol Biol Evol 29:249–260. https://doi.org/10.1093/molbev/msr190

Goldstein DB, Chikhi L (2002) Human migrations and population structure: what we know and why it matters. Annu Rev Genom Human Genet 3:129–152. https://doi.org/10.1146/annurev.genom.3.022502.103200

Grossman SR, Shlyakhter I, Shylakhter I et al (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327:883–886. https://doi.org/10.1126/science.1183863

Grossman SR, Andersen KG, Shlyakhter I et al (2013) Identifying recent adaptations in large-scale genomic data. Cell 152:703–713. https://doi.org/10.1016/j.cell.2013.01.035

Hahn M (2008) Toward a selection theory of molecular evolution. Evolution 62:255–265

Hamblin MT, Thompson EE, Di Rienzo A (2002) Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70:369–383. https://doi.org/10.1086/338628

Hancock AM, Rienzo AD (2008) Detecting the genetic signature of natural selection in human populations: models, methods, and data. Annu Rev Anthropol 37:197–217. https://doi.org/10.1146/annurev.anthro.37.081407.085141

Hancock AM, Alkorta-Aranburu G, Witonsky DB, Di Rienzo A (2010a) Adaptations to new environments in humans: the role of subtle allele frequency shifts. Philos Trans R Soc Lond Ser B Biol Sci 365:2459–2468. https://doi.org/10.1098/rstb.2010.0032

Hancock AM, Witonsky DB, Ehler E et al (2010b) Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proc Natl Acad Sci 107(Suppl 2):8924–8930. https://doi.org/10.1073/pnas.0914625107

Handley LJL, Manica A, Goudet J, Balloux F (2007) Going the distance: human population genetics in a clinal world. Trends Genet 23:432–439. https://doi.org/10.1016/j.tig.2007.07.002

Harris EE, Meyer D (2006) The molecular signature of selection underlying human adaptations. Am J Phys Anthropol Suppl 43:89–130. https://doi.org/10.1002/ajpa.20518

Hawks J, Wang ET, Cochran GM et al (2007) Recent acceleration of human adaptive evolution. Proc Natl Acad Sci 104:20753–20758. https://doi.org/10.1073/pnas.0707650104

Hermisson J, Pennings P (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169:2335–2352

Hernandez RD, Kelley JL, Elyashiv E et al (2011) Classic selective sweeps were rare in recent human evolution. Science 331:920–924. https://doi.org/10.1126/science.1198878

Hinds DA, Stuve LL, Nilsen GB et al (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079. https://doi.org/10.1126/science.1105436

Huerta-Sanchez E, DeGiorgio M, Pagani L et al (2013) Genetic signatures reveal high-altitude adaptation in a set of ethiopian populations. Mol Biol Evol 30:1877–1888. https://doi.org/10.1093/molbev/mst089

Huertas-Vazquez A, Plaisier CL, Geng R et al (2010) A nonsynonymous SNP within *PCDH15* is associated with lipid traits in familial combined hyperlipidemia. Hum Genet 127:83–89. https://doi.org/10.1007/s00439-009-0749-z

Ingram CJE, Mulcare CA, Itan Y et al (2009) Lactose digestion and the evolutionary genetics of lactase persistence. Hum Genet 124:579–591. https://doi.org/10.1007/s00439-008-0593-6

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320. https://doi.org/10.1038/nature04226

International HapMap Consortium, Frazer KA, Ballinger DG et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861. https://doi.org/10.1038/nature06258

International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in 822 diverse human populations. Nature 467:52–58

Johansson A, Gyllensten U (2008 Jun) Identification of local selective sweeps in human populations since the exodus from Africa. Hereditas 145(3):126–137. https://doi.org/10.1111/j.0018-0661.2008.02054.x

Jones BL, Raga TO, Liebert A et al (2013) Diversity of lactase persistence alleles in Ethiopia: signature of a soft selective sweep. Am J Hum Genet 93:538–544. https://doi.org/10.1016/j.ajhg.2013.07.008

Jorde LB, Watkins WS, Bamshad MJ et al (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66:979–988. https://doi.org/10.1086/302825

Kaessmann H, Wiebe V, Weiss G, Pääbo S (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat Genet 27:155–156. https://doi.org/10.1038/84773

Kayser M, Brauer S, Stoneking M (2003) A genome scan to detect candidate regions influenced by local natural selection in human populations. Mol Biol Evol 20:893–900. https://doi.org/10.1093/molbev/msg092

Kelley JL, Madeoy J, Calhoun JC et al (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res 16:980–989. https://doi.org/10.1101/gr.5157306

Kidd JM, Newman TL, Tüzün E et al (2007) Population stratification of a common *APOBEC* gene deletion polymorphism. PLoS Genet 3:e63. https://doi.org/10.1371/journal.pgen.0030063

Kim Y, Maruki T (2011) Hitchhiking effect of a beneficial mutation spreading in a subdivided population. Genetics 189:213–226. https://doi.org/10.1534/genetics.111.130203

Kimura R, Fujimoto A, Tokunaga K, Ohashi J (2007 Mar 14) A practical genome scan for population-specific strong selective sweeps that have reached fixation. PLoS One 2(3):e286. https://doi.org/10.1371/journal.pone.0000286.

Kimura R, Ohashi J, Matsumura Y, Nakazawa M, Inaoka T, Ohtsuka R, Osawa M, Tokunaga K (2008 Aug) Gene flow and natural selection in oceanic human populations inferred from genome-wide SNP typing. Mol Biol Evol 25(8):1750–1761. https://doi.org/10.1093/molbev/msn128

Krawczak M, Zschocke J (2003) A role for overdominant selection in phenylketonuria? Evidence from molecular data. Hum Mutat 21:394–397. https://doi.org/10.1002/humu.10205

Lamason RL, Mohideen M-APK, Mest JR et al (2005) *SLC24A5*, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310:1782–1786. https://doi.org/10.1126/science.1116238

Lappalainen T, Salmela E, Andersen PM, Dahlman-Wright K, Sistonen P, Savontaus ML, Schreiber S, Lahermo P, Kere J (2010 Apr) Genomic landscape of positive natural selection in Northern European populations. Eur J Hum Genet 18(4):471–478. https://doi.org/10.1038/ejhg.2009.184

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175–195

Li J, Li H, Jakobsson M et al (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? Mol Ecol 21:28–44. https://doi.org/10.1111/j.1365-294X.2011.05308.x

Lohmueller KE, Albrechtsen A, Li Y et al (2011a) Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet 7:e1002326. https://doi.org/10.1371/journal.pgen.1002326

Lohmueller KE, Bustamante CD, Clark AG (2011b) Detecting directional selection in the presence of recent admixture in African Americans. Genetics 187:823. https://doi.org/10.1534/genetics.110.122739

López Herráez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, Nandineni MR, Gross A, Scholz M, Stoneking M (2009 Nov 18) Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. PLoS One 4(11):e7888. https://doi.org/10.1371/journal.pone.0007888.

McVicker G, Gordon D, Davis C, Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet 5:e1000471. https://doi.org/10.1371/journal.pgen.1000471

Meyer D, Thomson G (2001) How selection shapes variation of the human major histocompatibility complex: a review. Ann Hum Genet 65:1–26

Mou C, Thomason HA, Willan PM et al (2008) Enhanced ectodysplasin-A receptor (*EDAR*) signaling alters multiple fiber characteristics to produce the East Asian hair form. Hum Mutat 29:1405–1411. https://doi.org/10.1002/humu.20795

Neel JV (1962) Diabetes mellitus: a “thrifty” genotype rendered detrimental by “progress”? Am J Hum Genet 14:353–362

Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420

Nielsen R, Williamson S, Kim Y et al (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15:1566–1575. https://doi.org/10.1101/gr.4252305

Nielsen R, Hellmann I, Hubisz M et al (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8:857–868. https://doi.org/10.1038/nrg2187

Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, Indap A, Bustamante CD, Clark AG (2009 May) Darwinian and demographic forces affecting human protein coding genes. Genome Res 19(5):838–849. https://doi.org/10.1101/gr.088336.108.

Oleksyk TK, Zhao K, De La Vega FM, Gilbert DA, O’Brien SJ, Smith MW (2008 Mar 5) Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. PLoS One 3(3):e1712. https://doi.org/10.1371/journal.pone.0001712.

O’Reilly PF, Birney E, Balding DJ (2008) Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome Res 18:1304–1313. https://doi.org/10.1101/gr.067181.107

Olds LC, Sibley E (2003) Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. Hum Mol Genet 12:2333–2340. https://doi.org/10.1093/hmg/ddg244

Payseur BA, Cutter AD, Nachman MW (2002) Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol Biol Evol 19:1143–1153

Perry GH, Dominy NJ, Claw KG et al (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39:1256–1260. https://doi.org/10.1038/ng2123

Pickrell JK, Coop G, Novembre J et al (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19:826–837. https://doi.org/10.1101/gr.087577.108

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20:R208–R215. https://doi.org/10.1016/j.cub.2009.11.055

Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 160:1179–1189

Przeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. Evolution 59:2312–2323

Quintana-Murci L, Barreiro LB (2010) The role played by natural selection on Mendelian traits in humans. Ann N Y Acad Sci 1214:1–17. https://doi.org/10.1111/j.1749-6632.2010.05856.x

Ramachandran S, Deshpande O, Roseman C (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci 102:15942–15947

Reed FA, Akey JM, Aquadro CF (2005) Fitting background-selection predictions to levels of nucleotide variation and divergence along the human autosomes. Genome Res 15:1211–1221. https://doi.org/10.1101/gr.3413205

Ronald J, Akey JM (2005) Genome-wide scans for loci under selection in humans. Hum Genomics 2:113–125

Sabeti PC, Reich DE, Higgins JM et al (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837. https://doi.org/10.1038/nature01140

Sabeti PC, Varilly P, Fry B et al (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918. https://doi.org/10.1038/nature06250

Santiago E, Caballero A (2005) Variation after a selective sweep in a subdivided population. Genetics 169:475–483. https://doi.org/10.1534/genetics.104.032813

Scheinfeldt LB, Tishkoff SA (2010) Living the high life: high-altitude adaptation. Genome Biol 11:133. https://doi.org/10.1186/gb-2010-11-9-133

Scheinfeldt LB, Soi S, Thompson S et al (2012) Genetic adaptation to high altitude in the Ethiopian highlands. Genome Biol 13:R1. https://doi.org/10.1186/gb-2012-13-1-r1

Schierup MH, Charlesworth D, Vekemans X (2000) The effect of hitch-hiking on genes linked to a balanced polymorphism in a subdivided population. Genet Res 76:63–73

Schroeder SA, Gaughan DM, Swift M (1995) Protection against bronchial asthma by *CFTR* ΔF508 mutation: a heterozygote advantage in cystic fibrosis. Nat Med 1:703–705. https://doi.org/10.1038/nm0795-703

Slatkin M, Wiehe T (1998) Genetic hitch-hiking in a subdivided population. Genet Res 71:155–160

Stajich JE, Hahn MW (2005) Disentangling the effects of demography and selection in human history. Mol Biol Evol 22:63–73. https://doi.org/10.1093/molbev/msh252

Stephan W (2010) Genetic hitchhiking versus background selection: the controversy and its implications. Philos Trans R Soc Lond Ser B Biol Sci 365:1245–1253. https://doi.org/10.1098/rstb.2009.0278

Storz JF, Payseur BA, Nachman MW (2004) Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Mol Biol Evol 21:1800–1811. https://doi.org/10.1093/molbev/msh192

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tang K, Thornton KR, Stoneking M (2007 Jul) A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol 5(7):e171. https://doi.org/10.1371/journal.pbio.0050171.

Tennessen JA, Akey JM (2011) Parallel adaptive divergence among geographically diverse human populations. PLoS Genet 7:e1002127. https://doi.org/10.1371/journal.pgen.1002127

Tennessen JA, Madeoy J, Akey JM (2010) Signatures of positive selection apparent in a small sample of human exomes. Genome Res 20:1327–1334. https://doi.org/10.1101/gr.106161.110

Tennessen JA, Bigham AW, O'Connor TD et al (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337:64–69. https://doi.org/10.1126/science.1219240

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? Genome Res 16:702–712. https://doi.org/10.1101/gr.5105206

Tishkoff SA, Varkonyi R, Cahinhinan N et al (2001) Haplotype diversity and linkage disequilibrium at human *G6PD*: recent origin of alleles that confer malarial resistance. Science 293:455–462. https://doi.org/10.1126/science.1061573

Tishkoff SA, Reed FA, Ranciaro A et al (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39:31–40. https://doi.org/10.1038/ng1946

Tishkoff SA, Reed FA, Friedlaender FR et al (2009) The genetic structure and history of Africans and African Americans. Science 324:1035–1044. https://doi.org/10.1126/science.1172257

Verrelli BC, McDonald JH, Argyropoulos G et al (2002) Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*. Am J Hum Genet 71:1112–1128. https://doi.org/10.1086/344345

Vigilant L, Stoneking M, Harpending H et al (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507

Vina MAF, Hollenbach JA, Lyke KE et al (2012) Tracking human migrations by the analysis of the distribution of HLA alleles, lineages and haplotypes in closed and open populations. Philos Trans R Soc Lond Ser B Biol Sci 367:820–829. https://doi.org/10.1098/rstb.2011.0320

Vitti JJ, Cho MK, Tishkoff SA, Sabeti PC (2012) Human evolutionary genomics: ethical and interpretive issues. Trends Genet 28:137–145. https://doi.org/10.1016/j.tig.2011.12.001

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:e72. https://doi.org/10.1371/journal.pbio.0040072

Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for *Homo sapiens*. Proc Natl Acad Sci 103:135–140. https://doi.org/10.1073/pnas.0509691102

Williamson SH, Hubisz MJ, Clark AG et al (2007) Localizing recent adaptive evolution in the human genome. PLoS Genet 3:e90. https://doi.org/10.1371/journal.pgen.0030090

Xu S, Li S, Yang Y et al (2011) A genome-wide search for signals of high-altitude adaptation in Tibetans. Mol Biol Evol 28:1003–1011. https://doi.org/10.1093/molbev/msq277

Zhang C, Bailey DK, Awad T, Liu G, Xing G, Cao M, Valmeekam V, Retief J, Matsuzaki H, Taub M, Seielstad M, Kennedy GC (2006 Sep 1) A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. Bioinformatics 22(17):2122–2128. https://doi.org/10.1093/bioinformatics/btl365

Zhong M, Lange K, Papp JC, Fan R (2010) A powerful score test to detect positive selection in genome-wide scans. Eur J Hum Genet 18:1148–1159. https://doi.org/10.1038/ejhg.2010.60

Zhong M, Zhang Y, Lange K, Fan R (2011) A cross-population extended haplotype-based homozygosity score test to detect positive selection in genome-wide scans. Stat Interface 4:51–63