10

Inferences about the structure and history of populations: coalescents and intraspecific phylogeography

JOHN WAKELEY
Department of Organismic and Evolutionary Biology, Harvard University, Cambridge

10.1 Introduction

Population geneticists and phylogeneticists view tree structures differently. To the phylogeneticist, tree structures are the objects of study and the branching patterns a tree displays are inherently significant. Phylogeneticists are interested in the relationships among species or other taxa, and these histories are tree-like structures. To the population geneticist, particularly to the student of coalescent theory, individual tree structures are usually not of interest. Instead attention is focused on the characteristics of populations or species, and intraspecific trees, or gene genealogies, are a stepping stone on the path to such knowledge. This difference in approach divides workers who study current and historical population structure into two groups: those who ascribe significance to single gene trees and those who focus on summary properties of gene trees over many loci. The purpose of this chapter is to give some perspective on this division and to suggest ways of identifying the domain of application of coalescents and intraspecific phylogeography in terms of the histories of populations or species. This is not meant to be divisive. In the not too distant future, we can hope that these complementary approaches will be unified, as models catch up with data and a science of population genomics is realized.

10.1.1 Population genetics history

Theoretical population genetics was born out of the tension between Biometricians (or Darwinians) and Mendelians in the early decades of last century. We often trace our field back to the famous paper of Fisher (1918) which settled this dispute; see Provine (1971). In short, the Biometricians, represented by W. F. R. Weldon and Karl Pearson, had for decades been measuring quantitative traits and considering such things as the correlation of traits between parents and offspring. They maintained that natural selection acted on these

The Evolution of Population Biology, ed. R. S. Singh and M. K. Uyenoyama. Published by Cambridge University Press.

continuous characters and that change in these was slow; discrete variation was unimportant to evolution. After the rediscovery of Mendel's laws in 1900, William Bateson, Hugo de Vries, and other Mendelians argued for the importance of discrete variations in evolution. Their views were directly opposed to those of the Biometricians; selection on continuous variation could not result in significant evolutionary steps, which were discontinuous. In hindsight we might say that the Biometricians' mistake was to confuse the continuity of traits with that of the underlying variation, and the Mendelians' error was to equate the mechanism of inheritance with that of evolution itself. In any case, it is clear that the two camps agreed only on one point: continuous variation and Mendelian inheritance were incompatible.

This fundamental conflict was resolved mathematically by Fisher (1918). Specifically, Fisher showed that continuous variation could be explained by the action of many Mendelian loci of small effect. In the decade or so after this remarkable start, the major results of this new branch of science, which was called theoretical population genetics, were laid down by Fisher (1930), Haldane (1932), and Wright (1931). Following the birth of theoretical population genetics, the mathematical theory was extended and the facts of genetics were reconciled with Darwin's theory of evolution. During the Modern Synthesis, these avenues of research were merged into the neo-Darwinian theory of evolution, providing a series of well-justified, more or less qualitative explanations of patterns of speciation, adaptation, and geographic variation. Two of the major architects of the Modern Synthesis were Dobzhansky (1937) and Mayr (1942). Our modern understanding of evolution is grounded in neo-Darwinism. During the next few decades, many workers contributed to the theory, although Malécot (1948) and Kimura (1955a,b) certainly stand out. By 1960 the mathematical theory of population genetics had developed a very high degree of sophistication, although for the most part, as Lewontin (1974) notes, this was in the absence of genetic data.

It wasn't until the mid 1960s that population genetics finally confronted genetic data (Harris 1966, Lewontin and Hubby 1966). Since then, we have seen a grand shift in population genetics from the forward-looking view of the classical theory of Fisher, Haldane, and Wright to the backward-looking view of the coalescent or genealogical approach; see Ewens (1990) for a review of this transformation. The modern approach focuses on inferences from samples of genetic data and, often to great advantage, recasts theoretical problems in terms of genealogies. Significant works along the path to this include Ewens (1972), which describes the distribution of the counts of alleles in a moderate-sized sample from a large population, and Watterson (1975), which describes the distribution of the number of polymorphic nucleotide sites in either a moderate or a large sample from a large population. The retrospective approach came fully to life in the early 1980s with the introduction of the coalescent process by Kingman (1982a,b,c), Hudson (1983b), and Tajima (1983). The present relative lack of concern for the structures of particular

gene genealogies traces back to the constant-size, single-population, neutral coalescent model described in these works, which is discussed in detail in Section 10.2 below.

10.1.2 Phylogenetics and intraspecific phylogeography

Charles Darwin's famous book contains just one figure: a hypothetical phylogenetic tree. Long before Darwin (1859) and Wallace (1858) put forward the idea of descent with modification, biologists had employed trees to depict the relationships among species and higher taxa. Tree structures are a natural way to represent such affinities, which are groups nested within other groups. Prior to Darwin and Wallace, however, trees had been employed strictly as convenient organizational tools to represent systematic affinities. For example, the classification system put forward by Linnaeus (1735) is a branching structure which delineates relationships, yet Linnaeus rejected the idea of evolution. When the idea of descent with modification gained acceptance as the explanation for biological diversity, these tree structures gained a new significance. They were no longer an expedient, but rather represented the actual histories of groups of species. The development of phylogenetics since Darwin and Wallace has been strongly influenced by the concept of trees as history. In addition phylogenetic theory and methodology have been shaped by the evolutionary idea that descendant species which trace back to a common ancestor will inherit any unique characteristics that ancestral species had evolved.

Until the last 30 years or so, the role of theory in phylogenetics and in population genetics could not have been more different. Although there is now a lot of overlap of approach, historical differences do persist. Theoretical population genetics has always been firmly grounded in traditional applied mathematics and probability theory. In this sense population genetics has many parallels with physics. The theoretical framework is mathematical and statistical, and there is broad acceptance of this framework and its attendant models within the field of biology.

In contrast, within the field of phylogenetics there has been widespread skepticism of such approaches, particularly statistical ones. This is most evident in the cladistic approach, which practitioners credit to Hennig (1965, 1966). This approach seeks to identify the phylogenetic tree which disagrees the least with the data at hand. The criterion for it is parsimony: pick the tree that requires the fewest character state changes. The tree is then considered a potentially true statement about history. It is a phylogenetic hypothesis which predicts what further study should uncover and which thus may be shown to be false. It is not viewed as an estimate of some unknown quantity. This approach is understandable if Hennig's view is accepted: that the phylogeneticist can directly observe (the results of) history through careful study of the morphology and development of a group of organisms, by identifying shared, uniquely derived characters, or synapomorphies. Sound arguments against the cladistic

approach have been made in response to seeing the blind application of the parsimony method to data which have not been subject to the careful prior study Hennig envisioned, and which are more labile than complex morphological features. Thus, with the introduction of model-based approaches, like the maximum likelihood method of Felsenstein (1981), the recent history of phylogenetics has been a progressive acceptance of the mathematical and statistical theory. However, this process of acceptance is still ongoing.

Coincident with the emergence of the backward-looking, genealogical approach to population genetics, phylogenetic methods began to be applied to intraspecific data. This was greatly facilitated by the nonrecombining nature of the first molecule examined – animal mitochondrial (mt) DNA – and the growing technical ability during the 1970s and 1980s to assay samples of mtDNA from natural populations. The result was a new and active subfield of evolutionary biology called intraspecific phylogeography, or just phylogeography (Avise *et al.* 1987, Avise 1989, 2000). A number of new methods of historical inference have resulted from this approach (Neigel *et al.* 1991, Neigel and Avise 1993, Templeton *et al.* 1995, Templeton 1998). The hallmark of phylogeography is that inferences are drawn from intraspecies or organismal gene trees which are reconstructed from data. The focus on gene trees as indicators of population structure, population history, and speciation has provided a much needed bridge between phylogenetics and population genetics (Hey 1994, Avise 2000). However, there is still a gulf between workers schooled in population genetics and those who favor traditional phylogenetics or cladistics. Bluntly put, the latter group tends to place too much emphasis on single gene genealogies whereas the former group places too little. Drawing conclusions from single genealogies can be problematic because each is only a single point in the space of all possible genealogies. Under some kinds of population histories, this will cause serious errors in inference. Conversely, focusing too much on the standard, structure-less, history-less coalescent model gives a picture of the utility of single gene trees that is too discouraging.

10.2 Gene genealogies and the coalescent

In the early 1980s, the ancestral process known as the coalescent was described. Kingman (1982a,b,c) provided a mathematical proof of the result. Hudson (1983b) and Tajima (1983) introduced this genealogical approach to population geneticists and derived many biologically relevant results. Nordborg (2001) provides a recent review; see also Hudson (1990) and Donnelly and Tavaré (1995). Kingman found a simple ancestral process to hold for samples from a wide variety of different types of populations, in the limit of large population size and providing that the genetic lineages in the population are exchangeable (Cannings 1974). Exchangeable lineages are ones whose predicted properties are unchanged if they are relabeled or permuted (Kingman

1982b, Aldous 1985). With the assumption that all variation is neutral, the familiar Wright–Fisher model (Fisher 1930, Wright 1931) of a population with nonoverlapping generations fits this criterion, as does the overlapping generation model of Moran (1958). Different populations will differ in how the actual population size is related to the effective population size that determines the rate of the coalescent process. The standard coalescent involves two very important assumptions besides exchangeability. For this model to hold, the population must be of constant effective size over time and there must be no population subdivision.

When time is measured in units of $2N_e$ generations for a population of diploid organisms, or in units of N_e generations for a population of haploid organisms, the time to a coalescent event is exponentially distributed with mean

$$E(t_k) = \frac{2}{k(k-1)} \tag{10.1}$$

where k is the number of ancestral lineages present. Under the coalescent model, each of the $\binom{k}{2}$ possible pairs of lineages coalesces with rate 1. Without recombination, which will be treated later, a sample of size n will go through exactly $n-1$ coalescent events to reach the common ancestor of the entire sample. Thus, every genealogy has $n-1$ coalescent intervals, beginning with the most recent, $k = n$, and ending with the most ancient, $k = 2$. Figure 10.1 shows an average coalescent genealogy; that is, with the lengths of the coalescent intervals drawn in proportion to Equation 10.1. The more recent coalescent intervals tend to be much shorter than the ancient ones, and on average the final coalescent interval represents more than half of the total time from the present back to the most recent common ancestor of the sample. Because the time scale of the coalescent process depends inversely on N_e, we expect genealogies to be longer when the effective size is larger.

As we trace the ancestry of the lineages back in time, because each pair that exists has the same rate of coalescence, when a common ancestor event happens each pair is equally likely to be the one that coalesces. The structure of trees under the coalescent is determined by this process of joining random pairs of lineages. The result is, if we think in forward time for the moment starting at the root of the tree, a random-bifurcating tree topology. This results from the fact that there is no structure to the coalescent process – that all lineages are exchangeable – and the resulting trees are likewise unstructured. Without intralocus recombination, all the sites at a single genetic locus will share the same genealogy. Loci that segregate independently of each other will have uncorrelated genealogies, both in terms of the coalescent times and topology. Considering topological structure, if we took a sample of three items, and labeled them A, B, and C, then each of the three possible rooted tree topologies – $((A, B), C)$, $((A, C), B)$, $((B, C), A)$ – is equally likely to occur. If

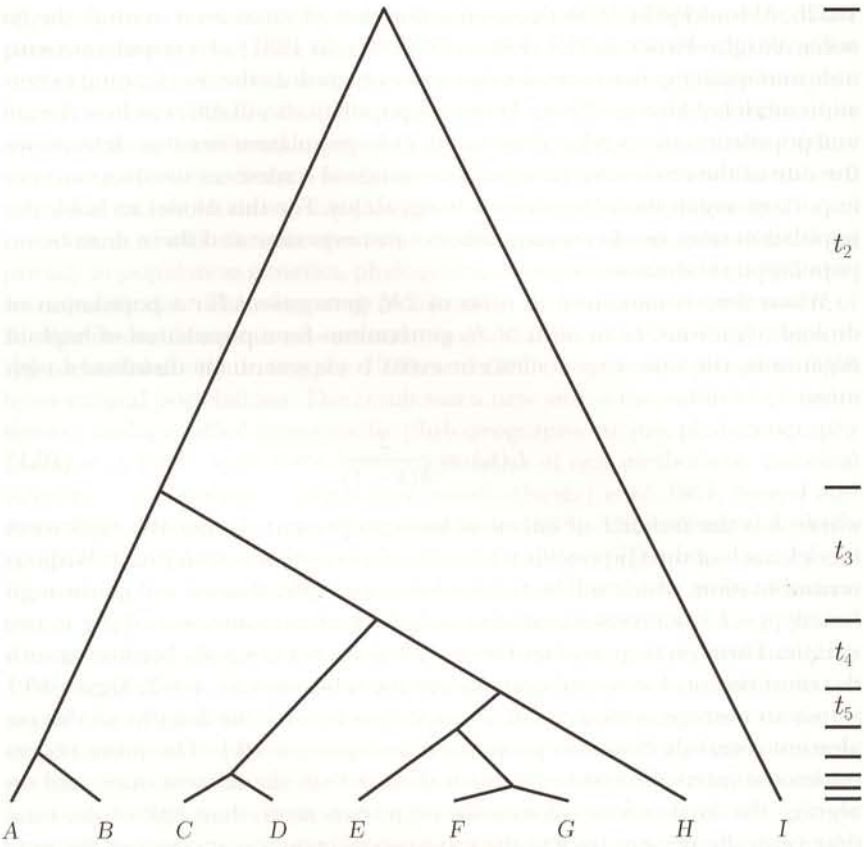

Figure 10.1. A hypothetical coalescent genealogy of a sample of size $N = 9$. The lengths of the coalescent intervals, t_n through t_2, are drawn in proportion to their expected values given by Equation 10.1.

we take a large sample of independently segregating loci, we expect to observe equal numbers of each of these three trees.

10.3 The axes of genealogical variation: tree size and branching pattern

As a starting point in talking about demographic history, we can take the standard, coalescent process as a null model. The underlying, exchangeable population genetic models, such as the Wright–Fisher model, are familiar to most biologists and their use as null models is not uncommon. This establishes predictions for what we should observe in a sample of sequences from a population. With reference to the discussion of the coalescent above, we are interested in two kinds of genealogical variation: (1) variation in the total

length of the tree, and (2) variation in the branching pattern. The length of a genealogy is the sum of the lengths of all its branches. Under the standard coalescent model, this is given by the sum of $n-1$, independent exponential times with different parameters. We expect this distribution to be realized when a large number of independent loci are sampled. The branching pattern of a genealogy specifies $2n-3$ partitions of the n sampled sequences, tips, or leaves of the tree. That is, each branch in the genealogy divides the members of the sample into two groups, the ones on either side of the branch. The genealogy or branching pattern at each sampled locus will be a random draw from the rather large universe of all possible random-bifurcating trees.

It is very important to note that our ability to observe the length and topology of genealogies is mediated by mutation. Even without any variation, genealogies will come in different sizes and shapes; we just won't know it. We rely on mutations occurring along the branches of the tree to produce the sequence polymorphisms that provide clues about history. The rate of mutation per locus is typically very small, somewhere around 10^{-4} to 10^{-6} per generation, and mutation events in different generations are independent. Therefore, the number of mutations that occur along a genetic lineage of length t will be Poisson distributed with expectation tu, where u is the mutation rate per generation. When time is rescaled as in the coalescent, this becomes $T\theta/2$, where $T = t/(2N_e)$ and $\theta = 4N_e u$. In the standard coalescent model, the parameter θ is equal to the expected number of nucleotide differences between two randomly chosen gene copies. The randomness of the mutation process is an important factor in determining among-locus variation in the observable indicator of tree length: the number of polymorphic sites in the sample. The letter S is used to denote the number of these segregating sites in a sample. Even when the genealogies at different loci are all identical in size there will be Poisson variation around the expectation due to the randomness of the mutation process. This imposes a lower bound on the variation in S among loci, namely that the variance will be equal to the mean.

Our ability to uncover genealogical topology also depends on mutation. We become aware of particular branches in the tree when mutations occur on them. When the mutation rate at each nucleotide site at a genetic locus is small, and recombination is absent or very unlikely, the infinite-sites mutation model of Watterson (1975) is a good approximation to the mutation process. Under this model, each time a new mutation occurs, it happens at a previously unmutated site. The assumption of no recombination guarantees that all sites in a sample of DNA sequences will share the same bifurcating topology, but this is not the most important aspect of Watterson's (1975) model. If each site mutates at most once in the history of the sample, then each polymorphism is the result of a single mutation event on some branch in the tree, and the partitions of the sample made by the branch and by the polymorphism are identical. Correlation in genealogical topologies among loci will be represented in sequence data by the repetition of such site frequency patterns at many loci.

10.4 The effects of population structure and population history on genealogies

This section describes the effects on the size and shape of genealogies of deviations from the assumptions of the standard coalescent model, particularly changes in effective size over time and two kinds of population subdivision. These effects are summarized in Figure 10.2. The thin lines in the figure represent population boundaries, and thin, dashed lines indicate incomplete barriers to the movement of individuals. The genealogies of these samples of size four are drawn using thick lines. For each historical scenario, (a) through (d), hypothetical genealogies are shown for samples from two independently segregating loci. This illustrates the effects of population structure and population history on the sizes and shapes of genealogies. Note that "shape" here refers only to topological structure and not to the relative lengths of different parts of a tree. In brief, changes in population size through time change the distribution of tree sizes by making the coalescence rate time dependent, but do not affect the topology of trees. Population subdivision alters the distribution of tree lengths, but it also can have dramatic effects on the shape of trees because it makes some common ancestor events much more likely than others.

10.4.1 Population growth

If one population is twice as big as another, the former has one half the rate of coalescence as the latter. On average, trees will be twice as big in the larger population as in the smaller one. When a single population has grown in size, the rate of coalescence responds proportionately. Looking back in time, the rate of coalescence will be low until the time of growth, then it will increase. The predictions of the standard coalescent for the relative sizes of ancient and recent coalescent intervals pictured in Figure 10.1 will no longer hold. Instead, the more recent intervals will be relatively longer and the more ancient intervals will be relatively shorter. If growth is rapid and relatively recent, genealogies will tend to be star shaped, that is, to have small internal branches (Slatkin and Hudson 1991). Population growth by itself will not alter the probabilities of genealogical topologies, because when a coalescent event occurs each pair of lineages still has an equal chance of being the one that coalesces.

If population growth is rapid enough, it is well approximated by a single abrupt change in population size. In this case, the ancestral process has two additional parameters: T_C, the time of change in population size measured in units of $2N_e$ (current effective size) generations, and $Q = N_{eA}/N_e$, the ratio of the ancestral and current effective population sizes. Between the present and time T_C, each pair of lineages coalesces with rate equal to one, whereas before T_C the rate is Q per pair of lineages. Of course, this model also describes

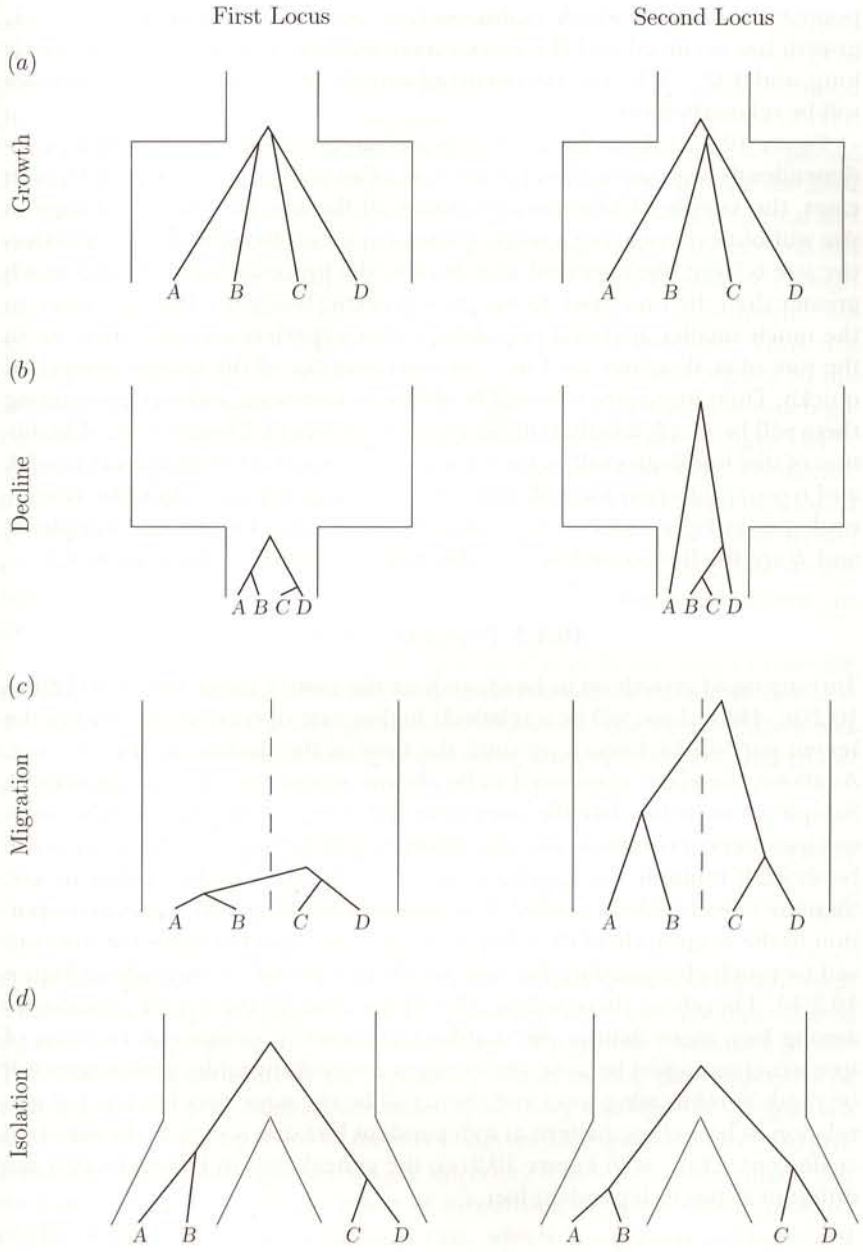

Figure 10.2. Two hypothetical genealogies for a sample of size four at two independently segregating loci under the four population models discussed in the text: (*a*) population growth, (*b*) population decline, (*c*) equilibrium migration, and (*d*) isolation without gene flow. Thin lines indicate population boundaries.

population decline, which is discussed in Section 10.4.2 below. If $Q < 1$, growth has occurred and the more recent coalescent times will be relatively long, and if $Q > 1$ decline has occurred and the most recent coalescent times will be relatively short.

Figure 10.2(a) shows the genealogies of samples from two hypothetical, independently segregating loci for the case of an abrupt growth event. In both cases, the sample of four lineages traces all the way back to the change in size without experiencing a single coalescent event. Because the recent effective size is large, the expected time back to the first coalescent event is much greater than the time back to the growth event. When the lineages arrive in the much smaller ancestral population, they experience a great increase in the rate of coalescence, and the common ancestor of the sample is reached quickly. Therefore, most trees will be about the same size, and variation among them will be much less than in Kingman's coalescent. However, the distribution of tree topologies will be the same as in the standard, constant-size model, and trees at different loci will differ in branching pattern. Thus, the two genealogies in Figure 10.2(a) have different structures. On the left, samples A and B are the first to coalesce, on the right it is B and C which are first.

10.4.2 Population decline

Turning rapid growth on its head, we have the case of rapid decline in Figure 10.2(b). Here there will be a relatively higher rate of coalescence during the recent part of the history, up until the time of the decline in effective size. As above, the event is assumed to be abrupt, simply for ease of explanation. Samples at some loci, like the one on the left in Figure 10.2(b), will trace back to a most recent common ancestor before reaching the event. These trees will be short. If multiple lineages trace their ancestry back to the decline in size, then the rate of coalescence for those remaining lineages decreases in proportion to the magnitude of the change in size. The ancient coalescent intervals will be much elongated in this case, which is depicted on the right in Figure 10.2(b). Therefore, there will be a lot of variation in the size of genealogies among loci, more than in the standard, constant-N_e coalescent. In terms of tree structure, again because the lineages are exchangeable, genealogies will be random-bifurcating trees and there will be the same very low level of correlation in branching pattern at independent loci that is seen in the standard coalescent. Thus, as in Figure 10.2(a), the genealogies in Figure 10.2(b) are different at two independent loci.

10.4.3 Equilibrium migration

Population subdivision introduces structure to genealogies, structure that may correlate with geography, and causes the tree topologies at different loci to be correlated. Subdivision will also affect variation in the sizes of genealogies

among loci, but the direction of this effect depends on whether migration can occur among subpopulations or demes, as this section supposes, or not, as in Section 10.4.4 below. For simplicity, assume that a population is subdivided into D demes and conforms to the symmetric island model of Wright (1931). The demes are of equal size, N, and the fraction of each deme that is replaced by migrants each generation is the same and equal to m. This is by far the most commonly employed model of a subdivided population in both empirical and theoretical studies. The term equilibrium migration refers to the fact that this constant-rate migration is supposed to have been ongoing for long enough that the effects of any prior history are erased. In Wright's island model, migrants are equally likely to come from any deme in the population. Thus, this model does not include explicit geography. Populations that adhere to the assumptions of the island model will not display the correlation between geography and genetic variation known as isolation by distance (Wright 1943). They will show different levels of polymorphism within vs. between demes, and powerful nonparametric tests to detect subdivision have been developed (Hudson *et al.* 1992). In the case of just two populations, the island model can be considered an explicit model of geography. This simple case is considered here in order to illustrate the effects of equilibrium migration on genealogies.

The parameters that determine the pattern of genetic variation in a sample of n_1 sequences from one deme and n_2 sequences from another are θ and $M = 4Nm$. If π_w and π_b are the average number of pairwise nucleotide differences within and between populations, respectively, then for the D-deme island model we have

$$E(\pi_w) = D\theta, \tag{10.2}$$

$$E(\pi_b) = D\theta\left(1 + \frac{1}{2M}\right) \tag{10.3}$$

(Li 1976). For the two-deme model, we put $D = 2$ in Equations 10.2 and 10.3. There are two surprising aspects of these equations. First, the expected value of π_w does not depend on the rate of migration (Slatkin 1987, Strobeck 1987). This is a special property of the symmetric island model: the tendencies of within-deme pairwise coalescence times to be short if neither of the pair is a migrant and to be long if one of them is a migrant average out perfectly to give Equation 10.2. If any asymmetries are introduced into the model, this result no longer holds. Second, the effect of subdivision depends on the product of the deme size and the migration rate, which is captured in the scaled migration rate M. As M grows large, the expectation of π_b converges on that of π_w, and the population will appear panmictic. This surprising result traces back to Wright (1931), and explains why populations that are obviously not panmictic sometimes show no evidence of subdivision. That is, M can be large even when the per-generation rate of migration, m, is small. Equations for the

variances of π_w and π_b both within and among loci can be found (Wakeley 1996a,b), and these both depend on the scaled migration rate. When M is large, the variances become those expected in a panmictic population, and as M decreases the variances of pairwise differences grow.

The predictions of Equations 10.2 and 10.3 can be extended to levels of polymorphism in larger samples: under equilibrium migration, levels of genetic variation will be larger on average for multi-deme samples than for single-deme samples. The effect of this will be greater when M is small. In the sample (n_1,n_2) from two demes, coalescent times among the n_1 sequences from deme one, and among the n_2 sequences from deme two, will tend to be shorter than coalescent times between sequences from different demes. This means that the topological structure of genealogies will no longer be the random-bifurcating trees predicted by the standard coalescent. There will be a tendency towards trees which have a branch that divides the sample exactly into the n_1 and n_2 sequences taken from each deme, for example trees in which the demic samples are reciprocally monophyletic. Again, this tendency will be more pronounced if the scaled migration rate between the two demes is small. Thus, the genealogies for two independent loci on the right and left of Figure 10.2(c) both show this kind of topology. In addition, variation in levels of polymorphism among loci will depend inversely on the scaled migration rate, M; for example, see Hey (1991). So, for the same average rate of polymorphisms under equilibrium migration, some loci will have very short and some very long histories. This is also displayed in Figure 10.2(c).

10.4.4 Isolation without gene flow

Equilibrium migration is just one of a multitude of possible explanations for the occurrence of subdivision. In fact, it is probably uncommon for a population to remain stably subdivided, both in the sizes of demes and in the rates and patterns of migration, for long enough to reach equilibrium. One of the earliest tenets to emerge from phylogeographic studies is that most species appear to have experienced dramatic shifts in demography over time and space (Avise 1989). Confining ourselves for the moment to models with discrete demes, the polar opposite of equilibrium migration is isolation and divergence without genetic exchange. This isolation model posits an ancestral population that splits into two descendant populations at some time, T_D, in the past and after that time the two populations do not exchange migrants. The isolation model can be compared with the migration model in Section 10.4.3 to illustrate the striking differences between equilibrium and nonequilibrium population subdivision.

In general, each population in the isolation model might be of a different size, and we would have $\theta_1 = 4N_1u$, $\theta_2 = 4N_2u$, and $\theta_A = 4N_Au$ as parameters (Wakeley and Hey 1997). However, for purposes of comparison with the equilibrium migration model of Section 10.4.3, we assume that $\theta_1 = \theta_2 = \theta_A$. In

this case, the average numbers of pairwise differences within and between demes have expected values

$$E(\pi_w) = \theta \tag{10.4}$$

$$E(\pi_b) = \theta(1 + T_D) \tag{10.5}$$

(Li 1977). Aside from a constant scaling factor (D), equilibrium migration and isolation without gene flow make identical predictions about average levels of genetic variation within and between demes where $T_D = 1/(2M)$. In other words, if π_w and π_b are measured from data, then both models could be fit and their parameters estimated, but π_w and π_b would not serve to distinguish between migration and isolation. The most obvious difference between the two models is in the interpretation of the pattern of polymorphism. Under the isolation model, genetic variation between demes in a sample is a snapshot for a particular T_D. If the population were sampled again at a later date, $T_D + T$, the level of divergence would be greater. Equation 10.3, in contrast, holds for all time, and represents a dynamic balance achieved between ongoing genetic drift and migration.

In addition to this difference in interpretation, variation in levels of genetic variation among loci will be different under migration and isolation even when the average levels are the same (Li 1976, 1977, Takahata and Nei 1985, Wakeley 1996a). The variances are larger under migration than under isolation, and the difference grows with $T_D = 1/(2M)$. This results from the fact that under migration, coalescent events between samples from different demes can occur at any time, mediated by migration, whereas under isolation there can be no interdeme coalescent events until the lineages trace back into the ancestral population. In the extreme of a very long divergence time in the isolation model ($T_D \gg 1$), difference between $E(\pi_b)$ and θT_D will be negligible. In this case the distribution of the number of segregating sites among loci will approach a Poisson distribution, with mean and variance equal to θT_D. In contrast, in the extreme of a very low migration rate in the migration model, the variance of the number of segregating sites among loci will be much greater than the mean (Wakeley 1996a). Thus, the trees for two independent loci under isolation in Figure 10.2(d) are more similar in size than those shown in Figure 10.2(c) for migration. Equilibrium migration and isolation without gene flow share the prediction that genealogical trees will tend towards reciprocal monophyly, and this is also displayed in Figure 10.2(d).

10.5 Domains of application: coalescents and phylogeography

The above discussion illustrates some general principles about the effect of population structure and population history on the sizes and shapes of genealogies. To summarize:

1. population growth/decline tends to decrease/increase variation in tree size among loci but does not affect variation in tree shape relative to the standard coalescent model,
2. both equilibrium and nonequilibrium population subdivision (migration vs. isolation above) alter the structure of genealogies such that genealogies at independently segregating loci will tend to share topological features, and
3. migration increases variation in tree size among loci whereas isolation decreases it.

This section investigates how the strengths of these trends depend on the parameters of a population. The goal is to identify population histories for which the analysis of single gene genealogies is likely to be fruitful and those for which it will be less useful to refer to any specific genealogy. Simulations are used to determine the distribution of tree size and shape among loci. The parameters are those discussed above in Section 10.4 and the quantities used to measure variation in the size and shape of genealogies are described below.

10.5.1 Measures of variation in tree size

The most straightforward measure of the size of a genealogy is the number of segregating sites, S. A sample from any population will have some expected value of S and some variance. For example, in the case of a sample of n sequences under the standard, constant size, unstructured coalescent with infinite-sites mutation,

$$E(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i} \tag{10.6}$$

$$V(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2} \tag{10.7}$$

(Watterson 1975). When we sample a large number of loci, we should find that the mean and variance among them would conform to Equations 10.6 and 10.7. This, of course, assumes that the sample size, n, and the mutation parameter, θ, are the same at every locus. However, this assumption is made only as a matter of convenience in comparing different population structures and histories below; it would be straightforward to allow for differences in θ and n among loci.

There are many ways in which we could compare levels of variation in S, our measure of tree size, among loci. The standardized measure,

$$\Omega = \frac{\widehat{V(S)} - \bar{S}}{\widehat{V(S)}} \tag{10.8}$$

will be used here, in which $\bar{S}$ is the average number of segregating sites and $\widehat{V(S)}$ is the observed variance of S among loci. Given a multilocus data set, Ω

is easy to compute. The expectation of Ω is given approximately by

$$E(\Omega) \approx \frac{V(S) - E(S)}{V(S)} \tag{10.9}$$

The number of segregating sites, S, is a compound random variable (see Section 10.3). Thus we can intuitively partition $V(S)$ into contributions due (1) to variation in tree size and (2) to variation in the mutation process. If there is no variation in the size of genealogies among loci, then all of the variation in S will be due to the Poisson mutation process and the expected value of Ω will be zero. Instead, if the variation in tree size among loci is much greater than the mean, then $V(S)$ will be large and Ω will be close to its upper bound of one. Thus, Ω is a normalized measure which can be compared under different assumptions about the population. Our null model, the standard coalescent, predicts a fairly high value of Ω, depending of course on θ and n. If $\theta = 10$ and $n = 20$, which are the values used in simulations below, Equation 10.9 gives $E(\Omega) = 0.82$.

10.5.2 Measures of correlation in branching pattern

There is also a multitude of ways we could compare genealogical topologies among loci. If we knew the true trees or if we were very confident about our trees reconstructed from data, then we could use a tree comparison metric like that of Robinson and Foulds (1981). Alternatively, if we are not confident about our reconstructed trees or do not wish to make explicit reference to them, we could use some measure of the correlation in haplotype patterns among loci such as coefficient of linkage disequilibrium (Lewontin and Kojima 1960). This measures gametic associations between alleles at two loci, but multilocus statistics are also possible (Smouse 1974). Here, because of the focus on simple two-deme models of subdivision, we will instead consider the co-occurrence of identical data partitions among loci, that is the observation of identical patterns of polymorphism among members of the sample at several loci. This presupposes that the same individuals were assayed at all genetic loci.

Assuming that the infinite-sites mutation model holds, each polymorphic site in a sample divides the members of the sample into two groups, ones which retain the ancestral base at the site and ones which have inherited the mutant base. As noted in Section 10.3, the one-to-one correspondence between mutation events and polymorphic sites in the sample, and the observation of a pattern in the data guarantee the existence of a branch in the genealogy of the sample, one that divides the sample exactly as the polymorphism does. For example, a mutation event on the shortest internal branch in the genealogy in Figure 10.1, the one which exists only during t_5, would make a polymorphic site at which samples E, F, and G would show the mutant base and samples A, B, C, D, H, and I would show the ancestral base.

In the standard coalescent model, we would not expect to see this pattern repeated at another, independent locus sequenced in the same individuals because the fraction of random-bifurcating trees that contain such a branch is very small. However, all genealogies contain n external branches, on which singleton polymorphisms can arise, so we would expect to see these partitions, i.e., all n kinds of singletons, repeated at many loci. Thus, there is a negative correlation between the allele frequency at a polymorphic site and the chance that the same pattern will be found at other loci.

In a sample from a subdivided population, we expect sites which divide the sequences along deme-sample lines to tend to be repeated at multiple loci. There might be a fairly low overall concordance of whole tree topologies among loci, because of the variability of within-deme patterns of common ancestry, but some branches would tend to be repeated. For the simple two-deme models considered here, these repeated branches will be the ones that divide the sample into the n_1 and n_2 sequences sampled from demes one and two. A statistic that will be sensitive to the co-occurrence of single partitions across loci is $max(p_i)$, in which p_i is the fraction of loci that show at least one polymorphic site with partition i. Singleton partitions are excluded in the calculation of $max(p_i)$ because all loci are expected to show these regardless of population structure and history. This measure will be sensitive to the effects of subdivision as it is modeled here. As the level of subdivision increases, the partition most frequently observed across loci will be the one that corresponds exactly to the two demes' samples, and $max(p_i)$ will approach one. We take the null distribution of $max(p_i)$ to be that found under the standard coalescent. This will depend on the sample size and on θ. For $\theta = 10$ and $n = 20$, used in the simulations below, the standard coalescent gives $max(p_i) \approx 0.04$.

10.5.3 Simulations of population structure and population history

The usual coalescent simulations were performed (Hudson 1990), adding a change in size, cf. Hudson (1990), or migration/isolation, cf. Wakeley (1996b), as indicated. The statistics Ω and $max(p_i)$ were computed for each simulation replicate. In addition to simulations under the standard coalescent model, a small set of parameter values was chosen to illustrate the effects of population structure and population history on the joint distribution of Ω and $max(p_i)$. The sample size was $n = 20$ when there was no structure, and $n_1 = n_2 = 10$ under migration and isolation. Only one case each of growth and decline is presented: ($\theta = 100.0$, $Q = 0.01$, $T_C = 0.1$) and ($\theta = 0.25$, $Q = 100.0$, $T_C = 0.1$). These were selected to represent extreme growth and extreme decline respectively, and the values of θ were chosen so that the average number of polymorphic sites per locus would be the same under both models. Several levels of subdivision were investigated for equilibrium migration and isolation without gene flow. Under migration these were $M = 0.5, 0.25, 0.01$ with $\theta = 5.0$, and under isolation they were

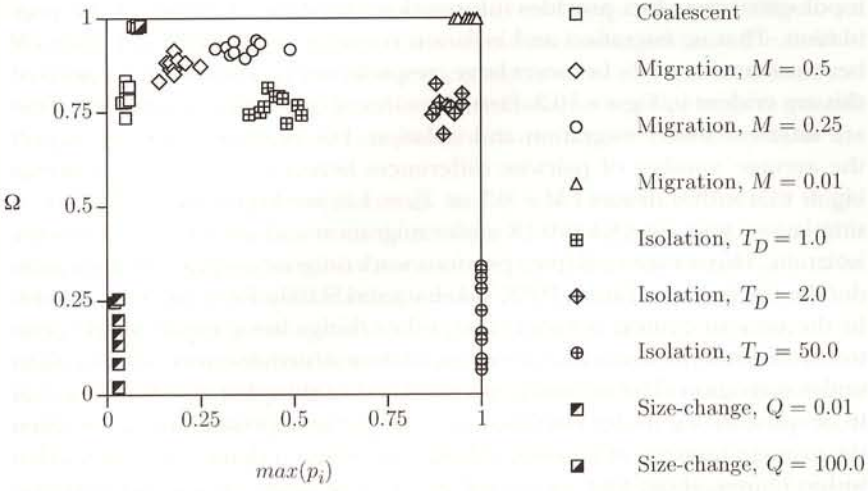

Figure 10.3. The results of the simulations described in the text. Each point in the scatterplot is the pair of $(max(p_i), \Omega)$ values for a single simulation replicate.

$T_D = 1.0, 2.0, 50.0$ with $\theta = 10.0$. These parameter sets were chosen in consideration of Equations 10.2 through 10.5, so that the expected numbers of pairwise differences within and between the two demes would be equivalent in the two models for three different levels of differentiation. One hundred independent loci were surveyed in the sampled individuals.

The results are shown in Figure 10.3. Only ten simulation replicates were performed for each set of parameters, as this was enough to distinguish the cases, and the results of all replicates are plotted in Figure 10.3. Simulations under the standard coalescent model cluster around the values $\Omega = 0.82$ and $max(p_i) \approx 0.04$ mentioned above. Under population growth and decline, the value of $max(p_i)$ is nearly unchanged from the constant-size case, but the value of Ω changes drastically. This accords well with the discussion in Section 10.4 above. The minor differences in $max(p_i)$ between these and the standard coalescent result from the fact that singleton polymorphisms are ignored in computing $max(p_i)$, and there are a lot more singletons under population growth than under population decline. This is essentially the same as the mutation rate effect on Ω that can be seen for the standard coalescent from Equations 10.6 and 10.7; as θ grows, so does the expected value of Ω. In sum, under this model of dramatic growth we expect the size of even a single genealogy to accurately represent the history of the population but, because there is no structure to the population, the topology of the tree contains little or no information about historical demography. Under decline, neither the size nor the shape of a single genealogy will be informative about history.

Subdivided populations vary both in Ω and in $max(p_i)$. Under both equilibrium and nonequilibrium subdivision, the repetition of genealogical

topologies across loci provides information about the structure of the population. That is, migration and isolation converge on $max(p_i) = 1$ when M becomes small and T_D becomes large, respectively. Two interesting aspects of this are evident in Figure 10.3. First, the rates of convergence to this extreme are different under migration and isolation. For example, when we expect the average number of pairwise differences between demes to be twice as big as that within demes ($M = 0.5$ or $T_D = 1.0$; see Equations 10.2 to 10.5), simulations give $max(p_i) \approx 0.18$ under migration and $max(p_i) \approx 0.45$ under isolation. This is expected from previous work on genealogical topologies under the two models (Tajima 1983, Takahata and Slatkin 1990, Wakeley 1996b). In the present context it means that, other things being equal, single gene trees will be more informative about population structure under isolation than under migration. The second point is related to this; that is, subdivision has to be quite strong under migration for $max(p_i)$ to approach one. Even when the average number of pairwise differences between demes is 50 times that within demes, about four out of 100 loci will not show the (n_1, n_2) partition that defines the samples. That equilibrium migration is a highly variable process can also be seen in values for Ω, which approach one as M decreases. In contrast, as T_D increases between two isolated demes, Ω decreases, but a very long divergence time is required for Ω to be close to zero.

The measures Ω and $max(p_i)$ appear to distinguish well among the models. In addition, they serve to illustrate how single gene trees might or might not be representative of population structure and population history in terms of the parameters of the models. The broad empty area of Figure 10.3, for lower values of Ω and intermediate values of $max(p_i)$, is an artifact of the simplicity of the models considered here. Populations that follow the isolation model but have a small value of θ_A relative to θ_1 and θ_2 can produce values in this range.

10.6 Conclusions

While reconstructing a genealogy is not a necessary step in population genetic inference, it can be quite informative under some circumstances. There is a difference of approach in this regard between workers who use coalescent techniques and those who practise intraspecific phylogeography. While this dichotomy is far from complete, it is real enough. Coalescent technicians do not usually make reference to particular gene trees. This is part of the culture of coalescents: that gene trees are unobservable random quantities which certainly shape genetic variation but whose branching patterns do not contain much information about population history. This view is most reasonable when populations conform to the standard coalescent model. When trees are referred to explicitly, it is typical to "integrate" over them in making inferences (Kuhner *et al.* 1995, Grifiths and Tavaré 1996). In contrast, the first step in a phylogeographic analysis is to reconstruct a gene tree from

data, and inferences are based upon this inferred tree. This sensibility about the significance of inferred trees was received and adapted from the field of phylogenetics. At the intraspecific level, roughly speaking, the circumstances favorable to using inferred gene trees are those in which random genetic drift is relatively unimportant compared with nonequilibrium factors like the splitting of populations.

Only the simplest nonequilibrium model was considered here: a single population that split into two isolated demes at some time in the past. This kind of history has the qualities necessary for the single-tree approach to be most fruitful; that is, small Ω and large $max(p_i)$. However, most of the branches in the genealogies under this model, those for the intrademe patterns of common ancestry, will be discordant among loci. A more ideal scenario for the single-gene-tree approach is the stepping-stone model of range expansion considered by Slatkin (1993), which is a history of multiple isolation events. If a single sample was taken from each subpopulation, then we might expect the population tree to be reproduced at many loci. Of course, this too will depend upon the population splits being separated enough in time for the effect of drift to be negligible. Otherwise, even without migration, a gene tree may be different from the population tree (Neigel and Avise 1986, Pamilo and Nei 1988). This will be an issue as well for continuously distributed populations that have undergone range expansions; the movement of individuals will have to be restricted for historical structure to be evident in gene tree topologies.

This treatment has assumed no recombination within loci and free recombination between loci. Intralocus recombination will decouple sites' histories. Multiple genealogies will be realized in the history of a single locus and these will be correlated along the sequence (Hudson 1983a, Kaplan and Hudson 1985). Restricted interlocus recombination will make genealogies across sampled loci correlated. Both of these processes should tend to increase $max(p_i)$. Intralocus recombination increases the number of chances a locus has to realize a given partition, and restricted recombination between loci will cause branches to be shared across loci. They should have opposite effects on Ω, though. Intralocus recombination will lower the variation in tree sizes because there will be more independence among sites. The increased correlation among loci caused by restricted interlocus recombination, conversely, will increase the variance of tree size. Intralocus recombination is quite problematic for inferred gene-tree approaches since the genealogy is no longer a bifurcating tree (Hein 1993). It also represents a significant computational hurdle to coalescent inference methods which make explicit use of linkage patterns (Grifiths and Marjoram 1996).

The entire field of population genetics will benefit from increased exchange between coalescents and phylogeography. There is growing overlap already. On the one hand, the importance of coalescent approaches is evident in Avise's (2000) book about phylogeography. On the other, one of the currently most used coalescent inference programs, GENETREE (Bahlo

and Grifiths 2000), produces an inferred genealogy. The future availability of multilocus genetic data will serve as a further bridge between these two approaches.

10.7 Acknowledgments

It has been my pleasure of the past few years to be a colleague to Dick Lewontin. I am thankful to him for the inspiration to do good work, and to Rama Singh for the invitation to contribute to this volume in Dick's honor. I also thank Monty Slatkin for comments on the manuscript. This work was supported by grant DEB-9815367 from the National Science Foundation.

REFERENCES

Aldous, D. J. (1985). Exchangeability and related topics. In A. Dold and B. Eckmann (eds) *École d'Été de Probabilités de Saint-Flour XII – 1983*, pp. 1–198. Vol. 1117 of *Lecture Notes in Mathematics*. Berlin: Springer-Verlag.

Avise, J. C. (1989). Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution* 43:1192–208.

Avise, J. C. (2000). *Phylogeography: the History and Formation of Species*. Cambridge, MA: Harvard University Press.

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel,J. E., Reeb, C. A., and Saunders, N. C. (1987). Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annu. Rev. Ecol. Syst.* 18:489–522.

Bahlo, M., and Grifiths, R. C. (2000). Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* 57:79–95.

Cannings, C. (1974). The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. *Adv. Appl. Prob.* 6:260–90.

Darwin, C. (1859). *On the Origin of Species*. London: Murray.

Dobzhansky, T. (1937). *Genetics and the Origin of Species*. New York: Columbia University Press.

Donnelly, P., and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* 29:401–21.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3:87–112.

Ewens, W. J. (1990). Population genetics theory – the past and the future. In S. Lessard (ed.) *Mathematical and Statistical Developments of Evolutionary Theory*, pp. 177–227. Amsterdam: Kluwer Academic Publishers.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368–76.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edin.* 52:399–433.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford: Clarendon.

Grifiths, R. C., and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *J. Comp. Biol.* 3:479–502.

Grifiths, R. C., and Tavaré, S. (1996). Monte Carlo inference methods in population genetics. *Math. Comput. Modelling* 23:141–58.

Haldane, J. B. S. (1932). *The Causes of Natural Selection.* London: Longmans Green.

Harris, H. (1966). Enzyme polymorphism in man. *Proc. R. Soc. Lond. Ser. B* 164:298–310.

Hein, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.* 36:396–405.

Hennig, W. (1965). Phylogenetic systematics. *Annu. Rev. Entomol.* 10:97–116.

Hennig, W. (1966). *Phylogenetic Systematics.* Urbana: University of Illinois Press.

Hey, J. (1991). A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theor. Popul. Biol.* 39:30–48.

Hey, J. (1994). Bridging phylogenetics and population genetics with gene tree models. In B. Schierwater, G. P. Wagner, and R. DeSalle (eds) *Molecular Ecology and Evolution: Approaches and Applications*, pp. 435–49. Basel, Switzerland: Birkhäuser Verlag.

Hudson, R. R. (1983a). Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* 23:183–201.

Hudson, R. R. (1983b). Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37:203–17.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In D. J. Futuyma and J. Antonovics (eds) *Oxford Surveys in Evolutionary Biology*, pp. 1–44. Vol. 7. Oxford: Oxford University Press.

Hudson, R. R., Boos, D. D., and Kaplan, N. L. (1992). A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.* 9:138–51.

Kaplan, N. L., and Hudson, R. R. (1985). The use of sample genealogies for studying a selectively neutral *m*-loci model with recombination. *Theor. Popul. Biol.* 28:382–96.

Kimura, M. (1955a). Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA* 41:144–50.

Kimura, M. (1955b). Stochastic processes and the distribution of gene frequencies under natural selection. *Cold Spring Harbor Symp. Quant. Biol.* 20:33–53.

Kingman, J. F. C. (1982a). The coalescent. *Stochastic Process. Appl.* 13:235–48.

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

Kingman, J. F. C. (1982c). On the genealogy of large populations. *J. Appl. Prob.* 19A:27–43.

Kuhner, M. K., Yamato, J., and Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* 140:1421–30.

Lewontin, R. C. (1974). *The Genetic Basis of Evolutionary Change.* New York: Columbia University Press.

Lewontin, R. C., and Hubby, J. L. (1966). A molecular aproach to the study of genic diversity in natural populations II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura. Genetics* 54:595–609.

Lewontin, R. C., and Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution* 14:450–72.

Li, W.-H. (1976). Distribution of nucleotide difference between two randomly chosen cistrons in a subdivided population: the finite island model. *Theor. Popul. Biol.* 10:303–8.

Li, W.-H. (1977). Distribution of nucleotide difference between two randomly chosen cistrons in a finite population. *Genetics* 85:331–7.

Linnaeus, K. (1735). *Systema Naturae.*

Malécot, G. (1948). *Les Mathématiques de l'Hérédité.* Paris: Masson. Extended translation: *The Mathematics of Heredity.* San Francisco: W. H. Freeman (1969).

Mayr, E. (1942). *Systematics and the Origin of Species.* New York: Columbia University Press.

Moran, P. A. P. (1958). Random processes in genetics. *Proc. Camb. Phil. Soc.* 54:60–71.

Neigel, J. E., and Avise, J. C. (1986). Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation, In E. Nevo and S. Karlin (eds) *Evolutionary Processes and Theory,* pp. 515–34. New York: Academic Press.

Neigel, J. E., and Avise, J. C. (1993). Application of a random walk model to geographic distribution of animal mitochodrial DNA variation. *Genetics* 135:1209–20.

Neigel, J. E., Ball, M., and Avise, J. C. (1991). Estimation of single generation migration distances from geographic variation in animal mitochodrial DNA. *Evolution* 45:423– 32.

Nordborg, M. (2001). Coalescent theory. In D. J. Balding, M. J. Bishop, and C. Cannings (eds) *Handbook of Statistical Genetics.* Chichester: John Wiley.

Pamilo, P., and Nei, M. (1988). The relationships between gene trees and species trees. *Mol. Biol. Evol.* 5:568–83.

Provine, W. B. (1971). *The Origins of Theoretical Population Genetics.* Chicago: University of Chicago Press.

Robinson, D. F., and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Math. Biosci.* 53:131–47.

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

Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* 47:264–79.

Slatkin, M., and Hudson, R. R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129:555–62.

Smouse, P. E. (1974). Likelihood analysis of recombination disequilibrium in multiple-locus gametic frequencies. *Genetics* 76:557–65.

Strobeck, C. (1987). Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* 117:149–53.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–60.

Takahata, N., and Nei, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* 110:325–44.

Takahata, N., and Slatkin, M. (1990). Genealogy of neutral genes in two partially isolated populations. *Theor. Popul. Biol.* 38:331–50.

Templeton, A. R. (1998). Nested clade analysis of phylogeographic data: testing hypotheses about gene flow and population history. *Mol. Ecol.* 7:381–97.

Templeton, A. R., Routman, E., and Phillips, C. (1995). Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum. Genetics* 140:767–82.

Wakeley, J. (1996a). The variance of pairwise nucleotide differences in two populations with migration. *Theor. Popul. Biol.* 49:39–57.

Wakeley, J. (1996b). Distinguishing migration from isolation using the variance of pairwise differences. *Theor. Popul. Biol.* 49:369–86.

Wakeley, J., and Hey, J. (1997). Estimating ancestral population parameters. *Genetics* 145:847–55.

Wallace, A. R. (1858). On the tendency of varieties to depart indefinitely from the original type. *Proc. Linn. Soc. Lond.* 3:53–62.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7:256–76.

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

Wright, S. (1943). Isolation by distance. *Genetics* 28:114–38.