THEORETICAL POPULATION BIOLOGY **49**, 39–57 (1996)
ARTICLE NO. 0002

The Variance of Pairwise Nucleotide Differences in Two Populations with Migration

JOHN WAKELEY*

Department of Integrative Biology University of California, Berkeley

Received August 16, 1994

The variances of three measures of pairwise difference are derived for the case of two populations that exchange migrants. The resulting expressions can be used to place standard errors on estimates of population genetic parameters. The three measures considered are the average number of intrapopulation nucleotide differences, the average number of interpopulation nucleotide differences, and the net number of nucleotide differences between the two populations. The expectations of these statistics are previously known and suggest that they might be used to the quantify the divergence between populations. However, the standard errors of all three statistics are shown to be quite large relative to their expectations. Thus, our ability to quantify divergence between populations with them is limited, at least using available data. An analysis of mitochondrial DNA sequences from grey-crowned babblers illustrates the application of the theory. The variances derived here for migration are compared to previously published results for two populations that have been completely isolated from one another for some length of time. All three variances are greater under migration than under isolation, suggesting that a test to distinguish these two demographic situations could be developed. © 1996 Academic Press, Inc.

1. INTRODUCTION

Understanding the demographic history of natural populations is of fundamental significance in population genetics. The multitudes of DNA sequence data currently being collected offer the hope of accurately quantifying important genetic and demographic parameters. Of particular interest are samples of multiple DNA sequences from single species, e.g., Edwards (1993), since the numbers of nucleotide differences and the genealogies of sequences contain information about population genetic history. This paper is concerned with the effects that population subdivision and migration have on the variance of pairwise nucleotide differences. The model of population subdivision considered here is a two-population

* Present address: Department of Biological Sciences, Rutgers University, Piscataway, NJ 08855-1059. E-mail: jwakeley@ddbj.nig.ac.jp.

0040-5809/96 $18.00
Copyright © 1996 by Academic Press, Inc.
All rights of reproduction in any form reserved.

version of the finite-island model (Kimura and Weiss, 1964; Maruyama, 1970). While this model is an over-simplification of the demography of most natural populations, it does provide a starting-point from which we can draw initial conclusions.

Using the notation of Takahata and Nei (1985), d_X, d_Y, d_{XY}, and d are measures of pairwise nucleotide difference within and between two populations called X and Y. The intrapopulation measures, d_X and d_Y, are defined as the average number of nucleotide differences between two randomly chosen sequences from within populations X and Y, respectively. The interpopulation measure, d_{XY}, is defined as the average number of differences between one sequence randomly chosen from population X and another sequence randomly chosen from population Y. Nei and Li's (1979) d, the net number of nucleotide differences between populations, is defined as

$$d = d_{XY} - \frac{(d_X + d_Y)}{2}. \tag{1}$$

When a number of sequences, n_X and n_Y, are sampled from populations X and Y, respectively, d_X and d_{XY} are given by

$$d_X = \frac{2}{n_X(n_X - 1)} \sum_{i=1}^{n_X - 1} \sum_{i'=i+1}^{n_X} k_{ii'} \tag{2}$$

and

$$d_{XY} = \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} k_{ij}, \tag{3}$$

where k represents the number of differences between a particular pair of sequences. The subscripts i and j denote sequences from populations X and Y, respectively, with primes indicating additional sequences from the same population. The expression for d_Y is identical to (2) but with i replaced by j and n_X replaced by n_Y. In computing d_X and d_Y, each sequence is compared to all others but not to itself.

The relationship between intrapopulation and interpopulation differences gives an indication of the degree of population subdivision (Slatkin, 1987; Strobeck, 1987). Specifically, d, as a measure of the excess number of substitutions, quantifies the extent of divergence between populations. However, a lack of knowledge about the variation in d_X, d_Y, d_{XY}, and d under particular genetic and demographic models has both discouraged their use (Berry and Kreitman, 1993) and made their interpretation difficult (Simmons *et al.*, 1989). What follows is a derivation of the variances of d_X, d_Y, d_{XY}, and d for the case of two populations with migration. The resulting expressions are then used to place standard errors on estimates of

these parameters made using mitochondrial DNA sequences from the grey-crowned babbler (Edwards, 1993). The variances of d_X, d_Y, d_{XY}, and d derived here for two populations that exchange migrants are also compared to the variances, given by Takahata and Nei (1985), for two isolated populations. The results show that the variances of these quantities all quite large and are greater under migration than under isolation.

2. Methods

2.1. *Assumptions and Expectations*

Two randomly mating populations, each of effective haploid size N or diploid size $N/2$, are considered. Generations are nonoverlapping and the mutation rate at a locus for which data are available is u per sequence per generation. Each haploid individual has probability, m, of having immigrated from the other population in the previous generation. It is assumed that $1/N$, u, and m are all much less than one. This last assumption means that terms involving $(1/N)^2$, for example, can be ignored relative to terms involving $1/N$. Mutuations occur according to the infinite sites model of Kimura (1969) with the restriction that there is no recombination (Watterson, 1975). It is assumed that the two populations have reached equilibrium with respect to the above processes. Takahata (1983) studied that rate of approach to equilibrium conditions for the finite island model and found that it is approximately equal to the mutation rate.

The expectations of d_X, d_Y, d_{XY}, and d can be evaluated as $E(k_{ii'})$, $E(k_{jj'})$, $E(k_{ij})$, and $E(k_{ij}) - [E(k_{ii'}) + E(k_{jj'})]/2$. As illustrated below, these quantities are calculated by considering samples of two sequences from the two populations. With the assumptions oulined above, $E(d_X) = E(d_Y) = 4Nu$, $E(d_{XY}) = 4Nu + u/m$, and $E(d) = u/m$ (Nei and Feldman, 1972; Li, 1976; Griffiths, 1981; Slatkin, 1987; Strobeck, 1987; Notohara, 1990; Hey, 1991). The symbols θ and M are used below to represent $4Nu$ and $2Nm$, respectively. Thus, $E(d_X) = E(d_Y) = \theta$, $E(d_{XY}) = \theta + \theta/(2M)$, and $E(d) = \theta/(2M)$.

2.2. *The Variance of Pairewise Differences*

The variances of d_X, d_Y, d_{XY}, and d can be calculated using (1), (2), and (3). Following Tajima (1983) and Takahata and Nei (1985), these variances can be written as

$$\mathrm{Var}(d_X) = \frac{1}{n_X(n_X-1)}\,[2E(k_{ii'}^2) + 4(n_X-2)\,E(k_{ii'}k_{ii''}) + (n_X-2)(n_X-3)\,E(k_{ii'}k_{i''i'''})] - [E(k_{ii'})]^2, \qquad (4)$$

$$\mathrm{Var}(d_{XY}) = \frac{1}{n_X n_Y}\left[E(k_{ij}^2) + (n_X + n_Y - 2)\,E(k_{ij}k_{i'j}) + (n_X - 1)(n_Y - 1)\,E(k_{ij}k_{i'j'})\right] - \left[E(k_{ij})\right]^2, \quad (5)$$

and

$$\mathrm{Var}(d) = \tfrac{1}{4}\left[\mathrm{Var}(d_X) + \mathrm{Var}(d_Y)\right] + \mathrm{Var}(d_{XY}) + \tfrac{1}{2}\,\mathrm{Cov}(d_X, d_Y) - \mathrm{Cov}(d_{XY}, d_X) - \mathrm{Cov}(d_{XY}, d_Y), \quad (6)$$

where

$$\mathrm{Cov}(d_X, d_Y) = E(k_{ii'}k_{jj'}) - E(k_{ii'})\,E(k_{jj'}) \quad (7)$$

and

$$\mathrm{Cov}(d_{XY}, d_X) = \frac{2}{n_X}E(k_{ii'}k_{ij}) + \frac{n_X - 2}{n_X}E(k_{ii'}k_{i''j}) - E(k_{ii'})\,E(k_{ij}). \quad (8)$$

The assumptions of equivalent effective population sizes and symmetric migration make populations X and Y interchangeable. Thus, $E(k_{ii'}k_{ij}) = E(k_{jj'}k_{ji})$, $E(k_{ii'}k_{ii''}) = E(k_{jj'}k_{jj''})$, and so on, so that the expressions for $\mathrm{Cov}(d_{XY}, d_Y)$ and $\mathrm{Var}(d_Y)$ are obtained from the expressions for $\mathrm{Cov}(d_{XY}, d_X)$ and $\mathrm{Var}(d_X)$ by replacing n_X with n_Y.

Each of the terms on the right-hand sides of (4), (5), (7), and (8) can be calculated based upon the historical relationships among sequences sampled from the two populations. The history of a sample sequences is described by the "coalescent" process (Kingman, 1982a, 1982b; Hudson, 1983; Tajima, 1983). Looking back into the past, the coalescent models the occurrence of successive common ancestors of pairs of sequences in a sample (coalescent events) until the single common ancestor of all the sequences is reached. Hudson (1990) gives a thorough review. The approach taken here involves using matrices of single-generation transition probabilities among all the possible states that the sequences in a sample might have occupied during their history to derive probability generating functions for the times to particular coalescent events. This is illustrated below for the case of two sequences.

2.3. *Calculating the Expectation and Variance of $k_{ii'}$ and k_{ij} from a Sample of Two Sequences*

In any particular generation in the past, two sequences can either be in the same or in different populations, or they can have coalesced into a common ancestral sequence. That there are these three, rather than six possible states, results from the assumptions that the two populations are of the same effective size and that migration is symmetric. Otherwise, the

identity of the population in which a sequence resides would also be important (Takahata and Slatkin, 1990). The three states for the two sequences will be called zero, one, and two, respectively. If the sequences are labelled A and B, these states can be represented as AB, $A \mid B$, and (AB), where the vertical bar indicates that the two sequences are in different populations and the parentheses signify that they have coalesced. Since, for two sequences, only a single absorbing state exists, the probability of absorption in state two is equal to one.

Let p_{ij} represent the single-generation probability of transition from state i to state j. Then, the transition matrix for this sample has entries $p_{00} = 1 - 2m - 1/N$, $p_{01} = 2m$, $p_{02} = 1/N$, $p_{10} = 2m$, $p_{11} = 1 - 2m$, and $p_{22} = 1$, with all others equal to zero. Because of the assumption that $1/N$, u, and m are all much less than one, the chance of more than one event happening in a single generation is negligible. Thus, p_{01} represents the event that, in a single generation, the sequences move from state zero, being in the same population, to state one, being in different populations. This happens with probability $2m$, one for each haploid individual. Let $\phi(t)$ be the probability that two sequences destined to coalesce are separated from their common ancestor by exactly t generations. For the two sequences, A and B,

$$\phi_{02}(t) = p_{00}\,\phi_{02}(t-1) + p_{01}\,\phi_{12}(t-1) \tag{9}$$

and

$$\phi_{12}(t) = p_{11}\,\phi_{12}(t-1) + p_{10}\,\phi_{02}(t-1), \tag{10}$$

for $t \geqslant 2$ and $\phi_{02}(0) = 0$, $\phi_{02}(1) = 1/N$, and $\phi_{12}(0) = \phi_{12}(1) = 0$.

Equations (9) and (10) can be used to obtain the probability generating functions for the time to common ancestry starting in states zero and one. Let $\Phi(s) = \sum_{t=0}^{\infty} s^t \phi(t)$ be the probability generating function of $\phi(t)$. Multiplying (9) and (10) by s^t and summing over all values of t gives

$$\Phi_{02}(s) = (1 - 2m - 1/N)\,s\Phi_{02}(s) + 2ms\Phi_{12}(s) + s/N \tag{11}$$

$$\Phi_{12}(s) = (1 - 2m)\,s\Phi_{12}(s) + 2ms\Phi_{02}(s), \tag{12}$$

which are then solved to give

$$\Phi_{02}(s) = \frac{(s/N)[1 - (1-2m)\,s]}{[1 - (1-2m)\,s][1 - (1 - 2m - 1/N)\,s] - (2ms)^2} \tag{13}$$

$$\Phi_{12}(s) = \frac{(2m/N)\,s^2}{[1 - (1-2m)\,s][1 - (1 - 2m - 1/N)\,s] - (2ms)^2}. \tag{14}$$

The moments of t are easily derived from these probability generating functions. If $\Phi'(s)$ and $\Phi''(s)$ are the first and second derivatives of $\Phi(s)$ with respect to s, then $E(t) = \Phi'(1)$ and $\mathrm{Var}(t) = \Phi''(1) + \Phi'(1) - [\Phi'(1)]^2$.

Differentiating (13) and (14) and putting in $s = 1$ gives $E_{02}(t) = 2N$ and $E_{12}(t) = 2N + 1/(2m)$ with variances

$$\mathrm{Var}_{02}(t) \approx 4N^2 + N/m \tag{15}$$

$$\mathrm{Var}_{12}(t) \approx 4N^2 + N/m + 1/(4m^2). \tag{16}$$

Thus the expected time, in generations, to common ancestry of two sequences sampled from the same population is equal to the total size of the two populations together. The expected time to common ancestry of two sequences sampled from different populations is the waiting time until they are in the same population plus this value. The variances of the coalescence time for two sequences depend on both the population size and the migration rate. These results for the times to common ancestry are previously known (Notohara, 1990; Hey, 1991) and are consistent with work on the expected number of differences separating sequences from a subdivided population (Nei and Feldman, 1972; Li, 1976; Griffiths, 1981; Slatkin, 1987; Strobeck, 1987).

Once the expectation and variance of the time to a particular coalescent event are known, the rules of random sums are used to obtain information about the numbers of changes during that interval. The number of mutations on a particular lineage in the history of a sample is the sum, over the length of that lineage in generations, of the number of mutation per generation. Since u is assumed to be small, the number of mutations in a gene in one generation is one with probability u and zero with probability $1-u$. If a number of lineages exists over a random length of time, t, generations, and k and k' are the numbers of changes on two particular lineages during that time, then $E(k) = E(k') = uE(t)$, $\mathrm{Var}(k) = \mathrm{Var}(k') = u(1-u)\,E(t) + u^2\,\mathrm{Var}(t)$, and $\mathrm{Cov}(k, k') = u^2\,\mathrm{Var}(t)$. Again because u is assumed small, $\mathrm{Var}(k) = uE(t) + u^2\,\mathrm{Var}(t)$ can be used as an approximation.

Using these rules, $E(k_{ii'}) = 2uE_{02}(t)$ and $\mathrm{Var}(k_{ii'}) = 2uE_{02}(t) + 4u^2\,\mathrm{Var}_{02}(t)$, with the corresponding formulas for $E(k_{ij})$ and $\mathrm{Var}(k_{ij})$, obtained by replacing zero in the subscripts with one. Thus, $E(k_{ii'}) = \theta$ and $E(k_{ij}) = \theta + \theta/(2M)$, with $\theta = 4Nu$ and $M = 2Nm$, as already mentioned. The variances are given by

$$\mathrm{Var}(k_{ii'}) = \theta\left(1 + \theta + \frac{\theta}{2M}\right) \tag{17}$$

$$\mathrm{Var}(k_{ij}) = \theta\left(1 + \theta + \frac{\theta}{2M}\right) + \frac{\theta}{2M}\left(1 + \frac{\theta}{2M}\right). \tag{18}$$

The expectations of $k_{ii'}^2$ and k_{ij}^2, which are needed in (4) and (5) to calculate $\mathrm{Var}(d_X)$ and $\mathrm{Var}(d_{XY})$, are easily derived from these expressions.

2.4. *Samples of More the Two Sequences*

The remaining seven quantities in Eqs. (4), (5), (7), and (8) are calculated from samples of more than two sequences. Specifically, $E(k_{ii'}k_{ij})$, $E(k_{ij}k_{i'j})$, and $E(k_{ii'}k_{ii''})$ are calculated from samples of three sequences and $E(k_{ij}k_{i'j'})$, $E(k_{ii'}k_{jj'})$, $E(k_{ii'}k_{i''j})$, and $E(k_{ii'}k_{i''i'''})$ are calculated from samples of four sequences from the two populations. In each case, the protocol is similar to that followed above. However, the transition matrices for samples of more than two sequences generally have more than one absorbing state. That is, there are typically several possible coalescent events. This means that several different histories of each sample need to be considered. The seven quantities above are calculated separately for each possible history, then averaged, weighted by the probability of each.

Thus, when more than one coalescent event is possible, the probability of each being the first coalescent event to occur among the sequences must be derived. These probabilities, π_{ik}, of absorption in state k, starting in state i, satisfy $\pi_{ik} = \sum_{j=0}^{n} p_{ij}\pi_{jk}$, where n is the total number of states the sample can assume. Once obtained, they are used to construct conditional single-generation transition matrices, one for each absorbing state, given fixation in that state. The single-generation probability of transition from state i to state j, given eventual fixation in state k is given by $p_{ij}^{(k)} = p_{ij}\pi_{jk}/\pi_{ik}$ (cf. Ewens, 1979, Eq. (2.126)). These conditional transition matrices define sets of recursion equations, analogous to (9) and (10) above, that are solved to give the probability generating functions for the times to each particular coalescent event.

For example, $E(k_{ii'}k_{ij})$ is computed from the sample AAB, of two sequences from one population and one from the other. In this case, it is important to distinguish both whether the first coalescent event is intrapopulational or interpopulational and whether the two ancestral sequences were last in the same or in different populations. Thus, there are four possible coalescent events for the sample AAB. In general, the number of possible histories of a sample of size n is just the number of absorbing states for that sample times the number of absorbing states for the $n-1$ ancestral sequences times the number for $n-2$, and so on. Since there is only one absorbing state for the sample of two considered above, there are four possible histories of the sample AAB. These are shown in Fig. 1. The probabilities of these four histories are simply the probabilities of the four coalescent events of the sample AAB. The expectation and variance of the length of branches spanning t_3 are obtained from the probability generation functions for the times to each of these four events. The lengths of the branches spanning t_2 are described by the results of the previous section.

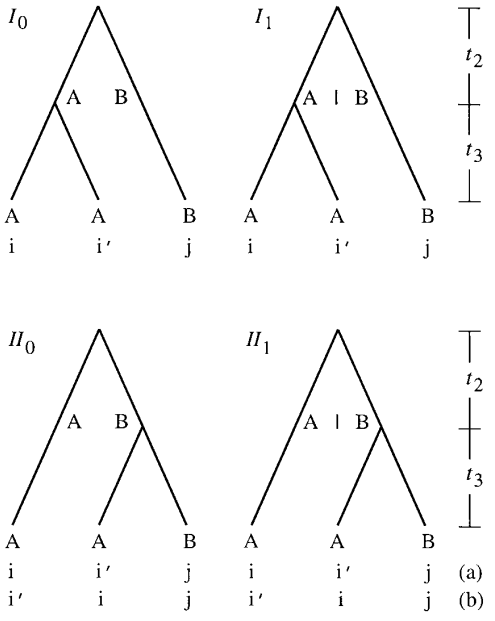


FIG. 1. The four possible histories of the sample *AAB*. *AB* at the $t_2 : t_3$ boundary indicates that the two ancestral sequences were in the same population and $A|B$ that they were in different populations. As discussed in the text, there are two equiprobable designations of the two *A*'s in the sample *AAB* as either i or i'.

In Calculating $E(k_{ii'}k_{ij})$ for each of the histories in Fig. 1, first $E(k_{ii'})$, $E(k_{ij})$, and $\mathrm{Cov}(k_{ii'}, k_{ij})$ are computed, then these quantities are combined to give the expectations of the product. The covariances and expectations are obtained by expressing $k_{ii'}$ and k_{ij}, in terms of the numbers of changes on the lineages spanning t_2 and t_3 in Fig. 1. For instance, if k_3 is the number of changes on a lineage spanning the interval t_3 in history I_0, then $E(k_{ii'}) = 2E(k_3)$. If k_3' is the number of changes on a different branch spanning t_3, then $\mathrm{Cov}(k_{ii'}, k_{ij}) = \mathrm{Var}(k_3) + 3\,\mathrm{Cov}(k_3, k_3')$ because there is no correlation between the numbers of changes on segments spanning t_3 and those on segments spanning t_2. For histories II_0 and II_1, values of $E(k_{ij})$ and $E(k_{ii'}k_{ij})$ differ depending on which of the sequences labelled *A* is designated i and which is designated i'. The two possible assignments, (a) and (b), shown in Fig. 1 are equiprobable and values of $E(k_{ij})$ and $E(k_{ii'}k_{ij})$ must be calculated for each case then averaged.

When four sequences are considered, these calculations become laborious. For example, in computing $E(k_{ij}k_{i'j'})$ and $E(k_{ii'}k_{jj'})$ from the sample *AABB*, of two sequences from each population, a total of 24 distinct histories must be considered. They are not reproduced here because of length considerations and since no new concepts need to be introduced. The details of the

derivation of all of $E(k_{ii'}k_{ij})$, $E(k_{ij}k_{i'j})$, $E(k_{ii'}k_{ii''})$, $E(k_{ij}k_{i'j'})$, $E(k_{ii'}k_{jj'})$, $E(k_{ii'}k_{i''j})$, and $E(k_{ii'}k_{i''i'''})$ can be found in (Wakeley, 1994) and are also available from the author upon request.

3. Results

Once all of the required quantities have been calculated, $\mathrm{Var}(d_X)$, $\mathrm{Var}(d_{XY})$, and $\mathrm{Var}(d)$ are obtained by simply substituting these expressions into Eqs. (4), (5), (7), and (8). The expressions for $\mathrm{Var}(d_X)$ and $\mathrm{Var}(d_{XY})$ are shown in the Appendix. The expression for $\mathrm{Var}(d)$ is substantially more complicated than either of these and is not shown. It is obtained by substituting the $\mathrm{Var}(d_X)$, $\mathrm{Var}(d_{XY})$, $\mathrm{Cov}(d_X, d_Y)$, and $\mathrm{Cov}(d_{XY}, d_Y)$, which are all shown in the Appendix, into Eq. (6).

As the expressions for all three of these variances are quite complicated, it is instructive to look at their behaviors under certain limiting conditions. For instance, when M is very large, these variances become

$$\mathrm{Var}(d_X) = \frac{(n_X+1)}{3(n_X-1)}\theta + \frac{2(n_X^2+n_X+3)}{9n_X(n_X-1)}\theta^2, \tag{19}$$

$$\mathrm{Var}(d_{XY}) = \frac{(2n_Xn_Y+n_X+n_Y+2)}{6n_Xn_Y}\theta + \frac{(2n_Xn_Y+n_X+n_Y+5)}{9n_Xn_Y}\theta^2, \tag{20}$$

and

$$\mathrm{Var}(d) = \frac{(n_X+n_Y-1)(n_X+n_Y-2)}{6n_X(n_X-1)\,n_Y(n_Y-1)}\theta\left(1+\frac{5}{3}\theta\right), \tag{21}$$

which are the values expected in a single population of size $2N$. Thus, (19) is identical to the variance derived by Tajima (1983) and (21) is the same as the corresponding expression given by Takahata and Nei (1985) for the case of two isolated populations as the time of separation between them goes to zero.

The effect of sample size on the variances of d_X, d_{XY}, and d is also of interest. When $n_X=2$, $\mathrm{Var}(d_X)$ reduces to (17) and when $n_X=n_Y=1$, $\mathrm{Var}(d_{XY})$ reduces to (18). Further, when the samples sizes, n_X and n_Y, are very large, the three variances become

$$\mathrm{Var}_{\mathrm{st}}(d_X) = \frac{\theta}{6(3+6M+2M^2)}\left[(6+13M+4M^2)+\frac{\theta}{6M(1+M)(1+2M)}\right.$$
$$\left.\times(3+4M)(6+27M+43M^2+34M^3+8M^4)\right], \tag{22}$$

$$\mathrm{Var}_{\mathrm{st}}(d_{XY}) = \frac{\theta}{6M(3+6M+2M^2)}\Bigg[(9+18M+15M^2+4M^3) + \frac{\theta}{6M(1+M)^2(1+2M)}(27+162M+459M^2 + 780M^3+843M^4+562M^5+208M^6+32M^7)\Bigg], \quad (23)$$

and

$$\mathrm{Var}_{\mathrm{st}}(d) = \frac{\theta}{2M(3+6M+2M^2)}\Bigg[(3+M)+\frac{\theta}{6M(1+2M)} \times (9+33M+54M^2+16M^3)\Bigg]. \quad (24)$$

These are the stochastic variances of d_X, d_{XY}, and d, that arise from the random nature of the history of any sample, as opposed to their sampling variances (Nei and Tajima, 1981; Tajima, 1983). The sampling variances of d_X, d_{XY}, and d are given by $\mathrm{Var}_{\mathrm{s}} = \mathrm{Var} - \mathrm{Var}_{\mathrm{st}}$.

When M, n_X, and n_Y, are all very large,

$$\mathrm{Var}(d_X) = \mathrm{Var}(d_{XY}) = \tfrac{1}{3}\theta + \tfrac{2}{9}\theta^2 \quad (25)$$

and $\mathrm{Var}(d)$ approaches zero, in agreement with the results of Tajima (1983) and Takahata and Nei (1985). Alternatively, when M is very small, and keeping only terms of order $1/M$ or larger, (22), (23), and (24) becomes $\theta^2/(6M)$, $\theta/(2M)[1+\theta/(2M)]$, and $\theta/(2M)[1+\theta/(2M)-\theta/6]$, respectively. These can be compared to (17) and (18) and the expression (not shown) for $\mathrm{Var}(d)$ when just two sequences are sampled from each population. The conclusion reached is that, when the level of divergence between the two populations is great, increasing the sample size can decrease $\mathrm{Var}(d_X)$ by about a factor of three, but has little effect on $\mathrm{Var}(d_{XY})$ and $\mathrm{Var}(d)$. In contrast, when M is large, comparing (20) and (21) with (25) shows that the effects of sample size on $\mathrm{Var}(d_{XY})$ and $\mathrm{Var}(d)$ can be substantial. Tables I and II illustrate these concepts.

Table I gives values of s_{d_X}, the standard error of d_X—the square root of $\mathrm{Var}(d_X)$—over a broad range of values of both $E(d_X)$ and $E(d)$ and for three different values of the sample size, n_X. Recall that $E(d_X) = \theta$ measures divergence within a single population and that $E(d) = \theta/(2M)$ measures divergence between populations. Several interesting trends emerge. First, for a given value of $E(d_X)$, s_{d_X} increases with $E(d)$, that is, as the amount of divergence between populations increases. This was first apparent in (17) for $\mathrm{Var}(k_{ii})$. As the migration rate between the two populations decreases,

TABLE I

The Standard Error, s_{d_X}, of d_X

		n_X		
$E(d_X)$	$E(d)$	2	10	1000
100.0	100.0	141.77	79.71	70.93
	10.0	105.36	55.84	49.81
	1.0	101.00	53.47	47.77
	0.1	100.55	53.23	47.56
	0.01	100.50	53.21	47.54
10.0	100.0	33.32	20.77	18.75
	10.0	14.49	8.20	7.31
	1.0	10.95	5.91	5.28
	0.1	10.54	5.68	5.08
	0.01	10.49	5.66	5.06
1.0	100.0	10.10	6.43	5.82
	10.0	3.46	2.16	1.95
	1.0	1.73	1.02	0.92
	0.1	1.45	0.85	0.77
	0.01	1.42	0.83	0.75
0.1	100.0	3.18	2.03	1.84
	10.0	1.05	0.67	0.61
	1.0	0.46	0.29	0.26
	0.1	0.35	0.22	0.20
	0.01	0.33	0.21	0.19

the variance of intrapopulation pairwise differences increases. Second, increasing the sample size, n_X, does decrease s_{d_X}, and the magnitude of this effect is similar over all values of $E(d)$. However, relatively little accuracy is gained by increasing n_X from 10 to 1000. Third, when $E(d_X)$ is large, s_{d_X} can be quite a bit smaller than $E(d_X)$ but when $E(d_X)$ is small, s_{d_X} is always larger than its expectation.

Table II gives values of s_d and $s_{d_{XY}}$, the standard errors of d and d_{XY}, over the same range of parameter values used in Table I. The values of $s_{d_{XY}}$ are given in parentheses after corresponding values of s_d. The expectation of d_{XY} is simply the sum of $E(d_X)$ and $E(d)$. Similarly to s_{d_X}, for a given value of $E(d_X)$, s_d and $s_{d_{XY}}$ both increase with increasing $E(d)$. That the same thing is not always true as $E(d_X)$ increases for a give value of $E(d)$ is a consequence of the fact that $E(d) = \theta/(2M)$; holding $E(d)$ constant and changing $\theta = E(d_X)$ means also changing M. Increasing the sample size does not have much of an effect on $s_{d_{XY}}$ but does greatly reduce s_d, except when $E(d)$ is greater than or equal to $E(d_X)$. Unlike s_{d_X} and $s_{d_{XY}}$, s_d continues to decrease with increasing n_X and n_Y even when a great number of

TABLE II

The Standard Errors, s_d and ($s_{d_{XY}}$), of d and d_{XY}

		$n_X = n_Y$		
$E(d_X)$	$E(d)$	2	10	1000
100.0	100.0	161.71(147.22)	112.64(127.45)	102.89(122.85)
	10.0	74.52(75.62)	21.63(56.32)	11.48(52.87)
	1.0	65.71(69.77)	11.90(50.93)	1.25(48.00)
	0.1	64.84(69.20)	10.97(50.42)	0.22(47.56)
	0.01	64.75(69.15)	10.87(50.37)	0.12(47.52)
10.0	100.0	105.75(103.37)	100.92(101.38)	99.92(100.89)
	10.0	16.48(15.18)	11.51(13.17)	10.52(12.70)
	1.0	7.64(7.95)	2.22(5.97)	1.19(5.61)
	0.1	6.74(7.35)	1.22(5.41)	0.13(5.11)
	0.01	6.66(7.29)	1.13(5.36)	0.02(5.06)
1.0	100.0	101.00(100.75)	100.52(100.55)	100.42(100.50)
	10.0	10.99(10.78)	10.50(10.58)	10.40(10.53)
	1.0	1.93(1.92)	1.37(1.68)	1.26(1.62)
	0.1	0.93(1.11)	0.28(0.87)	0.15(0.82)
	0.01	0.83(1.04)	0.15(0.79)	0.02(0.75)
0.1	100.0	110.55(100.52)	100.50(110.50)	100.49(100.50)
	10.0	10.54(10.51)	10.49(10.49)	10.48(10.49)
	1.0	1.45(1.45)	1.39(1.42)	1.38(1.42)
	0.1	0.37(0.42)	0.27(0.37)	0.25(0.36)
	0.01	0.19(0.27)	0.06(0.22)	0.03(0.20)

sequences have already been sampled. However, s_d is always greater than $E(d)$. Thus, even when the sampling variance is reduced to near zero, the standard error of d is still very large. Lastly, $s_{d_{XY}}$ is generally smaller than $E(d_{XY})$, except when $E(d_X)$ is small.

3.1. *An Example Using Mitochondrial DNA*

Edwards (1993) presented sequence data from the control region of mitochondrial DNA of the grey-crowned babbler (*Pomatostomus temporalis*) from 12 different populations in two subspecies and made estimates of gene flow between seven pairs of populations. Table III shows $(\hat{d}_X + \hat{d}_Y)/2$, $\hat{d}_{XY}$, and $\hat{d}$ and their associated standard errors for these seven population pairs. Also shown are values for several additional pairs of populations between which Edwards (1993) found no evidence of gene flow using the genealogical method of Slatkin and Maddison (1990). The standard errors shown in Table III were obtained by simply substituting values of $(\hat{d}_X + \hat{d}_Y)/2$ and $(\hat{d}_X + \hat{d}_Y)/(4\hat{d})$, which estimate θ and M, respectively, into the expressions

TABLE III

Values of $(\hat{d}_X + \hat{d}_Y)/2$, $\hat{d}_{XY}$, and $\hat{d}$ and Their Standard Errors Calculated from Edward's (1993) Babbler mtDNA Data

Population pairs	$(\hat{d}_X + \hat{d}_Y)/2$	$\hat{d}_{XY}$	$\hat{d}$
P. t. temporalis			
A–B	10.4(8.1)	13.2(7.4)	2.8(4.2, 3.2)
A–M	12.0(9.7)	20.0(12.1)	8.0(9.4, 8.6)
B–M	10.6(8.2)	13.9(7.8)	3.3(4.7, 3.7)
M–O	7.8(9.3)	30.1(23.8)	22.2(23.0, 22.5)
A–N	9.6(11.2)	34.7(27.1)	25.1(26.2, 25.4)
B–N	8.3(11.0)	37.0(30.4)	28.8(29.8, 29.0)
M–N	9.8(11.7)	38.3(30.4)	28.5(29.4, 28.8)
O–N	5.5(6.0)	17.5(13.4)	12.1(12.7, 12.4)
P. t. rubeculus			
G–H	5.8(4.5)	6.0(3.3)	0.20(0.74, 0.25)
D–F	3.4(3.5)	8.6(6.3)	5.2(5.8, 5.5)
F–K	6.8(5.6)	11.3(7.0)	4.5(5.3, 4.9)
D–E	3.7(4.2)	11.7(9.1)	8.0(8.6, 8.3)
G–I	4.5(4.1)	8.0(5.3)	3.5(4.4, 3.9)
E–I	3.8(4.5)	12.5(9.9)	8.7(9.5, 9.0)

Note. Standard errors of $(\hat{d}_X + \hat{d}_Y)/2$, $\hat{d}_{XY}$, and $\hat{d}$ are given in parentheses following each value. For $\hat{d}$, the second value in the parentheses is the standard error expected when $n_X = n_Y = 1000$.

for the variances derived here and taking the square root. The variance of $(d_X + d_Y)/2$ is given by $[\mathrm{Var}(d_X) + \mathrm{Var}(d_Y) + 2\,\mathrm{Cov}(d_X, d_Y)]/4$.

As the assumption of infinite sites was made throughout this work, the values of $\hat{d}_X$, $\hat{d}_{XY}$, and $\hat{d}$ shown in Table III are based on corrected distances. These were obtained using the method of Tamura and Nei (1993). Their method requires an estimate of the gamma distribution parameter, a, which quantifies the extent of rate variation among sites in the sequence. A value of $a = 0.19$ was obtained here from the entire data set by fitting a negative binomial distribution to the distribution of the inferred number of changes at each site (Uzzell and Corbin, 1971). The results below do not depend on the use of this or any other currently available distance correction. While the observed differences are, of course, smaller than the corrected values, the patterns shown in Table III are identical for both.

As expected, the standard errors of $\hat{d}$, which estimates $\theta/(2M)$, are quite large for Edward's (1993) data, greater than $\hat{d}$ for every population pair in Table III. While sample sizes for these populations range from 6 for population I to 20 for population K with a mean of about 14, increasing these to one thousand sequences per population would not affect this result. If this were done, the resulting standard errors of $\hat{d}$, which are shown in

Table III, would still be larger than $\hat{d}$ for every pair. Thus, as evident from Table II, much of the variance of $\hat{d}$ is due to stochastic factors rather than sampling. On the other hand, the standard errors of $(\hat{d}_X + \hat{d}_Y)/2$, and $\hat{d}_{XY}$ are smaller in comparison to the estimates of those parameters, indicating that θ and $\theta + \theta/(2M)$, at least, can be estimated with some confidence using these data. Of course, $\hat{d}$, as a measure of the extent of divergence between populations, is the parameter of interest here, so there results might be somewhat discouraging. Estimates of divergence between populations based on pairwise nucleotide differences at a single locus are not very accurate.

3.2. *Comparing Variances under Migration and Isolation*

While, perhaps, discouraging in one sense, the great magnitude of the variances of d_X, d_{XY}, and d for two populations with migration actually offers hope of distinguishing two important demographic scenarios. Takahata and Nei (1985) give expressions for $\mathrm{Var}(d_X)$, $\mathrm{Var}(d_{XY})$, and $\mathrm{Var}(d)$ for the case of two populations that have been completely isolated from each other for some length of time, T, measured in units of $2N$ generations. In this case, $E(d_X) = \theta$, $E(d_{XY}) = \theta + \theta T$, and $E(d) = \theta T$ (Kimura, 1969; Watterson, 1975; Li, 1977; Gillespie and Langley, 1979; Nei and Tajima, 1981; Takahata and Nei, 1985). Thus, when $T = 1/(2M)$, the expectations of all four pairwise difference measures are exactly the same under migration as under isolation, making them useless in distinguishing these two situations. In addition, Slatkin and Maddison (1989) and Takahata and Slatkin (1990) conclude that gene genealogies will also not serve to distinguish migration from isolation (but see Slatkin and Maddison, 1990).

Takahata and Nei (1985) showed that $\mathrm{Var}(d_X)$ in the isolation case is equal to the variance for a sample from a single, randomly mating population given in (19) and first derived by Tajima (1983). The expressions for $\mathrm{Var}(d_{XY})$ and $\mathrm{Var}(d)$ for the case of two isolated populations are not given here but are slightly different from those derived by Takahata and Nei (1985); some minor errors were found. Specifically, Takahata and Nei's (1985) F, $S2$, and $S1$ should be

$$F = \frac{1 - (1 + T)\, e^{-T}}{2(1 - e^{-T})}, \qquad S2 = \frac{1 - (1 + 3T)\, e^{-3T}}{3(1 - e^{-3T})},$$

and

$$S1 = \frac{2(1 - e^{-3T}) - 3Te^{-T}(1 + e^{-T})}{(1 - e^{-T})^2\,(2 + e^{-T})(1 + e^{-T} + e^{-2T})},$$

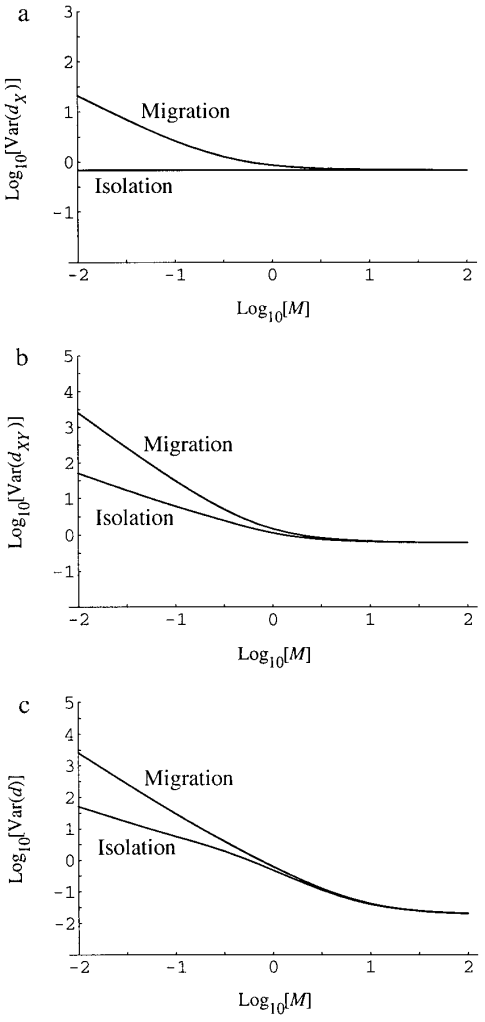


FIG. 2. Comparison of $\mathrm{Var}(d_X)$, graph (a), $\mathrm{Var}(d_{XY})$, graph (b), and $\mathrm{Var}(d)$, graph (c), under migration and under isolation. In all three graphs, θ is equal to one. Along the horizontal axes, $T = 1/(2M)$ so that the expectations of d_X, d_{XY}, and d are identical for the two models. As indicated, the upper curves describe the variances under migration and the lower curves describe the variances under isolation.

and I have used these corrected values. This changes the values of $\mathrm{Var}(d_{XY})$ and $\mathrm{Var}(d)$ only slightly and does not affect the conclusions of Takahata and Nei (1985).

Figure 2 plots $\mathrm{Var}(d_X)$, $\mathrm{Var}(d_{XY})$, and $\mathrm{Var}(d)$, under migration and under isolation when θ equals one. The equivalence of the expectations of d_X, d_{XY}, and d in these two cases when $T = 1/(2M)$ provides a convenient basis for comparing the variances, which are plotted as functions of $\log_{10}(M) = \log_{10}(1/2T)$). All three variances are generally greater under migration than under isolation. As the migration rate decreases or the time of separation increases, that is, as $M = 1/(2T)$ gets smaller, the variances of all three statistics become much greater under migration than under isolation. This dependence on $M = 1/(2T)$ is similar over a broad range of values of θ. However, when θ is smaller, smaller values of $M = 1/(2T)$ are needed for the differences between the variances to become apparent. The variance of d_X does not depend on T in the isolation case (Takahata and Nei, 1985). As $M = 1/(2T)$ increases, $\mathrm{Var}(d_X)$ in the migration case, and $\mathrm{Var}(d_{XY})$ and $\mathrm{Var}(d)$ under both migration and isolation approach the limiting values given in (19), (20), (21), respectively.

4. Discussion

Like the variances derived by Tajima (1983) and Takahata and Nei (1985), the expressions presented here for $\mathrm{Var}(d_X)$, $\mathrm{Var}(d_{XY})$, and $\mathrm{Var}(d)$ for a sample from two populations with migration include components due to both stochastic and sampling factors (Nei and Tajima, 1981). Lynch and Crease (1990) studied the partitioning of the sampling variances of d_X, d_{XY}, and d in subdivided populations into components attributable to various sources. These sampling variances can all be reduced to zero by simply increasing the sample sizes, n_X and n_Y. Stochastic variance, on the other hand, is the variance over all possible evolutionary histories and is unaffected by changes in sample size. The results presented here show that the stochastic components of $\mathrm{Var}(d_X)$, $\mathrm{Var}(d_{XY})$ and $\mathrm{Var}(d)$ for two populations with migration are so large that estimates of population genetic parameters using these measures are not very accurate.

The only way to decrease this stochastic component of the variance is to sample more loci. If we had sequence data for n loci, an improved estimate of $\theta/(2M)$ would be $\bar{d} = \sum_{i=1}^{n} d_i/n$. If the value of θ is the same for every locus and the histories of the loci independent, the variance of $\bar{d}$ would be $\mathrm{Var}(d_i)/n$. Thus, stochastic and sampling variances are defined relative to the sampling scheme. By adding another axis to our sampling scheme, former stochastic factors come under the realm of sampling. Or course, the assumptions of equal θ values and independence among loci will often be

violated. Thus, before the importance of multilocus sampling in quantifying genetic and demographic parameters can be properly assessed, appropriate multilocus models must be developed.

The results presented here demonstrate that the variances of pairwise nucleotide differences are generally greater under migration than under isolation and much greater as M or $1/(2T)$ decreases. Under isolation, interpopulation coalescent events can occur only prior to the time the populations were first separated. This restricts the range of possible coalescence times. For large T, it is unlikely that more than one common ancestor of the sequences sampled from each population will exist at the time of separation (Takahata and Nei, 1985; Takahata, 1989). In contrast, under migration, it is possible for a common ancestor of two sequences from different populations to exist in the very recent past, even when the migration rate is low. Since interpopulation coalescent events occur over a much broader range, the variance is larger.

The great difference between the variances under migration and under isolation suggests that a test to distinguish these two demographic situations could be developed. Because $\mathrm{Var}(d_X)$, $\mathrm{Var}(d_{XY})$ and $\mathrm{Var}(d)$ as derived here include variation over all possible histories, this would require data from multiple loci. Figure 2 implies that our ability to distinguish migration from isolation will be poor when the populations are not substantially diverged, but it might be quite good when the migration rate is low, or equivalently, when the time of separation is great. However, in constructing such a test, knowledge of the covariances between the expectations and the variances of pairwise nucleotide differences, as well as among the variances, is likely to be important. Thus, while the present results indicate that a test is possible, much work remains before one might be implemented effectively.

Appendix A

$\mathrm{Var}(d_X)$ for a sample of n_X sequences from population X. With the assumptions outlined in the body of the paper, the expression for $\mathrm{Var}(d_Y)$ is obtained simply by substituting n_Y for n_X:

$$\mathrm{Var}(d_X) = \frac{\theta}{6n_X(n_X-1)(3+6M+2M^2)}\Bigg[\, 6M + n_X(6+7M+4M^2)$$

$$+ n_X^2(6+13M+4M^2) + \frac{\theta}{6M(1+M)(1+2M)}\{6M(1+M)$$

$$\times (15 + 54M + 64M^2 + 16M^3) + n_X(18 + 123M + 327M^2 + 358M^3 + 160M^4 + 32M^5) + n_X^2(3 + 4M) \times (6 + 27M + 43M^2 + 34M^3 + 8M^4\} \Bigg].$$

$\mathrm{Var}(d_{XY})$ for a sample of n_X sequences from population X and n_Y sequences from population Y:

$$\mathrm{Var}(d_X) = \frac{\theta}{6n_X n_Y M(3 + 6M + 2M^2)} \Bigg[M^2(9 + 4M) + (n_X + n_Y) \times M(3 + M)(3 + 2M) + n_X n_Y(9 + 18M + 15M^2 + 4M^3) + \frac{\theta}{6M(1 + M)^2(1 + 2M)} \{ M^2(36 + 222M + 573M^2 + 706M^3 + 400M^4 + 80M^5) + (n_X + n_Y) M(27 + 153M + 390M^2 + 525M^3 + 374M^4 + 128M^5 + 16M^6) + n_X n_Y(27 + 162M + 459M^2 + 780M^3 + 843M^4 + 562M^5 + 208M^6 + 32M^7)\} \Bigg].$$

$\mathrm{Cov}(d_X, d_Y)$ and $\mathrm{Cov}(d_{XY}, d_X)$ as needed to calculate $\mathrm{Var}(d)$:

$$\mathrm{Cov}(d_X, d_Y) = \frac{\theta}{6(3 + 6M + 2M^2)} \Bigg[M(9 + 4M) + \frac{\theta}{6(1 + M)^2(1 + 2M)} \times (3 + 2M)(3 + 18M + 53M^2 + 56M^3 + 16M^4) \Bigg]$$

$$\mathrm{Cov}(d_{XY}, d_X) = \frac{\theta}{6n_X(3 + 6M + 2M^2)} \Bigg[2M(5 + 2M) + n_X(1 + M) \times (9 + 4M) + \frac{\theta}{6M(1 + M)(1 + 2M)} \times \{16M^3(4 + 7M + 2M^2) - 6(3 + 11M + 11M^2) + n_X(3 + 2M)(3 + 18M + 53M^2 + 56M^3 + 16M^4)\} \Bigg].$$

Acknowledgments

I am very grateful to Montgomery Slatkin for the guidance and encouragement he provided throughout this work. I am also thankful to Sarah Otto for detailed comments on earlier versions of the manuscript, to Naoyuki Takahata for helpful discussions of Takahata and Nei (1985), and to Scott Edwards for graciously supplying the babbler mtDNA data. This work was supported by NIH Grant GM40282 to M. Slatkin and the NIH Post-Graduate Training Program in Genetics Grant GM07127 at U.C. Berkeley.

References

Berry, A., and Kreitman, M. 1993. *Genetics* **134**, 869–893.
Edwards, S. V. 1993. *Proc. R. Soc. Lond. B* **252**, 177–185.
Ewens, W. J. 1979. "Mathematical Population Genetics," Springer-Verlag, Berlin.
Gillespie, J. H., and Langley, C. H. 1979. *J. Mol. Evol.* **13**, 27–34.
Griffiths, R. C. 1981. *J. Math. Biol.* **12**, 251–261.
Hey, J. 1991. *Theor. Popul. Biol.* **39**, 30–48.
Hudson, R. R. 1983. *Evolution* **37**, 203–217.
Hudson, R. R. 1990. *in* "Oxford Surveys in Evolutionary Biology" (D. J. Futuyma and J. Antonovics, Eds.), Vol. 7, Oxford Univ. Press, Oxford.
Kimura, M. 1969. *Genetics* **61**, 893–903.
Kimura, M., and Weis, G. H. 1964. *Genetics* **49**, 561–576.
Kingman, J. F. C. 1982a. *Stochastic Process. Appl.* **13**, 235–248.
Kingman, J. F. C. 1982b. *J. Appl. Probab. A* **19**, 27–43.
Li, W.-H. 1976. *Theor. Popul. Biol.* **10**, 303–308.
Li, W.-H. 1977. *Genetics* **85**, 331–337.
Lynch, M., and Crease, T. J. 1990. *Mol. Biol. Evol.* **7**, 377–394.
Maruyama, T. 1970. *Theor. Popul. Biol.* **1**, 273–306.
Nei, M., and Feldman, M. W. 1972. *Theor. Popul. Biol.* **3**, 460–465.
Nei, M., and Li, W.-H. 1979. *Proc. Nat. Acad. Sci.* **76**, 5269–5273.
Nei, M., and Tajima, F. 1981. *Genetics* **97**, 145–163.
Notohara, M. 1990. *J. Math. Biol.* **29**, 59–75.
Simmons, G. M., Kreitman, M. E., Quattlebaum, W. F., and Miyashita, N. (1989). *Evolution* **43**, 393–409.
Slatkin, M. 1987. *Theor. Popul. Biol.* **32**, 42–49.
Slatkin, M., and Maddison, W. P. 1989. *Genetics* **123**, 603–613.
Slatkin, M., and Maddison, W. P. 1990. *Genetics* **126**, 249–260.
Strobeck, C. 1987. *Genetics* **117**, 149–153.
Tajima, F. 1983. *Genetics* **105**, 437–460.
Takahata, N. 1983. *Genetics* **104**, 497–512.
Takahata, N. 1989. *Genetics* **122**, 957–966.
Takahata, N., and Nei, M. 1985. *Genetics* **110**, 325–344.
Takahata, N., and Slatkin, M. 1990. *Theor. Popul. Biol.* **38**, 331–350.
Tamura, K., and Nei, M. 1993. *Mol. Biol. Evol.* **10**, 512–526.
Uzzell, T., and Corbin, K. W. 1971. *Science* **172**, 1089–1096.
Wakeley, J. 1994. "Substitution Rate Variation among Sites and the Variance of Pairwise Nucleotide Differences." Ph.D. thesis, University of California, Berkeley.
Watterson, G. A. 1975. *Theor. Popul. Biol.* **7**, 256–276.