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

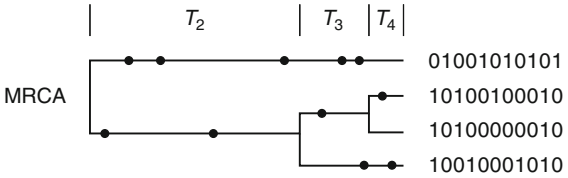

Fig. 1.1 A hypothetical gene genealogy of $n = 4$ sequences from the standard neutral coalescent model without recombination and assuming infinite-sites mutation. The coalescent intervals T_2, T_3, and T_4 are drawn in proportion to their expected values. Hypothetical DNA sequence data (or haplotypes) 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

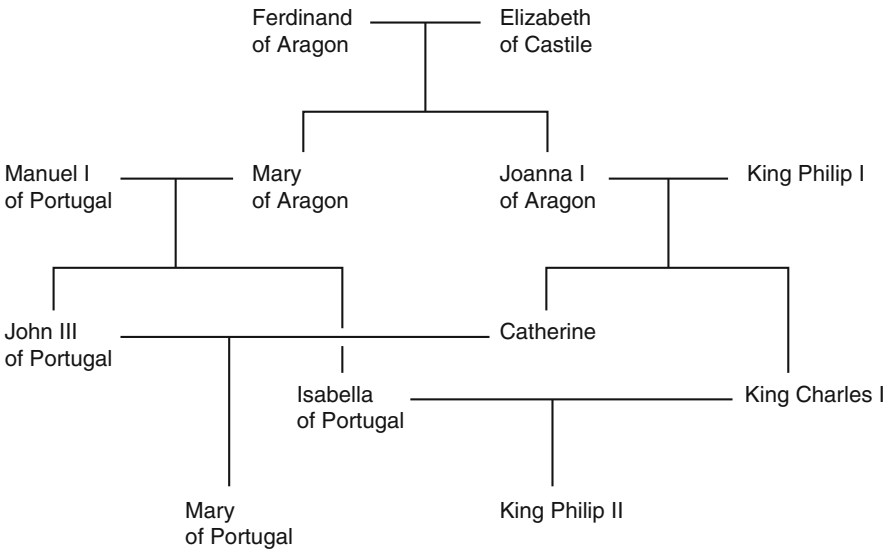

Fig. 1.2 Part of the pedigree of the Spanish Habsburg royal family, extracted from Fig. 1 in Alvarez et al. (2009) and redrawn

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

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

and

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

(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

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

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,

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

(Tajima 1983). Estimates of θ based on π are unbiased but inconsistent in the statistical sense because $\text{Var}[\pi]$ 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 $1 \le i \le n/2$. 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.

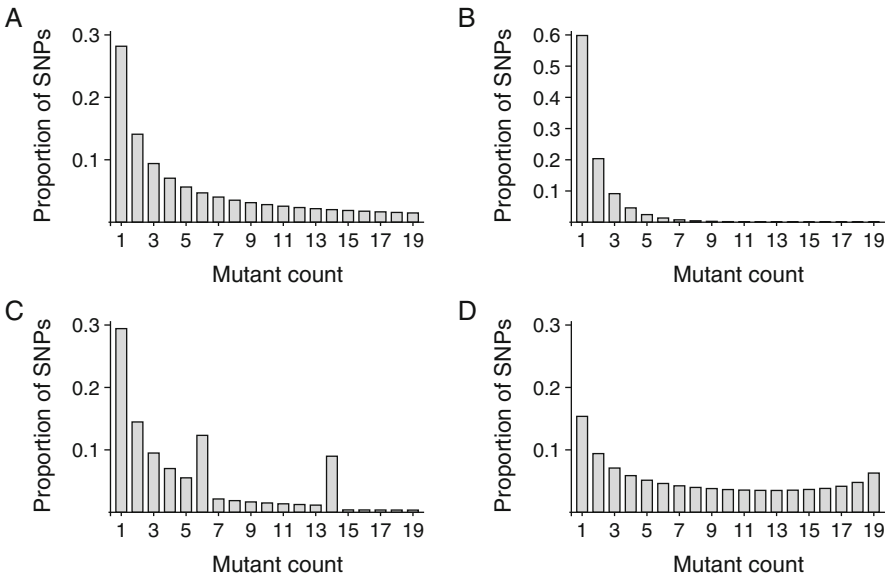

Fig. 1.4 Four site-frequency spectra illustrating the range of possible predictions of neutral population-genetic models. For a given SNP, the mutant count is the number of sequences that carry the mutant base out of a total sample of $n = 20$ sequences. The heights of bars give the proportion of all 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

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

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

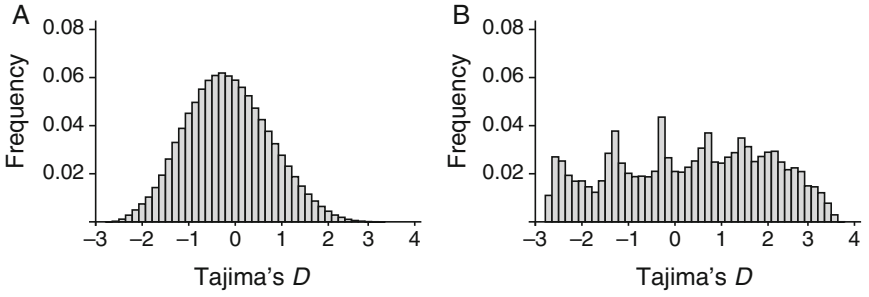

Fig. 1.5 (**a**) The distribution of Tajima's D among 10^6 data sets simulated under the standard neutral coalescent model at a hypothetical locus with $n = 20$ and $= 10$. (**b**) The corresponding distribution for a sample from a single subpopulation under the island migration model with $M = 1$ and 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.

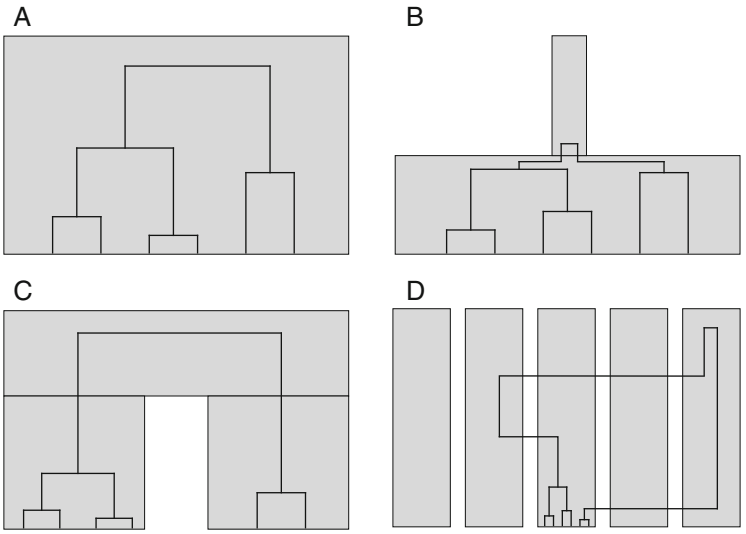

Fig. 1.6 Cartoon depictions of the four types of neutral population-genetic models for which results for expected site-frequency distributions are given in Fig. 1.4. Shaded blocks represent populations over time, with the present at the bottom and the past at the top and with widths in proportion to 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