Evolution, 54(4), 2000, pp. 1092–1101

THE EFFECTS OF SUBDIVISION ON THE GENETIC DIVERGENCE OF POPULATIONS AND SPECIES

JOHN WAKELEY
Department of Organismic and Evolutionary Biology, Harvard University, 288 Biological Laboratories, 16 Divinity Avenue, Cambridge, Massachusetts 02138;
E-mail: jwakeley@oeb.harvard.edu

Abstract.—An island model of migration is used to study the effects of subdivision within populations and species on sample genealogies and on between-population or between-species measures of genetic variation. The model assumes that the number of demes within each population or species is large. When populations (or species), connected either by gene flow or historical association, are themselves subdivided into demes, changes in the migration rate among demes alter both the structure of genealogies and the time scale of the coalescent process. The time scale of the coalescent is related to the effective size of the population, which depends on the migration rate among demes. When the migration rate among demes within populations is low, isolation (or speciation) events seem more recent and migration rates among populations seem higher because the effective size of each population is increased. This affects the probability of reciprocal monophyly of two samples, the chance that a gene tree of a sample matches the species tree, and relative likelihoods of different types of polymorphic sites. It can also have a profound effect on the estimation of divergence times.

Key words.—Coalescent, divergence, genealogy, migration, species.

Received August 6, 1999. Accepted February 24, 2000.

Population subdivision has been observed in many species. The resultant structuring of genetic variation is well documented and in certain cases understood from a theoretical standpoint. This paper follows some recent work (Wakeley 1998, 1999) on the coalescent in the island model of migration (Wright 1931, 1943). In these papers it was shown that the history of a sample of nonrecombining DNA sequences is relatively simple when the number of demes in the population is large. This allows population subdivision to be included in previously well-studied models of other phenomena. The purpose of the present work is to demonstrate how subdivision within populations and species shapes patterns of interpopulation and interspecies genetic variation. In other words, the usual assumption of migration and divergence models—that there is no subdivision within groups—will be relaxed. It is shown that the effect of intragroup subdivision depends on two things: the dependence of effective population size (and thus the coalescent time scale) on the migration rate and the distribution of the sample among demes. The results have implications for the estimation of divergence times, of the relative sizes of populations, and of phylogenetic relationships.

This work focuses on two models in particular: two-population equilibrium migration and two-population isolation without gene flow. Figure 1 depicts these and shows one possible genealogy of a sample of four homologous, nonrecombining DNA sequences under each model. Many results have been derived under both of these models, and this provides a firm basis for understanding the effects of (further) subdivision within each population. The results of interest here are the expectations of average pairwise differences within and between populations (Slatkin 1987; Strobeck 1987) and the probabilities of genealogical topologies under isolation and migration (Tajima 1983; Takahata and Slatkin 1990). Because few other results are available for two-population migration, the rest of the work will focus on isolation models and will consider two aspects of genealogical history: the numbers of segregating sites divided into shared, fixed, and exclusive polymorphisms (Wakeley and Hey 1997) and the probability of consistency between gene trees and species trees (Pamilo and Nei 1988; Takahata 1989). In the last case, multipopulation or multispecies models of isolation without gene flow will be considered.

Recent interest in gene genealogies has enhanced our ability to discuss evolutionary processes in direct relation to observable data. Simply put, demographic factors such as genetic drift, changes in population size, migration, vicariance, and selection influence the structure of genealogies, and this in turn determines what we see in genetic data. Although it has been known for some time that subdivision can increase the effective size (N_e) of a population (Wright 1943), we have so far lacked a theoretical framework for studying its effect for arbirary samples from a population. Nei and Takahata (1993) found a formula for N_e under the finite island model (Maruyama 1970; Latter 1973) that applies to sample of size two. Hoelzer (1997), in response to Moore (1995), used this result to argue that mitochondrial gene trees can be deeper than nuclear gene trees if the female migration rate is low. Later, Hoelzer et al. (1998) illustrated this phenomenon using simulations. However, their results are not easily compared to others because the model of genetic drift used was not a standard one (Hoelzer et al. 1999). The subdivided population model considered here allows such issues to be investigated easily and for some general conclusions to be drawn.

In all of what follows, a population is assumed to be composed of D demes connected by Wright's (1931) island-model migration. Each deme is of constant, diploid size N, but the results will also hold for haploid populations in which the deme size is $2N$. In both cases each deme contains $2N$ copies of any particular locus in the genome. Generations are nonoverlapping, and there is random mating within demes. The migration rate among demes is m, and this is the probability that a particular sequence came from one of the other $D - 1$

demes in the immediately previous generation. Thus, each deme receives $2Nm$ migrants every generation. The neutral mutation rate is u per sequence per generation, and it is assumed that every mutation occurs at a unique site (Kimura 1969; Watterson 1975). Such a population can be characterized by two parameters: $\theta = 4NDu$ and $M = 2NmD/(D - 1)$ (Wakeley 1998). A sample of n items (e.g., DNA sequences) drawn from d different demes within the population is designated $\mathbf{n} = (n_1, n_2, \ldots, n_d)$. The demes from which we have sampled are arbitrarily labeled 1 through d, and n_i is the sample size from the ith deme. Thus, $n = \sum_{i=1}^{d} n_i$. Superscripts are added to this sample notation below when more than one population is considered.

When the total number of demes in the population is large, the genealogical history of the sample can be broken into two parts: a recent ''scattering'' phase and a more ancient ''collecting'' phase (Wakeley 1998, 1999). During the scattering phase, coalescent events occur at rate $\binom{n_i}{2}/(2N)$ and migration events at rate $n_i m$ per generation within each deme, where $i = 1, \ldots, d$. Here a migration event means that some member of the sample came from one of the other $D - 1$ demes in the immediately previous generation. The chance that the deme it came from contains any lineages ancestral to the sample is always smaller than $(n - 1)/(D - 1)$, which is very close to zero when D is large relative to n. Thus, all migration events during the scattering phase of the history are to demes that do not contain any lineages ancestral to the sample. The scattering phase ends when all ancestral lineages are in separate demes. The number of these lineages, n', depends on the relative rates of migration and coalescence. The number of lineages has a maximum value of n when migration dominates so completely that the entire sample spreads out into different demes, and it achieves its minimum value of d when the common ancestor of each deme's sample is reached before any migration events occur.

A different process, called the collecting phase, takes over for the n' ancestral lineages. These will wander around the population until a migration event places one of the lineages into a deme that contains another lineage. Because the number of demes is assumed to be much larger than the sample size ($D \gg n$), again the great majority of migration events will be to demes unoccupied by any ancestral lineages. Thus, many migration events will occur before two lineages are found in the same deme and have the chance to coalesce. When two lineages are in the same deme, they will coalesce with probability $1/(2M + 1)$ or one of them will migrate, again almost certainly, to an unoccupied deme. Eventually, the required number $(n - 1)$ of collecting-phase coalescent events will occur and the common ancestor of the entire sample will be reached. Because migration events are required before coalescent events can happen, the time it takes to reach the common ancestor of the sample will be shorter if the migration rate is higher.

The duration of the scattering phase is of order N generations. It will be shorter if the migration rate is high, and will approach a within-deme coalescent as M goes to zero. The collecting phase is a coalescent process (Kingman 1982) when time is measured in units of

$$2N_e = 2ND\left(1 + \frac{1}{2M}\right) \tag{1}$$

generations (Wakeley 1999), so the duration of the collecting phase is at least of order ND generations. However, it will be longer when M is small. Thus, the ratio of the average length of the scattering phase to the average length of the collecting phase will always be less than $1/D$. For large values of D, so much of the history of the sample is spent in the collecting phase that the duration of the scattering phase is negligible.

This work uses the fact the collecting phase is a coalescent to add within-group subdivision to models of migration among populations and models of divergence of populations and species. First consider the multideme, but single-population sample $n_1 = n_2 = \ldots = n_d = 1$. This sample has no scattering phase or, more precisely, its scattering phase has only one possible outcome: $n' = n$. All the standard coalescent results for a single panmictic population (e.g., Hudson 1983; Tajima 1983; Tavare 1984) will apply to this sample when the effective population size is given by equation (1). For the general sample, $\mathbf{n} = (n_1, n_2, \ldots, n_d)$, the coalescent results will apply to the collecting-phase sample, whose size, n', is unknown. We must take the scattering phase into account, and any results will have to be averaged over all possible collecting-phase samples.

Let n_i' be the number of ancestral lineages of the sample from deme i at the end of the scattering phase. Again, each of these is in a separate deme and $1 \le n_i' \le n_i$. The distribution of n_i' is the same as the distribution of the number of alleles in the Ewens sampling formula (Ewens 1972; Karlin and McGregor 1972), but with infinite alleles mutation replaced by infinite demes migration:

$$P[n_i' \mid n_i] = \frac{S_{n_i}^{(n_i')}(2M)^{n_i'}}{(2M)_{(n_i)}} \tag{2}$$

(Wakeley 1998), where $x_{(r)} = x(x + 1) \ldots (x + r - 1)$, and $S_i^{(j)}$ is an unsigned Stirling number of the first kind (Abramowitz and Stegun 1964). We let $\mathbf{n}' = (n_1', n_2', \ldots, n_d')$ and note that events in different demes are independent. Then the joint distribution of all the Let n_i' is given by the product:

$$P[\mathbf{n}' \mid \mathbf{n}] = \prod_{i=1}^{d} P[n_i' \mid n_i]. \tag{3}$$

The collecting phase begins with $n' = \sum_{i=1}^{d} n_i'$ lineages, each in a separate deme, and with $d \le n' \le n$. Simulations indicate that this large-D approximation, including both the scattering and collecting phases, is valid even if the number of demes in the population is only three to four times larger than the sample size (Wakeley 1998).

Theory and Results

We begin with a pair of populations that conform either to the two-population equilibrium migration model or the two-population isolation model discussed above and shown in Figure 1. Each population is subdivided according the D-deme island model. We assume identical values of θ and M for both populations. This assumption could be relaxed, but it is a good starting point. The rate of migration between the

two populations is m_{12} in the migration model. That is, m_{12} is the probability that a member of a sample from a deme in one population came from the other population in the previous generation. When D is large, it is shown below that the appropriate between-population migration parameter is $M_{12} = 2NDm_{12}$. In the isolation model, the two populations are assumed to have separated at generation t in the past, which we measure in units of $2ND$ generations with the parameter $T = t/(2ND)$. Considering the coalescent process within each population, the appropriate way to scale time is in units of $2N_e$ generations. Then, from equation (1) we can see that the isolation event occurred at time $T/[1 + (1/2M)]$ on this new time scale. This means that, in terms of the number of coalescent events we expect to occur between the present and generation t in the past, the time of isolation appears more recent when M is small. Similarly, in the two-population migration model, the migration rate between populations will seem higher by a factor of $1 + 1/(2M)$.

Expected Coalescence Times for Pairs of Sequences

Consider a single, infinite-sites locus within which there is no recombination (Watterson 1975). We begin with the simplest possible sample from two subdivided populations: two sequences taken (1) from the same deme; (2) from different demes in the same population; or (3) from different populations. This two-sequence sampling scheme, together with the assumptions outlined above, defines three different coalescence times: t_w is the time to coalescence for a pair of sequences sampled from the same deme (and necessarily from the same population); t_b is the time to coalescence for a pair of sequences from different demes within one population; and t_{12} is the time to coalescence for a pair of sequences sampled one from each population. Again, time is measured in units of $2ND$ generations.

Assuming that D is large and using standard coalescent machinery (Kaplan et al. 1988; Notohara 1990; Wakeley 1998), we have the following set of equations for the expected values of these coalescence times under the two-population migration model:

$$E(t_w) = \frac{2M}{2M + 1}E(t_b), \tag{4}$$

$$E(t_b) = \frac{1}{2M + 2M_{12}} + \frac{2M}{2M + 2M_{12}}E(t_w) + \frac{2M_{12}}{2M + 2M_{12}}E(t_{12}), \quad \text{and} \tag{5}$$

$$E(t_{12}) = \frac{1}{2M_{12}} + E(t_b). \tag{6}$$

The solutions to these equations are:

$$E(t_w) = 2, \tag{7}$$

$$E(t_b) = 2\left(1 + \frac{1}{2M}\right), \quad \text{and} \tag{8}$$

$$E(t_{12}) = 2\left(1 + \frac{1}{2M}\right) + \frac{1}{2M_{12}}. \tag{9}$$

These are a special case of Slatkin and Voelm's (1991) results. The expectations under the two-population isolation model can similarly be obtained and are given by:

$$E(t_w) = 1, \tag{10}$$

$$E(t_b) = 1 + \frac{1}{2M}, \quad \text{and} \tag{11}$$

$$E(t_{12}) = 1 + \frac{1}{2M} + T. \tag{12}$$

The factor of two difference between these expectations under migration and isolation reflects the fact that to make these two models comparable in terms of pairwise coalescence times within and between populations, the isolation model population sizes must be twice those in the migration model and T must be equal to $1/(2M_{12})$ (see Wakeley 1996a).

The fact that $E(t_w)$ is independent of migration rate and is equal to the expected coalescence time in a panmictic population of the same total size is another instance of Slatkin's (1987) and Strobeck's (1987) well-known result. Looking at the two pairs of equations, (8) and (9) and (11) and (12), we see a parallel with previous results for migration and isolation (Nei and Feldman 1972; Li 1976, 1977; Gillespie and Langley 1979; Takahata and Nei 1985; Slatkin 1987; Strobeck 1987). That is, the samples in which just a single sequence is taken from each deme behave as if each population is panmictic with an effective size given by equation (1). Consequently, small values of M manifest themselves as large values of N_e for these samples, which makes the difference between $E(t_{12})$ and $E(t_b)$ seem smaller. In relation to the coalescent process within populations, the migration rate between populations appears greater in the migration model and the time of separation appears shorter in the isolation model when the among-deme migration rate is small. Shifting to equations (7) and (8) for the migration model or equations (10) and (11) for the isolation model, we see that $E(t_b)$ becomes identical to $E(t_w)$ as M increases. In this case, the demic structure of the populations disappears and the expectations converge on the values for two panmictic populations connected either by migration or by isolation.

Now define π_w to be the average number of within-deme pairwise differences; π_b to be the average number of between-deme, but within-population pairwise differences; and π_{12} to be the average number of between-population pairwise differences for a sample from the two populations. The expected values of these do not depend on the sample size, and are obtained by multiplying equations (7), (8), and (9) by θ. Thus, we can estimate the parameters of the migration model using

$$\hat{\theta} = \frac{\pi_w}{2}, \tag{13a}$$

$$\hat{M} = \frac{\pi_w}{2(\pi_b - \pi_w)}, \quad \text{and} \tag{13b}$$

$$\hat{M}_{12} = \frac{\pi_w}{4(\pi_{12} - \pi_b)} \tag{13c}$$

and those of the isolation model using

$$\hat{\theta} = \pi_w, \tag{14a}$$

$$\hat{M} = \frac{\pi_w}{2(\pi_b - \pi_w)}, \quad \text{and} \tag{14b}$$

$$\hat{T} = \frac{\pi_{12} - \pi_b}{\pi_w}. \tag{14c}$$

Formulae for the variances of $\hat{\theta}$, M, M_{12}, and T can also be derived. These will depend on the sample size (Tajima 1983; Takahata and Nei 1985) and will be inversely related to the migration parameters M and M_{12} (Wakeley 1996b).

The Hierarchy of Migration Rates

For the subdivision between populations to be evident, the large-D island model requires that the migration rate between populations, m_{12}, be much smaller than the migration rate among demes, m. This is reflected in the choice of parameters above: $M = 2NmD/(D - 1)$ and $M_{12} = 2NDm_{12}$. Postponing the assumption that D is large and letting $M_{12}^* = 2Nm_{12}$, we find that

$$E(t_w) = 2, \tag{15}$$

$$E(t_b) = 2\left[1 + \frac{1}{2(M + M_{12}^*)}\right], \quad \text{and} \tag{16}$$

$$E(t_{12}) = 2\left[1 + \frac{1}{2(M + M_{12}^*)}\right] + \frac{1}{2DM_{12}^*}\left(\frac{M - M_{12}^*}{M + M_{12}^*}\right). \tag{17}$$

Again, equations (15), (16), and (17) are a special case of Slatkin and Voelm's (1991) results. These equations suggest that $E(t_b)$ might be larger than or smaller than $E(t_{12})$, depending on the sign of the rightmost term in equation (17), which is positive when $M > M_{12}^*$ and negative when $M < M_{12}^*$. Thus, using observed values of π_w, π_b, and π_{12}, we could estimate θ and the combined migration parameter, $(M + M_{12}^*)$, and we could distinguish whether intrapopulation migration was stronger than interpopulation migration or vice versa.

However, as D grows $E(t_b)$ and $E(t_{12})$ in equations (16) and (17) become identical. They both become ''between-deme'' times when D is large, because in this case migration is not restricted between populations relative to between demes. The important parameter in shaping interpopulation variation is the total number of migrants each population receives each generation, $2NDm_{12}$. If this number is much larger than one, the effect of subdivision on between-population differentiation is negligible. Whatever fixed value of M_{12}^* we assume, the number of migrants entering each population, $2NDm_{12} = DM_{12}^*$, will go to infinity as D increases. Substituting M_{12} for DM_{12}^* in equations (16) and (17) and letting D go to infinity gives equations (8) and (9). When there are a large number of demes in each population, for the effect of between-population subdivision to be observable, the number of migrants each deme receives per generation from within its own population must be much larger than the number it receives from the other population.

Probabilities of Genealogies under Migration and Isolation

Takahata and Slatkin (1990) obtained genealogical tree probabilities for a sample of two sequences from each population under the two-population migration model assuming no subdivision within populations. Tajima (1983) did the same for the isolation model. These include the probabilities that the sample is monophyletic, paraphyletic, or polyphyletic, with respect to the two populations. The genealogies in Figure 1 show sample polyphyly. Here we focus on the probability of monophyly—that the pairs of sequences within each population have a common ancestor more recently than the common ancestor of the entire sample. This is sometimes called reciprocal monophyly (e.g., Avise 1994). Figure 2a shows the probability of monophyly under migration and isolation when there is no subdivision within the populations. If the migration rate is high or, correspondingly, the time of separation is short, the probability is close to 1/9, which is the value for a single panmictic population (Tajima 1983). As the migration rate between populations decreases or the time of separation increases, the chance of sample monophyly approaches one, but with slightly different profiles under migration and isolation.

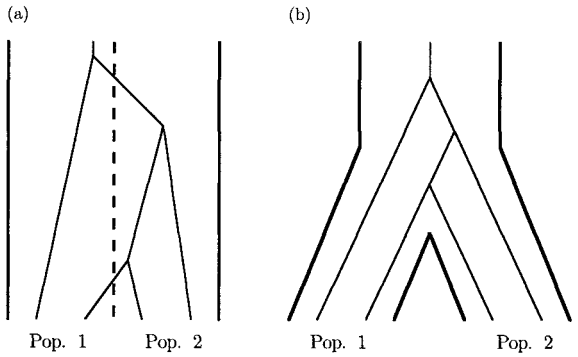

FIG. 1. The two-population migration model (a) and the two-population isolation model (b). The thick lines indicate population boundaries, with dashes showing that gene flow can occur. The thin lines trace ancestral lineages in one possible history of a sample of two sequences from each population.

To study the effects of subdivision within the populations on the probability of sample monophyly, we must consider whether the pairs of sequences from each population are sampled from the same or from different demes. Here we will focus on just two possibilities: that the pairs are sampled from the same deme within their respective populations and that the pairs are sampled from different demes within their respective populations. In the notation outlined above, now with superscripts to designate the two populations, these samples would be called $\{\boldsymbol{n}^{(1)} = (2), \boldsymbol{n}^{(2)} = (2)\}$ and $\{\boldsymbol{n}^{(1)} = (1, 1), \boldsymbol{n}^{(2)} = (1, 1)\}$. As discussed above, the second of these samples will have no scattering phase and will thus behave as if each population is panmictic with effective size, N_e, given by equation (1). For the first sample, the scattering phase will have to be considered.

Again, the collecting phase is a coalescent when time is measured in units of twice the effective population size given in equation (1). Scaling time by equation (1) means substituting $M_{12}[1 + (1/2M)]$ for Takahata and Slatkin's (1990) migration parameter, and $T/[1 + (1/2M)]$ for the time in Tajima's (1983) expressions for the probability of sample monophyly. The scattering phase for $\{\boldsymbol{n}^{(1)} = (2), \boldsymbol{n}^{(2)} = (2)\}$

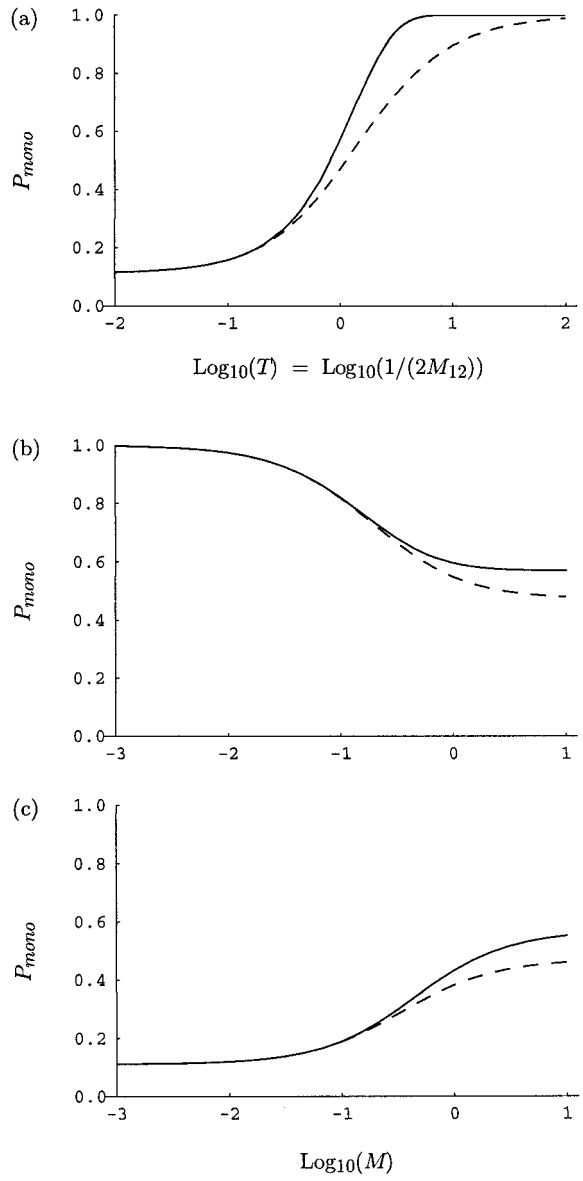

FIG. 2. The probability of monophyly for a sample of two sequences from each of two populations under isolation (———) and migration (– – –). (a) The case of two panmictic populations; (b) and (c) the case of two subdivided populations, but different samples: {(2), (2)} in (b), {(1,1), (1,1)} in (c). Along the horizontal axis, the average numbers of pairwise differences within and between populations are the same for both models, i.e., $T = 1/(2M_{12})$ and the population sizes under isolation are twice as large as those under migration (see text).

leads to the following characterization of the probability of monophyly, P_{mono}, for that sample:

$$P_{mono}\{(2), (2)\} = \left(\frac{1}{2M+1}\right)^2 + 2\left(\frac{1}{2M+1}\right)\left(\frac{2M}{2M+1}\right)P_{mono}\{(1), (1, 1)\} + \left(\frac{2M}{2M+1}\right)^2 P_{mono}\{(1, 1), (1, 1)\}. \quad (18)$$

The first term on the right represents the event that both samples coalesce during the scattering phase, in which case monophyly is necessarily achieved. The second and third terms represent one and two scattering-phase migration events, respectively. In both cases, the resulting P_{mono} is found in Takahata and Slatkin (1990) for migration and in Tajima (1983) for isolation, but with time measured according to equation (1).

Figures 2b and 2c show P_{mono} {(2), (2)} and P_{mono} {(1, 1), (1, 1)} over a range of values of the within-population migration rate, M, when $M_{12} = 0.25$ in the two-population migration model or, equivalently, $T = 2.0$ in the two-population isolation model. Thus, the values for the curves at the far right of Figure 2, which represent a high migration rate ($M = 10$), are essentially the same as the values of the curves for two panmictic populations shown in Figure 2a when $\log_{10}(T) = \log_{10}[1/(2M_{12})] = 0.30$. Figure 2 shows that, for particular values of the migration rate or the time of separation between populations, the probability of sample monophyly depends both on the migration rate among demes within populations and on the way the sample was taken.

When the pairs of sequences are each sampled from the same deme within their respective populations (Fig. 2b) decreasing the migration rate among demes increases the chance of sample monophyly. This is due mainly to the first term on the right of equation (18), which becomes one as M decreases, while the other two terms go to zero. When all four sequences are from different demes (Fig. 2c the probability of sample monophyly approaches the panmictic value of 1/9 as the migration rate decreases to zero, as the curves in Figure 2a do when T decreases. Because there is no scattering phase for this sample, Figure 2c simply reflects the dependence of the coalescent time scale within populations on M. When the migration rate among demes is low, the time of isolation of the two populations appears more recent or, alternatively, the interpopulation migration rate seems larger. Again, this is because genealogical time is measured in units of $2N_e$ generations, with the effective size given by equation (1)

Shared, Fixed, and Exclusive Polymorphisms under Isolation

Every polymorphic site in a sample from two populations falls into one of four categories: polymorphic exclusively in the sample from population 1, polymorphic exclusively in the sample from population 2, polymorphic in neither population, and polymorphic in both. Let the sample counts of these exclusive, fixed, and shared polymorphisms be called S_{x1}, S_{x2}, S_f, and S_s, respectively. Wakeley and Hey (1997) derived the expected values of S_{x1}, S_{x2}, S_f, and S_s under Watterson's (1975) infinite sites model of mutation for a general isolation model of two panmictic populations. The model allows the effective size of each of the three populations (ancestor and two descendents) to be different and specifies that they became isolated from one another t generations ago. The expected values, $E(S_{x1})$, $E(S_{x2})$, $E(S_f)$, and $E(S_s)$, are given by equations (12) through (16) in Wakeley and Hey (1997), and are not reproduced here.

Figure 3a shows the proportions of all segregating sites that are exclusive in population 1, shared by both, and fixed

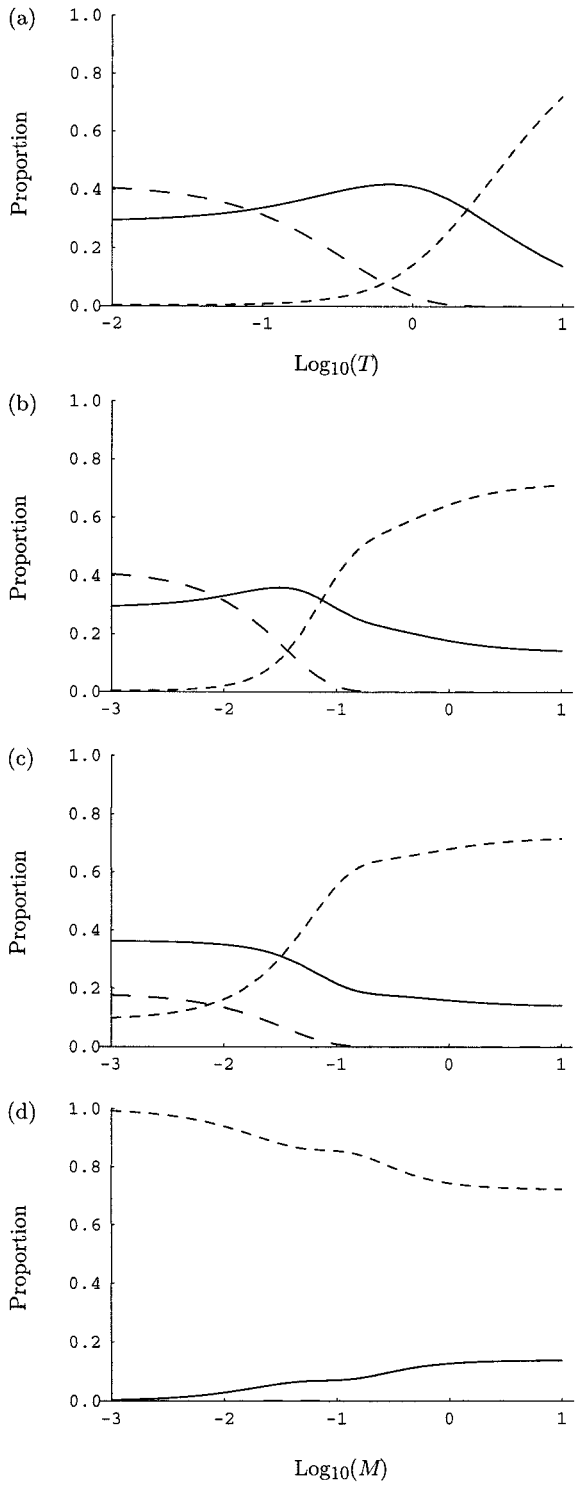

FIG. 3. The expected proportions of polymorphisms: $E(S_{x1})/E(S)$ (———), $E(S_s)/E(S)$ (– – –), and $E(S_f)/E(S)$ (- - - - -), for the isolation model, (a) without and (b), (c), and (d) with further subdivision. Panels (b), (c), and (d) are for three different samples: {(1, 1, 1, 1), (1, 1, 1, 1)}, {(2, 2), (2, 2)}, and {(4), (4)}, respectively.

between the two under this isolation model without subdivision. It is assumed that $\theta_1 = \theta_2 = \theta_A = 10$ and $n^{(1)} = n^{(2)} = 4$, and the values are plotted over a broad range of T. Because in this case $E(S_{x1})$ and $E(S_{x2})$ are the same, only three curves appear in Figure 3a. When the time of separation is great, the majority of segregating sites are fixed differences and essentially no shared polymorphisms are expected. As T increases, $E(S_{x1})$ converges on $\theta_1 \sum_{i=1}^{n-1} 1/i$ (Watterson 1975), which makes up an increasingly smaller fraction of all polymorphisms. At the other extreme, as T decreases to zero, the values become those expected in a single panmictic population, in which case fixed differences are very unlikely for samples of this size (Hey 1991). The reason for using proportions rather than the numbers themselves will become clear below.

If the populations in this isolation model are themselves subdivided, then the expressions for $E(S_{x1})$, $E(S_{x2})$, $E(S_f)$, and $E(S_s)$ in Wakeley and Hey (1997) will hold for the $n'^{(1)}$ and $n'^{(2)}$ sequences left in the two populations at the end of the scattering phase. Assume that we have sampled $\mathbf{n}^{(1)} = (n_1^{(1)}, n_2^{(1)}, \ldots, n_{d^{(1)}}^{(1)})$ and $\mathbf{n}^{(2)} = (n_1^{(2)}, n_2^{(2)}, \ldots, n_{d^{(2)}}^{(2)})$ sequences from $d^{(1)}$ and $d^{(2)}$ demes from populations 1 and 2. Then we have the general formula:

$$E(X) = \sum_{\mathbf{n}'^{(1)}} \sum_{\mathbf{n}'^{(2)}} P[\mathbf{n}'^{(1)} \,|\, \mathbf{n}^{(1)}] P[\mathbf{n}'^{(2)} \,|\, \mathbf{n}^{(2)}] E(X \,|\, n'^{(1)}, n'^{(2)}) \tag{19}$$

in which X could be any of S_{x1}, S_{x2}, S_f, and S_s, $P[\mathbf{n}'^{(i)} \,|\, \mathbf{n}^{(i)}]$ is given by equation (3), $n^{(i)} = \sum_{j=1}^{d^{(i)}} n'^{(i)}_j$, and the sums are taken over all possible values of $n'^{(1)}_j$ $(1 < j < d^{(1)})$ and $n'^{(2)}_j$ $(1 < j < d^{(2)})$. In words, equation (19) states that the expected values of S_{x1}, S_{x2}, S_f, and S_s are averages, taken over all possible outcomes of the scattering phase. We have already seen an example of such averaging in equation (18) above.

Figures 3b, 3c, and 3d show the same proportions as in Figure 3a, when T is fixed at 10 and still $\theta_1 = \theta_2 = \theta_A = 10$, but now over a range of values of the among-deme migration parameter, M. The three panels represent different possible samples of the four sequences from each population: Figure 3b, four sequences from four different demes in each population; 3c, two pairs of sequences from two different demes in each population; and 3d, all four sequences from a single deme in each population. In all three panels, at the far right the proportions are essentially the same as those at the far right in Figure 3a. This is to be expected because when $M = 10$, each population is very nearly panmictic.

Comparing Figure 3b to Figure 3a shows that decreasing M has the same effect as decreasing T, although the rates of change of the porportions differ. Because the sample represented in Figure 3b, {(1, 1, 1, 1), (1, 1, 1, 1)}, has no scattering phase, the only effect of changing M is to change the coalescent time scale. Skipping down to Figure 3d, we see a very different pattern. The proportion of fixed differences increases and the proportion of exclusive polymorphisms diminishes to zero with decreasing M, and few if any shared polymorphisms are expected at all. The genealogical structure of this sample, {(4), (4)}, is dominated by the scattering phase when M is small. In this case, the most likely configuration at the end of the scattering phase is a single ancestral lineage

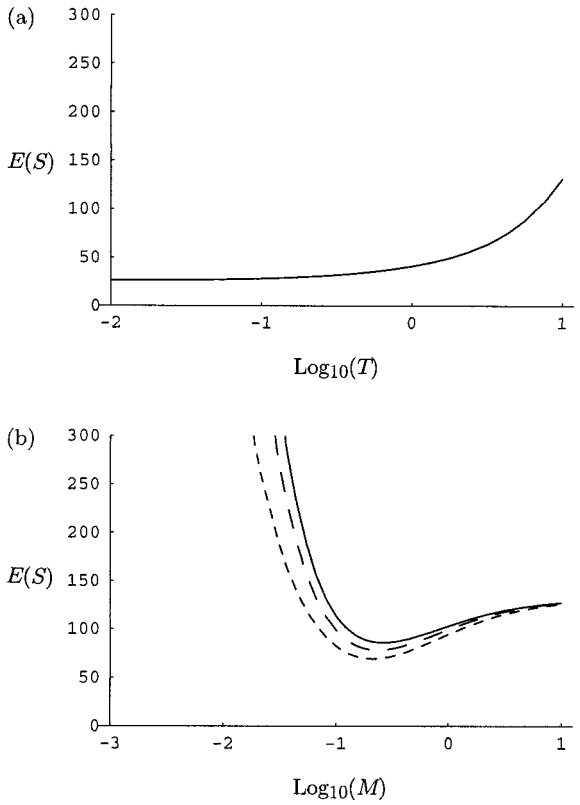

FIG. 4. The expected total number of polymorphisms, $E(S)$, for (a) two panmictic populations, and (b) two subdivided populations. The three different sampling schemes used in Figure 3 are all shown in (b): {(1, 1, 1, 1), (1, 1, 1, 1)}, (———), {(2, 2), (2, 2)} (– – –), and {(4), (4)} (- - - - -).

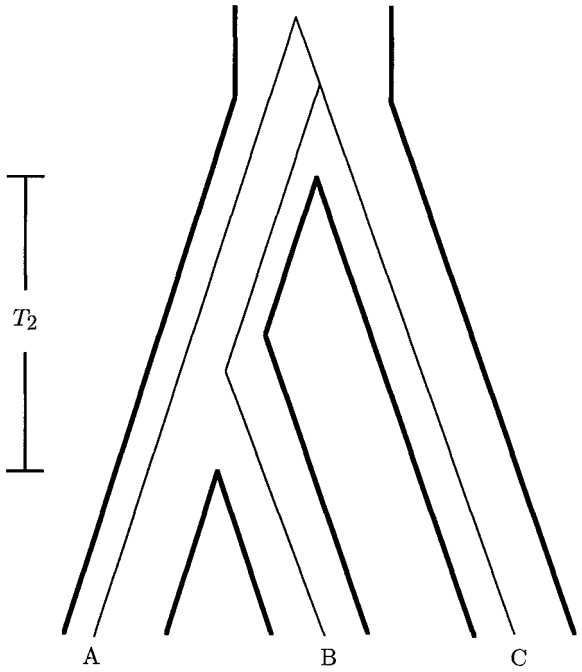

FIG. 5. A case where the gene tree and the species tree are not the same. As in Figure 1, thick lines indicate species boundaries and the thin lines trace ancestral lineages the history of the sample.

in each population, and the situation at the left of Figure 3d is obtained. The sample used to generate Figure 3c {(2, 2), (2, 2)}, might seem as if it should produce an intermediate pattern, but it is closer to Figure 3b. The reason for this is that, even though this sample has a scattering phase, the limiting situation when M is small is two collecting-phase lineages in each population. Thus, the values at the far left of Figure 3c are those expected in a panmictic population when two pairs of sequences are compared—rather than two sets of four as in Figures 3a and 3b—and this sample is expected to show all three types of polymorphisms: shared, fixed, and exclusive.

Figure 4 shows the total expected number of segregating sites, $E(S)$, for the same samples and parameters as in Figure 3. These differ strikingly under the two-population isolation model without further subdivision (Fig. 4a), and the isolation model for two subdivided populations (Fig. 4b). In Figure 4a, $E(S)$ starts off on the left very close to the value expected for a sample of eight sequences from a single panmictic population. As T increases and more fixed differences are expected, $E(S)$ increases as well. Figure 4b shows that when there is subdivision within each population we expect essentially the opposite trend: as M decreases, $E(S)$ increases. In contrast to Figure 3, the sampling scheme has little effect. Figure 4b simply illustrates the dependence of the effective population size on the migration rate among demes. Regardless of the strong scattering-phase effects on genealogical topologies shown in Figure 3, the duration of the collecting phase, i.e., the depth of the genealogy, grows according to equation (1) as M decreases.

Gene Trees and Species Trees with Subdivision

Because gene divergences must predate species divergences, there is always a chance that the genealogy of a sample of genes will be inconsistent with the phylogeny of the species from which it was taken. Pamilo and Nei (1988) studied the probability that a gene tree has the same topology as a species tree when a single sample is taken from each species. They considered cases of three, four, and five species. Here I will focus on just three and investigate the effect of subdivision within the species on the probability that the gene tree is inconsistent with the species tree when one sequence is taken from each species. The results will allow some inferences about what will happen in larger samples. Takahata (1989) has studied the behavior of these larger samples under the assumption of panmixia within species. Note that what follows applies to the genealogy only and does not consider whether appropriate genetic variation has accrued as Wu (1991, 1992) and Hudson (1992) have done.

Figure 5 shows the genealogy of a sample of one sequence from each of three species. The gene tree will be inconsistent with the species tree in Figure 5 if lineages A and B do not coalesce during the interval T_2 and then one of them coalesces with lineage C before the common ancestor of all three is reached. In Figure 5, lineages B and C are the first to coalesce. The probability of inconsistency under panmixia is given in Pamilo and Nei (1988), and under the D-deme island model within species it becomes:

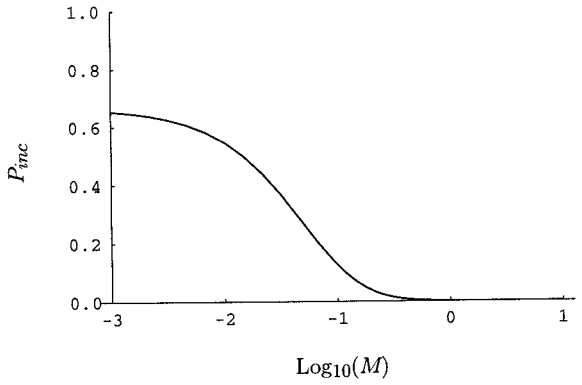

FIG. 6. The probability of inconsistency of the gene tree with the species tree for the sample in Figure 5, with $T_2 = 10$, against the among-deme migration rate.

$$P_{inc} = \frac{2}{3}\exp\left[-T_2 \Big/ \left(1 + \frac{1}{2M}\right)\right], \qquad (20)$$

where T_2 is the internode length measured in units of $2ND$ generations. As in Pamilo and Nei (1988), increasing the internode length decreases the chance of inconsistency, but now this can be opposed by decreasing M.

Figure 6 plots equation (20), when $T_2 = 10$, over the same range of M used in Figures 2–4. Under intrapopulation panmixia, $T_2 = 10$ virtually guarantees that the topologies will be the same. Strictly speaking, M here refers to migration only in the common ancestral species to A and B; subdivision within the other species is not required for these results to hold. When M is large, the probability of inconsistency between the gene tree and the species tree is very low. However, as M gets smaller, the probability of inconsistency increases, eventually reaching two-thirds. Two-thirds is the probability that A and B are not the first to coalesce in a single randomly mating population (Tajima 1983). In other words, when M is very small, it as if the species are not diverged at all. When $M = 0.1$, there is a 13% chance that the gene tree will not be the same as the species tree. Limited migration makes coalescence times longer within species, which causes divergence times to appear relatively shorter.

When more than one sequence is taken from each species, the probability of inconsistency is defined in terms of the first interspecific coalescence (Takahata 1989). If this is between an A and a B, the gene tree is consistent with the species tree, otherwise it is inconsistent. Takahata (1989) found that the probability of consistency increases with sample size. The larger the sample sizes from A and B, the more likely it is that multiple ancestral lineages will exist at the time of their split, and the greater the number of these, the more likely it is that there will be an interspecific coalescent event in the interval T_2. This will still be true if the species are themselves subdivided, but there will be an additional complication: how the sample is spread among demes will matter. This is easiest to picture when M is small, in which case each deme's sample will most likely be represented by a single lineage at the start of the collecting phase. Taking larger samples from a fixed number of demes will have little effect on the number of collecting-phase lineages when M is small, and thus little effect on the probability of consistency between the gene tree and the species tree. The probability of consistency will be increased most strongly by increasing the number of demes sampled.

DISCUSSION

Two unifying principles characterize the results presented here. When a population or species is subdivided into demes, migration: determines the historical distribution of a sample among demes; and alters the time scale of the coalescent process. These two closely related phenomena are at once obviously true and confoundingly difficult to illustrate for general models of subdivision and migration. The model used here sacrifices some of its generality by assuming island-type migration among a large number of demes, and in doing so distills (1) and (2) into the scattering phase and the collecting phase, respectively. The scattering phase, which is marked by coalescent events within demes and migration events to demes empty of ancestral lineages, is brief and ends when each remaining lineage is in a separate deme. The collecting phase takes these lineages through a coalescent process whose length depends (inversely) on the among-deme migration rate. Potentially lost in this simplification are the ability to address situations where the assumption of a large number of demes might not hold and to study biologically interesting aspects of geographic variation, such as isolation by distance. Gained is an easy framework for investigating other important processes within the context of a subdivided population.

The effect of subdivision on average pairwise differences; genealogical tree probabilities; shared, fixed, and exclusive polymorphisms; and the consistency probability of gene trees and species trees depends on the both the scattering phase and the collecting phase. The scattering phase modifies genealogical topologies and the collecting phase determines the time scale of the coalescent. When the migration rate among demes, M, is large, the scattering phase and the collecting phase conspire to make the history of all samples look like the usual, single-population coalescent (Kingman 1982; Hudson 1983; Tajima 1983), regardless of the distribution of the sample among demes. The population is essentially panmictic in this case. When M is small, however, the interaction between the scattering and collecting phases and the sampling scheme can produce a variety of effects.

On the one hand, the histories of multideme samples will be much longer when M is small. This can be seen in the expectations of pairwise difference (e.g., eq. 7) and in the relative dimunition of interpopulation and interspecies differences shown in Figures 2c, 3b, 3c, and 6 (and cf. eqs. 8 and 9). On the other hand, because of the scattering phase, some samples will not be affected by this new time scale when certain measures are considered. For example, the high probability of monophyly at the left of Figure 2b and the high porportion of fixed differences at the left of Figure 3d are the result of a low-M scattering phase. When other measures, which depend on the depth of the genealogy, are considered, the long collecting phase again becomes apparent (e.g., see Fig. 4b).

To illustrate some of the practical consequences of this work, consider the estimation of the divergence time for a

pair of species. Assume that two species separated 10 million years ago, that they each number 50,000 individuals, and have generation times of five years. The generation time and population size of the ancestral species is assumed to be identical to that of its descendents, and when subdivision exists, the migration rate among demes is assumed to be identical in all three species. We require three new parameters: the time of the split in years, τ, the generation time in years, g, and the mutation rate per year at the locus under study, μ. Thus, the old parameters t and u are equivalent to τ/g and μg, respectively, and we can rewrite any results for expectations of pairwise differences in these terms. Further, we have $\tau = 10^7$ and $2NDg = 5 \times 10^5$, both in years. Lastly, we assume that a good estimate of μ is available as a molecular clock.

A naive way to estimate the divergence time, τ, would be to use the number of differences between a pair of sequences, one sampled from each species, or the average numbers of differences between pairs of sequences in a larger sample, π_{12}, together with our molecular clock for the locus. In other words, we would use $\tau = \pi_{12}/(2\mu)$ as an estimate of the divergence time. Of course, it is well known that this will overestimate the divergence time even without any subdivision. The panmictic result is:

$$E(\pi_{12}) = 2\mu(\tau + 2NDg), \tag{21}$$

which can also be obtained from equation (12) by letting M increase to infinity. Thus, in our example, we would overestimate the divergence time by 500,000 years, or 5% of the total. Now, equation (12) (multiplied by θ) shows that subdivision within the ancestral species will further increase the degree of overestimation:

$$E(\pi_{12}) = 2\mu\left[\tau + 2NDg\left(1 + \frac{1}{2M}\right)\right]. \tag{22}$$

If $M = 0.5$, the time would by overestimated by one million years, and if $M = 0.1$ it would be overestimated by three million years, or 30% of the total.

Edwards (1997) suggested that this problem of overestimation due to subdivision within a common ancestor could be corrected using methods developed for the divergence of panmictic species. For panmictic species, Nei and Li (1979) showed that the net number of nucleotide differences gives an unbiased estimate of divergence when the ancestral and descendent population sizes are the same. The net number of nucleotide differences is defined to be:

$$d = \pi_{12} - \frac{\pi_1 + \pi_2}{2} \tag{23}$$

(Nei and Li 1979), which should not be confused with the same symbol used above for the number of sampled demes. In equation (23), π_1 and π_2 are the average numbers of differences between pairs of sequences from populations 1 and 2, respectively, without regard to demic structure. Under the panmictic model, the expectation of $(\pi_1 + \pi_2)/2$ is equal to θ or $4NDg$, so $E(d) = 2\mu\tau$ and, if our estimate of the mutation rate is correct, $\hat{\tau} = d/(2\mu)$ will be an unbiased estimate of the divergence time.

For the subdivided population model used here, the expectation of $(\pi_1 + \pi_2)/2$ will be equal to $E(\pi_w)$ when all sequences are sampled from a single deme within each population, and it will be equal to $E(\pi_b)$ when every sequence comes from a separate deme. Only in the latter case can using equation (23) provide an unbiased estimate of the divergence time. In the former case, we would overestimate the divergence time by NDg/M years. For all other samples, the expectation of $(\pi_1 + \pi_2)/2$ will be intermediate between $E(\pi_w)$ and $E(\pi_b)$, so $\tau = d/(2\mu)$ will overestimate τ by some amount. The exact relationship amount π_w, π_b, π_1, and π_2, which would determine the extent of overestimation using equation (23), is somewhat complicated because π_w and π_b are defined as averages over the entire sample, whereas π_1 and π_2 are averages within each population. However, as an example, if five sequences are samped from each of two demes within each population, for a total of 10 sequences from each population, then $(\pi_1 + \pi_2)/2$ is equal to $(4\pi_w + 5\pi_b)/9$, and using $M = 0.1$ from above, the divergence time would still be overestimated by 11%, or more than one million years, compared to the situation in which the ancestor is panmictic. In contrast, $\hat{\tau} = (\pi_{12} - \pi_b)/(2\mu)$ would give an unbiased estimate of the divergence time, again assuming that μ is without error.

Other methods of estimating divergence times, notably those of Takahata (1986) and Takahata et al. (1995), may not be subject to these biases when subdivision is ignored. The reason is that these methods, in addition to the usual assumption of neutral infinite-sites mutations, assume only that the genealogical process in the common ancestor is a coalescent. Because this is in fact the case under the model of subdivision used here, estimates of the divergence time and the ancestral effective population size will be accurate regardless of whether the ancestor was subdivided. In the case of a subdivided ancestor, we can interpret the effective size of the ancestor in terms of equation (1). A related point is that, although we can sometimes distinguish a panmictic ancestor from a subdivided one when the number of demes is small (Wakeley 1996a), it may not be possible when the number of demes in the ancestral population is large.

Acknowledgments

I am very grateful to S. Edwards and M. Hare for helpful discussions and for comments on an earlier version of the manuscript. This work was supported by National Science Foundation grant DEB-9815367.

Literature Cited

Abramowitz, M., and I. A. Stegun. 1964. Handbook of mathematical functions. Dover, New York.

Avise, J. C. 1994. Molecular markers, natural history and evolution. Chapman and Hall, New York.

Edwards, S. V. 1997. The relevance of microevolutionary processes to higher level molecular systematics. Pp. 251–278 *in* D. Mindell, ed. Avian molecular evolution and systematics. Academic Press, London.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. Theort. Pop. Biol. 3:87–112.

Gillespie, J. H., and C. H. Langley. 1979. Are evolutionary rates really variable? J. Mol. Evol. 13:27–34.

Hey, J. 1991. The structure of genealogies and the distribution of

fixed differences between DNA sequences from natural populations. Genetics 128:831–840.
Hoelzer, G. A. 1997. Inferring phylogenies from mtDNA variation: mitochondrialgene trees versus nuclear-gene trees revisited. Evolution 51:622–626.
Hoelzer, G. A., J. Wallman, and D. J. Melnick. 1998. The effects of social structure, geographical structure, and population size on the evolution of mitochondrial DNA. II. Molecular clocks and the lineage sorting period. J. Mol. Evol. 47:21–31.
———. 1999. Erratum: the effects of social structure, geographical structure, and population size on the evolution of mitochondrial DNA. II. Molecular clocks and the lineage sorting period. J. Mol. Evol. 48:628–629.
Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203–217.
———. 1992. Gene trees, species trees and the segregation of ancestral alleles. Genetics 131:509–512.
Kaplan, N. L., T. Darden, and R. R. Hudson. 1988. Coalescent process in models with selection. Genetics 120:819–829.
Karlin, S., and J. McGregor. 1972. Addendum to a paper of W. Ewens, Theoret. Pop. Biol. 3:113–116.
Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to the steady flux of mutations. Genetics 61:893–903.
Kingman, J. F. C. 1982. The coalescent. Stochastic Process. Appl. 13:235–248.
Latter, B. D. H. 1973. The island model of population differentiation: a general solution. Genetics 73:147–157.
Li, W.-H. 1976. Distribution of nucleotide difference between two randomly chosen cistrons in a subdivided population: the finite island model. Theoret. Pop. Biol. 10:303–308.
———. 1977. Distribution of nucleotide difference between two randomly chosen cistrons in a finite population. Genetics 85: 331–337.
Maruyama, T. 1970. Effective number of alleles in a subdivided population. Theoret. Pop. Biol. 1:273–306.
Moore, W. S. 1995. Inferring phylogenies from mtDNA variation: mitochondrialgene trees versus nuclear-gene trees. Evolution 49: 718–726.
Nei, M., and M. W. Feldman. 1972. Identity of genes by descent within and between populations under mutation and migration pressure. Theoret. Pop. Biol. 3:460–465.
Nei, M., and W.-H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. 76:5269–5273.
Nei, M., and N. Takahata. 1993. Effective population size, genetic diversity, and coalescence time in subdivided populations. J. Mol. Evol. 37:240–244.
Notohara, M. 1990. The coalescent and the genealogical process in geographically structured population. J. Math. Biol. 29:59–75.
Pamilo, P., and M. Nei. 1988. The relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.
Slatkin, M. 1987. The average number of sites separating DNA sequences drawn from a subdivided population. Theoret. Pop. Biol. 32:42–49.
Slatkin, M., and L. Voelm. 1991. F_{ST} in a hierarchical island model. Genetics 127:627–629.
Strobeck, C. 1987. Average number of nucleotide differences in an sample from a single subpopulation: a test for population subdivision. Genetics 117:149–153.
Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.
Takahata, N. 1986. An attempt to estimate the effective size of the ancestral species common to two extant speices from which homologous genes are sequenced. Genet. Res. Camb. 48: 187–190.
———. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122:957–966.
Takahata, N., and M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344.
Takahata, N., and M. Slatkin. 1990. Genealogy of neutral genes in two partially isolated populations. Theoret. Pop. Biol. 38: 331–350.
Takahata, N., Y. Satta, and J. Klein. 1995. Divergence time and population size in the lineage leading to modern humans. Theoret. Pop. Biol. 48:198–221.
Tavaré, S. 1984. Lines-of-descent and genealogical processes, and their application in population genetic models. Theoret. Pop. Biol. 26:119–164.
Wakeley, J. 1996a. Distinguishing migration from isolation using the variance of pairwise differences. Theoret. Pop. biol. 49: 369–386.
———. 1996b. The variance of pairwise nucleotide differences in two populations with migration. Theoret. Pop. Biol. 49:39–57.
———. 1998. Segregating sites in Wright's island model. Theoret. Pop. Biol. 53:166–175.
———. 1999. Non-equilibrium migration in human history. Genetics 153:1863–1871.
Wakeley, J., and J. Hey. 1997. Estimating ancestral population parameters. Genetics 145:847–855.
Watterson, G. A. 1975. On the number of segregating sites in genetical model without recombination. Theoret. Pop. Biol. 7: 256–276.
Wright, S. 1931. Evolution in Mendelian populations. Genetics 16: 97–159.
———. 1943. Isolation by distance. Genetics 28:114–138.
Wu, C.-I. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics 127:429–435.
———. 1992. Reply to Richard R. Hudson. Genetics 131:513.

Corresponding Editor: J. Neigel