7

Coalescent Theory

M. Nordborg

Department of Genetics, Lund University, Sweden

The coalescent process is a powerful modeling tool for population genetics. The allelic states of all homologous gene copies in a population are determined by the genealogical and mutational history of these copies. The coalescent approach is based on the realization that the genealogy is usually easier to model backward in time, and that selectively neutral mutations can then be superimposed afterwards. A wide range of biological phenomena can be modeled using this approach.

Whereas almost all of classical population genetics considers the future of a population given a starting point, the coalescent considers the present, while taking the past into account. This allows the calculation of probabilities of sample configurations under the stationary distribution of various population genetic models, and makes full likelihood analysis of polymorphism data possible. It also leads to extremely efficient computer algorithms for generating simulated data from such distributions, data which can then be compared with observations as a form of exploratory data analysis.

7.1 INTRODUCTION

The stochastic process known as 'the coalescent' has played a central role in population genetics since the early 1980s, and results based on it are now used routinely to analyze DNA sequence polymorphism data. In spite of this, there is no comprehensive textbook treatment of coalescent theory. For biologists, the most widely used source of information is probably Hudson's seminal review (Hudson, 1990), which, along with a few other book chapters (Donnelly and Tavaré 1995; Hudson, 1993; Li, 1997) and various unpublished lecture notes, is all that is available beyond the primary literature. Furthermore, since the field is very active, many relevant results are not generally available because they have not yet been published. They may be due to appear sometime in the indefinite future in a mathematical journal or obscure conference volume, or they may simply never have been written down. As a result of all this, there is a considerable gap between the theory that is available, and the theory that is being used to analyze data.

Handbook of Statistical Genetics. Edited by D.J. Balding et al.

The present chapter is intended as an up-to-date introduction suitable for a wider audience. The focus is on the stochastic process itself, and especially on how it can be used to model a wide variety of biological phenomena. I consider a basic understanding of coalescent theory to be extremely valuable – even essential – for anyone analyzing genetic polymorphism data from populations, and will try to defend this view throughout. First of all, such an understanding can in many cases provide an intuitive feeling for how informative polymorphism data is likely to be (the answer is typically 'Not very'). When intuition is not enough, the coalescent provides a simple and powerful tool for exploratory data analysis through the generation of simulated data. Comparison of observed data with data simulated under various assumptions can give considerable insight. However, the reader is also encouraged to study the complementary chapter by Stephens (this volume), in which more sophisticated methods of inference are described.

7.2 THE COALESCENT

The word 'coalescent' is used in several ways in the literature, and it will also be used in several ways here. Hopefully, the meaning will be clear from the context. The coalescent – or, perhaps more appropriately, the coalescent approach – is based on two fundamental insights, which are the topic of the next subsection. The subsection after that describes the stochastic process known as the coalescent, or sometimes Kingman's coalescent in honor of its discoverer (Kingman, 1982a; 1982b; 1982c). This process results from combining the two fundamental insights with a convenient limit approximation.

The coalescent will be introduced in the setting of the Wright–Fisher model of neutral evolution, but it applies more generally. This is one of the main topics for the remainder of the chapter. First of all, many different neutral models can be shown to converge to Kingman's coalescent. Second, more complex neutral models often converge to coalescent processes analogous to Kingman's coalescent.

The coalescent was described by Kingman (1982a; 1982b; 1982c), but it was also discovered independently by Hudson (1983) and by Tajima (1983). Indeed, arguments anticipating it had been used several times in population genetics (reviewed by Tavaré, 1984).

7.2.1 The Fundamental Insights

The first insight is that since selectively neutral variants by definition do not affect reproductive success, it is possible to separate the neutral mutation process from the genealogical process. In classical terms, 'state' can be separated from 'descent'.

To see how this works, consider a population of N clonal organisms that reproduce according to the neutral Wright–Fisher model, that is to say, generations are discrete, and each new generation is formed by randomly sampling N parents with replacement from the current generation. The number of offspring contributed by a particular individual is thus binomially distributed with parameters N (the number of trials) and $1/N$ (the probability of being chosen), and the joint distribution of the numbers of offspring produced by all N individuals is symmetrically multinomial. Now consider the random genealogical relationships (i.e. 'who begat whom') that result from reproduction in this setting. These can be represented graphically, as shown in Figure 7.1. Going forward in time, lineages

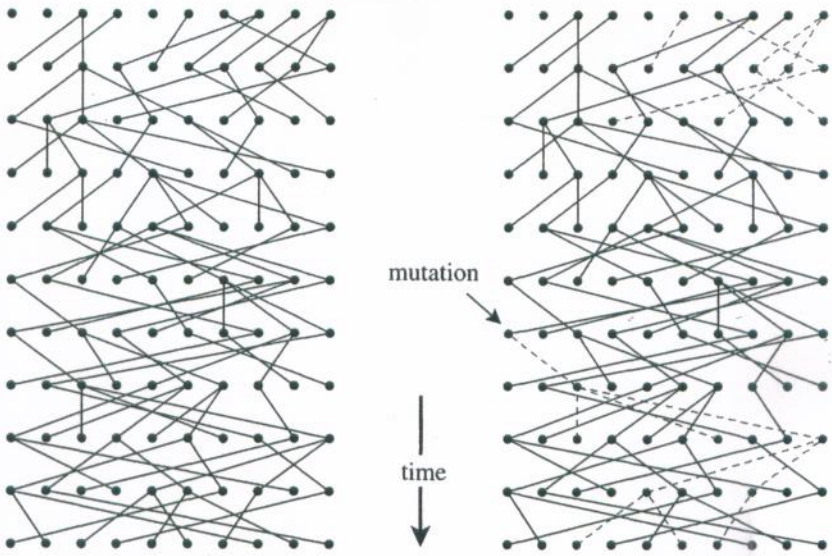

Figure 7.1 The neutral mutation process can be separated from the genealogical process. The genealogical relationships in a particular 10-generation realization of the neutral Wright–Fisher model (with population size $N = 10$) are shown on the left. On the right, allelic states of have been superimposed (so-called 'gene dropping').

branch whenever an individual produces two or more offspring, and end when there is no offspring. Going backward in time, lineages coalesce whenever two or more individuals were produced by the same parent. They never end. If we trace the ancestry of a group of individuals back through time, the number of distinct lineages will decrease and eventually reach one, when the most recent common ancestor (MRCA) of the individuals in question is encountered. None of this is affected by neutral genetic differences between the individuals.

As a consequence, the evolutionary dynamics of neutral allelic variants can be modeled through so-called 'gene dropping' ('mutation dropping' would be more accurate): given a realization of the genealogical process, allelic states are assigned to the original generation in a suitable manner, and the lines of descent then simply followed forward in time, using the rule that offspring inherit the allelic state of their parent unless there is a mutation (which occurs with some probability each generation). In particular, the allelic states of any group of individuals (for instance, all the members of a given generation) can be generated by assigning an allelic state to their MRCA and then 'dropping' mutations along the branches of the genealogical tree that leads to them. Most of the genealogical history of the population is then irrelevant (cf. Figures 7.1 and 7.2).

The second insight is that it is possible to model the genealogy of a group of individuals backward in time without worrying about the rest of the population. It is a general consequence of the assumption of selective neutrality that each individual in a generation can be viewed as 'picking' its parent at random from the previous generation. It follows that the genealogy of a group of individuals may be generated by simply tracing the lineages back in time, generation by generation, keeping track of coalescences between lineages, until eventually the MRCA is found. It is particularly easy to see how this is done for the Wright–Fisher model, where individuals pick their parents independently of each other.

In summary, the joint effects of random reproduction (which causes 'genetic drift') and random neutral mutations in determining the genetic composition of a group of clonal

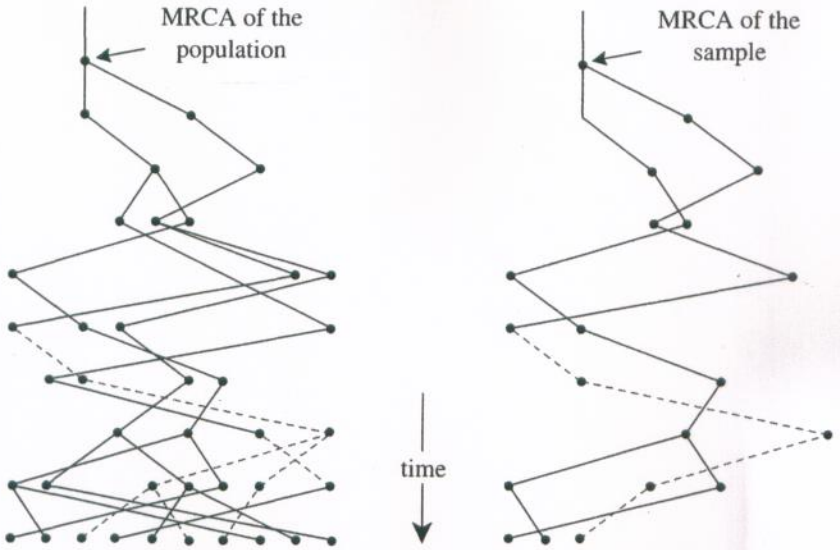

Figure 7.2 The genetic composition of a group of individuals is completely determined by the group's genealogy and the mutations that occur on it. The genealogy of the final generation in Figure 7.1 is shown on the left, and the genealogy of a sample from this generation is shown on the right. These trees could have been generated backward in time without generating the rest of Figure 7.1.

individuals (such as a generation or a sample thereof) may be modeled by first generating the random genealogy of the individuals backward in time, and then superimposing mutations forward in time. This approach leads directly to extremely efficient computer algorithms (cf. the 'classical' approach which is to simulate the *entire*, usually very large population forward in time for a long period of time, and then to look at the final generation). It is also mathematically elegant, as Section 7.2.2 will show. However, its greatest value may be heuristic: the realization that the pattern of neutral variation observed in a population can be viewed as the result of random mutations on a random tree is a powerful one, which profoundly affects the way we think about data.

In particular, we are almost always interested in biological phenomena that affect the genealogical process, but do not affect the mutation process (e.g. population subdivision). From the point of view of inference about such phenomena, the observed polymorphisms are only of interest because they contain information about the unobserved underlying genealogy. Furthermore, the underlying genealogy is only of interest because it contains information about the evolutionary process that gave rise to it. In statistical terms, almost all inference problems that arise from polymorphism data can be seen as 'missing data' problems.

It is crucial to understand this, because no matter how many individuals we sample, there is still only a *single* underlying genealogy to estimate. It could of course be that this single genealogy contains a lot of information about the interesting aspect of the evolutionary process, but if it does not, then our inferences will be as good as one would normally expect from a sample of size one!

Another consequence of the above is that it is usually possible to understand how model parameters affect polymorphism data by understanding how they affect genealogies. For this reason, I will focus on the genealogical process and only discuss the neutral mutation process briefly toward the end of the chapter.

7.2.2 The Coalescent Approximation

The previous subsection described the conceptual insights behind the coalescent approach. The sample genealogies central to this approach can be conveniently modeled using a continuous-time Markov process known as the coalescent (or Kingman's coalescent, or sometimes 'the n-coalescent' to emphasize the dependence on the sample size). We will now describe the coalescent and show how it arises naturally as a large-population approximation to the Wright–Fisher model. Its relationship to other models will be discussed later.

Figure 7.2 is needlessly complicated because the identity (i.e. the horizontal position) of all ancestors is maintained. In order to superimpose mutations, all we need to know is which lineage coalesces with which, and when. In other words, we need to know the topology, and the branch lengths. The topology is easy to model: because of neutrality, individuals are equally likely to reproduce; therefore all lineages must be equally likely to coalesce. It is convenient to represent the topology as a sequence of coalescing equivalence classes: two members of the original sample are equivalent at a certain point in time if and only if they have a common ancestor at that time (see Figure 7.3). But what about the branch lengths, that is, the coalescence times?

Follow two lineages back in time. We have seen that offspring pick their parents randomly from the previous generation, and that, under the Wright–Fisher model, they do so independently of each other. Thus, the probability that the two lineages pick the same parent and coalesce is $1/N$, and the probability that they pick different parents and remain distinct is $1 - 1/N$. Since generations are independent, the probability that they remain distinct more than t generations into the past is $(1 - 1/N)^t$. The expected coalescence time is N generations. This suggests a standard continuous-time diffusion approximation, which is good as long as N is reasonably large (see Neuhauser, this volume). Rescale time so that one unit of scaled time corresponds to N generations. Then the probability

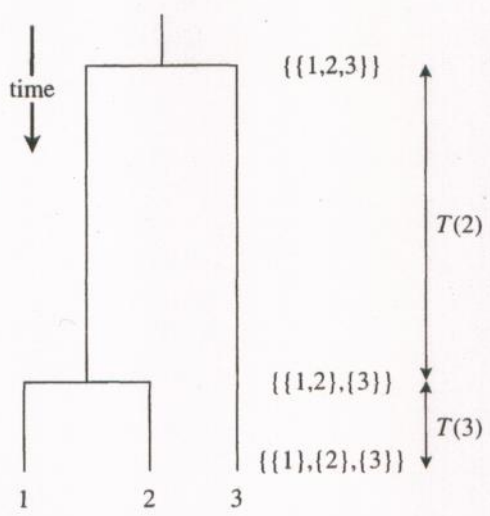

Figure 7.3 The genealogy of a sample can be described in terms of its topology and branch lengths. The topology can be represented using equivalence classes for ancestors. The branch lengths are given by the waiting times between successive coalescence events.

that the two lineages remain distinct for more than τ units of scaled time is

$$\left(1-\frac{1}{N}\right)^{[N\tau]} \to \mathrm{e}^{-\tau}, \tag{7.1}$$

as N goes to infinity ($[N\tau]$ is the largest integer less than or equal to $N\tau$). Thus, in the limit, the coalescence time for a pair of lineages is exponentially distributed with mean 1.

Now consider k lineages. The probability that none of them coalesce in the previous generation is

$$\prod_{i=0}^{k-1}\frac{N-i}{N}=\prod_{i=1}^{k-1}\left(1-\frac{i}{N}\right)=1-\frac{\binom{k}{2}}{N}+O\left(\frac{1}{N^2}\right), \tag{7.2}$$

and the probability that more than two do so is $O(1/N^2)$. Let $T(k)$ be the (scaled) time till the first coalescence event, given that there are currently k lineages. By the same argument as above, $T(k)$ is in the limit exponentially distributed with mean $2/[(k(k-1))]$. Furthermore, the probability that more than two lineages coalesce in the same generation can be neglected. Thus, under the coalescent approximation, the number of distinct lineages in the ancestry of a sample of (finite) size n decreases in steps of one back in time, so $T(k)$ is the time from k to $k-1$ lineages (see Figure 7.3).

In summary, the coalescent models the genealogy of a sample of n haploid individuals as a random bifurcating tree, where the $n-1$ coalescence times $T(n), T(n-1), \ldots, T(2)$ are mutually independent, exponentially distributed random variables. Each pair of lineages coalesces independently at rate 1, so the total rate of coalescence when there are k lineages is 'k choose 2'. A concise (and rather abstract) way of describing the coalescent is as a continuous-time Markov process with state space $\mathcal{E}_n$ given by the set of all equivalence relations on $\{1, \ldots, n\}$, and infinitesimal generator $Q = (q_{\xi\eta})_{\xi,\eta\in\mathcal{E}_n}$ given by

$$q_{\xi\eta} := \begin{cases} -k(k-1)/2, & \text{if } \xi = \eta, \\ 1, & \text{if } \xi \prec \eta, \\ 0, & \text{otherwise,} \end{cases} \tag{7.3}$$

where $k := |\xi|$ is the number of equivalence classes in ξ, and $\xi \prec \eta$ if and only if η is obtained from ξ by coalescing two equivalence classes of ξ.

It is worth emphasizing just how efficient the coalescent is as a simulation tool. In order to generate a sample genealogy under the Wright–Fisher model as described in Section 7.2.1, we would have to go back in time some N generations, checking for coalescences in each of them. Under the coalescent approximation, we simply generate $n-1$ independent exponential random numbers and, independently of these, a random bifurcating topology.

What do typical coalescence trees look like? Figure 7.4 shows four examples. It is clear that the trees are extremely variable, both with respect to topology and branch lengths. This should come as no surprise considering the description of the coalescent just given: the topology is independent of the branch lengths; the branch lengths are independent, exponential random variables; and the topology is generated by randomly picking lineages to coalesce (in this sense all topologies are equally likely).

Note that the trees tend to be dominated by the deep branches, when there are few ancestors left. Because lineages coalesce at rate 'k choose 2', coalescence events occur

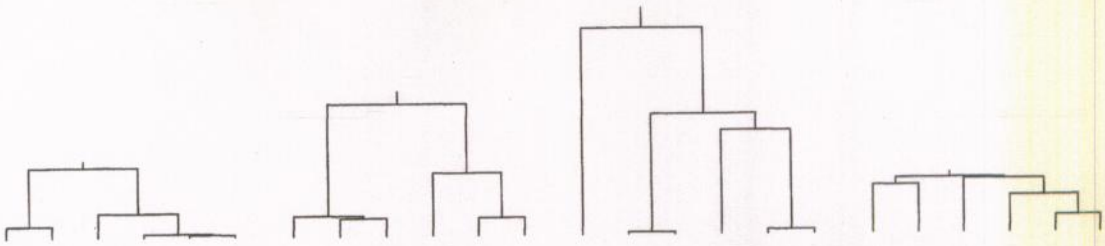

Figure 7.4 Four realizations of the coalescent for $n = 6$, drawn on the same scale (the labels 1–6 should be assigned randomly to the tips).

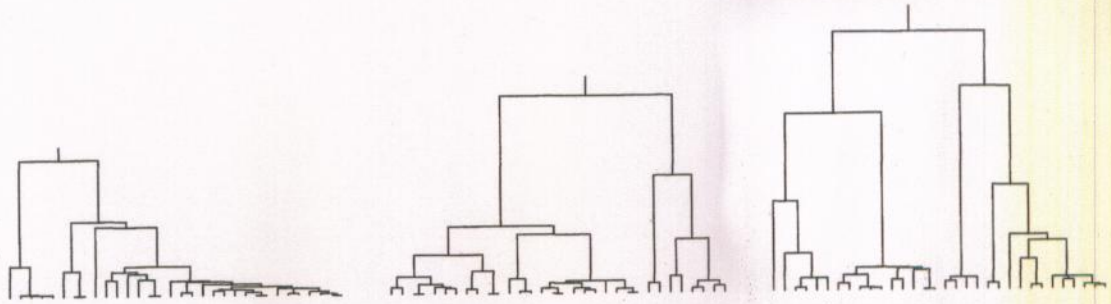

Figure 7.5 Three realizations of the coalescent for $n = 32$, drawn on the same scale (the labels 1–32 should be assigned randomly to the tips).

much more rapidly when there are many lineages (intuitively speaking, it is easier for lineages to find each other then). Indeed, the expected time to the MRCA (the height of the tree) is

$$\mathrm{E}\left[\sum_{k=2}^{n} T(k)\right] = \sum_{k=2}^{n} \mathrm{E}[T(k)] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\left(1 - \frac{1}{n}\right), \tag{7.4}$$

while $\mathrm{E}[T(2)] = 1$, so the expected time during which there are only two branches is greater than half the expected total tree height. Furthermore, the variability in $T(2)$ accounts for most of the variability in tree height. The dependence on the deep branches becomes increasingly apparent as n increases, as can be seen by comparing Figures 7.4 and 7.5.

The importance of realizing that there is only a single underlying genealogy was emphasized above. As a consequence of the single genealogy, sampled gene copies from a population must almost always be treated as dependent, and increasing the sample size is therefore often surprisingly ineffective (the point is well made by Donnelly, 1996). Important examples of this follow directly from the basic properties of the coalescent. Consider first the MRCA of the population. One might think that a large sample is needed to ensure that the deepest split is included, but it can be shown (this and related results can be found in Saunders et al., 1984) that the probability that a sample of size n contains the MRCA of the whole population is $(n-1)/(n+1)$. Thus even a small sample is likely to contain it and the total tree height will quickly stop growing as n increases. Second, the number of distinct lineages decreases rapidly as we go back in time. This severely limits inferences about ancient demography (e.g. Nordborg, 1998). Third, since increasing the sample size only adds short twigs to the tree (cf. Figure 7.5), the expected total branch

length of the tree, $T_{\text{tot}}(n)$ grows very slowly with n. We have

$$\mathrm{E}[T_{\text{tot}}(n)] = \mathrm{E}\left[\sum_{k=2}^{n} kT(k)\right] = \sum_{k=1}^{n-1} \frac{2}{k} \sim 2(\gamma + \log \mathrm{n}), \tag{7.5}$$

as $n \to \infty$ ($\gamma \approx 0.577\,216$ is Euler's constant). Since the number of mutations that are expected to occur in a tree is proportional to $\mathrm{E}[T_{\text{tot}}(n)]$, this has important consequences for estimating the mutation rate, as well as for inferences that depend on estimates of the mutation rate. Loosely speaking, it turns out that a sample of n copies of a gene often has the statistical properties one would expect of a random sample of size $\log n$, or even of size 1 (which is not much worse than $\log n$ in practice).

7.3 GENERALIZING THE COALESCENT

This section will present ideas and concepts that are important for generalizing the coalescent. The following sections will then illustrate how these can be used to incorporate greater biological realism.

7.3.1 Robustness and Scaling

We have seen that the coalescent arises naturally as an approximation to the Wright–Fisher model, and that it has convenient mathematical properties. However, the real importance of the coalescent stems from the fact that it arises as a limiting process for a wide range of neutral models, *provided time is scaled appropriately* (Kingman, 1982b; 1982c; Möhle, 1998b; 1999). It is thus robust in this sense.

This is best explained through an example. Recall that the number of offspring contributed by each individual in the Wright–Fisher model is binomially distributed with parameters N and $1/N$. The mean is thus 1, and the variance is $1 - 1/N \to 1$, as $N \to \infty$. Now consider a generalized version of this model in which the mean number of offspring is still 1 (as it must be for the population size to remain constant), but the limiting variance is σ^2, $0 < \sigma^2 < \infty$ (perhaps giants step on 90% of the individuals before they reach reproductive age). It can be shown that this process also converges to the coalescent, provided time is measured in units of N/σ^2 generations. We could also measure time in units of N generations as before, but then $\mathrm{E}[T(2)] = 1/\sigma^2$ instead of $\mathrm{E}[T(2)] = 1$, and so on. Either way, the expected coalescence time for a pair of lineages is N/σ^2 generations. The intuition behind this is clear: increased variance in reproductive success causes coalescence to occur faster (at a higher rate). In classical terms, 'genetic drift' operates faster. By changing the way we measure time, this can be taken into account, and the standard coalescent process obtained.

The remarkable fact is that a very wide range of biological phenomena (overlapping generations, separate sexes, mating systems – several examples will be given below) can likewise be treated as a simple linear change in the time scale of the coalescent. This has important implications for data analysis. The good news is that we may often be able to justify using the coalescent process even though 'our' species almost certainly does not reproduce according to a Wright–Fisher model (few species do). The bad news is that biological phenomena that can be modeled this way will never be amenable to inference based on

polymorphism data alone. For example, σ^2 in the model above could never be estimated from polymorphism data unless we had independent information about N (and vice versa).

Of course, we could not even estimate N/σ^2 without external data. It is important to realize that all parameters in coalescent models are scaled, and that only scaled parameters can be directly estimated from the data. In order to make any kind of statement about unscaled quantities, such as population numbers, or ages in years or generations, external information is needed. This adds considerable uncertainty to the analysis. For example, an often used source of external information is an estimate of the neutral mutation probability per generation. Roughly speaking, this estimate is obtained by measuring sequence divergence between species, and dividing by the estimated species divergence time (Li, 1997). The latter is in turn obtained from the fossil record and a rough guess of the generation length. It should be clear that it is not appropriate to treat such an estimate as a known parameter when analyzing polymorphism data. However, it should be noted that interesting conclusions can often be drawn directly from scaled parameters (e.g. by looking at relative values). Such analyses are likely to be more robust, given the robustness of the coalescent.

Because the generalized model above converges with the same scaling as a Wright–Fisher model with a population size of N/σ^2, it is sometimes said that it has an 'effective population size', $N_e = N/\sigma^2$. Models that scale differently would then have other effective population sizes. Although convenient, this terminology is unfortunate for at least two reasons. First, the classical population genetics literature is full of variously defined 'effective population sizes', only some of which are effective population sizes in the sense used here. For example, populations that are subdivided or vary in size cannot in general be modeled as a linear change in the time scale of the coalescent. Second, the term is inevitably associated with real population sizes, even though it is simply a scaling factor. To be sure, N_e is always a function of the real demographic parameters, but there is no direct relationship with the total population size (which may be smaller as well as much, much larger). Indeed, as we shall see in Section 7.7, it is now clear that N_e must vary between chromosomal regions in the same organism!

7.3.2 Variable Population Size

Real populations vary in size over time. Although the coalescent is not robust to variation in the population size in the sense described above (i.e. there is no 'effective population size'), it is nonetheless easy to incorporate changes in the population size, at least if we are willing to assume that we know what they were – that is, if we assume that the variation can be treated deterministically. Since a rigorous treatment of these results can be found in the review by Donnelly and Tavaré (1995), also in Neuhauser, Chapter 6, this volume, I will try to give an intuitive explanation.

Imagine a population that evolves according to the Wright–Fisher model, but with a different population size in each generation. If we know how the size has changed over time, we can trace the genealogy of a sample precisely as before. Let $N(t)$ be the population size t generations ago. Going back in time, lineages are more likely to coalesce in generations when the population is small than in generations when the population is large. In order to describe the genealogy by a continuous-time process analogous to the coalescent, we must therefore allow the rate of coalescence to change over time. However, since the time scale used in the coalescent directly reflects the rate of coalescence, we may instead let this scaling change over time. In the standard coalescent, t generations

ago corresponds to t/N units of coalescence time, and τ units of coalescence time ago corresponds to $[N\tau]$ generations. When the population size is changing, we find instead that t generations ago corresponds to

$$g(t) := \sum_{i=1}^{t} \frac{1}{N(i)} \tag{7.6}$$

units of coalescence time, and τ units of coalescence time ago corresponds to $[g^{-1}(\tau)]$ generations (g^{-1} denotes the inverse function of g). It is clear from equation (7.6) that many generations go by without much coalescence time passing when the population size is large, and, conversely, that much coalescence time passes each generation when the population is small. Let $N(0)$ go to infinity, and assume that $N(t)/N(0)$ converges to a finite number for each t, to ensure that the population size becomes large in every generation. It can be shown that the variable population size model converges to a coalescent process with a *nonlinear* time scale in this limit (Griffiths and Tavaré, 1994). The scaling is given by equation (7.6). Thus, a sample genealogy from the coalescent with variable population size can be generated by simply applying g^{-1} to the coalescence times of a genealogy generated under the standard coalescent.

An example will make this clearer. Consider a population that has grown exponentially, so that, backwards in time, it shrinks according to $N(t) = N(0)e^{-\beta t}$ (note that this violates the assumption that the population size be large in every generation – this turns out not to matter greatly). Then

$$g(t) \approx \int_0^t \frac{1}{N(s)} ds = \frac{e^{\beta t} - 1}{N(0)\beta} \tag{7.7}$$

and

$$g^{-1}(\tau) \approx \frac{\log(1 + N(0)\beta\tau)}{\beta}. \tag{7.8}$$

The difference between this model and one with a constant population size is shown in Figure 7.6. When the population size is constant, there is a linear relationship between real and scaled time. The genealogical trees will tend to look like those in Figures 7.4 and 7.5. When the population size is changing, the relationship between real and scaled time is nonlinear, because coalescences occur very slowly when the population was large, and more rapidly when the population was small. Genealogies in an exponentially growing population will tend to have most coalescences early in the history. Since all branches will then be of roughly equal length, the genealogy is said to be 'starlike'.

Models of exponential population growth have often been used in the context of human evolution (e.g. Rogers and Harpending, 1992; Slatkin and Hudson 1991). Marjoram and Donnelly (1997) have pointed out that some of the predictions from such models (e.g. the starlike genealogies) depend crucially on exponential growth from a *very* small size – unrealistically small for humans. However, other predictions are more robust. For example, the argument in the previous paragraph explains why it may be reasonable to ignore growth altogether when modeling human evolution, even though growth has clearly taken place: if the growth was rapid and recent enough, no scaled time would pass, and no coalescence occur. In classical terms, exponential growth stops genetic drift.

Finally, it should be pointed that it is not entirely clear how general the nonlinear scaling approach to variable population sizes is. It relies, of course, on knowing the

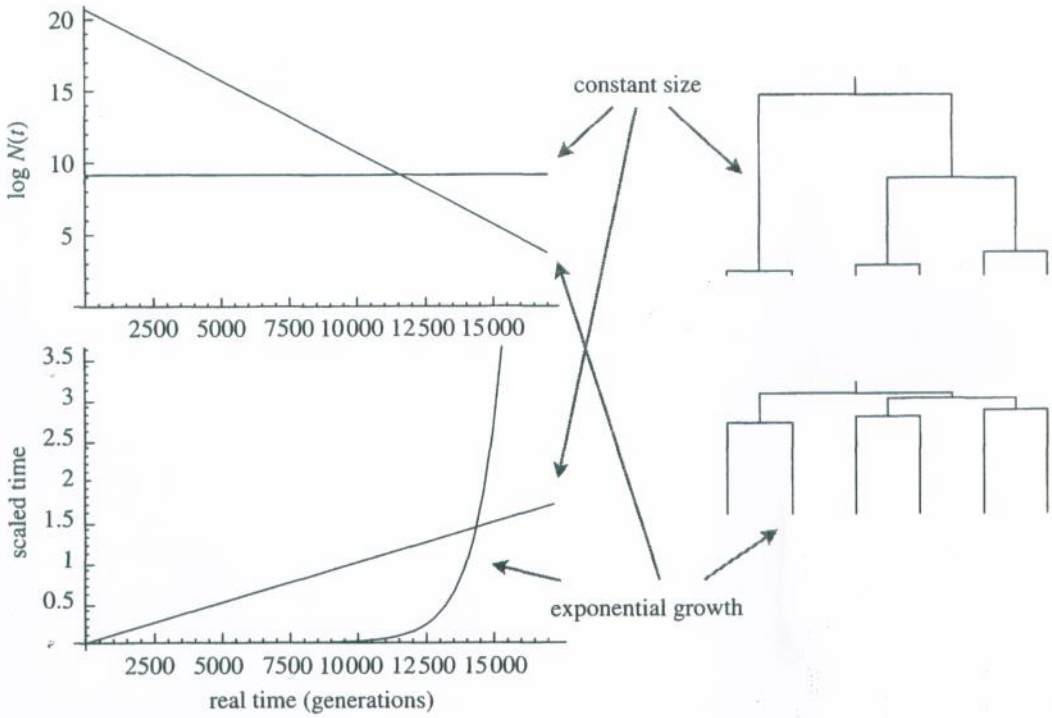

Figure 7.6 Variable population size can be modeled as a standard coalescent with a nonlinear time scale. Here, a constant population is compared to one that has grown exponentially. As the latter population shrinks backward in time, the scaled time begins to run faster, reflecting the fact that coalescences are more likely to have taken place when the population was small. Note that the trees are topologically equivalent and differ only in the branch lengths.

historical population sizes, but it also requires assumptions about the type of density regulation (Marjoram and Donnelly, 1997).

7.3.3 Population Structure on Different Time Scales

Real populations are also often spatially structured, and it is obviously important to be able to incorporate this in our models. However, structured models turn out to be even more important than one might have expected from this, because many biological phenomena can be thought of as analogous to population structure (Nordborg, 1997; Rousset, 1999a). Examples range from the obvious, like age structure, to the more abstract, like diploidy and allelic classes.

The following model, which may be called the 'structured Wright–Fisher model', turns out to be very useful in this context. Consider a clonal population of size N, as before, but let it be subdivided into patches of fixed sizes N_i, $i \in \{1, \ldots, M\}$, so that $\sum_i N_i = N$. In every generation, each individual produces an effectively infinite number of propagules. These propagules then migrate among the patches independently of each other, so that with probability m_{ij}, $i, j \in \{1, \ldots, M\}$, a propagule produced in patch i ends up in patch j. We also define the the 'backward migration' probability, b_{ij}, $i, j \in \{1, \ldots, M\}$, that a randomly chosen propagule in patch i after dispersal was produced in patch j; it is easy to show that

$$b_{ij} = \frac{N_j m_{ji}}{\sum_k N_k m_{ki}}. \tag{7.9}$$

The next generation of adults in each patch is then formed by random sampling from the available propagules.

Thus the number of offspring a particular individual in patch i contributes to the next generation in patch j is binomially distributed with parameters N_j and $b_{ji}N_i^{-1}$. The joint distribution of the numbers of offspring contributed to the next generation in patch j by *all* individuals in the current generation is multinomial (but no longer symmetric).

Just like the unstructured Wright–Fisher model, the genealogy of a finite sample in this model can be described by a discrete-time Markov process. Lineages coalesce in the previous generation if and only if they pick the same parental patch, and the same parental individual within that patch. A lineage currently in i and a lineage currently in j 'migrate' (backward in time) to k and coalesce there with probability $b_{ik}b_{jk}N_k^{-1}$.

It is also possible to approximate the model by a continuous-time Markov process. The general idea is to let the total population size, N, go to infinity with time scaled appropriately, precisely as before. However, we now also need to decide how M, N_i, and b_{ij} scale with N. Different biological scenarios lead to very different choices in this respect, and it is often possible to utilize convergence results based on separation of time scales (Möhle, 1998a; Nordborg, 1997; Nordborg, 1999; Nordborg and Donnelly, 1997; Wakeley, 1999). This important technique will be exemplified in what follows.

7.4 GEOGRAPHICAL STRUCTURE

Genealogical models of population structure have a long history. The classical work on identity coefficients (see Rousset, this volume) concerns genealogies when $n = 2$, and the coalescent was also quickly used for this purpose (for early work see Slatkin, 1987; Strobeck, 1987; Tajima, 1989a; Takahata, 1988).

Since geographical structure is reviewed by Rousset (this volume), we will mainly use it to introduce some of the scaling ideas that are central to the coalescent. The discussion will be limited to the structured Wright–Fisher model (which is a matrix migration model when viewed as a model of geographic subdivision). Most coalescent modeling has been done in this setting (reviewed in Wilkinson-Herbots, 1998 and Hudson, 1998). For time-scale approximations different from the ones discussed below, see Takahata (1991) and Wakeley (1999). An important variant of the model considers isolation: gene flow which stopped completely at some point in the past, for example due to speciation (e.g. Wakeley, 1996). For an attempt at modeling continuous environments, see Barton and Wilson (1995).

7.4.1 The Structured Coalescent

Assume that M, $c_i := N_i/N$, and $B_{ij} := 2Nb_{ij}$, $i \neq j$, all remain constant as N goes to infinity. Then, with time measured in units of N generations, the process converges to the so-called 'structured coalescent', in which each pair of lineages in patch i coalesces independently at rate $1/c_i$, and each lineage in i 'migrates' (backward in time) independently to j at rate $B_{ij}/2$ (Herbots, 1994; Notohara, 1990; Wilkinson-Herbots, 1998). The intuition behind this is as follows (an excellent discussion of how the scaled parameters should be interpreted can be found in Neuhauser, this volume). By assuming that B_{ij} remains constant, we assure that the backward per-generation probabilities of *leaving* a patch (b_{ij}, $i \neq j$), are $O(1/N)$. Similarly, by assuming that c_i remains constant, we assure that all

per-generation coalescence probabilities are $O(1/N)$. Thus, in any given generation, the probability that all lineages remain in their patch, without coalescing, is $1 - O(1/N)$. Furthermore, the probabilities that more than two lineages coalesce, that more than one lineage migrates, and that lineages both migrate and coalesce, are all $O(1/N^2)$ or smaller. In the limit $N \to \infty$, the only possible events are pairwise coalescences within patches, and single migrations between patches.

These events occur according to independent Poisson processes, which means the following. Let k_i denote the number of lineages currently in patch i. Then the waiting time till the first event is exponentially distributed with rate given by the sum of the rates of all possible events, that is,

$$h(k_1, \ldots, k_M) = \sum_i \left(\frac{\binom{k_i}{2}}{c_i} + \sum_{j \neq i} k_i \frac{B_{ij}}{2} \right). \tag{7.10}$$

When an event occurs, it is a coalescence in patch i with probability

$$\frac{\binom{k_i}{2}/c_i}{h(k_1, \ldots, k_M)}, \tag{7.11}$$

and a migration from i to j with probability

$$\frac{k_i B_{ij}/2}{h(k_1, \ldots, k_M)}. \tag{7.12}$$

In the former case, a random pair of lineages in i coalesces, and k_i decreases by one. In the latter case, a random lineage moves from i to j, k_i decreases by one, and k_j increases by one. A simulation algorithm would stop when the MRCA is found, but note that this single remaining lineage would continue migrating between patches if followed further back in time.

Structured coalescent trees generally look different from standard coalescent trees. Whereas variable population size only altered the branch lengths of the trees, population structure also affects the topology. If migration rates are low, lineages sampled from the same patch will tend to coalesce with each other, and a substantial amount of time can then pass before migration allows the ancestral lineages to coalesce (see Figure 7.7). Structure will often increase the mean and, equally importantly, the variance in time to the MRCA considerably (discussed in the context of human evolution by Marjoram and Donnelly, 1997).

7.4.2 The Strong-Migration Limit

It is intuitive that weak migration, which corresponds to strong population subdivision, can have a large effect on genealogies. Conversely, we would expect genealogies in models with strong migration to look much like standard coalescent trees. This intuition turns out to be correct, except for one important difference: the scaling changes. Strong migration is thus one of the phenomena that can be modeled as a simple linear change in the time scale of the coalescent. It is important to understand why this happens.

Formally, the strong-migration limit means that $\lim_{N\to\infty} Nb_{ij} = \infty$ because the per-generation migration probabilities, b_{ij}, are not $O(1/N)$. Since the coalescence probabilities *are* $O(1/N)$, this means that, for large N, migration will be much more likely than

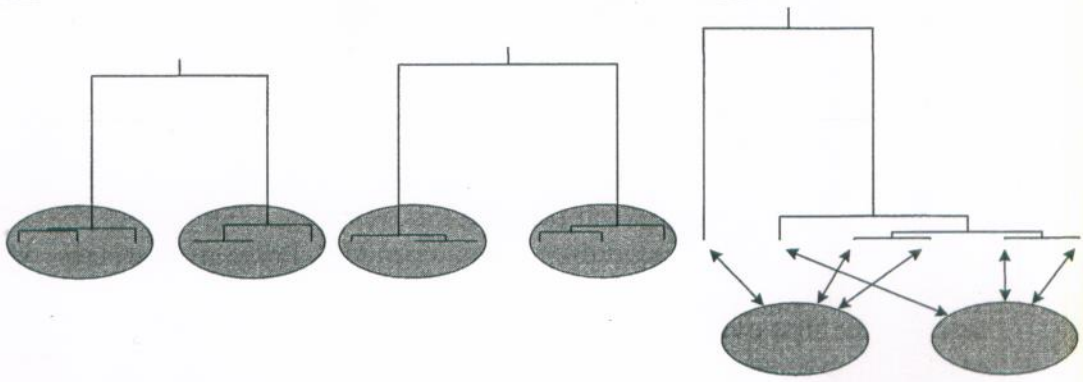

Figure 7.7 Three realizations of the structured coalescent in a symmetric model with two patches, and $n = 3$ in each patch (labels should be assigned randomly within patches). Lineages tend to coalesce within patches – but not always, as shown by the rightmost tree.

coalescence. As $N \to \infty$, there will in effect be infinitely many migration events between coalescence events. This is known as separation of time scales: migration occurs on a faster time scale than does coalescence. However, coalescences can of course still only occur when two lineages pick a parent in the same patch. How often does this happen? Because lineages jump between patches infinitely fast on the coalescence time scale, this is determined by the stationary distribution of the migration process (strictly speaking, this assumes that the migration matrix is ergodic). Let π_i be the stationary probability that a lineage is in patch i. A given pair of lineages then co-occur in i a fraction π_i^2 of the time. Coalescence in this patch occurs at rate $1/c_i$. Thus the total rate at which pairs of lineages coalesce is $\alpha := \sum_i \pi_i^2/c_i$. Pairs coalesce independently of each other just as in the standard model, so the total rate when there are k lineages is $\binom{k}{2}\alpha$. If time is measured in units of $N_e = N/\alpha$ generations, the standard coalescent is retrieved (Nagylaki, 1980; Notohara, 1993).

It can be shown that $\alpha \geq 1$, with equality if and only if $\sum_{j\neq i} N_i b_{ij} = \sum_{j\neq i} N_j b_{ji}$ for all i. This condition means that, going forward in time, the number of emigrants equals the number of immigrants in all populations, a condition known as 'conservative migration' (Nagylaki, 1980). Thus we see that, unless migration is conservative, the effective population size with strong migration is smaller than the total population size. The intuitive reason for this is that when migration is nonconservative, some individuals occupy 'better' patches than others, and this increases the variance in reproductive success among individuals. The environment has 'sources' and 'sinks' (Pulliam, 1988; Rousset, 1999b). Conservative migration models (like Wright's island model) have many simple properties that do not hold generally (Nagylaki, 1982, 1998; Nordborg, 1997; Rousset, 1999a).

7.5 SEGREGATION

Because everything so far has been done in an asexual setting, it has not been necessary to distinguish between the genealogy of an organism and that of its genome. This becomes necessary in sexual organisms. Most obviously, a diploid organism that was produced sexually has two parents, and each chromosome came from one of them. The genealogy

of the genes is thus different from the genealogy (the pedigree) of the individuals: the latter describes the *possible* routes the genes could have taken (and is largely irrelevant – cf. Figure 7.9). This is simply Mendelian segregation viewed backwards in time, and it is the topic of this section. It is usually said that diploidy can be taken into account by simply changing the scaling from N to $2N$; it will become clear from what follows why, and in what sense, this is true.

The other facet of sexual reproduction, genetic recombination, turns out to have much more important effects. Genetic recombination causes ancestral lineages to branch, so that the genealogy of a sample can no longer be represented by a single tree: instead it becomes a collection of trees, or a single, more general type of graph. Recombination will be ignored until Section 7.6 (it makes sense to discuss diploidy first).

Sex takes many forms: I will first consider organisms that are hermaphroditic and therefore potentially capable of fertilizing themselves (this includes most higher plants and many mollusks), and thereafter discuss organisms with separate sexes (which includes most animals and many plants).

7.5.1 Hermaphrodites

The key to modeling diploid populations is the realization that a diploid population of size N can be thought of as a haploid population of size $2N$, divided into N patches of size 2. In the notation of the structured Wright–Fisher model above, $M = N$, $N_i = 2$, and $c_i = 2/N$. Thus, in contrast to the assumptions for the structured coalescent, both M and c_i depend on N. This leads to a convenient convergence result based on separation of time scales (Nordborg and Donnelly, 1997; for a formal proof, see Möhle, 1998a), that can be described as follows (cf. Figure 7.8).

If time is scaled in units of $2N$ generations, then each pair of lineages 'coalesces' into the same individual at rate 2. Whenever this happens, there are two possibilities: either the two lineages pick the same of the two available (haploid) parents, or they pick different ones. The former event, which occurs with probability $\frac{1}{2}$, results in a real coalescence,

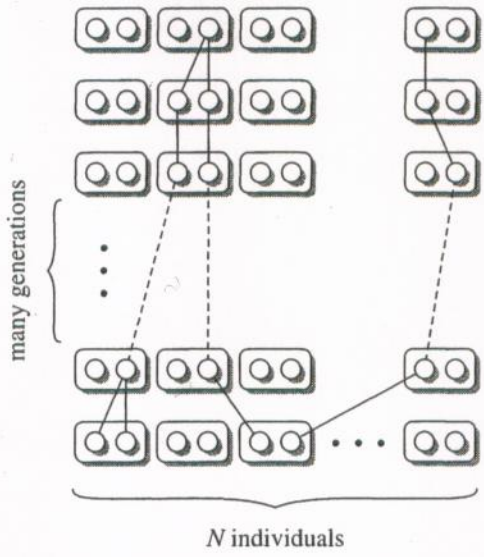

Figure 7.8 The coalescent with selfing. On the coalescent time scale, lineages within individuals instantaneously coalesce (probability F), or end up in different individuals (probability $1 - F$).

whereas the latter event, which also occurs with probability $\frac{1}{2}$, simply results in the two distinct lineages temporarily occupying the same individual. Let S be the probability that a fertilization occurs through selfing, and $1 - S$ the probability that it occurs through outcrossing. If the individual harboring two distinct lineages was produced through selfing (probability S), then the two lineages must have come from the same individual in the previous generation, and again pick different parents with probability $\frac{1}{2}$ or coalesce with probability $\frac{1}{2}$. If the individual was produced through outcrossing, the two lineages revert to occupying distinct individuals. Thus the two lineages will rapidly either coalesce or end up in different individuals. The probability of the former outcome is

$$\frac{S/2}{S/2+1-S} = \frac{S}{2-S} =: F, \tag{7.13}$$

and that of the latter $1 - F$. Thus each time a pair of lineages coalesces into the same individual, the total probability that this results in a coalescence event is $\frac{1}{2} \times 1 + \frac{1}{2} \times F = (1 + F)/2$, and since pairs of lineages coalesce into the same individual at rate 2, the rate of coalescence is $1 + F$. On the chosen time scale, all states that involve two or more pairs occupying the same individual are instantaneous.

Thus, the genealogy of a random sample of gene copies from a population of hermaphrodites can be described by the standard coalescent if time is scaled in units of

$$2N_e = \frac{2N}{1+F} \tag{7.14}$$

generations (cf. Pollak, 1987). If individuals are obligate outcrossers, $F = 0$, and the correct scaling is $2N$.

It should be pointed out that a sample from a diploid population is not a random sample of gene copies, because both copies in each individual are sampled. This is easily taken into account. It follows from the above that the two copies sampled from the same individual will instantaneously coalesce with probability F, and end up in different individuals with probability $1 - F$. The number of distinct lineages in a sample of $2n$ gene copies from n individuals is thus $2n - X$, where X is as a binomially distributed random variable with parameters n and F. This corresponds to the well-known increase in the frequency of homozygous individuals predicted by classical population genetics. Note that this initial 'instantaneous' process has much nicer statistical properties than the coalescent, and that most of the information about the degree of selfing comes from the distribution of variability within and between individuals (Nordborg and Donnelly, 1997).

7.5.2 Males and Females

Next consider a diploid population that consists of N_m breeding males and N_f breeding females so that $N = N_m + N_f$. The discussion will be limited to *autosomal* genes, that is, genes that are not sex-linked. With respect to the genealogy of such genes, the total population can be thought of as a haploid population of size $2N$, divided into two patches of size $2N_m$ and $2N_f$, respectively, each of which is further divided into patches of size 2, as in the previous section. Clearly, a lineage currently in a male has probability $\frac{1}{2}$ of coming from a male in the previous generation, and probability $\frac{1}{2}$ of coming from a female. Within a sex, all individuals are equally likely to be chosen. The model looks a lot like a structured Wright–Fisher model with $M = 2$, $c_m = N_m/N$, $c_f = N_f/N$, and

$b_{mf} = b_{fm} = \frac{1}{2}$, the only difference being that two distinct lineages in the same individual must have come from individuals of different sexes in the previous generation, and thus do not migrate independently of each other. However, because states involving two distinct lineages in the same individual are instantaneous, this difference can be shown to be irrelevant. Pairs of lineages in different individuals (regardless of sex) coalesce in the previous generation if and only if both members of the pair came from: (a) the same sex; (b) the same diploid individual within that sex; and (c) the same haploid parent within that individual. This occurs with probability

$$\frac{1}{4} \times \frac{1}{N_m} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{N_f} \times \frac{1}{2} = \frac{N_m + N_f}{8N_m N_f}, \tag{7.15}$$

or, in the limit $N \to \infty$, with time measured in units of $2N$ generations, and c_m and c_f held constant, at rate $\alpha = (4c_m c_f)^{-1}$ (in accordance with the strong-migration limit result above). Alternatively, if time is measured in units of

$$2N_e = 2N/\alpha = \frac{8N_m N_f}{N_m + N_f} \tag{7.16}$$

generations, the standard coalescent is obtained (cf. Wright, 1931). Note that if $N_m = N_f = N/2$, the correct scaling is again the standard one of $2N$.

7.6 RECOMBINATION

In the era of genomic polymorphism data, the importance of modeling recombination can hardly be overemphasized (see also Hudson, this volume). When viewed backward in time, recombination (in the broad sense that includes phenomena such as gene conversion and bacterial conjugation in addition to crossing over) causes the ancestry of a chromosome to spread out over many chromosomes in many individuals. The lineages branch, as illustrated in Figure 7.9. The genealogy of a sample of recombining DNA sequences can thus no longer be represented by a single tree: it becomes a graph instead. Alternatively, since the genealogy of each point in the genome (each base pair, say) *can* be represented by a tree, the genealogy of a sample of sequences may be envisioned as a 'walk through tree space'.

7.6.1 The Ancestral Recombination Graph

As was first shown by Hudson (1983), incorporating recombination into the coalescent framework is in principle straightforward. The following description is based on the elegant 'ancestral recombination graph' of Griffiths and Marjoram (1996; 1997), which is closely related to Hudson's original formulation (for different approaches, see Simonsen and Churchill, 1997; Wiuf and Hein, 1999b).

Consider first the ancestry of a single ($n = 1$) chromosomal segment from a diploid species with two sexes and an even sex ratio. As shown in Figure 7.9, each recombination event (depicted here as crossing over at a point – we will return to whether this is reasonable below) in its ancestry means that a lineage splits into two, when going backward in time. Recombination spreads the ancestry of the segment over many chromosomes, or

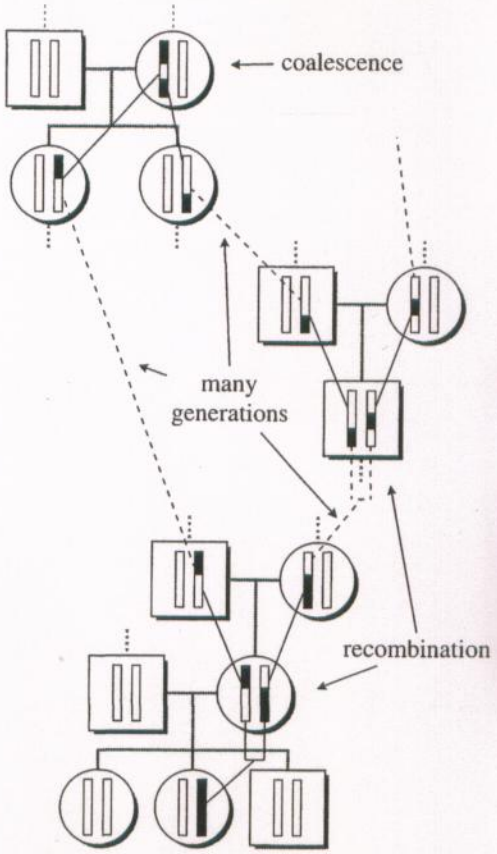

Figure 7.9 The genealogy of a DNA segment (colored black) subject to recombination both branches and coalesces. Note also that the genealogy of the sexually produced *individuals* (the pedigree) is very different from the genealogy of their *genes*.

rather over many 'chromosomal lineages'. However, as also shown in Figure 7.9, these lineages will coalesce in the normal fashion, and this will tend to bring the ancestral material back together on the same chromosome (Wiuf and Hein, 1997).

To model this, let the per-generation probability of recombination in the segment be r, define $\rho := \lim_{N\to\infty} 4Nr$, and measure time in units of $2N$ generations. Then the (scaled) time till the first recombination event is exponentially distributed with rate $\rho/2$ in the limit as N goes to infinity. Furthermore, once recombination has created two or more lineages, we find that these lineages undergo recombination independently of one another, and that simultaneous events can be neglected. This follows from standard coalescent arguments analogous to those presented for migration above. The only thing that may be slightly nonintuitive about recombination is that *the lineages we follow never recombine with each other* (the probability of such an event is vanishingly small): they always recombine with the (infinitely many) nonancestral chromosomes.

Each recombination event increases the number of lineages by one, and because lineages recombine independently, the total rate of recombination when there are k lineages is $k\rho/2$. Each coalescence event decreases the number of lineages by one, and the total rate of coalescence when there are k lineages is $k(k-1)/2$, as we have seen previously. Since lineages are 'born' at a linear rate, and 'die' at a quadratic rate, the number of lineages is guaranteed to stay finite and will even hit one occasionally (there will then temporarily be a single ancestral chromosome again Wiuf and Hein, 1997).

A sample of n lineages behaves in the same way. Each lineage recombines independently at rate $\rho/2$, and each pair of lineages coalesces independently at rate 1. The number of lineages *will* hit one occasionally. The segment in which this first occurs is known as the 'Ultimate' MRCA, because, as we shall see, each point in the sample may well have a younger MRCA[1].

The genealogy of a sample of n lineages back to the Ultimate MRCA can thus be described by a branching and coalescing graph (an 'ancestral recombination graph') that is analogous to the standard coalescent. A realization for $n = 6$ is shown in Figure 7.10.

What does a lineage in the graph look like? For each point in the segment under study, it must contain information about *which* (if any) sample members it is ancestral to. It is convenient to represent the segment as a (0,1) interval (this is just a coordinate system that can be translated into base pairs or whatever is appropriate). An ancestral lineage can then be represented as a set of elements of the form {interval, labels}, where the intervals are those resulting from all recombinational breakpoints in the history of the sample (Fisher's 'junctions' (Fisher, 1965) for aficionados of classical population genetics) and the labels denote the descendants of that segment (using the 'equivalence class' notation introduced previously). An example of this notation is given in Figure 7.10. Note that pieces of a given chromosomal lineage will often be ancestral to no one in the sample. Indeed, recombination in a nonancestral piece may result in an entirely nonancestral lineage!

So far nothing has been said about where or how recombination breakpoints occur. This has been intentional, to emphasize that the ancestral recombination graph does not depend on (most) details of recombination. It is possible to model almost any kind of recombination (including, for example, various forms of gene conversion) in this framework. But of course the graph has no meaning unless we interpret the recombination events somehow. To proceed, we will assume that each recombination event results in crossing over at a point, x, somewhere in (0,1). How x is chosen is again up to the modeler: it could be a fixed point; it could be a uniform random variable; or it could be drawn from some other distribution (perhaps centered around a 'hotspot'). In any case, a breakpoint needs to be generated for each recombination event in the graph. We also need to know which branch in the graph carries which recombination 'product' (remember that we are going backward in time). With breaks affecting a point, a suitable rule is that the left branch carries the material to the 'left' of the breakpoint (i.e. in $(0, x)$), and the right branch carries the material to the 'right' (i.e. in $(x, 1)$).

Once recombination breakpoints have been added to the graph, it becomes possible to extract the genealogy for any given point by simply following the appropriate branches. Figure 7.10 illustrates how this is done. An ancestral recombination graph contains a number of embedded genealogical trees, each of which can be described by the standard coalescent, but which are obviously not independent of each other. An alternative way of viewing this process is thus as a 'walk through tree space' along the chromosome (Wiuf and Hein 1999). The strength of the correlation between the genealogies for linked points depends on the scaled genetic distance between them, and goes to zero as this distance goes to infinity. The number of embedded trees equals the number of breakpoints plus

[1] The recent claims that human mtDNA may have recombined (Eyre-Walker, Smith and Maynard Smith, 1999; Hagelberg, Goldman, Liò, Whelan, Schiefenhövel, Clegg and Bowden, 1999) have led to the conclusion that recombination would imply that mitochondrial Eve never existed. This is false: Eve must still have existed, but she would not have the significance she is normally given. But then Eve without recombination does not have the significance she is normally given either – *plus ça change?*

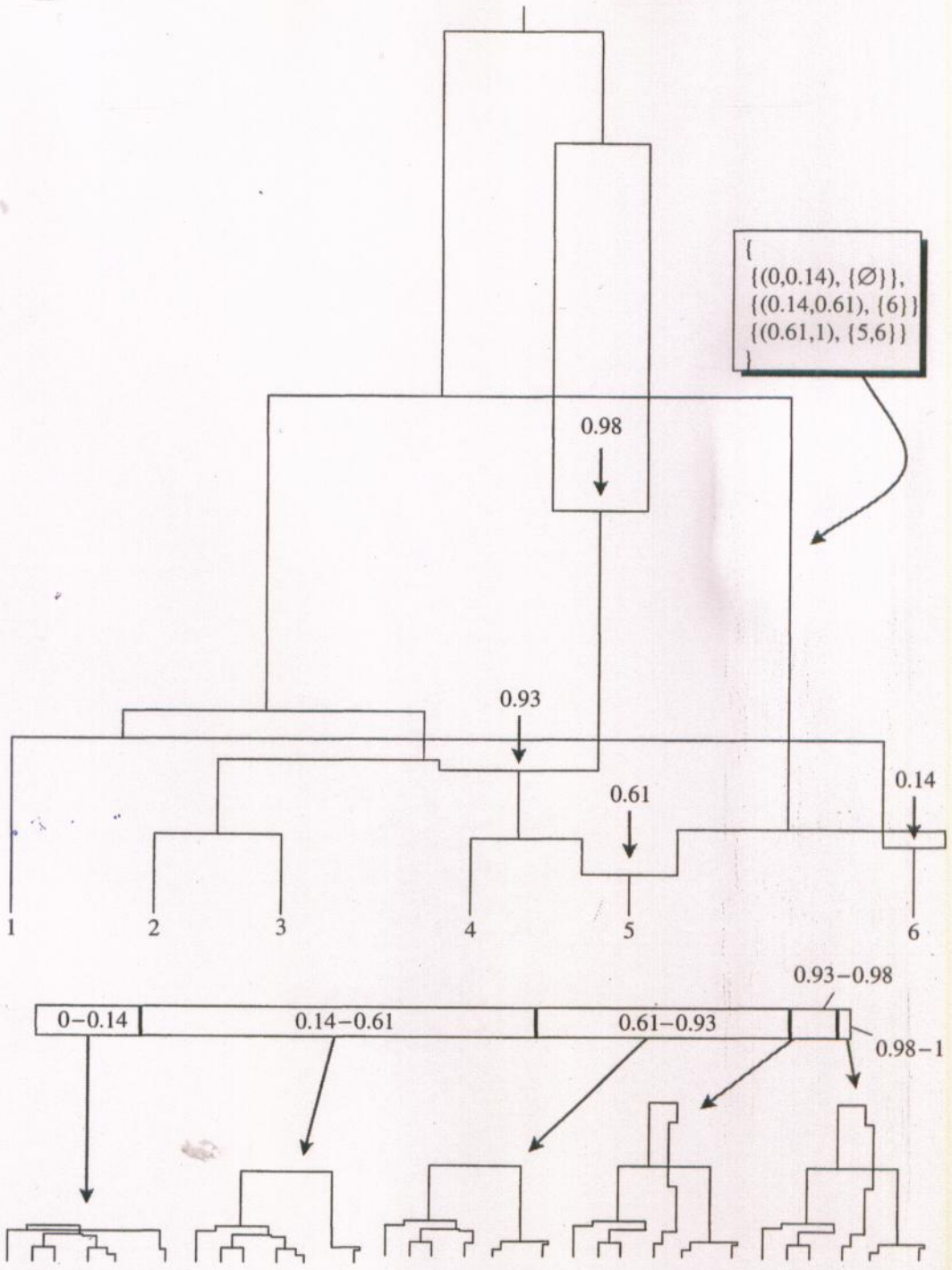

Figure 7.10 A realization of an ancestral recombination graph for $n = 6$. There were four recombination events, which implies $6 + 4 - 1 = 9$ coalescence events. Each recombination was assumed to lead to crossing over at a point, which was chosen randomly in $(0, 1)$. Four breakpoints (or 'junctions') implies five embedded trees, which are shown underneath. The tree for a particular chromosomal point is extracted from the graph by choosing the appropriate path at each recombination event. I have followed the convention that one should 'go left' if the point is located 'to the left' of (is less than) the breakpoint. Note that the two rightmost trees are identical. The box illustrates notation that may be used to represent ancestral lineages in the graph. The lineage pointed to is ancestral to: no (sampled) segment for the interval $(0, 0.14)$; segment 6 for the interval $(0.14, 0.61)$; and segments 5 and 6 for the interval $(0.61, 1)$.

one, but many of these trees may (usually will) be identical (cf. the two rightmost trees in Figure 7.10). Note also that the embedded trees vary greatly in height. This means that some pieces will have found their MRCA long before others. Indeed, it is quite possible for every piece to have found its MRCA long before the Ultimate MRCA. A number of interesting results concerning the number of recombination events and the properties of the embedded trees are available (see Griffiths and Marjoram, 1996; 1997; Hudson, 1983; 1987; Hudson and Kaplan, 1985; Kaplan and Hudson, 1985; Pluzhnikov and Donnelly, 1996; Simonsen and Churchill, 1997; Wiuf and Hein, 1999a,b; and Hudson this volume).

7.6.2 Properties and Effects of Recombination

It probably does not need to be pointed out that the stochastic process just described is extremely complicated. At least I have found that whereas it is possible to develop a fairly good intuitive understanding of the random trees generated by the standard coalescent, the behavior of the random recombination graphs continues to surprise. It may therefore be worth questioning first of all whether it is necessary to incorporate recombination. It would seem reasonable that recombination could be ignored if it is sufficiently rare in the segment studied (e.g. if the segment is very short). But what is 'sufficiently rare'? Consider a pair of segments. The probability that they coalesce before either recombines is

$$\frac{1}{1+2(\rho/2)} = \frac{1}{1+\rho} \tag{7.17}$$

(cf. equation (7.11)). In order for recombination not to matter, we would need to have $\rho \approx 0$. It is thus the *scaled* recombination rate that matters, not the per-generation recombination probability. Estimates based on comparing genetic and physical maps indicate that the average per-generation per-nucleotide probability of recombination is of roughly the same order of magnitude as the average per-generation per-nucleotide probability of mutation (which can be estimated in various ways). This means that the scaled mutation and recombination rates will also be of the same order of magnitude, and, thus, that recombination can be ignored when mutation can be ignored. In other words, as long we restrict our attention to segments short enough not to be polymorphic, we do not need to worry about recombination!

Of course, both recombination and mutation rates vary widely over the genome, so regions where recombination can be ignored almost certainly exist. Unfortunately, whereas direct estimates of recombination probabilities (genetic distances) are restricted to large scales, estimates of the recombination rate from polymorphism data are extremely unreliable (Griffiths and Marjoram, 1996; Hudson, 1987; Hudson and Kaplan, 1985; Wakeley, 1997; Wall, 2000). The latter problem is unavoidable. The main reason is the usual one that there is only a single realization of the underlying genealogy. Thus, for example, numerous recombination events in a particular region of a gene do not necessarily mean that it is a recombinational hotspot: it could just be that that region has a deep enough genealogy for multiple recombination events to have had time to occur. This is the same problem that affects estimates of the mutation rate.

However, there are also problems peculiar to recombination (see also Hudson, this volume). It is important to realize that most recombination events are undetectable (Hudson and Kaplan, 1985). Recombination in sequence data has often been inferred by identifying 'tracts' that have obviously moved from one sequence to another. The presence of such

tracts is actually indicative of *low* rather than of high recombination rates (Maynard Smith, 1999). Even a moderate amount of recombination will wipe out the tracts. Recombination can then only be 'detected' through the 'four-gamete test' (Hudson and Kaplan, 1985): the four linkage configurations *AB*, *Ab*, *aB*, and *ab* for two linked loci can only arise through recombination or repeated mutation (which is more likely is debatable (Eyre-Walker et al., 1999; Templeton et al., 2000)). Recombination events can clearly only be detected if there is sufficient polymorphism. However, many recombination events can *never* be detected even with infinite amounts of polymorphism (Griffiths and Marjoram, 1997; Hudson and Kaplan, 1985; Nordborg, 2000). Consider, for example, the two rightmost trees in Figure 7.10. These trees are identical. This means that the recombination event that gave rise to them cannot possibly leave any trace.

The phenomenon of undetectable breakpoints turns out to have special relevance for models with inbreeding. The 'forward' intuition that corresponds to undetectable recombination events is that these events took place in homozygous individuals. Inbreeding increases the frequency of homozygous individuals, and can therefore have a considerable effect on the recombination graph. It turns out that this effect can also be modeled as a scaling change, but this time of the recombination rate. Thus, for example, the recombination graph in a partially selfing hermaphrodite reduces to the standard recombination graph if we introduce an 'effective recombination rate', $\rho_e := \rho(1 - F)$ (Nordborg, 2000). Recombination breaks up haplotypes much less efficiently in inbreeders.

So far, we have only discussed the problems associated with recombination. It must be remembered that recombination is the only thing that allows us to get around the 'single underlying genealogy'. Unlinked loci will, with respect to most questions, provide independent samples. Of course this also applies within a segment: if ρ were infinite, then each base pair would be an independent locus (Pluzhnikov and Donnelly, 1996). High rates of recombination are thus an enormous advantage for many purposes.

Finally, it should be noted that since crossing over is mechanistically tied to gene conversion, there is reason to question the applicability of the simple model used above at the intragenic scale (Andolfatto and Nordborg, 1998; Nordborg, 2000). However, the ancestral recombination graph is quite general, and more realistic recombination models have been developed (Wiuf, 2000; Wiuf and Hein, 2000). Models of other kinds of recombination, such as bacterial transformation (Hudson, 1994) and intergenic gene conversion (Bahlo, 1998), also exist.

7.7 SELECTION

The coalescent depends crucially on the assumption of selective neutrality, because if the allelic state of a lineage influences its reproductive success, it is not possible to separate 'descent' from 'state'. Nonetheless, it turns out that it is possible to circumvent this problem, and incorporate selection into the coalescent framework. Two distinct approaches have been used. The first is an elegant extension of the coalescent process, known as the 'ancestral selection graph' (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997). The genealogy is generated backward in time, as in the standard coalescent, but it contains branching as well as coalescence events. The result is a genealogical graph that is superficially similar to the one generated by recombination. Mutations are then superimposed forward in time, and, with knowledge of the state of each branch, the

graph is 'pruned' to a tree by preferentially removing bad branches (i.e. those carrying selectively inferior alleles). In a sense, the ancestral selection graph allows the separation of descent from state by including 'potential' descent: lineages that might have lived, had their state allowed it.

The second approach is based on two insights. First, a polymorphic population may be thought of as subdivided into *allelic classes* within which there is no selection. Second, if we know the historical sizes of these classes, then they may be modeled as analogous to patches, using the machinery described above. Lineages then 'mutate between classes' rather than 'migrate between patches'. This approach was pioneered in the context of the coalescent by Kaplan et al. (1988). Knowing the past class sizes is the same as knowing the past allele frequencies, so it is obviously not possible to study the dynamics of the selectively different alleles themselves using this approach. However, it is possible to study the effects of selection on the underlying genealogical structure, which is relevant if we wish to understand how linked neutral variants are affected.

It is not entirely clear how the two approaches relate to each other. Since the second approach requires knowledge of the past allele frequencies, it may be viewed as some kind of limiting (strong selection) or, alternatively, conditional version of the selection graph (Nordborg, 1999). However, whereas the second approach would be most appropriate for very strong, deterministic selection, the selection graph requires all selection coefficients to be $O(1/N)$. This is an area of active research.

The ancestral selection graph is described by Neuhauser (this volume), and will not be discussed here. The second approach, which might be called the 'conditional structured coalescent', will be illustrated through three simple but very different examples.

7.7.1 Balancing Selection

By 'balancing selection' is meant any kind of selection that tends to maintain two or more alleles in the population. The effect of such selection on genealogies has been studied by a number of authors (Hey, 1991; Hudson and Kaplan, 1988; Kaplan et al., 1988; Kaplan et al., 1991; Nordborg, 1997; 1999; Takahata, 1990; Vekemans and Slatkin, 1994) (although the following treatment, which incorporates the ancestral recombination graph, has not previously been published). We will limit ourselves to the case of two alleles, A_1 and A_2, maintained at constant frequencies p_1 and $p_2 = 1 - p_1$ by strong selection. Alleles mutate to the other type with some small probability v per generation, and we define the scaled rate $\nu := 4Nv$. Reproduction occurs according to a diploid Wright–Fisher model, as for the recombination graph above.

Consider a segment of length ρ that contains the selected locus. Depending on the allelic state at the locus, the segment belongs to either the A_1 or the A_2 allelic class. Say that it belongs to the A_1 allelic class. Trace the ancestry of the segment a single generation back in time. It is easy to see that its creation involved an $A_2 \rightarrow A_1$ mutation with probability

$$\frac{vp_2}{vp_2 + (1-v)p_1} = \frac{\nu}{4N} \times \frac{p_2}{p_1} + O\left(\frac{1}{N^2}\right) \tag{7.18}$$

(cf. equation (7.9)), and involved recombination with probability $r = \rho/(4N)$. Thus the probability that neither happens is $1 - O(1/N)$, and the probability of two events, for example both mutation and recombination, is $O(1/N^2)$ and can be neglected. If nothing happens, then the lineage remains in the A_1 class. If there was a mutation, the lineage

'mutates' to the A_2 allelic class. If there was a recombination event, we have to know the genotype of the individual in which the event took place.

Because the lineage we are following is A_1, we know that the individual must have been either an A_1A_1 homozygote or an A_1A_2 heterozygote. What fraction of the A_1 alleles was produced by each genotype? In general, this will depend on their relative fitness as well as their frequencies. Let x_{ij} be the frequency of A_iA_j individuals, and w_{ij} their relative fitness. Then the probability that an A_1 lineage was produced in a heterozygote is

$$\frac{w_{12}x_{12}/2}{w_{12}x_{12}/2 + w_{11}x_{11}}. \tag{7.19}$$

If we can ignore the differences in fitness, and assume Hardy–Weinberg equilibrium (see Nordborg, 1999, for more on this), equation (7.19) simplifies to

$$\frac{p_1 p_2}{p_1 p_2 + p_1^2} = p_2. \tag{7.20}$$

Thus the probability that an A_1 lineage 'meets' and recombines with an A_2 segment is equal to the frequency of A_2 segments, which is intuitive. The analogous reasoning applies to A_2 lineages, which recombine with A_1 segments with probability p_1, and with members of their own class with probability p_2. It should be noted that the above can be made rigorous using a model that treats genotypes as well as individuals as population structure (Nordborg, 1999).

What happens when the lineage undergoes recombination? If it recombines in a homozygote, then both branches remain in the A_1 allelic class. However, if it recombines in a heterozygote, then one of the branches (the one *not* carrying the ancestry of the selected locus) will 'jump' to the A_2 allelic class. The other branch remains in the A_1 allelic class.

When more than two lineages exist, coalescences may occur, but only within allelic classes (remember that since mutation is $O(1/N)$ it is impossible for lineages to mutate and coalesce in the same generation).

If time is measured in units of $2N$ generations, and we let N go to infinity, the model converges to a coalescent process with the following types of events:

- each pair of lineages in the A_i allelic class coalesces independently at rate $1/p_i$;
- each lineage in A_i recombines with a segment in class j at rate ρp_j;
- each lineage in A_i mutates to A_j, $j \neq i$, at rate $\nu p_j/p_i$.

The process may be stopped either when the Ultimate MRCA is reached, or when all points have found their MRCA.

This model has some very interesting properties. Consider a sample that contains both types of alleles. Since coalescence is only possible within allelic classes, the selected locus (in the strict sense of the word, i.e. the 'point' in the segment where the selectively important difference lies) cannot coalesce without at least one mutation event. If mutations are rare, then this may have occurred a very long time ago. In other words, the polymorphism may be ancient. All coalescences will occur within allelic classes before mutation allows the final two lineages to coalesce. The situation is similar to strong population subdivision (see Figure 7.7). However, this is only true for the locus itself: linked pieces may 'recombine away' and coalesce much earlier. This will usually result in a

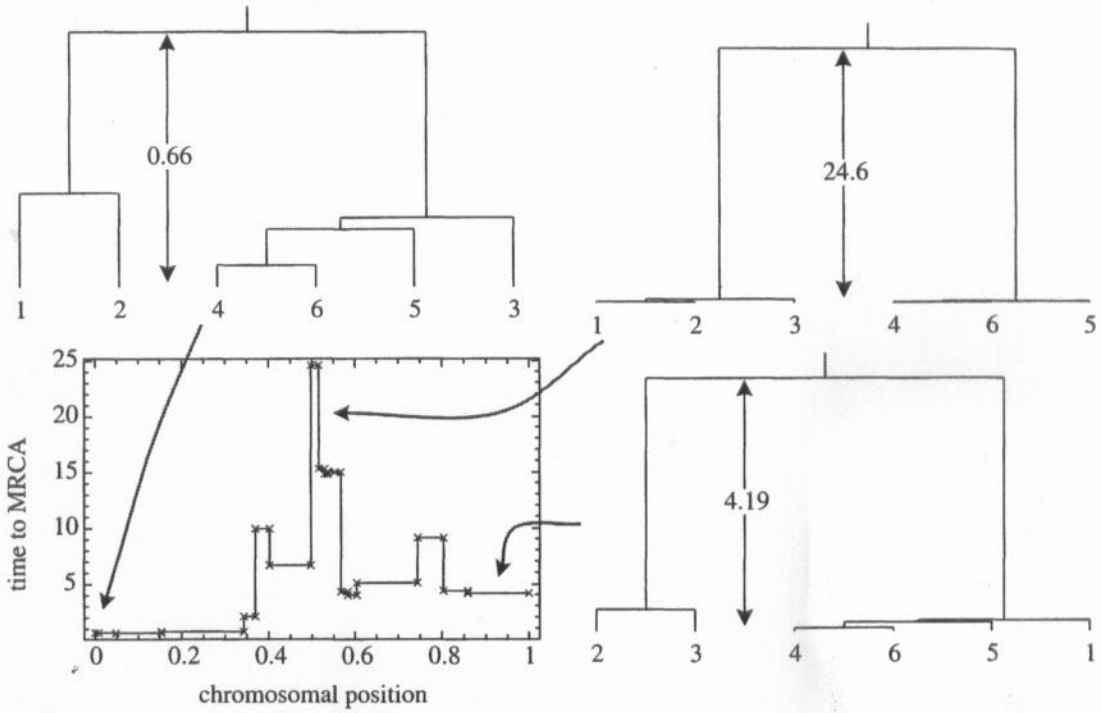

Figure 7.11 Selection will have a local effect on genealogies. A realization of the coalescent with recombination and strong balancing selection is shown. Lineages 1–3 belong to one allelic class, and lineages 4–6 to the other. The selected locus is located in the middle of the region. The plot shows how the time to the MRCA varies along the chromosome (the crosses denote cross-over points). The three extracted trees exemplify how the topology and branch lengths are affected by linkage to the selected locus. Note that the trees are not drawn to scale (the numbers on the arrows give the heights).

local increase in the time to MRCA centered around the selected locus, as illustrated in Figure 7.11. Because the expected number of mutations is proportional to the height of the tree, this may lead to a 'peak of polymorphism' (Hudson and Kaplan, 1988).

7.7.2. Selective Sweeps

Next consider a population in which favorable alleles arise infrequently at a locus, and are rapidly driven to fixation by strong selection. Each such fixation is known as a 'selective sweep' for reasons that will become apparent. This process can be modeled using the framework developed above, if we know how the allele frequencies have changed over time. Of course we do not know this, but if the selection is strong enough, it may be reasonable to model the increase in frequency of a favorable allele deterministically (Kaplan et al., 1989).

Consider a population that is currently not polymorphic, but in which a selective sweep recently took place. During the sweep, there were two allelic classes just as in the balancing selection model above. The difference is that these classes changed in size over time. In particular, the class corresponding to the allele that is currently fixed in the population will *shrink* rapidly back in time. The genealogy of the selected locus itself (in the 'point' sense used above) will therefore behave as if it were part of a population that has expanded from a very small size (cf. Figure 7.6). Indeed, unlike 'real' populations, the allelic class *will* have grown from a size of 1. A linked point must have grown in the

same way, unless recombination in a heterozygote took place between the point and the selected locus. Whether this happens or not will depend on how quickly the new allele increased. Typically, it depends on the ratio r/s, where s is the selective advantage of the new allele, and r is the relevant recombination probability.

The result of such a fixation event is thus to cause a local 'genealogical distortion', just like balancing selection. However, whereas the distortion in the case of balancing selection looks like population subdivision, the distortion caused by a fixation event looks like population growth. Close to the selected site, coalescence times will have a tendency to be short, and the genealogy will have a tendency to be starlike (cf. Figure 7.6). Note that a single recombination event in the history of the sample can change this, and that the variance will consequently be enormous (note the variance in time to MRCA in Figure 7.11). Shorter coalescence times mean less time for mutations to occur, so a local reduction in variability is expected. This is obvious: when the new allele sweeps through the population and fixes, it causes linked neutral alleles to 'hitchhike' along and also fix (Maynard Smith and Haigh, 1974). Repeated selective sweeps can thus decrease the variability in a genomic region (Kaplan et al., 1989; Simonsen et al., 1995). Because each sweep is expected to affect a bigger region the lower the rate of recombination is, this has been proposed as an explanation for the correlation between polymorphism and local rate of recombination that is observed in many organisms (Begun and Aquadro, 1992; Nachman, 1997; Nachman et al., 1998).

7.7.3 Background Selection

We have seen that selection can affect genealogies in ways reminiscent of strong population subdivision and of population growth. It is often difficult to distinguish statistically between selection and demography for precisely this reason (Fu and Li, 1993; Tajima, 1989b). It is also possible for selection to affect genealogies in a way that is completely undetectable, that is, as a linear change in time scale. This appears to be the case for selection against deleterious mutations, at least under some circumstances (Charlesworth et al., 1995; Hudson and Kaplan, 1994; 1995; Nordborg, 1997; Nordborg et al., 1996).

The basic reason for this is the following. Strongly deleterious mutations are rapidly removed by selection. Looking backward in time, this means that each lineage that carries a deleterious mutation must have a nonmutant ancestor in the near past. On the coalescent time scale, lineages in the deleterious allelic class will 'mutate' (backward in time) to the 'wild-type' allelic class instantaneously. The process looks like a strong-migration model, with the wild-type class as the source environment, and the deleterious class as the sink environment: the presence of deleterious mutations increases the variance in reproductive success. The resulting reduction in the effective population size is known as 'background selection' (Charlesworth et al., 1993).

More realistic models with multiple loci subject to deleterious mutations, recombination, and several mutational classes turn out to behave similarly. The strength of the background selection effect at a given genomic position will depend strongly on the local rate of recombination, which determines how many mutable loci influence a given point. Thus, deleterious mutations have also been proposed as an explanation for the correlation between polymorphism and local rate of recombination referred to above (Charlesworth et al., 1993). The 'effective population size' would thus depend on the mutation, selection, and recombination parameters in each genomic region.

same way, unless recombination in a heterozygote took place between the point and the selected locus. Whether this happens or not will depend on how quickly the new allele increased. Typically, it depends on the ratio r/s, where s is the selective advantage of the new allele, and r is the relevant recombination probability.

The result of such a fixation event is thus to cause a local 'genealogical distortion', just like balancing selection. However, whereas the distortion in the case of balancing selection looks like population subdivision, the distortion caused by a fixation event looks like population growth. Close to the selected site, coalescence times will have a tendency to be short, and the genealogy will have a tendency to be starlike (cf. Figure 7.6). Note that a single recombination event in the history of the sample can change this, and that the variance will consequently be enormous (note the variance in time to MRCA in Figure 7.11). Shorter coalescence times mean less time for mutations to occur, so a local reduction in variability is expected. This is obvious: when the new allele sweeps through the population and fixes, it causes linked neutral alleles to 'hitchhike' along and also fix (Maynard Smith and Haigh, 1974). Repeated selective sweeps can thus decrease the variability in a genomic region (Kaplan et al., 1989; Simonsen et al., 1995). Because each sweep is expected to affect a bigger region the lower the rate of recombination is, this has been proposed as an explanation for the correlation between polymorphism and local rate of recombination that is observed in many organisms (Begun and Aquadro, 1992; Nachman, 1997; Nachman et al., 1998).

7.7.3 Background Selection

We have seen that selection can affect genealogies in ways reminiscent of strong population subdivision and of population growth. It is often difficult to distinguish statistically between selection and demography for precisely this reason (Fu and Li, 1993; Tajima, 1989b). It is also possible for selection to affect genealogies in a way that is completely undetectable, that is, as a linear change in time scale. This appears to be the case for selection against deleterious mutations, at least under some circumstances (Charlesworth et al., 1995; Hudson and Kaplan, 1994; 1995; Nordborg, 1997; Nordborg et al., 1996).

The basic reason for this is the following. Strongly deleterious mutations are rapidly removed by selection. Looking backward in time, this means that each lineage that carries a deleterious mutation must have a nonmutant ancestor in the near past. On the coalescent time scale, lineages in the deleterious allelic class will 'mutate' (backward in time) to the 'wild-type' allelic class instantaneously. The process looks like a strong-migration model, with the wild-type class as the source environment, and the deleterious class as the sink environment: the presence of deleterious mutations increases the variance in reproductive success. The resulting reduction in the effective population size is known as 'background selection' (Charlesworth et al., 1993).

More realistic models with multiple loci subject to deleterious mutations, recombination, and several mutational classes turn out to behave similarly. The strength of the background selection effect at a given genomic position will depend strongly on the local rate of recombination, which determines how many mutable loci influence a given point. Thus, deleterious mutations have also been proposed as an explanation for the correlation between polymorphism and local rate of recombination referred to above (Charlesworth et al., 1993). The 'effective population size' would thus depend on the mutation, selection, and recombination parameters in each genomic region.

It should be pointed out that, unlike the many limit approximations presented in this chapter, the idea that background selection can be modeled as a simple scaling is not mathematically rigorous. However, we would rather hope that selection against deleterious mutations can be taken care of this way, because given that amino acid sequences are conserved over evolutionary time, practically all of population genetics theory would be in trouble otherwise!

7.8 NEUTRAL MUTATIONS

Not much has been said about the neutral mutation process because it is trivial from a mathematical point of view. Once we know how to generate the genealogy, mutations can be added afterwards according to a Poisson process with rate $\theta/2$, where θ is the scaled per-generation mutation probability. Thus, if a particular branch has length τ units of scaled time, the number of mutations that occur on it will be Poisson-distributed with mean $\tau\theta/2$ (and they occur with uniform probability along the branch). It is also possible to add mutations while the genealogy is being created, instead of afterwards. This can in some circumstances lead to much more efficient algorithms (see, for example, the 'urn scheme' described by Donnelly and Tavaré, 1995), although from the point of view of simulating samples, all coalescent algorithms are so efficient that such fine-tuning does not matter. However, it can matter greatly for the kinds of inference methods described by Stephens (this volume).

It should be noted that the mutation process is just as general as the recombination process. Almost any neutral mutation model can be used. A useful trick is so-called 'Poissonization': let mutation events occur according to a simple Poisson process with rate $\theta/2$, but once an event occurs, determine the *type* of event through some kind of transition matrix which includes mutation back to self (i.e. there was no mutation). This allows models where the mutation probability depends on the current allelic state.

The only restriction is that in order to interpret samples generated by the coalescent as samples from the relevant stationary distribution (which incorporates demography, migration, selection at linked sites, etc.), we need to be able to choose the type of the MRCA from the stationary distribution of the mutation process (alone, since demography, for example, does not affect samples of size $n = 1$). In many cases, such as the infinite-alleles model (each mutation gives rise to a new allele) or the infinite-sites model (each mutation affects a new site), the state of the MRCA does not matter, since all we are interested in is the number of mutational changes.

7.9 CONCLUSION

7.9.1 The Coalescent and 'Classical' Population Genetics

The differences between coalescent theory and 'classical' population genetics have frequently been exaggerated or misunderstood. First, the basic models do not differ. The coalescent is essentially a diffusion model of lines of descent. This can be done forward in time, for the whole population (e.g. Griffiths, 1980), but it was realized in the early 1980s that it is easier to do it backward in time. Second, the coalescent is not limited to finite

samples. Everything above has been limited to finite samples because it is mathematically much easier, but it is likely that all of it could be extended to the whole population. Of course, it is essential for the independence of events that the number of lineages be finite, but in the whole-population coalescent the number of lineages becomes finite infinitely fast (it is an 'entrance boundary', e.g. Griffiths, 1984). Third, classical population genetics is not limited to the whole population. A sample of size $n = 6$ from a K-allele model, say, could be obtained either through the coalescent, or by first drawing a population from the stationary distribution found by Wright (1949), and then drawing six alleles conditional on this population. Note, however, that it would be rather more difficult (read 'impossible') to use the second approach for most models. Fourth, the coalescent is in no sense tied to sequence data: any mutation model can be used. The impression that it is came about doubtless because models for sequence evolution such as the infinite-sites model are indeed impossibly hard to analyze using classical methods (Ethier and Griffiths, 1987).

I would argue that the real difference is more philosophical. As has been pointed out by Ewens (1979; 1990), essentially all of classical population genetics is 'prospective', looking forward in time. Another way of saying this is that it is conditional: given the state in a particular generation, what will happen? This approach is fine when modeling is done to determine 'how evolution might work' (which is what most classical population genetics was about). It is usually not suitable for statistical analysis of data, however. Wright considered how 'heterozygosity' would decay from the same starting point in infinitely many identical populations, that is to say, he took the expectation over evolutionary realizations. Data, alas, come from a single time-slice of one such realization. The coalescent forces us to acknowledge this, and allows the utilization of modern statistical methods, such as the calculation of likelihoods for samples.

7.9.2 The Coalescent and Phylogenetics

If the differences between coalescent theory and classical population genetics have sometimes been exaggerated, the differences between coalescent theory and phylogenetics have not always been fully appreciated. The central role played by trees in both turns out to be very misleading.

To be able to compare them, we need to model speciation. This has usually been done using an 'isolation' model in which randomly mating populations split into two completely isolated ones at fixed times in the past. The result is a 'species tree', within which we find 'gene trees' (see Figure 7.12). The model is quite simple: lineages will tend to coalesce within their species, and can only coalesce with lineages from other species back in the ancestral species.

Molecular phylogenetics attempts to estimate the species tree by estimating the genealogy of homologous sequences from the different species, that is, by estimating the gene tree. The species tree is assumed to exist and is treated as a model parameter.

In addition, the standard methods rely on all branches in the species tree being very long compared to within-species coalescence times. This means that the coalescent can be ignored: regardless of how we sample, all (neutral) gene genealogies will rapidly coalesce within their species, and thereafter have the same topology as the species tree. Furthermore, the variation in the branch lengths caused by different coalescence times in the ancestral species will be negligible compared to the lengths of the interspecific branches. There is no need to sample more than one individual per species, and recombination is completely irrelevant. Gene trees perfectly reflect the species tree.

Figure 7.12 A gene tree within a species tree.

It is of course widely acknowledged that gene trees and species trees may differ (see Avise, 1994; Avise and Ball, 1990; Hey, 1994; Hudson, 1992; Li, 1997; Nei, 1987; Takahata, 1989; Wu, 1991; phylogenetic methods are discussed by Huelsenbeck, this volume; and by Penny, this volume). Nonetheless, phylogenetic methods would not work unless the interspecific branches usually were long enough for the gene trees to reflect the species tree closely. Indeed, in many situations, the problem is the opposite: the branches are so long that repeated mutations have erased much of the phylogenetic information.

Phylogenetic inference can thus be viewed as a 'missing data' problem just like population genetic inference: polymorphisms contain information about an unobserved genealogy, which in turn provides information about an evolutionary model. However, note that in phylogenetics, there is relatively little doubt about what the right model is (it is typically an isolation model that gives rise to a species tree, as in Figure 7.12). Furthermore, because of the long branch lengths, the gene genealogies, although random variables with a coalescent distribution under the model, can be treated as parameters (which, among other things, means that we do not need to know the sizes of ancestral populations to estimate divergence times). None of this is true when analyzing population genetic data (which, strictly speaking, means any data for which the 'long branch' assumption above is not fulfilled). Unfortunately, the considerable success and popularity of phylogenetics (coupled with the ready availability of user-friendly software) has sometimes led to the inappropriate application of phylogenetic methods. It is important to remember that a genealogical tree from a population (or several populations that have not been isolated for a very long time) does not have an obvious interpretation: it certainly contains information about the process that gave rise to it, but usually less than we would hope (for a simple example, see Nordborg, 1998).

7.9.3 Prospects

An important issue which has been almost completely neglected in this review is the coalescent conditional on information in the sample. Rather than looking at all possible evolutionary realizations, we may be interested in those likely to have given rise to a particular sample. This is relevant when attempting to estimate the age of particular mutations, and also for so-called linkage disequilibrium mapping (e.g. Griffiths and Tavaré, 1998; Slatkin, 1996; Slatkin and Rannala, 1997a,b; Wiuf and Donnelly, 1999; see also Stephens, this volume; Hudson, this volume).

A theme of this review has been the versatility and generality of the coalescent model. In particular, the structured approach could easily be used to model many more situations. However, it is certainly not the case that everything has been solved when it comes to extending the coalescent. A very important example that no one knows how to deal with is the infinite-sites model with deleterious mutations. Each mutation is weakly selected, but the total selection pressure is deterministic so the genetic information is preserved. The ancestral selection graph cannot be used, because selection is too strong on most genotypes. The conditional approach cannot be used because selection is too weak on other genotypes (and, furthermore, the number of allelic classes is infinitely large). To study this model, one is forced to rely on 1970s techniques and simulate entire (very small) populations forward in time (Hudson and Kaplan, 1995; Nordborg et al., 1996). This is unfortunate, given that weak selection on a very large number of sites is likely to have shaped most genomes.

Although data may force us to abandon the notion that there are parts of the genome not affected by natural selection, the importance of the coalescent as a null model will continue.

Acknowledgments

This work was supported by grants from the Swedish Natural Sciences Research Council (NFR B-AA/BU 12026) and the Erik Philip-Sörensen Foundation. I wish to thank David Balding, Bengt Olle Bengtsson, Malia Fullerton, Jenny Hagenblad, Maarit Jaarola, Martin Lascoux, Claudia Neuhauser, François Rousset, Matthew Stephens, and Torbjörn Säll for comments on the manuscript.

REFERENCES

Andolfatto, P. and Nordborg, M. (1998). The effect of gene conversion on intralocus associations. *Genetics* **148**, 1397–1399.

Avise, J.C. (1994). *Molecular Markers, Natural History and Evolution*. Chapman & Hall, New York.

Avise, J.C. and Ball, R.M. (1990). In *Oxford Surveys in Evolutionary Biology*, 7, D. Futuyama and J. Antonovics, eds. Oxford University Press, Oxford, pp. 45–67.

Bahlo, M. (1998). Segregating sites in a gene conversion model with mutation. *Theoretical Population Biology* **54**, 243–256.

Barton, N.H. and Wilson, I. (1995). Genealogies and geography. *Proceedings of the Royal Society London B* **349**, 49–59.

Begun, D.J. and Aquadro, C.F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520.

Charlesworth, B., Morgan, M.T. and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303.

Charlesworth, D., Charlesworth, B. and Morgan, M.T. (1995). The pattern of neutral molecular variation under the background selection model. *Genetics* **141**, 1619–1632.

Donnelly, P. (1996). In *Variation in the Human Genome*, Ciba Foundation Symposium No. 197. Wiley, Chichester, pp. 25–50.

Donnelly, P. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics* **29**, 401–421.

Ethier, S.N. and Griffiths, R.C. (1987). The infinitely many sites model as a measure valued diffusion. *Annals of Probability* **5**, 515–545.

Ewens, W.J. (1979). *Mathematical Population Genetics*. Springer-Verlag, Berlin.

Ewens, W.J. (1990). In *Mathematical and Statistical Development of Evolutionary Theory*, S. Lessard, ed. Kluwer Academic, Dordrecht, pp. 177–227.

Eyre-Walker, A., Smith, N.H. and Maynard Smith, J. (1999). How clonal are human mitochondria? *Proceedings of the Royal Society London B* **266**, 477–483.

Fisher, R.A. (1965). *Theory of Inbreeding*, 2nd edition, Oliver and Boyd, Edinburgh.

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

Griffiths, R.C. (1980). Lines of descent in the diffusion approximation of neutral Wright–Fisher models. *Theoretical Population Biology* **17**, 37–50.

Griffiths, R.C. (1984). Asymptotic line-of-descent distributions. *Journal of Mathematical Biology* **21**, 67–75.

Griffiths, R.C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology* **3**, 479–502.

Griffiths, R.C. and Marjoram, P. (1997). In *Progress in Population Genetics and Human Evolution*, P. Donnelly and S. Tavaré, eds. Springer-Verlag, New York, pp. 257–270.

Griffiths, R.C. and Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London B* **344**, 403–10.

Griffiths, R.C. and Tavaré, S. (1998). The age of a mutant in a general coalescent tree. *Stochastic Models* **14**, 273–295.

Hagelberg, E., Goldman, N., Liò, P., Whelan, S., Schiefenhövel, W., Clegg, J.B. and Bowden, D.K. (1999). Evidence for mitochondrial DNA recombination in a human population of island Melanesia. *Proceedings of the Royal Society London B* **266**, 485–492.

Herbots, H.M. (1994). Stochastic models in population genetics: genealogy and genetic differentiation in structured populations. PhD thesis, University of London.

Hey, J. (1991). A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theoretical Population Biology* **39**, 30–48.

Hey, J. (1994). In *Molecular Ecology and Evolution: Approaches and Applications*, B. Schierwater, B. Streit, G.P. Wagner and R. DeSalle, eds. Birkhäuser, Basel, pp. 435–449.

Hudson, R.R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 183–201.

Hudson, R.R. (1987). Estimating the recombination parameter of a finite population model without selection. *Genetical Research, Cambridge* **50**, 245–250.

Hudson, R.R. (1990). In *Oxford Surveys in Evolutionary Biology*, Vol. 7, D. Futuyma and J. Antonovics, eds. Oxford University Press, Oxford, pp. 1–43.

Hudson, R.R. (1992). Gene trees, species trees and the segregation of ancestral alleles. *Genetics* **131**, 509–512.

Hudson, R.R. (1993). In *Mechanisms of Molecular Evolution*, N. Takahata and A.G. Clark, eds. Japan Scientific Societies Press, Tokyo, pp. 23–36.

Hudson, R.R. (1994). Analytical results concerning linkage disequilibrium in models with genetic transformation and conjugation. *Journal of Evolutionary Biology* **7**, 535–548.

Hudson, R.R. (1998). Island models and the coalescent process. *Molecular Ecology* **7**, 413–418.

Hudson, R.R. and Kaplan, N.L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.

Hudson, R.R. and Kaplan, N.L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840.

Hudson, R.R. and Kaplan, N.L. (1994). In *Non-neutral Evolution: Theories and Molecular Data*, G.B. Golding, ed. Chapman & Hall, New York, pp. 140–153.

Hudson, R.R. and Kaplan, N.L. (1995). Deleterious background selection with recombination. *Genetics* **141**, 1605–1617.

Kaplan, N.L. and Hudson, R.R. (1985). The use of sample genealogies for studying a selectively neutral *m*-loci model with recombination. *Theoretical Population Biology* **28**, 382–396.

Kaplan, N.L., Darden, T. and Hudson, R.R. (1988). The coalescent process in models with selection. *Genetics* **120**, 819–829.

Kaplan, N.L., Hudson, R.R. and Iizuka, M. (1991). The coalescent process in models with selection, recombination and geographic subdivision. *Genetical Research, Cambridge* **57**, 83–91.

Kaplan, N.L., Hudson, R.R. and Langley, C.H. (1989). The 'hitch-hiking' effect revisited. *Genetics* **123**, 887–899.

Kingman, J.F.C. (1982a). The coalescent. *Stochastic Processes and their Applications* **13**, 235–248.

Kingman, J.F.C. (1982b). In *Exchangeability in Probability and Statistics*, G. Koch and F. Spizzichino, eds. North-Holland, Amsterdam, pp. 97–112.

Kingman, J.F.C. (1982c). In *Essays in Statistical Science: Papers in Honour of P.A.P. Moran*, J. Gani and E.J. Hannan, eds. Applied Probability Trust, Sheffield, pp. 27–43. *Journal of Applied Probability*, special volume **19A**.

Krone, S.M. and Neuhauser, C. (1997). Ancestral processes with selection. *Theoretical Population Biology* **51**, 210–237.

Li, W.-H. (1997). *Molecular Evolution*. Sinauer, Sunderland, MA.

Marjoram, P. and Donnelly, P. (1997). In *Progress in Population Genetics and Human Evolution*, P. Donnelly and S. Tavaré, eds. Springer-Verlag, New York, pp. 107–131.

Maynard Smith, J. (1999). The detection and measurement of recombination from sequence data. *Genetics* **153**, 1021–1027.

Maynard Smith, J. and Haigh, J. (1974). The hitchhiking effect of a favourable gene. *Genetical Research, Cambridge* **23**, 23–35.

Möhle, M. (1998a). A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Advances in Applied Probability* **30**, 493–512.

Möhle, M. (1998b). Robustness results for the coalescent. *Journal of Applied Probability* **35**, 438–447.

Möhle, M. (1999). Weak convergence to the coalescent in neutral population models. *Journal of Applied Probability* **36**, 446–460.

Nachman, M.W. (1997). Patterns of DNA variability at *X*-linked loci in *Mus domesticus*. *Genetics* **147**, 1303–1316.

Nachman, M.W., Bauer, V.L., Crowell, S.L. and Aquadro, C.F. (1998). DNA variability and recombination rates at *X*-linked loci in humans. *Genetics* **150**, 1133–1141.

Nagylaki, T. (1980). The strong-migration limit in geographically structured populations. *Journal of Mathematical Biology* **9**, 101–114.

Nagylaki, T. (1982). Geographical invariance in population genetics. *Journal of Theoretical Biology* **99**, 159–172.

Nagylaki, T. (1998). The expected number of heterozygous sites in a subdivided population. *Genetics* **149**, 1599–1604.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Neuhauser, C. and Krone, S.M. (1997). The genealogy of samples in models with selection. *Genetics* **145**, 519–534.

Nordborg, M. (1997). Structured coalescent processes on different time scales. *Genetics* **146**, 1501–1514.

Nordborg, M. (1998). On the probability of Neanderthal ancestry. *American Journal of Human Genetics* **63**, 1237–1240.

Nordborg, M. (1999). In *Statistics in Molecular Biology and Genetics*, IMS Lecture Notes Monograph Series **33**, F. Seillier-Moiseiwitsch, ed. Institute of Mathematical Statistics, Hayward, CA, pp. 56–76.

Nordborg, M. (2000). Linkage disequilibrium, gene trees, and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**, 923–929.

Nordborg, M. and Donnelly, P. (1997). The coalescent process with selfing. *Genetics* **146**, 1185–1195.

Nordborg, M., Charlesworth, B. and Charlesworth, D. (1996). The effect of recombination on background selection. *Genetical Research, Cambridge* **67**, 159–174.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured populations. *Journal of Mathematical Biology* **29**, 59–75.

Notohara, M. (1993). The strong-migration limit for the genealogical process in geographically structured populations. *Journal of Mathematical Biology* **31**, 115–122.

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

Pollak, E. (1987). On the theory of partially inbreeding finite populations. I. Partial selfing. *Genetics* **117**, 353–360.

Pulliam, H.R. (1988). Sources, sinks, and population regulation. *American Naturalist* **132**, 652–661.

Rogers, A.R. and Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology Evolution* **9**, 552–569.

Rousset, F. (1999a). Genetic differentiation in populations with different classes of individuals. *Theoretical Population Biology* **55**, 297–308.

Rousset, F. (1999b). Genetic differentiation within and between two habitats. *Genetics* **151**, 397–407.

Saunders, I.W., Tavaré, S. and Watterson, G.A. (1984). On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability* **16**, 471–491.

Simonsen, K.L. and Churchill, G.A. (1997). A Markov chain model of coalescence with recombination. *Theoretical Population Biology* **52**, 43–59.

Simonsen, K.L., Churchill, G.A. and Aquadro, C.F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429.

Slatkin, M. (1987). The average number of sites separating DNA sequences drawn from a subdivided population. *Theoretical Population Biology* **32**, 42–49.

Slatkin, M. (1996). Gene genealogies within mutant allelic classes. *Genetics* **143**, 579–587.

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

Slatkin, M. and Rannala, B. (1997a). Estimating the age of alleles by use of intraallelic variability. *American Journal of Human Genetics* **60**, 447–458.

Slatkin, M. and Rannala, B. (1997b). The sampling distribution of disease-associated alleles. *Genetics* **147**, 1855–1861.

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. (1989a). DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. *Genetics* **123**, 229–240.

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

Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genetical Research, Cambridge* **52**, 213–222.

Takahata, N. (1989). Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* **122**, 957–966.

Takahata, N. (1990). A simple genealogical structure of strongly balanced allelic lines and trans-species polymorphism. *Proceedings of the National Academy of Sciences (USA)* **87**, 2419–2423.

Takahata, N. (1991). Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. *Genetics* **129**, 585–595.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetic models. *Theoretical Population Biology* **26**, 119–164.

Templeton, A.R., Clark, A.G., Weiss, K.M., Nickerson, D.A., Boerwinkle, E. and Sing, C.F. (2000). Recombinational and mutational hotspots within the human lipoprotein lipase gene. *American Journal of Human Genetics* **66**, 69–83.

Vekemans, X. and Slatkin, M. (1994). Gene and allelic genealogies at a gametophytic self-incompatibility locus. *Genetics* **137**, 1157–1165.

Wakeley, J. (1996). Distinguishing migration from isolation using the variance of pairwise differences. *Theoretical Population Biology* **49**, 369–386.

Wakeley, J. (1997). Using the variance of pairwise differences to estimate the recombination rate. *Genetical Research, Cambridge* **69**, 45–48.

Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics* **153**, 1863–1871.

Wall, J.D. (2000). A comparison of estimators of the population recombination rate. *Molecular Biology Evolution* **17**, 156–163.

Wilkinson-Herbots, H.M. (1998). Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology* **37**, 535–585.

Wiuf, C. (2000). A coalescence approach to gene conversion. *Theoretical Population Biology* **57**, 357–367.

Wiuf, C. and Donnelly, P. (1999). Conditional genealogies and the age of a neutral mutant. *Theoretical Population Biology* **56**, 183–201.

Wiuf, C. and Hein, J. (1997). On the number of ancestors to a DNA sequence. *Genetics* **147**, 1459–1468.

Wiuf, C. and Hein, J. (1999a). The ancestry of a sample of sequences subject to recombination. *Genetics* **151**, 1217–1228.

Wiuf, C. and Hein, J. (1999b). Recombination as a point process along sequences. *Theoretical Population Biology* **55**, 248–259.

Wiuf, C. and Hein, J. (2000). The coalescent with gene conversion. *Genetics* **155**, 451–462.

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

Wright, S. (1949). In *Genetics, Palaeontology, and Evolution*, G.L. Jepson, G.G. Simpson and E. Mayr, eds. Princeton University Press, Princeton, NJ, pp. 365–389.

Wu, C.-I. (1991). Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* **127**, 429–435.