Annu. Rev. Ecol. Evol. Syst. 2003. 34:99–125
doi: 10.1146/annurev.ecolsys.34.011802.132359

First published online as a Review in Advance on September 29, 2003

THE EFFECTS OF GENETIC AND GEOGRAPHIC STRUCTURE ON NEUTRAL VARIATION

Brian Charlesworth, Deborah Charlesworth, and Nicholas H. Barton

Institute for Cell, Animal, and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom; email: Brian.Charlesworth@ed.ac.uk, Deborah.Charlesworth@ed.ac.uk, N.H.Barton@ed.ac.uk

Key Words genetic drift, effective population size, migration, recombination, coalescent process

■ **Abstract** Variation within a species may be structured both geographically and by genetic background. We review the effects of such structuring on neutral variants, using a framework based on the coalescent process. Short-term effects of sex differences and age structure can be averaged out using fast timescale approximations, allowing a simple general treatment of effective population size and migration. We consider the effects of geographic structure on variation within and between local populations, first in general terms, and then for specific migration models. We discuss the close parallels between geographic structure and stable types of genetic structure caused by selection, including balancing selection and background selection. The effects of departures from stability, such as selective sweeps and population bottlenecks, are also described. Methods for distinguishing population history from the effects of ongoing gene flow are discussed. We relate the theoretical results to observed patterns of variation in natural populations.

INTRODUCTION

The neutral theory of molecular evolution has transformed evolutionary genetics by providing a null model against which alternative hypotheses can be tested (Kimura 1983). There is a rapid accumulation of data on natural variation at the DNA level, from microsatellites to single nucleotide polymorphisms, much of which is presumed to be nearly neutral. The action of selection at a particular site in the genome can cause deviations from the patterns predicted by neutral theory at nearby sites, which are not themselves the direct targets of selection (Kreitman 2000). This is triggering efforts to conduct genome-wide searches for evidence of selection (Akey et al. 2002, Harr et al. 2002). However, many other evolutionary processes can cause departures from the predictions of classical neutral theory. Although this complicates tests for the action of selection, it also provides opportunities to make inferences about such processes.

It is hard to disentangle all the processes that have shaped the properties of samples from natural populations. In particular, neutral variants are affected by both geographic structure and genetic structure (i.e., the division of the gene pool into different genetic backgrounds, defined by loci that influence fitness). Here, we emphasize the close analogy between the different kinds of structures and their similar effects on neutral variation. Rather than reviewing the numerous specific theoretical and empirical results, we aim to provide a general framework for understanding the factors that shape neutral variation. This is essential for correctly interpreting data. There are dangers in merely constructing plausible scenarios that are consistent with what we see, without considering alternative possibilities. Similarly, elaborate statistical methods can be misleading if the basic assumptions (such as a single ancestral population or a lack of selection at linked sites) are incorrect.

MODELS OF GENETIC DRIFT

The process of random genetic drift acting on neutral variants can be viewed in several ways, each appropriate in different contexts. One useful descriptor is the increase in relatedness between alleles in a population, i.e., inbreeding (Wright 1931). The degree of inbreeding can be quantified by the probability of identity by descent, the chance that two distinct alleles trace back to the same ancestral gene (Cotterman 1940; Malécot 1941, 1969). Common ancestry is the inevitable outcome of genetic drift in a finite population because drift consists of fluctuations in the frequencies of alternative alleles, eventually leading to fixation (Fisher 1922, Wright 1931). Neutral variability within a population can be quantified by the probability that a pair of homologous genes or nucleotide sites have the same allelic state (Kimura 1983).

Coalescence Theory

In recent years, genetic drift has increasingly been studied in terms of the coalescence of gene lineages (Donnelly & Tavaré 1995, Hudson 1990, Slatkin & Veuille 2002). If a sample of n homologous genes is taken from a population, we can trace allelic lineages back in time. (By gene we simply mean a defined region of DNA within which recombination is assumed to be negligible.) Consider a standard Wright-Fisher random mating, discrete-generation population of N breeding individuals; the next generation is formed by random draws with replacement from the pool of $2N$ genes of the parental generation (Fisher 1922, Wright 1931). The chance that any pair of genes come from a common ancestral gene in the previous generation (i.e., they coalesce) is $1/(2N)$. If n is small compared with N, at most one coalescent event will occur in any generation. The probability that t generations elapse before the first coalescence is approximated by an exponential distribution with mean $2N/(n[n-1]/2) = 4N/(n[n-1])$. Once this event has occurred, the time to the next event follows an independent exponential distribution with mean

$4N/([n-1][n-2])$, and so on until all the genes have coalesced into a single common ancestor (Hudson 1990).

Following the ancestry of the genes in samples allows the development of powerful statistical methods for making inferences about evolutionary processes (Donnelly & Tavaré 1995, Hudson 1990, Slatkin & Veuille 2002). These inferences are based on observed variability (i.e., mutations that have occurred since the common ancestor of the sample). Given a distribution of genealogies, it is easy to superimpose any chosen model of mutation and compare the results with data. For example, under the infinite sites model of mutation, where at most a single mutation segregates in the sample at a given nucleotide site (Kimura 1983), the number of differences between a pair of sequences follows a Poisson distribution whose mean is proportional to the time since their last common ancestor (Hudson 1990).

Much classical work on identity probabilities carries over to the coalescent process. For example, under the infinite alleles mutation model (Kimura 1983), each allele is assumed to have a probability μ of mutating to a different allele. We can then determine the probability of identity in allelic state, f, between two genes, from the probability distribution, $\psi(t)$, of their time to coalescence, because they are identical only if neither gene has mutated since they shared a common ancestor (Hudson 1990, Wilkinson-Herbots 1998)

$$f = \sum_{t=1}^{\infty} (1-\mu)^{2t} \psi(t). \tag{1}$$

For the Wright-Fisher model, $\psi(t) \approx 2N \exp -t/(2N)$.

Mathematically, this expression is the moment generating function of the distribution of coalescence times for a pair of genes, from which the mean and higher moments can be obtained by standard methods (Hudson 1990). Thus, classical results concerning f are directly related to this distribution.

The Effects of Population Structure on Neutral Variation

Drift is caused by random variation in the reproductive contributions of individual genes, which depends on the demographic, spatial, and genetic structuring of populations. Alleles at a locus may come from different local populations, or they may differ in their background genotypes (for example, the numbers of deleterious mutations that are present, or the genotypes at a locus subject to balancing selection). They may also be in different sexes, age classes, or stage classes. Multigene families also represent a form of genetic structure (Ohta 2002). Stable genetic or geographic structure sometimes increases variability because lineages may remain associated with different genetic backgrounds or different places for a long time (Wright 1943). Conversely, fluctuations in the structure of a population can greatly reduce variability, as in the case of a population bottleneck (Maruyama & Fuerst 1985) or a selective sweep associated with the spread of a favorable mutation (Maynard Smith & Haigh 1974). The effects of structure depend on how long

genes remain associated with particular genetic backgrounds or demes, relative to the timescales of coalescence and fluctuations in structure.

Analysis of movement of lineages between locations or genetic backgrounds requires a theoretical framework known as the structured coalescent (Hey 1991). In the past few years, this has been applied to many problems in evolutionary genetics and widely used to make inferences about evolutionary processes (Slatkin & Veuille 2002). We first describe the effects of different types of structure, treating demographic, geographic, and genetic structure in parallel, and then considering their joint effects.

GENERAL CONSIDERATIONS AND SOME SIMPLIFICATIONS

Much of the power of the coalescent process to generate useful results about samples of alleles from a population comes from the simple properties of the probability distribution of coalescence times for a sample of alleles under the Wright-Fisher model (Donnelly & Tavaré 1995, Hudson 1990). With population structure, the probabilities of coalescence of alleles may depend on their sources, so that there is no longer homogeneity of coalescence time distributions for all alleles. There are two ways to deal with structured populations: one is to find simplifications that avoid some of the complexities, and the other is to look for general but useful properties. These two approaches can, of course, be combined, as we show below.

Short Timescales and the Effective Size of a Population

An important example of the first approach is when alleles at a neutral locus move rapidly between different local populations, age classes, or genotypes at other loci. When viewed over a longer timescale, these rapid movements simply change the rate of genetic drift without causing any significant differentiation between the different compartments that define the population structure (Nagylaki 1980, Nordborg 1997, Nordborg & Krone 2002). This simplifies the study of drift in populations structured by sex, stage, or age, compared with previous methods (Chesser et al. 1993, Wang & Caballero 1999), largely eliminating the need to worry about such structure in the context of a geographically or genetically structured population (Laporte & Charlesworth 2002).

To see this, we note that, to a very good approximation, the rate of genetic drift in more complex situations than the Wright-Fisher model is inversely proportional to the effective population size, N_e (Caballero 1994, Crow & Kimura 1970, Laporte & Charlesworth 2002, Wright 1931). The effects of geographic structuring on genetic differentiation between *demes* (defined as geographical units within which mating is effectively random), depend on both the N_e values of demes and the pattern of gene flow among them (Maruyama 1977, Wright 1951).

It has long been known that subdivision causes little noticeable neutral genetic differentiation between populations, unless the products of migration rates and the effective population sizes of demes are around 1 or less (Maruyama 1977, Nagylaki 1980, Wright 1951). For geographic structure to cause significant genetic differentiation, an allele sampled from deme i with effective population size N_{ei} must have been in this deme for a time on the order of N_{ei} generations. Short-term processes, such as movements of alleles between sexes or age classes, will generally reach their equilibrium states much more quickly (Figure 1). We can then treat the coalescence of alleles within demes (or stably maintained genotypes), and the migration of alleles between demes, as taking place on a long timescale, separate from the short-timescale processes involving age and sex classes (and some others to be described later) (Figure 1). If the discrepancy between these two timescales is large, an allele sampled from deme or genotype i can simply be treated as having a fixed probability α_{ir} of being in class r (defined by age and/or sex), whose value is determined by the matrix describing the flow of genes among classes (Laporte & Charlesworth 2002, Nordborg 1997, Nordborg & Krone 2002).

We can use this principle to determine the probability, P_i, that two alleles sampled from the ith deme coalesced in the previous time period. For a population with nonoverlapping generations, the relevant time period is the previous generation (and the generation time is 1). For age- and stage-structured populations that reproduce over discrete time intervals, the generation time t_i is larger than one time interval (Charlesworth 2001, Nordborg & Krone 2002). In many cases, the chance (P_{ir}) that two alleles which both trace their ancestry back to a particular class r coalesce in this class is independent of the classes from which they were originally sampled. P_i then takes the simple form

$$P_i = \sum_r \alpha_{ir}^2 P_{ir}. \tag{2}$$

The intuitive interpretation of this is that genes can only coalesce once they trace back to the same class (Figure 1); α_{ir}^2 is the equilibrium probability that two genes from deme i both come from class r.

We can then define N_{ei} as the reciprocal of twice the per-generation probability of coalescence, or half the expected number of generations to coalescence: $N_{ei} = 1/(2t_i P_i)$. These expressions can be used to obtain effective population sizes for different modes of inheritance, such as X- versus Y-linked genes, and autosomal versus cytoplasmic genes (Charlesworth 2001, Laporte & Charlesworth 2002, Nordborg & Krone 2002).

Long Timescales and Migration

We can now add the slow process of migration. Migration between demes requires us to specify the probabilities for all possible kinds of migration between classes. We can write m_{irjs} for the probability that a gene sampled from an individual of class r in deme i originated from an individual of class s in deme j in the previous

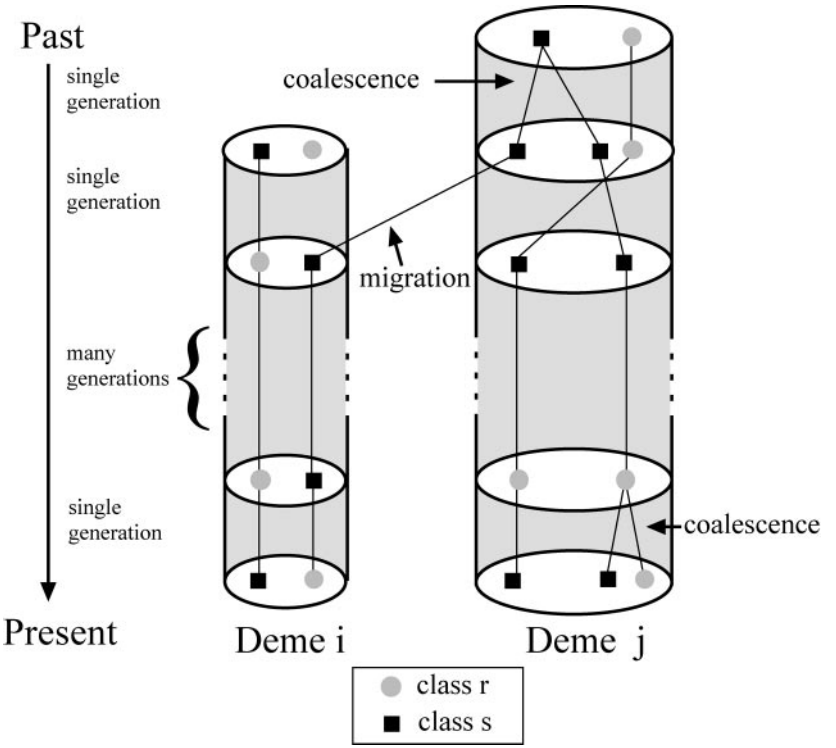

Figure 1 The lines show the ancestry of alleles present in two classes (*r* and *s*, indicated by the *square* or *circle* symbols) of individuals in two different demes (*i* and *j*, indicated in gray). The classes could be sexes, age classes, or genotypes. An allele in one class can be inherited from a parent in either class, or, more rarely, from a different deme (via migration). The example shows two coalescent events and one migration event; alleles in demes *i* and *j* eventually have a common ancestor (coalesce) in deme *j*. In this example, movement between the classes is much faster than movement between demes (i.e., the migration rate between demes is low). The chance that two alleles from the *i*th deme, which trace their ancestry back to a given class, coalesce in this class is therefore independent of the classes from which they were found in the sample (i.e., in the present generation of the figure).

time period. Applying the fast timescale approximation, we can average over all possible classes (weighting class *r* by α_{ir}) and derive a net migration probability m_{ij} for demes *i* and *j* (Laporte & Charlesworth 2002). Combining this with Equation 2, the simplification provided by the separation of timescales enables us to obtain a more general version of the standard equilibrium equation for identity probabilities under the infinite alleles model in a geographically structured population (Malécot 1969, Nagylaki 1982).

$$f_{ij} \approx (1-\mu)^2\left(\sum_{kl} m_{ik}m_{jl}f_{kl} + \delta_{ij}(1-f_{ii})P_i\right), \tag{3}$$

where f_{ij} is the probability of identity of alleles sampled from demes i and j, respectively; δ_{ij} is 1 when $i=j$, and zero otherwise.

From Equation 1, this gives the moment generating functions from which the probability distributions of times to coalescences of pairs of alleles can be determined (Wilkinson-Herbots 1998). In particular, the mean coalescence time, T_{ij}, for a pair of alleles sampled from demes i and j (Laporte & Charlesworth 2002, Nagylaki 1998a) is given by

$$T_{ij} \approx 1 + \sum_{kl} m_{ik}m_{jl}T_{kl} - \delta_{ij}P_iT_{ii} \tag{4}$$

However, one can only obtain simple analytic expressions for the T_{ij} with certain specific models of population structure (see below).

The Invariance Principle

There is, however, a powerful general result that can usefully be combined with results derived by this approach. We can define a weighted mean coalescent time for pairs of alleles sampled within demes as

$$T_0 = \sum_i v_i^2 / \sum_i v_i^2 P_i, \tag{5a}$$

where v_i is the ith component of the leading left eigenvector of the migration matrix $\{m_{ij}\}$ (Laporte & Charlesworth 2002, Nagylaki 1998a). (v_i measures the equilibrium probability that an allele sampled from the population will be found in deme i at some time in the past.) Using this definition, Equation 5a gives

$$T_0 = 1/\sum_i v_i^2 P_i. \tag{5b}$$

The weighted mean within-deme expected coalescence time is therefore independent of the migration structure of the population (Laporte & Charlesworth 2002, Nagylaki 1998a). This is a generalization of Maruyama's geographical invariance principle (Maruyama 1971, 1977; Nagylaki 1982).

As discussed above, the expected values of observable properties of genetic variability based on differences between pairs of alleles, such as nucleotide site diversity under the infinite sites model, depend on mean coalescence times (Hudson 1990). It is therefore attractive to use the invariant just derived to characterize the properties of a species with spatially structured populations because it suggests that the appropriately weighted mean diversity among pairs of alleles sampled from the same deme should depend only on the mutation process and T_0.

However, there is no general way to measure v_i or P_i. In the absence of gross asymmetries in the migration process, migration should spread genes fairly evenly among demes, so that v_i should be approximately equal to k_i, the proportion of

the total breeding population represented by deme i. When the flow of migrant individuals into each deme exactly balances the outward flow (conservative migration), v_i equals k_i (Maruyama 1971, 1977; Nagylaki 1982). This applies to the most frequently analyzed models of population structure, such as the island and stepping-stone models (Malécot 1969, Weiss & Kimura 1965, Wright 1943). It also applies to "genetic migration" processes such as crossing over and unbiased gene conversion, where the flow is between different genotypes (Nordborg 1997), but it is violated in many biologically significant situations, e.g., with predominantly unidirectional migration from a source to a sink population (Nordborg 1997), or in a metapopulation with frequent local extinction and recolonization of demes (Pannell & Charlesworth 2000, Whitlock & Barton 1997).

If the demographic factors that influence effective population size do not vary among populations, the P_i values for each deme must be approximately inversely proportional to their numbers of breeding individuals. We can then write $P_i = 1/(2c k_i N_T)$, where N_T is the total number of breeding individuals in the species and c is the ratio of effective population size to census population size (for ways to calculate c, see Caballero 1994, Crow & Kimura 1970, Laporte & Charlesworth 2002). If these simplifying proportionalities hold, the mean within-deme coalescence time is the same as that for a panmictic population, i.e., $2cN_T$ (Nagylaki 1998a).

Geographic Differentiation: F_{ST}

The most widely used measures of between-population differentiation relative to within-population variation are Wright's F_{ST} (Wright 1943, 1951) and related measures such as G_{ST} (Nei 1973) and θ (Weir & Cockerham 1984). Although F_{ST} has classically been defined in terms of variance in allele frequencies (Nagylaki 1998b; Wright 1943, 1951) or identity probabilities of alleles (Nagylaki 1998b), it can also conveniently be defined in terms of mean coalescence times (Hudson et al. 1992, Slatkin 1991, Takahata 1991). Slightly different measures have been suggested by different authors; these become equivalent for a large number of demes. We shall use the definition

$$F_{ST} = (T_T - T_S)/T_T, \tag{6}$$

where T_T is the expected coalescence time for a pair of alleles sampled randomly from the population as a whole and T_S is the expected time for pairs of alleles sampled within demes (Hudson et al. 1992).

For a given set of demes, T_T involves all possible pairs of demes from which pairs of alleles can be drawn, with a specified weight for each deme pair, and similarly for T_0. If the weights used are the v values in Equation 5a, $T_S = T_0$. The theoretical value of F_{ST} can be related to empirical estimates obtained from genetic data, with appropriate corrections for sampling biases (Hudson et al. 1992; Weir & Cockerham 1984). The relation is particularly close for data on nucleotide site diversity, given the proportionality between coalescence time and divergence in this case (see above). For microsatellite loci that obey the stepwise mutation

model, the analogue of F_{ST} based on the variances in number of repeats within and between populations (R_{ST}) can be used with Equation 6 (Goldstein & Schloetterer 1999).

Long Timescales: N_{eT}

Species are often spread over large geographic ranges, or consist of very many demes. Coalescence for a pair of alleles sampled randomly from the species as a whole is usually much slower than the timescale of migration between local subpopulations. The effective size of the whole population, N_{eT}, can be defined by setting the expected time to coalescence for a random pair of genes, T_T, equal to $2N_{eT}$. T_T must obviously exceed T_S because alleles sampled from two different populations can only coalesce once their ancestors find themselves in the same population. Unfortunately, no useful general formula for T_T exists, although some insights can be obtained from calculating the effects of the variance in the long-term contributions of demes to the species' gene pool (Whitlock & Barton 1997). In the following sections, we show that the long timescale approximation provides a powerful way of analyzing models of migration and population history. Over sufficiently long times, ancestral lineages become dispersed uniformly over the species' range, and so the coalescent process for a pair of alleles approximates that in a single panmictic population.

STABLE POPULATION STRUCTURE

The Island Model

The island model assumes d demes, each with the same effective size N_e and a probability m that a gene sampled from a given deme comes from a different, randomly chosen deme in the previous generation. A pair of alleles sampled from different demes waits an average of $(d-1)/(2m)$ generations to trace back to the same deme; from Equation 5, once they are in the same deme, they wait an average of $T_0 = 2dN_e$ generations to coalesce (Slatkin 1991). Their time to coalescence is therefore exponentially distributed, with mean $T_0 + (d-1)/(2m)$. The mean coalescence time of two genes sampled randomly from the species as a whole is

$$T_T = T_0 + \frac{(d-1)^2}{2dm}. \tag{7}$$

This provides a measure of the total amount of genetic diversity in the species as a whole; the difference $T_T - T_0$ measures the absolute amount of population differentiation. The corresponding expression for F_{ST} reduces to Wright's classical formula (Wright 1943, 1951) when there are many demes.

Equations 6 and 7 imply that N_{eT} is increased to $1/(1-F_{ST})$ times the panmictic value (Wright 1943). However, we stress again that N_{eT} is not always increased; reduced T_0, due to departures from the conservative migration assumption,

lowers N_{eT} compared with panmixia (Nordborg 1997, Pannell & Charlesworth 2000, Wakeley & Aliacar 2001, Whitlock & Barton 1997). In such situations, diversity is reduced because demes vary greatly in their contributions to the next generation, and alleles coalesce to common ancestors faster than in the above model (Whitlock & Barton 1997). Nonetheless, the results from the island model provide a useful framework for examining the properties of subdivided populations. For example, using N_e values given by Equation 2 for different modes of inheritance, and values of m under different assumptions about sex-specific migration rates, we can derive illuminating conclusions about the expected patterns of genetic differentiation for genes with different modes of inheritance (Laporte & Charlesworth 2002). Because the product N_em determines the extent of population differentiation, differences in mode of inheritance (e.g., lower N_e for Y-linked versus autosomal genes) and sex differences in migration rates between the sexes (e.g., the possible lower migration rates for human males than females) affect the relative levels of genetic differentiation (Laporte & Charlesworth 2002).

For the island model, Wakeley (1998, 1999) has used the separation of timescales between local migration and drift in the whole population to derive an elegant description of the whole genealogical process for a sample of n alleles. The coalescent process for a set of alleles sampled in an arbitrary way from a large set of demes can be divided into two phases. The first is the scattering phase, during which two or more alleles sampled from the same deme either coalesce within that deme or switch demes. This phase ends when the ancestors of every allele in the original sample are in different demes; from then on, the process enters the collecting phase, which follows a standard coalescent process with effective size N_{eT} and the ancestral allele number remaining when the collecting phase starts (Wakeley 1998, 1999). The scattering phase will usually be short compared to the collecting phase, so that the sampling properties of a collection of alleles can be well approximated by assuming that mutations are placed onto the gene tree during the collecting phase only (Wakeley 1999). This result provides a powerful way of examining models of migration and population history. Below, we show that a large two-dimensional population shows similar qualitative behavior.

Spatial Structure

With arbitrary deme sizes and migration rates, there is no general expression for the distribution of coalescence times. However, numerical values can readily be found by solving recursions such as Equation 3, and there are analytical results for identity probabilities (Malécot 1969, Maruyama 1977, Weiss & Kimura 1965, Wilkinson-Herbots 1998). We next use these to give expressions for the overall effective population size and distribution of coalescence times in spatially structured populations. It is mathematically convenient to work in continuous space, rather than assuming discrete demes as in Equations 3 and 4. There are, however, two difficulties in analyzing spatial continua. First, the interactions between neighboring individuals that are required to regulate population density make truly

continuous models intractable (Felsenstein 1975). Second, there is strong genetic differentiation over small spatial scales in a two-dimensional population that must be taken into account. Nevertheless, it is possible to smooth out local fluctuations by defining appropriate effective density and dispersal rates; over all but very local scales, continuous models show the same behavior as their demic counterparts (Barton et al. 2002).

One-Dimensional Habitats

Consider a discrete-generation population of N_T adults with equal probabilities of reproductive success, evenly distributed around a circular habitat of length L; the variance of distance moved in each generation is σ^2. An ancestral lineage makes a random walk through this habitat, with an approximately Gaussian distribution of distance moved and with a variance increasing through time as $\sigma^2 t$. Each gene can be considered to draw its ancestor t generations back from a pool of approximately $2\rho\sigma\sqrt{t}$ individuals, where $\rho = N_T/L$ is the population density (Barton & Wilson 1995, Wright 1943). The probability of coalescence time t for two nearby genes declines with $1/\sqrt{t}$ (Barton & Wilson 1995, Wright 1943).

The mean coalescence time between genes separated by a distance x is

$$T_x = 2N_T + x(L - x)/(2\sigma^2) \tag{8a}$$

(Nagylaki 1978, Wilkinson-Herbots 1998). Two genes sampled from the same point ($x = 0$) have mean coalescence time $T_0 = 2N_T = 2\rho L$, consistent with the invariance principle (Equations 5). However, we emphasize that these results for mean coalescence times may be misleading because the distribution is highly skewed (Barton & Wilson 1995): There is a 1% chance that coalescence occurs more recently than $0.00008\,(4\rho\sigma)^2$ and a 1% chance that it occurs after $3000\,(4\rho\sigma)^2$ generations.

The mean coalescence time for a randomly chosen pair of genes ($2N_{eT}$) is

$$T_T = 2N_{eT} = 2\rho L + L^2/(12\sigma^2). \tag{8b}$$

This expression can be combined with that for T_0 to obtain an analogue of F_{ST} for alleles from populations separated by a given distance (Slatkin 1993, Wilkinson-Herbots 1998). For a pair of alleles sampled at random over all the demes, this approaches 1 for a large number of demes, consistent with the classical result that genetic differentiation in linear habitats can be very high (Maruyama 1977; Weiss & Kimura 1965; Wilkins & Wakeley 2002; Wright 1943, 1946). This implies that species living in long one-dimensional habitats may have much higher total nucleotide sequence diversities than species with comparable population sizes that occupy two-dimensional habitats.

Two-Dimensional Habitats

Most organisms are distributed across a two-dimensional habitat. By the same argument as before, each gene can be considered to draw its ancestor one generation

back from a Gaussian distribution covering an area of $4\pi\sigma^2$ and hence from a pool of $4\pi\rho\sigma^2$ individuals. Wright called this the neighborhood size, which we write as N_b (Wright 1946). It can be thought of as the number of individuals within one generation's dispersal range. The probability of coalescence time t for two nearby genes is now expected to decline approximately as $1/(2N_bt)$. There is no explicit formula for the probability distribution, but it can be calculated numerically (Barton & Wilson 1995). There is an appreciable probability, inversely proportional to N_b, that two nearby genes share very recent ancestry; this falls away very rapidly with distance. In contrast, in a one-dimensional population, relatedness decreases smoothly with distance, and genetic differentiation between locations separated by short distances ($<\sigma$) is negligible (Barton et al. 2002).

In a finite range of area L^2, with population density ρ, the mean coalescence time for a pair of randomly sampled genes is

$$T_T = 2\rho L^2 + L^2 \ln(KL/\sigma)/(2\pi\sigma^2), \tag{9}$$

where K is a constant of order 1, which depends on local population structure (Barton et al. 2002). The two terms on the right-hand side of Equation 9 now both increase with L^2 for a given population density. This implies that, even over very large distances initially separating two alleles, there will be significant coalescence over both short and long timescales. The ratio of the two terms is proportional to N_b. For small N_b, the second term dominates, implying that the overall genetic differentiation between populations does not differ greatly from that for the island model, given similar population sizes and migration rates (compare Equations 7 and 9) (Malécot 1969; Maruyama 1977; Wright 1943, 1946).

Stable Genetic Structure: Deleterious Mutations

We now turn from geographic to genetic structure, but for the moment keep the assumption that population structure remains stable. An important type of genetic structuring of a population is caused by the presence of deleterious alleles maintained in the population by recurrent mutation at many loci (Charlesworth et al. 1993). The effects of these deleterious alleles on neutral variability can be investigated using the framework developed above (Nordborg 1997).

If selection against deleterious mutations is sufficiently strong, and if there are no fitness interactions between loci, allele frequencies can be treated by standard deterministic theory: At locus i, the frequency q_i of deleterious alleles is u_i/t_i, where t_i is the selective disadvantage of heterozygotes, and u_i is the mutation rate from the wild-type allele A_i to the mutant allele a_i. We assume that recombination between a given neutral site and locus i takes place at rate r_i. Gametes carrying A_i and a_i can be regarded as two classes; the rate of gene flow between them depends on the mutation and recombination rates. The effect of a single selected locus on the mean time to coalescence of a pair of alleles at the neutral locus can easily be obtained using Equation 2 (Nordborg 1997). This can be extended to m loci, assuming that they are independent of each other (i.e., in linkage equilibrium) (Hudson &

Kaplan 1995, Nordborg et al. 1996, Nordborg 1997, Santiago & Caballero 1998). The mean coalescent time is

$$T \approx 2N_e \exp - \sum_{i=1}^{m} q_i / a_i^2, \tag{10}$$

where $a_i = 1 + r_i(1 - t_i)/t_i$.

By substituting plausible mutation, selection, and recombination parameters into this equation, it is possible to show that such background selection can account for much of the observed relationship between recombination rate and variability in *Drosophila melanogaster* (Hudson & Kaplan 1995, Charlesworth 1996). This does not, of course, prove that background selection is the sole cause of this relationship, and the issue of how to explain the relationship between recombination rates and levels of genetic diversity remains open (Andolfatto 2001, Lercher & Hurst 2002, Nachman 2001).

More detailed results can be obtained when there is no recombination, as is appropriate for Y chromosomes, asexual species, and highly self-fertilizing species. The flow of neutral alleles between the multilocus genotypic classes can then be described by a matrix similar to that used for flow among sexes and/or age classes (Charlesworth et al. 1995, Gordo et al. 2002). If selection is sufficiently strong relative to $1/N_e$, the system behaves according to the fast timescale approximation. The expected time to coalescence is then $2f_0N_e$, where f_0 is the frequency of the zero-mutation class (Charlesworth et al. 1993, Hudson & Kaplan 1994). If selection is weak, the fast timescale approximation fails, and the equivalent of Equation 4 must be used (Gordo et al. 2002). The effect of background selection on diversity is then much smaller. This approach can even be used to give accurate predictions of expected coalescence times when selection is so weak in relation to drift that *Muller's ratchet* is operating, i.e., the frequency distribution of mutational classes is unstable, with successive losses of the classes with the fewest deleterious mutations (Gordo et al. 2002). It can also be used to construct statistical tests for the departures from neutral expectations of variant frequencies caused by weak background selection, by simulating genealogies of sets of alleles (Charlesworth et al. 1995, Gordo et al. 2002).

Balancing Selection

The flow of genes by recombination between genotypes maintained by long-continued balancing selection can be treated in a similar way to migration between demes but using the long timescale approximation (Hudson 1990, Hudson & Kaplan 1988, Nordborg 1997, Takahata & Satta 1998). For example, consider two selected alleles, *A* and *a*, which have been maintained at equal frequencies in a discrete-generation panmictic population with effective size N_e, for much longer than the standard coalescence time. The probability that a neutral allele sampled from a gamete carrying *A* derives from an *a* background in the previous generation is then $r/2$, where r is the frequency of recombination between the neutral and

selected loci. This is equivalent to the migration rate between two populations, each of size N_e. Substitution into Equation 7 shows that the expected coalescence time between a random pair of alleles at the neutral locus is increased by $1/(2r)$, or by $1/(4N_e r)$ relative to the panmictic coalescence time $2N_e$ (Hudson 1990, Nordborg 1997). Because recombination acts like a conservative migration process, the coalescence time for a pair of neutral alleles sampled from the same allelic class is simply the panmictic value.

We can use Equation 6 to obtain a measure of the extent of genetic differentiation between the allelic classes at the selected locus. This is closely related to measures of linkage disequilibrium between variants at the neutral and selected loci, and between the neutral sites themselves (Charlesworth et al. 1997, Strobeck 1983). For neutral sites for which $4N_e r \ll 1$, the expected diversity and linkage disequilibrium will thus be much higher than for more distant sites. The stochasticity of the coalescent process, however, implies considerable random fluctuation around these expectations, so that peaks of diversity and linkage disequilibrium may be hard to detect in practice (Nordborg 2000).

Sometimes balancing selection acts at several genes [e.g., linked coadapted loci such as the components of the Brassica self-incompatibility system (Sato et al. 2002)], and/or on multiple sites in a gene [e.g., MHC loci (Takahata & Satta 1998)]. Balancing selection at multiple sites can potentially cause greatly increased diversity at tightly linked neutral loci (Barton & Navarro 2002) because many genetic backgrounds are maintained and can diverge from each other. The effect on neutral diversity depends on the number of distinct genotypic classes, which increases exponentially with the number of selected polymorphic sites. But with many selected sites, either linked polymorphisms come into strong linkage disequilibrium, with only two common haplotypes segregating (Kelly & Wade 2000), or random fluctuations reduce variation below the predictions with stable genotype frequencies (Navarro & Barton 2002).

There is evidence for increased neutral diversity due to balancing selection for plant self-incompatibility alleles, a classical example of frequency-dependent selection. *S* alleles are expected to remain polymorphic for long evolutionary times (Vekemans & Slatkin 1994), and very similar alleles have recently been found in related species of Brassica (Sato et al. 2002). *S* alleles indeed have extremely high diversity at synonymous sites (Charlesworth & Awadalla 1998, Richman et al. 1996) and in the introns, which are presumably not under balancing selection (Nishio et al. 1997). High variability is also found in MHC genes (Takahata & Satta 1998) and for some loci encoding allozymes (Filatov & Charlesworth 1999).

The theoretical results can be extended to populations at equilibrium under a system of regular inbreeding, such that the inbreeding coefficient at neutral loci is F. Alleles at the selected locus move between genotypic classes on a fast timescale and so can be treated as if they are in equilibrium. Using this result, the rate of flow between classes at the selected locus is reduced by a factor $(1 - F)$ (Dye & Williams 1997; Nordborg 1997, 2000), enhancing the increase in coalescence time accordingly. We also have to take into account the fact that, by a similar

argument, inbreeding decreases the effective population size by a factor of $1/(1 + F)$ (Laporte & Charlesworth 2002, Nordborg 1997, Nordborg & Donnelly 1997). For high inbreeding coefficients, the region over which balancing selection causes increased diversity and linkage disequilibrium, compared with the strictly neutral case, is much wider than with random mating (Charlesworth et al. 1997; Nordborg 1997, 2000).

Combinations of Forces

In the real world, many of the processes we have been discussing act jointly to influence levels of neutral genetic variability within and between demes. These joint effects can be studied with the fast and slow timescale approach. For example, under the island model, background selection reduces expected coalescent times within demes (T_0) without altering the increased expected coalescence time for alleles from different demes (Charlesworth et al. 1997, Nordborg 1997). This reduces the denominator of Equation 6, increasing F_{ST}. The same principle applies to almost any factor that reduces local N_e values while leaving migration rates unchanged, and can be expected to apply to a range of migration models (Charlesworth et al. 1997).

With balancing selection and population subdivision, the increase in mean coalescence time between selected alleles relative to the within-class coalescent time is identical in form to the panmictic case (see above), with the appropriate change in effective population size to N_{eT} (Charlesworth et al. 1997, Nordborg 1997). When background selection acts together with balancing selection, coalescent times are reduced within allelic classes at the selected locus (especially in inbreeding populations, with their low effective recombination rates), but between-class coalescent times are unaffected, so that it should be easier to detect the effects of balancing selection (Charlesworth et al. 1997, Nordborg 1997).

Balancing selection causes reduced differentiation between demes compared with neutral loci (Schierup et al. 2000) because an incoming migrant allele that is not already present in a deme has an increased chance of establishment, so that its effective migration rate is increased. The few available relevant empirical observations are consistent with this prediction. In the fungus *Schizopyllum commune*, the incompatibility loci are much less differentiated than a reference locus (James et al. 1999). Such differences in diversity patterns may allow balancing selection to be detected, particularly now that data from multiple loci can be compared (Akey et al. 2002, Baer 1999, Bamshad et al. 2002).

A different situation exists when different alleles at a locus are favored in different demes (local selection) (Charlesworth et al. 1997). This reduces the effective rate of migration at neutral sites linked to the selected locus because migrants into a population carrying a locally deleterious allele are selected against. For the simple case of a biallelic locus and two demes, with symmetrical migration and strong selection against locally maladapted alleles, the effective migration rate at a neutral site is approximately $mr^*/(s + r^*)$, where m is the migration rate,

s is the selection coefficient against an allele in the "wrong" deme, and r^* is the effective recombination rate between the selected and neutral sites (Charlesworth et al. 1997). This can be substituted into Equations 6 and 7 to find the expected equilibrium level of differentiation among demes. A similar result applies when there is a selective disadvantage to heterozygotes at a selected locus, as can arise with gene flow between partially reproductively isolated populations (Barton & Bengtsson 1986).

The reduced effective migration rate due to these effects leads to increased F_{ST} or other measures of population differentiation. The increase depends on the relative values of the recombination rates and selection coefficients, whereas the effect of balancing selection depends on $N_e r$. Increased variability can therefore extend over a much wider section of genome than with balancing selection (Charlesworth et al. 1997). Examples of increased F_{ST} due to local selection are starting to be discovered (Wilding et al. 2001).

FLUCTUATING STRUCTURE

Bottlenecks and Hitchhiking

A drastic population bottleneck causes an episode of enhanced genetic drift. Seen forward in time, allele frequencies change randomly, and some alleles become fixed in the population. Looking backward in time, we see lineages suddenly coalesce at the bottleneck, such that only a few lineages in the ancestral population contribute to our sample of alleles (Maruyama & Fuerst 1985, Tajima 1989). As variability recovers, each mutation is represented in one lineage, and so there will be an excess of rare variants. However, if several allelic lineages survive the bottleneck, samples could contain several haplotypes, each with different sets of ancestral mutations (Maruyama & Fuerst 1985, Tajima 1989); this produces an excess of high-frequency polymorphisms. This mixture of different patterns, corresponding to mutations that arose before and after the bottleneck, makes it difficult to infer a bottleneck simply from the spectrum of allele frequencies, although one would usually expect some detectable departure from neutral expectation (Charlesworth et al. 1993, Tajima 1989). Population bottlenecks can also dramatically increase levels of linkage disequilibrium, offering a potentially powerful way of detecting their effects (Stumpf & Goldstein 2003).

The fixation of a new favorable mutation has a similar effect to a bottleneck (Barton 2000, Maynard Smith & Haigh 1974). Sites tightly linked to the successful mutation share their ancestry, all tracing back to one founding haplotype. Looking back in time, sites that are not tightly linked may recombine away before coalescence occurs, and so may show a more diverse ancestry. Thus, a hitchhiking event will remove all variation in a narrow region (a selective sweep), whereas the evolution of nearby sites is analogous to a population bottleneck combined with immigration. The region of reduced variation is inversely proportional to the time taken for the mutation to fix in the population, i.e., of order $s/\log(2N)$, where s

is the selection coefficient (Barton 2000, Kim & Stephan 2002). The effects of multiple selective sweeps distributed randomly over the genome have also been modeled (Kaplan et al. 1989, Stephan 1995). These results can be used to ask how frequently selective sweeps must occur in order to account for the observed relations between variability and recombination rate (Nachman 2001, Stephan 1995).

In addition to reduced variability, selective sweeps may be detected from the associated distortions of allele frequency distributions and effects on linkage disequilibrium. Several statistical tests for these have been proposed (Braverman et al. 1995, Fay & Wu 2000, Kim & Stephan 2002, Kreitman 2000, Simonsen et al. 1995). Although some examples of significant departure from neutrality have been detected in regions of low recombination in *Drosophila* (Langley et al. 2000), reduced variability is often observed without such departures (Jensen et al. 2002, Langley et al. 2000).

The contribution of selective sweeps to the association between low recombination and reduced variability thus remains unresolved. However, unexpectedly low variability in regions of normal recombination may indicate the signature of recent adaptive evolution. For example, some microsatellite loci in *D. melanogaster* show reduced variation in non-African populations relative to African populations, suggesting that selective sweeps are associated with adaptation to life outside Africa (Harr et al. 2002). This local reduction was confirmed by finding that sequence variation near these anomalous microsatellites is also reduced (Harr et al. 2002).

General Effects of Fluctuating Structure

We have discussed the most drastic changes in the structure of a population: a sudden reduction in population size or the increase of a single mutant allele. Less severe fluctuations also influence neutral diversity but are harder to describe simply because so many situations are possible. In the context of spatial structure, much attention has been given to metapopulations, in which local populations may go extinct and then be recolonized (Pannell & Charlesworth 2000, Wakeley & Aliacar 2001). More generally, random changes in deme size due to demographic stochasticity, chaotic population dynamics, or extrinsic environmental variability all tend to reduce genetic diversity (Whitlock & Barton 1997).

For similar reasons, balancing selection may sometimes reduce diversity. Imagine a polymorphism maintained by balancing selection for very long periods of time, but with allele frequencies fluctuating as environmental conditions change. Neutral sites very close to sites maintained by selection will have increased diversity (see above). However, neutral sites somewhat farther away will be associated with a given selected allele for less time, and so will be less influenced by the net change in background frequency; the effect is equivalent to a partial selective sweep (Barton 2000). Even if diversity is increased in a narrow genomic region, it may be reduced over a wider region (Sved 1983).

In addition, a locus under balancing selection at a stable frequency may only recently have reached its equilibrium owing to the spread of a new variant to an

intermediate frequency. The pattern of variability associated with the new allele will then resemble that for a selective sweep, with greatly reduced variability at closely linked sites and strong linkage disequilibrium, as is observed for the *Sod* gene in *D. melanogaster* (Hudson et al. 1994). Other examples of balancing selection, such as chromosomal inversions in *Drosophila* (Andolfatto et al. 2001) and human loci associated with malaria resistance (Currat et al. 2002, Sabeti et al. 2002), also suggest recent increases in frequency.

Even in constant environmental conditions, deme sizes fluctuate through the random reproductive success of their members, and frequencies of genetic backgrounds fluctuate through genetic drift. These fluctuations can allow weakly selected genes to affect neutral diversity: If $N_e s$ is of order 1, selection is strong enough to have appreciable effects but does not fully determine the genetic structure. The coalescent approach can be extended to model directional selection on a small number of linked selected sites (Fearnhead 2001, Neuhauser & Krone 1997). Simulations show that the joint effects on genealogies of drift and selection are surprisingly small in this case, unless $N_e s$ is very large (Przeworski et al. 1999, Williamson & Orive 2002). There can, however, be large net effects of many weakly selected sites, such as synonymous variants in coding sequences (Comeron & Kreitman 2002, Tachida 2000). Calculations based on the structured coalescent also show that, even with quite strong selection relative to drift (e.g., $N_e s$ of order 10), stochastic fluctuations at a locus subject to balancing selection can substantially reduce its effects on linked neutral diversity (Barton & Etheridge 2003).

Population Structure and Genetic Diversity in Inbreeding Species

Populations reproducing by a regular system of inbreeding, such as self-fertilization, provide a good example of the ways in which multiple forces act on genetic variability within and between populations. As already mentioned, inbreeding reduces local N_e values (Laporte & Charlesworth 2002, Nordborg & Donnelly 1997), the maximal effect being to halve N_e with complete inbreeding. A highly inbred hermaphroditic population should thus have equal N_e values for the nuclear and organelle genomes, assuming maternal transmission of organelles, because both values will be half the nuclear N_e for a random mating hermaphrodite population with a comparable population size. Comparing a dioecious outcrosser with a selfing hermaphrodite with similar population size, the organelle N_e is twice as high in the selfer. Relative neutral diversity values can thus be predicted for panmictic populations.

Another effect of inbreeding is a reduction of the effective frequency of recombination throughout the genome (Dye & Williams 1997, Nordborg 2000). This increases the importance of processes such as background selection and selective sweeps, which will further reduce local N_e (Charlesworth et al. 1993, Hedrick 1980); this even affects the organelle genomes (Charlesworth et al. 1993). Finally,

several differences from outcrossers lead to increased isolation between demes in inbreeders (Baker 1953), both because self-pollination prevents female gametes from outcrossing (Lloyd 1979) and because inbreeding species also generally evolve reduced pollen or sperm output (Lloyd 1979), which lowers gene flow. Many inbreeders are weedy colonizing species and may undergo frequent bottlenecks during founder events, further reducing within-deme diversity (Schoen & Brown 1991).

As discussed above, isolation may lead to high between-deme diversity, so that total genetic diversity in a selfing species may be high, possibly even higher than that for an otherwise comparable outcrosser, despite the factors reducing selfers' local N_e values. On the other hand, situations with population turnover, including extinction and recolonization, can greatly reduce species' genetic diversity (Pannell & Charlesworth 2000, Wakeley & Aliacar 2001, Whitlock & Barton 1997), and this may be more prevalent in inbreeding than outcrossing species (Ingvarsson 2002). Thus, no general prediction can be made about relative species-wide diversity in inbreeders and outbreeders.

Adaptation to specific local environmental conditions is also common in plant populations (Baker 1953, Linhart & Grant 1996) and may cause local selective sweeps, although we are not aware of any well-documented example. In addition to reduced within-population diversity, the effect of local adaptation in retarding genetic exchange can be very strong in inbreeders because of their low effective recombination rates. Inbreeding populations are therefore even more isolated (relative to comparable outbreeding ones) than their reduced pollen movement would predict. Large allele frequency differences, and high F_{ST} values, may thus be seen at marker loci (Charlesworth et al. 1997). This may account for the strong differentiation at multiple loci observed between some highly inbreeding populations such as *Hordeum spontaneum* (Volis et al. 2002).

In inbreeders, it may thus be particularly difficult to identify loci that are the targets of selection and distinguish them from genes whose diversity is affected by selection at other loci. If a large proportion of loci are differentiated between populations, this may imply either severe isolation long enough for sequences to have diverged or lesser isolation plus some local adaptation. Haplotype structure in species-wide samples of sequences is therefore not good evidence for balancing selection maintaining variability within demes at the locus, or even at a nearby gene.

Many of the factors just outlined predict higher F_{ST} values in inbreeders than outbreeders. Because many processes (see above) affect variability within demes and differentiation between them, it is impossible to predict the quantitative effect of inbreeding on F_{ST}. However, some qualitative predictions can be tested with data on natural populations. There is clear evidence for reduced allozyme diversity in inbreeding species, both species-wide, and (more markedly), within populations (Hamrick & Godt 1990). Because some allozymes may be subject to balancing selection, silent nucleotide site diversity should also be compared using DNA sequence data. Few comparisons have so far been made for orthologous genes from natural populations of related inbreeding and outcrossing species, but these

studies also find low diversity within inbreeding plant populations (Baudry et al. 2001, Charlesworth & Pannell 2001), and inbreeding *Caenorhabditis* species have low diversity species-wide, compared with the dioecious outcrossing congener *C. remanei* (Graustein et al. 2002). The high diversity often found between inbreeding populations suggests that extinction/recolonization cannot explain their low within-deme diversity because this should reduce diversity in the entire set of populations. The allozyme results for other inbreeding plants suggest the same conclusion, but it is difficult to exclude other possibilities (Charlesworth & Pannell 2001).

POPULATION HISTORY VERSUS POPULATION STRUCTURE

Distinguishing Complete Isolation from Partial Gene Flow

An important problem in reconstructing the history of populations within a species, and the history of closely related species, is to distinguish whether populations have been completely isolated from each other since their separation from a common ancestor or whether limited gene flow continues between them (Avise 1999). In either case, substantial genetic differentiation between the populations may coexist with shared polymorphisms. For example, DNA sequence polymorphism data indicate a surprising amount of shared polymorphism among some groups of *Drosophila* sibling species (Kliman et al. 2000, Machado et al. 2002).

Attempts to infer historical events are difficult because of the stochasticity of mutations (which provide the information used) and especially because of the huge number of possible outcomes of the genealogical process (the evolutionary variance) (Hudson 1990). This variance must be taken into account when using genealogies reconstructed from sequences obtained from the same species, which are the basis for inferences about the populations' histories (Slatkin & Veuille 2002). For two populations recently separated from a common ancestor, Wakeley & Hey (1997) provided statistics based on the numbers of fixed nucleotide differences, polymorphisms shared between the two populations, and polymorphisms that are unique to each of the populations. For a single locus, the expectations of these statistics can be related to the rate of gene flow, the time since divergence, and the relative population sizes of the populations and their common ancestor. Tests can be done using simulations of the distribution over loci of test statistics for a null model of no gene flow (Kliman et al. 2000, Wakeley 1996, Wang et al. 1997) or by calculating the likelihood of a set of nucleotide sequence data for a nonrecombining locus in two populations (Nielsen & Wakeley 2001).

Inferring Population History

These considerations imply that we need a large number of markers and a clear pattern across a majority of neutral marker loci to infer the history of a set of populations. With molecular markers, the first requirement can be met for many species

in the wild. The second is unlikely to be true within a species, where few variants are diagnostic of populations or races; it may be seen in hybrid zones, where there is a clear geographic pattern so that the hypothesis of isolation is supported by the markers having different alleles at a consistent boundary (Barton & Hewitt 1985). However, the locations of such boundaries are not strong evidence for the locations of ancestral populations; an apparently clear division of a nonrecombining genome into two geographically distinct clades can occur quite frequently by chance in a subdivided population (Irwin 2002). Furthermore, selection may affect some markers in regions of environmental change, and inconsistent marker behavior is indeed often observed in boundary regions (Martinsen et al. 2001, Shaw 2002).

Much of the neutral diversity in outcrossing species is expected to be found within any local population, except for completely isolated populations. In other words, most of the ancestry of a sample of genes is in the distant past and has spread throughout the species' entire ancestral range. Phylogenetic inferences about outcrossing populations will therefore be extremely difficult because shared polymorphisms will be common, and few markers will have patterns that reflect the populations' histories (Wall 2000). A recent simulation study illustrates the low ability to infer even simple historical patterns of population splitting and to test alternative hypotheses (Knowles & Maddison 2002).

In inferring relationships between closely related species, such as humans, chimpanzees and gorillas, multiple independently inherited markers are essential. Indeed, when many genes are studied, variation among gene trees can be used to estimate the effective population size of the common ancestor (Chen & Li 2001, Wakeley & Hey 1997). Nuclear genes have therefore been added to earlier work on mitochondrial sequences on human populations (Chen & Li 2001, Yang 2002). For studying populations within species, mitochondrial sequences (and sometimes chloroplast sequences for plants) have often been used for phylogeographic studies. Because these genomes recombine rarely, or possibly not at all (McVean 2001), they provide only a single outcome of the evolutionary history, with different loci differing only by the mutational variance, so that independent data can be obtained only from nuclear genes.

When population history is simple, or the histories of different species have enough in common, useful inferences may be possible. For example, the general picture of decreasing diversity with latitude suggests recolonization after the last Ice Age in northern Europe (Hewitt 2001), and a similar pattern suggests northward spread of the cactus Lophocereus in Baja, California (Nason et al. 2002). Genetic distances based on studies of multiple allozyme loci are useful in inferring the origins of populations, for example to show multiple recent evolution of inbreeding populations of the plant *Eichhornia paniculata* from outcrossing ones (Husband & Barrett 1993).

The ancestry of human populations is of wide interest and has been inferred in various ways (Harpending & Rogers 2000). One approach uses diversity differences to suggest that Africans form a large and diverse source population from

which less diverse populations were derived via founder events and bottlenecks. Although a gradient of human diversity with distance from Africa is well established for many loci (Harpending & Rogers 2000), it is difficult, even with the large datasets now available, to exclude the alternative that low diversity outside Africa reflects a past history of isolated small populations, during which ancestral variants were lost (Takahata & Satta 2002).

The *Annual Review of Ecology, Evolution, and Systematics* is online at http://ecolsys.annualreviews.org

LITERATURE CITED

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12:1805–14

Andolfatto P. 2001. Adaptive hitchhiking effects on genome variability. *Curr. Opin. Genet. Dev.* 11:635–41

Andolfatto P, Depaulis F, Navarro A. 2001. Inversion polymorphisms and nucleotide variability in *Drosophila*. *Genet. Res.* 2001:1–8

Avise JC. 1999. *Phylogeography: The History and Formation of Species.* Cambridge, MA: Harvard Univ. Press. 458 pp.

Baer CF. 1999. Among-locus variation in F_{st}: Fish, allozymes and the Lewontin-Krakauer test revisited. *Genetics* 152:653–59

Baker HG. 1953. Race formation and reproductive method in flowering plants. *SEB Symp.* 7:114–45

Bamshad MJ, Mummidi S, Gonzalez E, Ahuja SS, Dunn DM, et al. 2002. A strong signature of balancing selection in the 5′ cis- regulatory region of CCR5. *Proc. Natl. Acad. Sci. USA* 99:10539–44

Barton NH. 2000. Genetic hitchhiking. *Philos. Trans. R. Soc. London Ser. B* 355:1553–62

Barton NH, Bengtsson BO. 1986. The barrier to genetic exchange between hybridising populations. *Heredity* 56:357–76

Barton NH, Depaulis F, Etheridge AM. 2002. Neutral evolution in spatially continuous populations. *Theor. Popul. Biol.* 61:31–48

Barton NH, Etheridge AM. 2003. The effect of selection on genealogies. *Genetics.* In press

Barton NH, Hewitt GM. 1985. Analysis of hybrid zones. *Annu. Rev. Ecol. Syst.* 16:113–48

Barton NH, Navarro A. 2002. Extending the coalescent to multilocus systems: the case of balancing selection. *Genet. Res.* 79:129–39

Barton NH, Wilson I. 1995. Genealogies and geography. *Philos. Trans. R. Soc. London Ser. B* 349:49–59

Baudry E, Kerdelhué C, Innan H, Stephan W. 2001. Species and recombination effects on DNA variability in the tomato genus. *Genetics* 158:1725–35

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchiking effect on the site frequency spectrum of DNA polymorphism. *Genetics* 140:783–96

Caballero A. 1994. Developments in the prediction of effective population size. *Heredity* 73:657–79

Charlesworth B. 1996. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genet. Res.* 68:131–50

Charlesworth B. 2001. The effect of life-history and mode of inheritance on neutral genetic variability. *Genet. Res.* 77:153–66

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–303

Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in

subdivided populations. *Genet. Res.* 70:155–74

Charlesworth D, Awadalla P. 1998. The molecular population genetics of flowering plant self-incompatibility polymorphisms. *Heredity* 81:1–9

Charlesworth D, Charlesworth B, Morgan MT. 1995. The pattern of neutral molecular variation under the background selection model. *Genetics* 141:1619–32

Charlesworth D, Pannell JR. 2001. Mating systems and population genetic structure in the light of coalescent theory. In *Integrating Ecology and Evolution in a Spatial Context. British Ecological Society Special Symposium 2000*, ed. J Silvertown, J Antonovics, pp. 73–95. Oxford: Blackwell

Chen FC, Li WH. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* 68:444–46

Chesser RK, Rhodes OE, Sugg DW, Schnabel A. 1993. Effective sizes for subdivided populations. *Genetics* 135:1221–32

Comeron JM, Kreitman M. 2002. Population, evolutionary and genomic consequences of interference selection. *Genetics* 161:389–410

Cotterman CW. 1940. *A calculus for statistico-genetics*. PhD thesis. Ohio State Univ., Columbus. 115 pp.

Crow JF, Kimura M. 1970. *An Introduction to Population Genetics Theory*, New York: Harper & Row. 591 pp.

Currat M, Trabuchet G, Rees D, Perrin P, Harding RM, et al. 2002. Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the beta(S) Senegal mutation. *Am. J. Hum. Genet.* 70:207–23

Donnelly P, Tavaré S. 1995. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* 29:410–21

Dye C, Williams BG. 1997. Multigenic drug resistance among inbred malaria parasites. *Proc. R. Soc. London Ser. B* 264:61–67

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–13

Fearnhead P. 2001. Perfect simulation from population genetic models with selection. *Theor. Popul. Biol.* 59:263–79

Felsenstein J. 1975. A pain in the torus: some difficulties with the model of isolation by distance. *Am. Nat.* 109:359–68

Filatov DA, Charlesworth D. 1999. DNA polymorphism, haplotype structure and balancing selection in the Leavenworthia *PgiC* locus. *Genetics* 153:1423–34

Fisher RA. 1922. On the dominance ratio. *Proc. R. Soc. Edinburgh* 52:312–41

Goldstein DB, Schloetterer C, eds. 1999. *Microsatellites. Evolution and Applications.* Oxford: Oxford Univ. Press

Gordo I, Navarro A, Charlesworth B. 2002. Muller's ratchet and the pattern of variation at a neutral locus. *Genetics* 161:835–48

Graustein A, Gaspar JM, Walters JR, Palopoli MF. 2002. Levels of DNA polymorphism vary with mating system in the nematode genus Caenorhabditis. *Genetics* 161:99–107

Hamrick JL, Godt MJ. 1990. Allozyme diversity in plant species. In *Plant Population Genetics, Breeding, and Genetic Resources*, ed. AHD Brown, MT Clegg, AL Kahler, BS Weir, pp. 43–63. Sunderland, MA: Sinauer

Harpending H, Rogers AR. 2000. Genetic perspectives on human origins and differentiation. *Annu. Rev. Genomics Hum. Genet.* 1:361–85

Harr B, Kauer M, Schlötterer C. 2002. Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* 99:12949–54

Hedrick PW. 1980. Hitch-hiking: a comparison of linkage and partial selfing. *Genetics* 94:791–808

Hewitt GM. 2001. Speciation, hybrid zones and phylogeography—or seeing genes in space and time. *Mol. Ecol.* 10:537–49

Hey J. 1991. A multidimensional coalescent process applied to multiallelic selection models and migration models. *Theor. Popul. Biol.* 39:30–48

Hudson RR. 1990. Gene genealogies and the coalescent process. *Oxford Surv. Evol. Biol.* 7:1–45

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ. 1994. Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. *Genetics* 136:1329–40

Hudson RR, Boos DD, Kaplan NL. 1992. A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.* 9:138–51

Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and recombination. *Genetics* 120:831–40

Hudson RR, Kaplan NL. 1994. Gene trees with background selection. In *Non-neutral Evolution: Theories and Molecular Data*, ed. B Golding, pp. 140–53. London: Chapman & Hall

Hudson RR, Kaplan NL. 1995. Deleterious background selection with recombination. *Genetics* 141:1605–17

Husband BC, Barrett SCH. 1993. Multiple origins of self-fertilization in tristylous *Eichhornia paniculata* (Pontederiaceae): inferences from style morph and isozyme variation. *J. Evol. Biol.* 6:591–608

Ingvarsson PK. 2002. A metapopulation perspective on genetic diversity and differentiation in partially self-fertilizing plants. *Evolution* 56:2368–73

Irwin DE. 2002. Phylogeographic breaks without geographic barriers to gene flow. *Evolution* 56:2383–94

James TY, Porter D, Hamrick JL, Vilgalys R. 1999. Evidence for limited intercontinental gene flow in the cosmopolitan mushroom *Schizophyllum commune*. *Evolution* 53:1665–77

Jensen MA, Charlesworth B, Kreitman M. 2002. Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. melanogaster*. *Genetics* 160:493–507

Kaplan NL, Hudson RR, Langley CH. 1989. The "hitch-hiking" effect revisited. *Genetics* 123:887–99

Kelly JK, Wade MJ. 2000. Molecular evolution near a two-locus balanced polymorphism. *J. Theor. Biol.* 204:83–102

Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking on a recombining chromosome. *Genetics* 160:765–77

Kimura M. 1983. *The Neutral Theory of Molecular Evolution.* Cambridge, UK: Cambridge Univ. Press. 367 pp.

Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M, et al. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex of species. *Genetics* 156:1913–31

Knowles LL, Maddison WP. 2002. Statistical phylogeography. *Mol. Ecol.* 11:2623–35

Kreitman M. 2000. Methods to detect selection in populations with applications to humans. *Annu. Rev. Genomics Hum. Genet.* 1:539–59

Langley CH, Lazzaro BP, Phillips W, Heikkinen E, Braverman JM. 2000. Linkage disequilibria and the site frequency spectra in the su(s) and su(wa) regions of the *Drosophila melamogaster* X chromosome. *Genetics* 156:1837–52

Laporte V, Charlesworth B. 2002. Effective population size and population subdivision in demographically structured populations. *Genetics* 162:501–19

Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends. Genet.* 18:337–40

Linhart YB, Grant MC. 1996. Evolutionary significance of local genetic differentiation in plants. *Annu. Rev. Ecol. Syst.* 27:237–77

Lloyd DG. 1979. Some reproductive factors affecting the selection of self-fertilization in plants. *Am. Nat.* 113:67–79

Machado CA, Kliman RM, Markert JA, Hey J. 2002. Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Mol. Biol. Evol.* 19:472–88

Malécot G. 1941. Etudes mathématiques des populations 'mendéliennes.' *Ann. Univ. Lyon Sci. Sect. C* 206:153–55

Malécot G. 1969. *The Mathematics of Heredity.* San Francisco: WF Freeman. 88 pp.

Martinsen GD, Whitham TG, Turek RJ, Keim P. 2001. Hybrid populations selectively filter gene introgression between species. *Evolution* 55:1325–35

Maruyama T. 1971. An invariant property of a subdivided population. *Genet. Res.* 18:81–84

Maruyama T. 1977. *Lecture Notes in Biomathematics. 17. Stochastic Problems in Population Genetics.* Berlin: Springer-Verlag. 245 pp.

Maruyama T, Fuerst PA. 1985. Population bottlenecks and non-equilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. *Genetics* 111:675–89

Maynard Smith J, Haigh J. 1974. The hitchhiking effect of a favourable gene. *Genet. Res.* 23:23–35

McVean GAT. 2001. What do patterns of genetic variability reveal about mitochondrial recombination? *Heredity* 87:613–20

Nachman MW. 2001. Single nucleotide polymorphisms and recombination rate in humans. *Trends. Genet.* 17:481–84

Nagylaki T. 1978. A diffusion model for geographically structured populations. *J. Math. Biol.* 6:375–82

Nagylaki T. 1980. The strong-migration limit in geographically structured populations. *J. Math. Biol.* 9:101–14

Nagylaki T. 1982. Geographical invariance in population genetics. *J. Theor. Biol.* 99:159–72

Nagylaki T. 1998a. The expected number of heterozygous sites in a subdivided population. *Genetics* 149:1599–604

Nagylaki T. 1998b. Fixation indices in subdivided populations. *Genetics* 148:1325–32

Nason JD, Hamrick JL, Fleming TH. 2002. Historical vicariance and postglacial colonization effects on the evolution of genetic structure in Lophocereus, a sonoran desert columnar cactus. *Evolution* 56:2214–26

Navarro A, Barton NH. 2002. The effects of multilocus balancing selection on neutral variability. *Genetics* 161:849–63

Nei M. 1973. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA* 70:3321–23

Neuhauser C, Krone SM. 1997. The genealogy of samples in models with selection. *Genetics* 145:519–34

Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158:885–96

Nishio T, Kusaba M, Sakamoto K, Ockendon D. 1997. Polymorphism of the kinase domain of the *S*-locus receptor kinase gene (*SRK*) in *Brassica oleracea* L. *Theor. Appl. Genet.* 95:335–42

Nordborg M. 1997. Structured coalescent processes on different timescales. *Genetics* 146:1501–14

Nordborg M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* 154:923–29

Nordborg M, Charlesworth B, Charlesworth D. 1996. The effect of recombination on background selection. *Genet. Res.* 67:159–74

Nordborg M, Donnelly P. 1997. The coalescent process with selfing. *Genetics* 146:1185–95

Nordborg M, Krone SM. 2002. Separation of time-scales and convergence to the coalescent in structured populations. In *Modern Developments in Population Genetics. The Legacy of Gustave Malécot*, ed. M Slatkin, M Veuille, pp. 194–232. Oxford, UK: Oxford Univ. Press

Ohta T. 2002. Usefulness of the identity coefficients for inferring evolutionary forces. In *Modern Developments in Theoretical Population Genetics*, ed. M Slatkin, M Veuille, pp. 37–51. Oxford, UK: Oxford Univ. Press

Pannell JR, Charlesworth B. 2000. Effects of metapopulation processes on measures of genetic diversity. *Philos. Trans. R. Soc. London Ser. B* 355:1851–64

Przeworski M, Charlesworth B, Wall JD. 1999. Genealogies and weak purifying selection. *Mol. Biol. Evol.* 16:246–52

Richman AD, Uyenoyama MK, Kohn JR. 1996. S-allele diversity in a natural population of ground cherry *Physalis crassifolia* assessed by RT-PCR. *Heredity* 76:497–505

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–37

Santiago E, Caballero A. 1998. Effective size and polymorphism of linked neutral loci in populations under selection. *Genetics* 149:2105–17

Sato T, Nishio T, Kimura R, Kusaba M, Suzuki G, et al. 2002. Coevolution of the S-locus genes SRK, SLG and SP11/SCR in *Brassica oleracea and B. rapa*. *Genetics* 162:931–40

Schierup MH, Vekemans X, Charlesworth D. 2000. The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genet. Res.* 76:51–62

Schoen DJ, Brown AHD. 1991. Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. *Proc. Natl. Acad. Sci. USA* 88:4494–97

Shaw KL. 2002. Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. *Proc. Natl. Acad. Sci. USA* 99:16122–27

Simonsen KL, Churchill GA, Aquadro CF. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* 141:413–29

Slatkin M. 1991. Inbreeding coefficients and coalescence times. *Genet. Res.* 58:167–75

Slatkin M. 1993. Isolation by distance in equilibrium and nonequilibrium populations. *Evolution* 47:264–79

Slatkin M, Veuille M, eds. 2002. *Modern Developments in Theoretical Population Genetics.* Oxford, UK: Oxford Univ. Press. 264 pp.

Stephan W. 1995. An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. *Mol. Biol. Evol.* 12:959–62

Strobeck C. 1983. Expected linkage disequilibrium for a neutral locus linked to a chromosomal rearrangement. *Genetics* 103:545–55

Stumpf MPH, Goldstein DB. 2003. Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Curr. Biol.* 13:1–8

Sved JA. 1983. Does natural selection increase or decrease variability at linked loci? *Genetics* 105:239–40

Tachida H. 2000. DNA evolution under weak selection. *Gene* 261:3–9

Tajima F. 1989. The effect of change in population size on DNA polymorphism. *Genetics* 123:597–601

Takahata N. 1991. Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. *Genetics* 129:585–95

Takahata N, Satta Y. 1998. Footprints of intragenic recombination at *HLA* loci. *Immunogenetics* 47:430–41

Takahata N, Satta Y. 2002. Out of Africa with regional interbreeding? Modern human origins. *Bioessays* 24:871–75

Vekemans X, Slatkin M. 1994. Gene and allelic genealogies at a gametophytic self-incompatibility locus. *Genetics* 137:1157–65

Volis S, Mendlinger S, Turuspekov Y, Esnazarov U. 2002. Phenotypic and allozyme variation in mediterranean and desert populations of wild barley, *Hordeum spontaneum* Koch. *Evolution* 56:1403–15

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

Wakeley J. 1998. Segregating sites in Wright's island model. *Theor. Popul. Biol.* 53:166–74

Wakeley J. 1999. Nonequilibrium migration in human history. *Genetics* 153:1863–71

Wakeley J, Aliacar N. 2001. Gene genealogies in a metapopulation. *Genetics* 159:893–905

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

Wall JD. 2000. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* 154:1271–79

Wang JL, Caballero A. 1999. Developments in predicting the effective size of subdivided populations. *Heredity* 82:212–26

Wang RL, Wakeley J, Hey J. 1997. Gene

flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. *Genetics* 147:1091–106

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–70

Weiss GH, Kimura M. 1965. A mathematical analysis of the stepping stone model of genetic correlation. *J. Appl. Prob.* 2:129–49

Whitlock MC, Barton NH. 1997. The effective size of a subdivided population. *Genetics* 146:427–41

Wilding CS, Butlin RK, Grahame J. 2001. Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *J. Evol. Biol.* 14:611–19

Wilkins JF, Wakeley J. 2002. The coalescent in a continuous, finite, linear population. *Genetics* 161:873–88

Wilkinson-Herbots HM. 1998. Genealogy and subpopulation differentiation under various models of population structure. *J. Math. Biol.* 37:535–85

Williamson SM, Orive ME. 2002. The genealogy of a sequence subject to purifying selection at multiple sites. *Mol. Biol. Evol.* 19:1376–84

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

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

Wright S. 1946. Isolation by distance under diverse systems of mating. *Genetics* 31:39–59

Wright S. 1951. The genetical structure of populations. *Ann. Eugen.* 15:323–54

Yang Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics* 162:1811–23