Estimating Ancestral Population Parameters

John Wakeley and Jody Hey

Department of Biological Sciences, Rutgers University, Piscataway, New Jersey 08855-1059

Manuscript received September 2, 1996
Accepted for publication December 6, 1996

ABSTRACT

The expected numbers of different categories of polymorphic sites are derived for two related models of population history: the isolation model, in which an ancestral population splits into two descendents, and the size-change model, in which a single population undergoes an instantaneous change in size. For the isolation model, the observed numbers of shared, fixed, and exclusive polymorphic sites are used to estimate the relative sizes of the three populations, ancestral plus two descendent, as well as the time of the split. For the size-change model, the numbers of sites segregating at particular frequencies in the sample are used to estimate the relative sizes of the ancestral and descendent populations plus the time the change took place. Parameters are estimated by choosing values that most closely equate expectations with observations. Computer simulations show that current and historical population parameters can be estimated accurately. The methods are applied to DNA data from two species of Drosophila and to some human mitochondrial DNA sequences.

HISTORICAL events such as the formation of two species from a common ancestor or drastic changes in population size manifest themselves in the DNA of organisms by structuring the genealogies of nucleotide sites. Consider a situation where a single ancestral population splits into two descendent populations, and after the split no genetic exchange occurs between the two. Figure 1 depicts this isolation model and shows examples of two possible genealogical histories of a site in the sample. The branches in Figure 1 represent ancestral lineages of the sampled sequences. If the per-site mutation rate is small, which we will assume is true, then each branch presents opportunities for the creation of a particular kind of polymorphic site. When a mutation has occurred on one of these ancestral lineages, it appears as a polymorphic site that divides the sample into two groups: one that shows the ancestral nucleotide and one that shows the new, mutant nucleotide.

For example, a mutation on the long internal branch of the genealogy in Figure 1a will divide the sample into two groups that correspond exactly to the two population samples. This type of polymorphism is commonly referred to as a fixed difference (HEY 1991). In contrast, the genealogy in Figure 1b does not allow for the possibility of a fixed difference because there is no branch that divides the sample appropriately. Instead, a mutation on the smallest internal branch of that genealogy yields a different type of polymorphism: one that is shared by both populations. A mutation on any other branch than these two in either Figure 1a or b will produce a site that is polymorphic in only one of the two population samples. Thus, Figure 1 illustrates the relationship between population history and classes of segregating sites, as mediated through sites' genealogies. If the time of separation of the two descendent populations is short, then genealogies will likely resemble the one in Figure 1b and shared polymorphisms may appear in the data. If the time of separation is long, the most probable genealogies will, like Figure 1a, have an internal branch on which fixed differences can accumulate.

For any given time of separation, every possible genealogy will have an associated probability. However, in the absence of recombination, all sites in a particular sample will share the same genealogy. The particular one observed will be a single draw from the universe of possibilities. As single observations, individual genealogies are not likely to contain enough information to make accurate and general statements about population-level processes. On the other hand, if there is recombination or if multiple loci are sampled, then different sites may have different genealogical histories. In the case of a sample from two populations, some sites' genealogies may resemble the one in Figure 1a and others may be like the one in Figure 1b. In large data sets, many of the possible genealogies will be realized in the histories of sites in the sample and will be represented in proportion to the relative likelihood of observing each.

A similar picture can be drawn of the size-change model, which is like the isolation model but with only a single descendent population. Here, polymorphic sites can be partitioned according to the frequencies of mutant and nonmutant bases, as, for example, TAJIMA (1989b) and FU and LI (1993) have done. Again the

Corresponding author: John Wakeley, Nelson Biological Labs, P.O. Box 1059, Rutgers University, Busch Campus, Piscataway, NJ 08855-1059. E-mail: jwakeley@rci.rutgers.edu

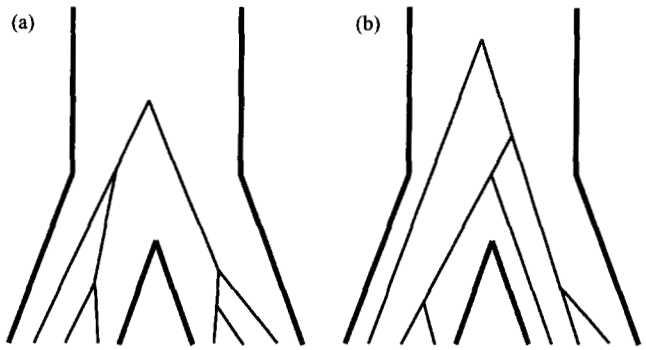

FIGURE 1.—Two possible genealogies of a sample of three sequences from each of two isolated populations. Thick lines represent population boundaries, and thin lines trace the ancestral lineages up into the past.

numbers of each kind of polymorphic site observed in a sample of DNA will depend on the genealogies of sites, which, in turn, depend on the time and magnitude of the change in population size. For instance, if the population has recently grown in size, sites' genealogies will tend to have longer terminal and shorter internal branches relative to the genealogy of a sample from a constant-sized population (SLATKIN and HUDSON 1991). This will result in an excess of sites where the mutant base is in frequency $1/n$ in a sample of n sequences and a dearth of middle-frequency polymorphic sites.

We adopt general and easily interpreted versions of the isolation and size-change models. Before generation t in the past, there was a single, panmictic population of size N_A. Exactly at t, the ancestral population either split into two descendent populations (isolation model) or simply changed size (size-change model). The descendent populations are also panmictic, but in the isolation model there is no gene flow between them. The sizes of the descendent populations are N_1 and N_2 (isolation) or just N_1 (size-change), and no restrictions are put on the relative sizes of N_1, N_2, and N_A. All populations conform to the commonly used Wright-Fisher model (FISHER 1930; WRIGHT 1931). Generations are nonoverlapping and N_1, N_2, and N_A remain constant over time except at t, where there might be a change in population size. All variation is assumed to be neutral and mutations occur according to the infinite sites model with mutation rate u per sequence per generation. Four parameters, then, describe the isolation model: $\theta_1 = 4N_1u$, $\theta_2 = 4N_2u$, $\theta_A = 4N_Au$, and $\tau = 2ut$. The size-change model is characterized by three parameters: θ_1, θ_A and τ.

The isolation and size-change models form the basis of many current studies in population genetics. The isolation model has been considered both as a null model of species formation (HEY 1994) and as a model of the divergence of populations (TAKAHATA and NEI 1985). As used here, it involves four parameters and thus represents a generalization of past implementations, *e.g.*, those of TAKAHATA and NEI (1985) and HUDSON *et al.* (1987). The size-change model has been applied to recent human evolution (ROGERS and HARPENDING 1992). However, a model of exponential growth has also been suggested (SLATKIN and HUDSON 1991) and may be more realistic than the instantaneous size-change model. Clearly, both the isolation and size-change models are simple models. Whether or not they are too simple to describe the history of most populations and species is an empirical question that deserves attention.

Our purpose here is to show how we can glean more information from DNA data to estimate both current and historical population parameters. This is a starting point, from which other questions might spring and be addressed, and to which other factors, such as migration and selection, might be added. We begin by deriving the expected values of the various partitions of polymorphic sites. These then form the basis of a method of estimating the parameters of the isolation and size-change models.

THEORY AND METHODS

The segregating sites in a sample of sequences from two populations can be partitioned into four mutually exclusive categories that correspond to different aspects of genealogical history. The first comprises sites that are polymorphic in population 1, but monomorphic in population 2. Next are sites that are polymorphic in population 2, but monomorphic in population 1. Call the numbers of each of these types of exclusive polymorphic sites S_{X1} and S_{X2}. The third are sites at which a polymorphism is shared across population boundaries, *i.e.*, where the same two bases appear in both populations' samples. Let the number of shared polymorphic sites be called S_S. Fourth, there are sites showing fixed differences between the two populations, the number of which are referred to as S_f.

Segregating sites can be also classified as polymorphic in one population, either 1 or 2, regardless as to whether they are polymorphic in the other population. Call the numbers of polymorphic sites counted in this manner S_1 and S_2. Finally, we can simply count the total number of polymorphic sites in the entire sample, and the number of these is referred to as S. These different categories of sites are related in the following way:

$$S_1 = S_{X1} + S_S, \tag{1}$$

$$S_2 = S_{X2} + S_S, \tag{2}$$

$$S = S_{X1} + S_{X2} + S_S + S_f. \tag{3}$$

Single-population expectations: To calculate the expectations of S_{X1}, S_{X2}, S_f and S_S, we use (1)–(3) and start with the simplest case. Consider two samples of sequences taken from a single, randomly mating, diploid population of effective size N. Let the numbers of sequences sampled be n_1 and n_2. WATTERSON (1975) showed that

$$E(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i}, \tag{4}$$

where $\theta = 4Nu$ and $n = n_1 + n_2$. In fact, Equation 4 applies to any randomly taken sample, so that

$$E(S_1) = \theta \sum_{i=1}^{n_1-1} \frac{1}{i} \quad \text{and} \quad E(S_2) = \theta \sum_{i=1}^{n_2-1} \frac{1}{i}. \tag{5}$$

Then, only one more quantity is required in order to know the expectations of all four mutually exclusive partitions of segregating sites in a single population. The expectation of S_f can be derived by considering the number of sites that divide the sample into n_1 and n_2 sequences. The expected number of these is $\theta(1/n_1 + 1/n_2)/\delta$, where δ is 2, if $n_1 = n_2$ and 1 otherwise (TAJIMA 1989b; FU and LI 1993; FU 1995). The chance that these n_1 and n_2 bases are distributed among the two subsamples as a fixed difference is related to the hypergeometric distribution and is just $\delta/\binom{n}{n_1}$. Thus,

$$E(S_f) = \frac{\theta\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\binom{n}{n_1}} \tag{6}$$

is the expected number of fixed differences in a sample of $n = n_1 + n_2$ sequences from a single, randomly mating population.

Then, using (1)–(3),

$$E(S_{X1}) = \theta\left[\sum_{i=1}^{n-1}\frac{1}{i} - \sum_{i=1}^{n_2-1}\frac{1}{i} - \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\binom{n}{n_1}}\right], \tag{7}$$

$$E(S_{X2}) = \theta\left[\sum_{i=1}^{n-1}\frac{1}{i} - \sum_{i=1}^{n_1-1}\frac{1}{i} - \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\binom{n}{n_1}}\right], \tag{8}$$

$$E(S_S) = \theta\left[\sum_{i=1}^{n_1-1}\frac{1}{i} + \sum_{i=1}^{n_2-1}\frac{1}{i} - \sum_{i=1}^{n-1}\frac{1}{i} + \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\binom{n}{n_1}}\right] \tag{9}$$

are the expected numbers of polymorphisms exclusive to populations 1 and 2, and of polymorphisms shared between 1 and 2.

Two isolated populations: Under the isolation model, Equations 6–9 give the expectations of the four mutually exclusive partitions of segregating sites in the ancestral population. However, the numbers of distinct ancestors of the presently sampled n_1 and n_2 sequences, which existed at the time the two populations separated, are unknown. These numbers, called n_1' and n_2', must be considered random quantities that follow some probability distribution. TAKAHATA and NEI (1985) derived the following expression for the probability that, at generation t in the past, there are n_1' ancestors of n_1 sequences sampled at the present:

$$P_{n_1 n_1'}(T_1) = 1 - \sum_{i=2}^{n_1} \frac{e^{-\binom{i}{2}T_1}}{\prod_{j=2, j\neq i}^{n_1}\left[1 - \frac{i(i-1)}{j(j-1)}\right]}, \quad \text{if } n_1' = 1 \tag{10}$$

$$P_{n_1 n_1'}(T_1) = \frac{1}{\binom{n_1'}{2}} \sum_{i=n_1'}^{n_1} \frac{\binom{i}{2}e^{-\binom{i}{2}T_1}}{\prod_{j=n_1', j\neq i}^{n_1}\left[1 - \frac{i(i-1)}{j(j-1)}\right]}, \quad \text{if } 2 \le n_1' \le n_1. \tag{11}$$

In (10) and (11), T_1, which is equivalent to τ/θ_1, is the time of separation measured in units of $2N_1$ generations. The equations for population 2 differ from these only by a change of subscripts. Thus, the probability that the ancestors of the presently sampled n_1 and n_2 sequences numbered n_1' and n_2' at generation t is equal to $P_{n_1 n_1'}(\tau/\theta_1)P_{n_2 n_2'}(\tau/\theta_2)$.

The expectations of S_{X1}, S_{X2}, S_S and S_f are derived by considering every possible ancestral sample at time t and weighting by the probability of each. This is most clearly seen for shared polymorphic sites because these can result only from mutations that occurred before the time of separation of the populations. The average of (9) is taken over all possible relevant ancestral sample sizes:

$$E(S_S) = \theta_A \sum_{n_1'=2}^{n_1}\sum_{n_2'=2}^{n_2} P_{n_1 n_1'}(\tau/\theta_1)P_{n_2 n_2'}(\tau/\theta_2) \times \left[\sum_{i=1}^{n'-1}\frac{1}{i} + \sum_{i=1}^{n_2'-1}\frac{1}{i} - \sum_{i=1}^{n'-1}\frac{1}{i} + \frac{\left(\frac{1}{n_1'} + \frac{1}{n_2'}\right)}{\binom{n'}{n_1'}}\right], \tag{12}$$

where $n' = n_1' + n_2'$.

Every mutation that occurs in either of the descendent populations, given that $n_1' > 1$ or $n_2' > 1$, appears as an exclusive polymorphism in the data. The expected number of these in population 1 is simply

$$E(S_{X1}\text{ after }t) = \theta_1\left[\sum_{i=1}^{n_1-1}\frac{1}{i} - \sum_{n_1'=2}^{n_1} P_{n_1 n_1'}(\tau/\theta_1)\sum_{i=1}^{n_1'-1}\frac{1}{i}\right]. \tag{13}$$

In words, $E(S_{X1}\text{ after }t)$ is equal to the expected number of segregating sites in a sample of n_1 sequences, regardless of time, minus the expected number of these that would have occurred before time t in the past. Equation 7 helps in deriving the expectation of S_{X1} before t:

$$E(S_{X1}\text{ before }t) = \theta_A \sum_{n_1'=2}^{n_1}\sum_{n_2'=1}^{n_2} P_{n_1 n_1'}(\tau/\theta_1)P_{n_2 n_2'}(\tau/\theta_2) \times \left[\sum_{i=1}^{n'-1}\frac{1}{i} - \sum_{i=1}^{n_2'-1}\frac{1}{i} - \frac{\left(\frac{1}{n_1'} + \frac{1}{n_2'}\right)}{\binom{n'}{n_1'}}\right]. \tag{14}$$

Note that in (14) when $n_2' = 1$ the middle term in the brackets is defined to be equal to zero. Again, the equation that applies to S_{X2} is obtained simply by switching subscripts. Of course, $E(S_{X1}) = E(S_{X1}\text{ before }t) + E(S_{X1}\text{ after }t)$ and similarly for $E(S_{X2})$, but these full equations are not reproduced here in the interest of space.

Like exclusive polymorphisms, fixed differences can result from mutations that occurred before the two populations separated as well as those that occurred after the split. $E(S_f\text{ before }t)$ is calculated similarly to (12) and (14) above,

$$E(S_f\text{ before }t) = \theta_A \sum_{n_1'=1}^{n_1}\sum_{n_2'=1}^{n_2} P_{n_1 n_1'}(\tau/\theta_1) \times P_{n_2 n_2'}(\tau/\theta_2)\frac{\left(\frac{1}{n_1'} + \frac{1}{n_2'}\right)}{\binom{n'}{n_1'}}. \tag{15}$$

In addition, if there was only a single common ancestral sequence of either population sample at the time the two separated, then fixed differences might have accumulated after the split. In Figure 1a, this is true for one population, but not the other. HEY (1991) calculated the expected length of time during which such fixed differences might have accumulated. Considering both populations, and in the notation used here, the expected number is given by

$$E(S_f \text{ after } t) = \tau - \frac{\theta_1}{2}\sum_{i=2}^{n_1}\frac{1-e^{-\binom{i}{2}\tau/\theta_1}}{\binom{i}{2}\prod_{j=2,j\neq i}^{n_1}\left[1-\frac{i(i-1)}{j(j-1)}\right]} - \frac{\theta_2}{2}\sum_{i=2}^{n_2}\frac{1-e^{-\binom{i}{2}\tau/\theta_2}}{\binom{i}{2}\prod_{j=2,j\neq i}^{n_2}\left[1-\frac{i(i-1)}{j(j-1)}\right]}, \quad (16)$$

and, again, $E(S_f) = E(S_f \text{ before } t) + E(S_f \text{ after } t)$.

Site frequencies: Let $z_{1,i}$ be the number of polymorphic sites at which the mutant base is found in i copies in the sample of n_1 sequences from population 1. Likewise, $z_{2,i}$ represents mutations of size i in the sample from population 2. Using the same sort of approach, it is possible to derive the expectations of these quantities. These are especially important for the size-change model because shared, fixed, and exclusive polymorphisms are defined only when there are two populations. Again, mutations can be separated into two groups: those that occurred before the population split and those that occurred after. Then, $E(z_{1,i} \text{ before } t)$ is given by

$$E(z_{1,i} \text{ before } t) = \sum_{n_1'=2}^{n_1} P_{n_1 n_1'}(\tau/\theta_1)\sum_{k=1}^{n_1'-1} P(k \to i \mid n_1, n_1')E(z_{A,k}), \quad (17)$$

where $P(k \to i \mid n_1, n_1')$ is the probability that a mutation of size k in the sample n_1' grows to size i in the sample n_1 and $E(z_{A,k})$ is the expected number of mutations of size k in the sample n_1' at the moment the two populations split apart.

The expectation of $z_{A,k}$ is equal to θ_A/k (TAJIMA 1989b; FU 1995). $P(k \to i \mid n_1, n_1')$ is represented by the Polya-Eggenberger distribution; for example, see JOHNSON and KOTZ (1977), section 4.2. In words, $P(k \to i \mid n_1, n_1')$ is the probability that $(i - k)$ mutant lines are added when n_1' lineages become n_1 by the random selection and then bifurcation of lineages. It follows that

$$P(k \to i \mid n_1, n_1') = \binom{n_1 - n_1'}{i-k}\frac{k^{[i-k]}(n_1'-k)^{[n_1-n_1'-i+k]}}{n_1'^{[n_1-n_1']}}, \quad (18)$$

where $x^{[r]} = x(x+1)(x+2)\cdots(x+r-1)$.

The expectation of $z_{1,i}$ after t is calculated similarly to (13):

$$E(z_{1,i} \text{ after } t) = \theta_1\left[\frac{1}{i} - \sum_{n_1'=2}^{n_1} P_{n_1 n_1'}(\tau/\theta_1)\sum_{k=1}^{n_1'-1} P(k \to i \mid n_1, n_1')\frac{1}{k}\right], \quad (19)$$

and overall, *i.e.*, $E(z_{1,i} \text{ before } t) + E(z_{1,i} \text{ after } t)$,

$$E(z_{1,i}) = \frac{\theta_1}{i} + (\theta_A - \theta_1)\sum_{n_1'=2}^{n_1} P_{n_1 n_1'}(\tau/\theta_1)\times\sum_{k=1}^{n_1'-1} P(k \to i \mid n_1, n_1')\frac{1}{k}. \quad (20)$$

Again, the expression for $E(z_{2,i})$ is gotten simply by changing subscripts. Equation 20 is the decomposition of $E(S_1)$ into site frequencies; $E(z_{1,i})$ is taken without regard to polymorphism in population 2. Thus, (20), when summed over all possible frequencies, $i = 1$ to $i = n - 1$, is equivalent to TAJIMA's (1989a) Equation 9.

Of course in data, without an outgroup, we cannot distinguish between mutations represented by i copies and those represented by $n_1 - i$ copies, because we do not know which is the ancestral base. Let η_i be the number of polymorphic sites with frequency i/n_1 in the sample, where now $i \leq n_1/2$. Then

$$E(\eta_i) = \frac{E(z_{1,i}) + E(z_{1,n_1-i})}{\delta}, \quad (21)$$

where δ is two if $i = n_1 - i$ and one otherwise (FU 1995).

Jointly polymorphic sites can also be distinguished by their frequencies. Let z_{ij} be the number of polymorphic sites at which the mutant nucleotide has frequency i/n_1 in the sample from population 1 and frequency j/n_2 in the sample from population 2. Then

$$E(z_{ij}) = \theta_A\sum_{n_1'=2}^{n_1}\sum_{n_2'=2}^{n_2} P_{n_1 n_1'}(\tau/\theta_1)P_{n_2 n_2'}(\tau/\theta_2)\times\sum_{k_1=1}^{n_1'-1}\sum_{k_2=1}^{n_2'-1} P(k_1 \to i \mid n_1, n_1') \times P(k_2 \to j \mid n_2, n_2')\times P(k_1, k_2 \mid n_1, n_2)\frac{1}{k_1+k_2}, \quad (22)$$

where

$$P(k_1, k_2 \mid n_1, n_2) = \frac{\binom{n_1'}{k_1}\binom{n_2'}{k_2}}{\binom{n_1'+n_2'}{k_1+k_2}} \quad (23)$$

is the probability that a mutant of size $(k_1 + k_2)/(n_1' + n_2')$ in the ancestral sample has k_1 copies in the sample n_1' and k_2 copies in the sample n_2'.

Estimating population parameters: The theory outlined here provides a framework for parameter estimation. The isolation model has four parameters and, correspondingly, we can partition the segregating sites in a sample from two populations into four mutually exclusive categories. The expected values, $E(S_{X1})$, $E(S_{X2})$, $E(S_S)$, and $E(S_f)$, are given by (12) – (16), and, although complicated, are simply functions of the four parameters, θ_1, θ_2, θ_A, and τ. By equating observed values of S_{X1}, S_{X2}, S_S, and S_f with these expectations, we can solve numerically to find the values of θ_1, θ_2, θ_A, and τ that most closely equate the expected and observed values. Similarly, counts of site frequency patterns can be used to estimate the parameters of the size-change model. Assume that we have taken a sample of seven sequences from a population that has undergone a rapid change in population size. There are three possible site frequency patterns and three parameters: θ_1, θ_A, and τ. A sample size of seven was chosen so that the number of possible site frequencies would be the same as the number of parameters. The expectations of η_1, η_2, and η_3 are given by (21), so again we can equate observed and expected values and solve numerically to estimate the unknown parameters.

It is important to note that recombination within a sequence does not affect the expected numbers of the various types of segregating sites in a sample but that it does affect

the variance (TAJIMA 1993; PLUZHNIKOV and DONELLY 1996; WAKELEY 1997). Thus, these methods of estimation can be used on sequences that have undergone recombination. The effect of recombination is to lower the variances of the numbers of segregating sites, making observed values of S_{X1}, S_{X2}, S_f, and S_S, or of η_1, η_2, and η_3 tend to be closer to their expected values. Thus recombination should improve the quality of parameter estimates (TAJIMA 1993; PLUZHNIKOV and DONELLY 1996; WAKELEY 1997). If samples of the same size are taken from multiple loci, these methods can be used directly on the combined data and are expected to perform better the more recombination occurs between loci. The methods could also easily be modified for use on multilocus samples of different sizes. When multiple loci are used, the parameters estimated are the total parameters for all loci, the sum of single-locus values.

Recombination within and among loci has a similar effect on the correlations between the various classes of polymorphic sites, *i.e.*, it decreases them. Preliminary simulations showed that strong correlations (especially between S_f and S_S) associated with little or no recombination significantly decreased the accuracy of these methods. As with the variances, lower correlations lead to better parameter estimates. The Drosophila DNA data used below to illustrate the estimation of isolation model parameters show clear evidence of recombination both within and between loci. However, the human mitochondrial DNA data to which the size-change model is fit do not undergo recombination. In this case, we adopt another strategy for decreasing the correlations among classes of polymorphic sites: taking subsamples of size seven from a larger data set and averaging the site frequency patterns.

SIMULATIONS AND RESULTS

Computer simulations were done to demonstrate the effectiveness of this method of parameter estimation. The isolation model was simulated using the routine "make_tree" given in HUDSON (1990), but with three populations (ancestral plus two descendent), and three different population sizes, rather than one. The usual "coalescent" process proceeded independently in each of the two descendents until generation t, in the past, when the remaining sequences were united in the ancestral population. One set of values of θ_1, θ_2, and θ_A was chosen to illustrate the estimation procedure, and simulations were done over a range of the scaled time parameter, τ. Five thousand replicate data sets were generated for each set of parameter values. For each data set, S_{X1}, S_{X2}, S_S, and S_f were counted and then equated with the expectations derived above. These four equations were solved numerically using a modified NEWTON-RAPHSON method; see, for example, chapter 9 of PRESS *et al.* (1992). This gave estimates of θ_1, θ_2, θ_A, and τ for each replicate.

Preliminary simulations showed that the estimation is effective only when there is some representation in the data of the range of possible genealogies. Single genealogies do not contain enough information. Under the isolation model, for instance, without recombination there can be either fixed differences or shared polymorphisms, but there can never be both. Thus, the presence of a shared polymorphism in such a sample determines that the number of fixed differences is zero. This strong negative correlation causes the method to fail, and only disappears when there is considerable recombination or when we have samples of multiple independent loci. To insure that a number of genealogies would be represented in the data from each replicate, samples of 10 independent loci were simulated and the estimation was done only when both shared and fixed polymorphisms were observed. Within each locus no recombination was allowed.

Figure 2 shows the results of these simulations. The two descendent parameters, θ_1 and θ_2, are estimated with a fairly high degree of accuracy. To illustrate, for the case of $\tau = 40$ per locus, the standard error of $\hat{\theta}_1$ is only ~10% larger than when WATTERSON's (1975) estimator is used to estimate a single-population θ from identical data, *i.e.*, 20 sequences from each of 10 loci with $\theta = 20$ per locus. In this same case, the standard error of $\hat{\theta}_2$ is only ~5% larger than that of WATTERSON's (1975) estimator. This is to be expected since, for $\tau = 40$, the chance of within population monophyly is 64% for the sample from population 1 and 81% for the sample from population 2. As the time of separation decreases, a greater number of ancestral lineages is expected, which means the data will contain less information about the descendent populations and the standard error will increase.

Figure 2c shows that θ_A is estimated with somewhat less accuracy than θ_1 and θ_2. This is due in part to uncertainty about the configuration of the ancestral sample, but also results from the particular choice of parameters. We expect to have relatively more information about the ancestral population when the time of separation is short, but even when $\tau = 10$ per locus, the most probable ancestral sample is $n_1' = 3$ and $n_2' = 3$. The standard error of θ_A in this case, shown in Figure 2, is about the same as when WATTERSON's (1975) estimator is used with 10 sequences sampled from a single locus with $\theta = 100$. The time parameter, τ, is estimated quite accurately, except when the time of separation is long. In this case, τ tends to be underestimated, and, correspondingly, θ_A tends to be overestimated.

Just one set of parameter values was chosen to illustrate the effectiveness of using site frequencies to estimate θ_1, θ_A, and τ in the size-change model. These were $\theta_1 = 21.70$, $\theta_A = 0.00$, and $\tau = 4.77$ at a single locus with $n = 69$. These are the values estimated for one of the human mitochondrial datasets analyzed below. For each replicate, site frequencies were averaged over 1000 random subsamples of seven sequences from the simulated sample of 69 sequences. The average ± 1 SE of the estimates of the parameters over 10,000 simulation replicates were $\hat{\theta}_1 = 20.4(\pm 29.1)$, $\hat{\theta}_A = 0.25(\pm 0.64)$, and $\hat{\tau} = 4.6(\pm 1.2)$. This level of error is higher than in estimating the parameters of the isolation model, probably due to the fact that here just a single locus was used. Figure 3 depicts the distributions of the estimates of the three parameters and shows that

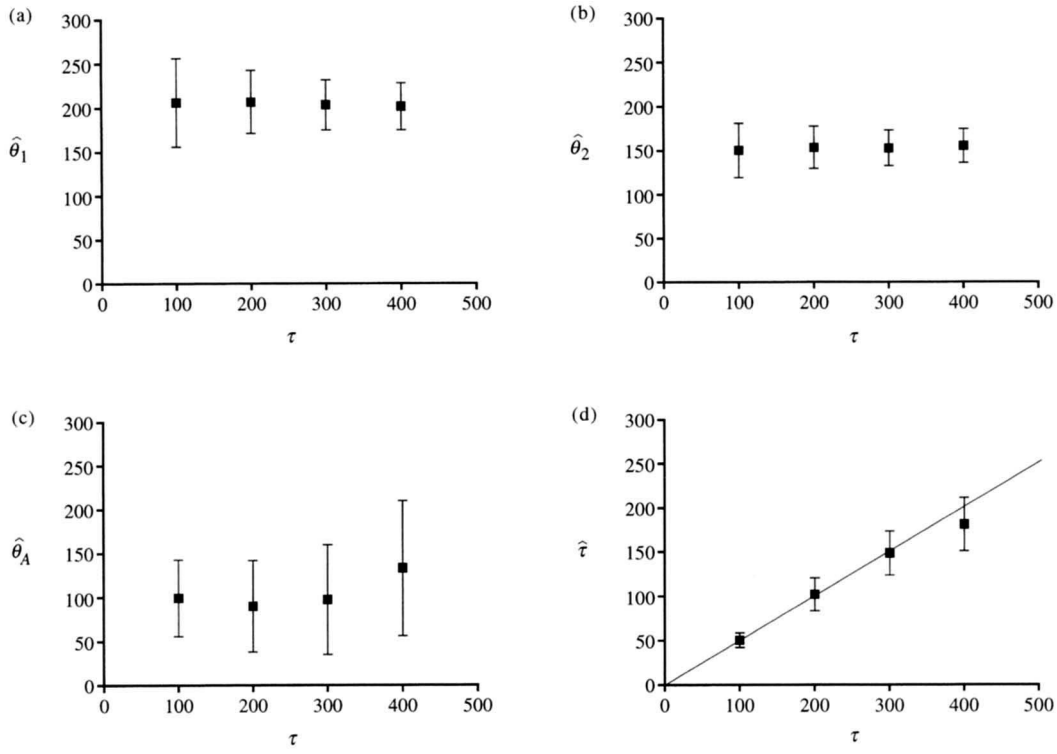

FIGURE 2.—Simulation results of total parameter estimates for a sample of 20 sequences from each population when $\theta_1 = 20$, $\theta_2 = 15$, and $\theta_A = 10$ per locus and data are from 10 independent loci. Four values of τ were used as follows: 10, 20, 30, and 40 per locus. Solid black boxes plot the means of the estimated parameters over all replicates that showed both shared and fixed polymorphic sites. Out of a total of 5000 replicates, the numbers of times this occurred were 3343, 4862, 3345, and 1455 for τ equal to 10, 20, 30, and 40 per locus, respectively. Bars, 1 SE of the estimates over replicates.

estimates cluster mainly around the true parameter values. Figure 3b shows the extreme L-shape of the distribution of $\widehat{\theta_A}$. While the mean quoted above indicates bias in estimating θ_A, fully 75% of the estimates were smaller than 10^{-6}.

APPLICATION TO DNA DATA

We used some previously reported DNA sequence data from two closely related species of Drosophila to illustrate the estimation of θ_1, θ_2, θ_A, and τ in the isolation model. Species 1 was *D. simulans* and species 2 was *D. mauritiana*. These two species separated only ~770,000 years ago and data from three *X*-linked loci were previously obtained. Since *X*-linked loci have three-fourths the effective population size of autosomal loci, estimates of θ_1, θ_2, and θ_A will be correspondingly lower. KLIMAN and HEY (1992) sequenced 1878 bp of the *period* locus and HEY and KLIMAN (1992) sequenced of 999 bp the *zeste* locus and 1114 bp of the *yolk protein 2* locus in the same six isofemale lines from each species. There were a total of 56 polymorphisms exclusive to *D. simulans* and 47 exclusive to *D. mauritiana*, 11 shared between the two, and six fixed differences. Thus, $S_{X1} = 56$, $S_{X2} = 47$, $S_S = 11$, and $S_f = 6$. Table 1 shows the results of solving for the parameters that give the best fit to these numbers.

From the estimates in Table 1 and assuming that the mutation rate has remained constant over time, it appears that the ancestor of *D. simulans* and *D. mauritiana* had an effective population size that was intermediate between those of its two descendents. Further, the population size of *D. mauritiana* is estimated to be about three-quarters that of *D. simulans*. The two species are estimated to have split apart 9.0 mutational units in the past, but this is not the customary measure of time in population genetics; time is typically measured in units of $2N$ generations. For *D. simulans* this is estimated as 0.60 whereas for *D. mauritiana* it is estimated to be 0.78, which reflects the difference in effective population size of these two species.

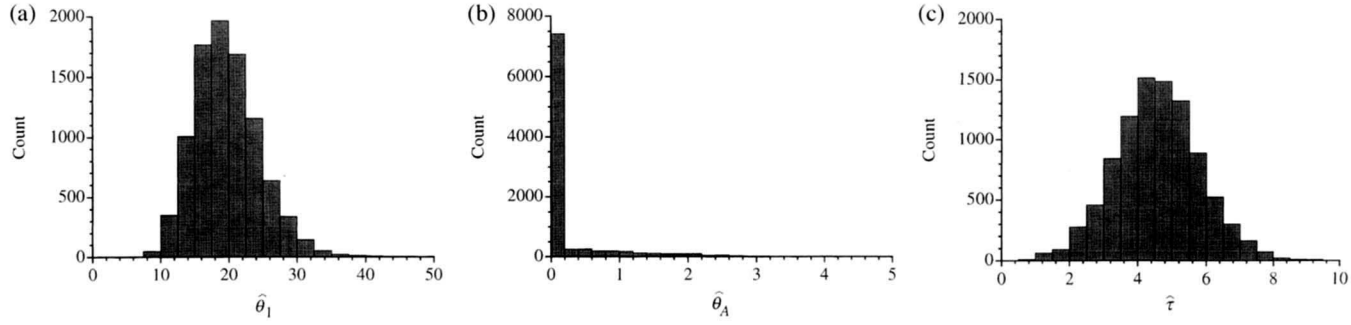

FIGURE 3.—Simulation results for the size-change model with $n = 69$, $\theta_1 = 21.70$, $\theta_A = 0.00$, and $\tau = 4.77$. Histograms plot values of parameter estimates over 9235 replicates; ~7% of the time, the numerical solving routine failed to converge.

TABLE 1

Estimates of population parameters for *D. simulans* and *D. mauritiana*

Parameter estimates: total		Corresponding expectations
$\hat{\theta}_1 = 30.2\ (0.0076)$		$E(S_{x1}) = 56.0$
$\hat{\theta}_2 = 23.0\ (0.0058)$	$\Longleftrightarrow$	$E(S_{x2}) = 47.0$
$\hat{\theta}_A = 28.6\ (0.0072)$		$E(S_x) = 11.0$
$\tau = 18.0\ (0.0045)$		$E(S_f) = 6.0$

Values in parentheses are number per site.

A dataset of human mitochondrial DNA serves to illustrate the use of site frequencies in estimating current *vs.* historical population sizes. In estimating θ_1 and θ_A, it is important to note that because mitochondria are haploid and maternally inherited in humans, their effective population size is about one-fourth that of autosomal loci. DiRienzo and Wilson (1991) sequenced part of the control region in 111 humans: 69 from Sardinia and 42 from the Middle East. They suggested that the unimodal distribution of pairwise differences for these populations resulted from recent growth in population size in both Sardinia and the Middle East. Rogers and Harpending (1992) later used these distributions to estimate the parameters of a model of instantaneous growth identical to the present size-change model with $\theta_1 > \theta_A$. They estimated that $\theta_1 = 17.55$, $\theta_A = 0.76$, and $\tau = 3.99$ for Sardina and $\theta_1 = 3117.40$, $\theta_A = 0.00$, and $\tau = 7.34$ for the Middle East (Rogers and Harpending 1992). Fitting expectations, given by Equation 21 above, to the averages of η_1, η_2, and η_3 over 100,000 random subsamples of seven sequences, we obtain $\theta_1 = 21.70$, $\theta_A = 0.00$, and $\tau = 4.77$ for Sardina and $\theta_1 = 21.94$, $\theta_A = 0.00$, and $\tau = 8.51$ for the Middle East. Thus, our analysis also supports a relatively recent and rapid expansion for these two populations.

Figure 4 shows the distributions of pairwise differences (a and b) and the average site frequency counts (c and d) for DiRienzo and Wilson's (1991) Sardinian and Middle Eastern data. Also shown are the expected distributions and counts for each, given the different estimates made here *vs.* those of Rogers and Harpending (1992). Figure 4a shows that, for Sardinia, both our estimates and those of Rogers and

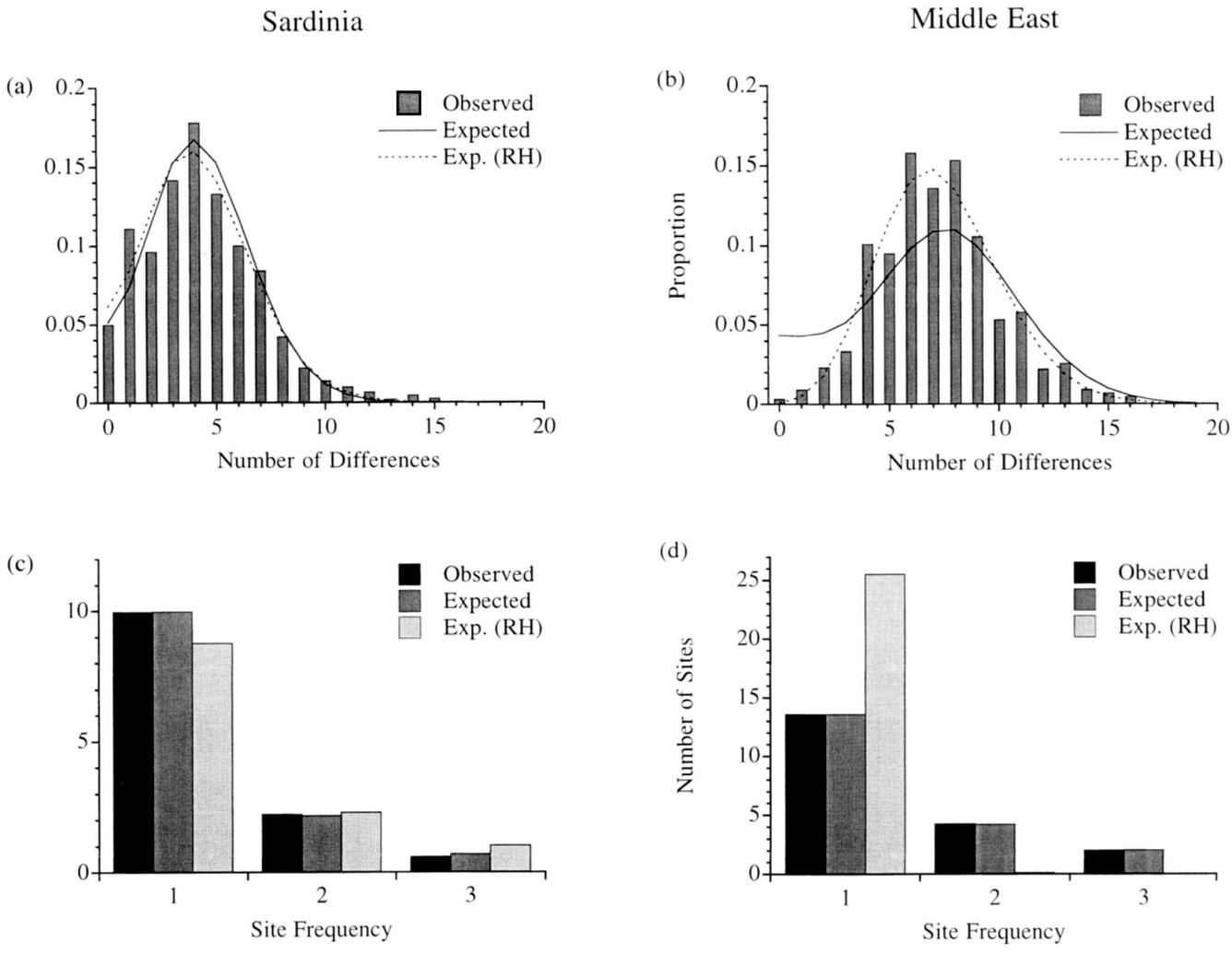

Figure 4.—Results for Sardinia and the Middle East. Expected values in a and b were obtained by simulating 10,000 replicate datasets with our parameter estimates and those of Rogers and Harpending (1992), designated (RH).

HARPENDING (1992) give similar predictions for the distribution of pairwise differences and that these fit the observed data well. Figure 4c shows that, as expected, the estimates made here provide a nearly perfect fit to the site frequency distribution, but ROGERS and HARPENDING's (1992) numbers fit well also. A different situation is found for the Middle East, shown in Figure 4, b and d. Our estimates based on site frequencies do not reproduce the distribution of pairwise differences and ROGERS and HARPENDING's (1992) estimates based on the distribution of pairwise differences predict a very different pattern of site frequencies than what is observed.

DISCUSSION

Polymorphic sites in a sample of DNA sequences can be partitioned into categories that correspond to components of genealogical history and from which population parameters can be estimated. The methods of estimation presented here depend on a decoupling of different sites' histories, so that many of the possible genealogies are realized in the sample. For example, the accurate estimation of θ_1, θ_2, θ_A, and τ requires that both shared and fixed differences be observed. However, this is not possible if all sites in the sample share the exact same history because any single genealogy allows for the creation of only one of these two kinds of polymorphic sites. This introduces a strong negative correlation between S_S and S_f in the isolation model, which decreases only when there is recombination in the sequences or when multiple loci are studied.

The Drosophila data analyzed above show ample evidence of recombination. Applying the "four gamete" test of HUDSON and KAPLAN (1985) to the *period* locus sequences, a minimum of seven and nine recombination events are inferred to have occurred in *D. simulans* and *D. mauritiana*, respectively (KLIMAN and HEY 1992). In addition, two other loci were used together with *period* in the example above. Thus, the numbers of polymorphic sites used in the estimation routines probably reflect population history rather than the correlations imposed by particular genealogical structures. Simulations show that when this is true the resulting estimates are close to their true values.

Another method of estimating θ_A and τ in the isolation model was developed by TAKAHATA (1986) and extended to the case of three species by TAKAHATA *et al.* (1995). Those methods require multiple loci with a sample size of one from each species. This precludes the observance of ancestral polymorphisms and serves to distinguish those methods from the one we have developed here. When species are distantly related enough that shared polymorphisms are rare, TAKAHATA's (1986) and TAKAHATA *et al.*'s (1995) will be the methods of choice. Of course, they will also work when shared polymorphisms and fixed differences are both likely to be observed, but in such cases it may be preferable to use the method developed here (provided that samples of more than one sequence are available) because it extracts the information contained in ancestral polymorhphisms.

When only shared polymorphisms are observed, as will often be the case for very recently diverenged populations or species, especially when there is no recombination, the following method could be used to extract information about the common ancestor, independent of the descendents. In this case, the minimum interpopulation pairwise differences, *i.e.*, the smallest k_{ij}, where k_{ij} is the number of differences between sequence i from population 1 and sequence j from population 2, will give a reasonable estimate of τ (TAKAHATA and NEI 1985). The estimate will, of course, be somewhat larger than the true value of τ, but the magnitude of this bias is small when there are two or more ancestral lineages, as required to observe shared polymorphisms. The average number of interpopulation pairwise differences,

$$d_{12} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} k_{ij}, \tag{24}$$

can also be computed, and under the isolation model this has expectation $\tau + \theta_A$. Then θ_A is estimated simply by $d_{12} - \min(k_{ij})$. Simulations results (not shown) demonstrate that these estimators of τ and θ_A, while slightly biased, have very low standard errors. SATTA *et al.* (1991) used $\min(k_{ij})$ to estimate evolutionary rates in Mhc loci.

The reciprocal disagreement, shown in Figure 4, between the Middle East data and the expectations for the size-change model from parameter estimates made here using site frequencies and by ROGERS and HARPENDING (1992) using pairwise differences, is interesting and deserves further study. It implies a lack of fit between the size-change model and the Middle Eastern data. The hypervariable segment of the control region sequenced by DIRIENZO and WILSON (1991)

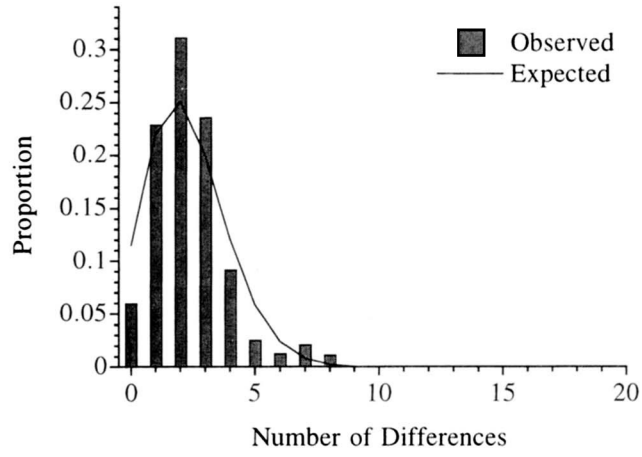

FIGURE 5.—Observed and expected distributions of pairwise differences when only sites that are inferred to have changed just once are analyzed.

displays a great deal of variation in substitution rate among nucleotide sites (WAKELEY 1993), and this might explain the discrepancies. Figure 5 shows the result of redoing our analysis after excluding 22 sites determined to have changed more than once by the method of WAKELEY (1993). In this case, the parameter estimates from average site frequencies are $\theta_1 = 24.46$, $\theta_A = 0.00$, and $\tau = 2.46$. In Figure 5, as in Figure 4c, simulations with these parameter values were used to compute the expected distribution of pairwise differences. The improvement in the fit between the (new) observed distribution of pairwise differences and our expectations implies that multiple changes at some sites may explain the apparent lack of fit of the size change model. However, throwing out data (in this case 22 out of 60 polymorphic sites) is probably not the ideal approach. Better would be a full accounting of rate variation in the development of methods similar to the ones proposed here.

Comments of two anonymous reviewers improved the manuscript. This work was supported by National Institutes of Health grant GM-17745-01 to J.W. and National Science Foundation grant DEB-9306625 to J.H.

LITERATURE CITED

DIRIENZO, A., and A. C. WILSON, 1991 Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc. Natl. Acad. Sci. USA **88:** 1597–1601.

FISHER, R. A., 1930 *The Genetical Theory of Natural Selection.* Clarendon, Oxford.

FU, X.-Y., 1995 Statistical properties of segregating sites. Theoret. Pop. Biol. **48:** 172–197.

FU, X.-Y., and W.-H. LI, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

HEY, J., 1991 The structure of genealogies and the distribution of fixed differences between DNA sequences from natural populations. Genetics **128:** 831–840.

HEY, J., 1994 Bridging phylogenetics and population genetics with gene tree models, pp. 435–499 in *Molecular Ecology and Evolution: Approaches and Applications,* edited by B. SCHIERWATER, G. P. WAGNER and R. DESALLE. Birkhäuser Verlag, Basel, Switzerland.

HEY, J., and R. M. KLIMAN, 1992 Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. Mol. Biol. Evol. **10:** 804–822.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology,* Vol. 7, edited by D. J. FUTUYMA and J. ANTONOVICS. Oxford University Press, Oxford.

HUDSON, R. R., and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

JOHNSON, N. L., and S. KOTZ, 1977 *Urn Models and Their Application.* Wiley, New York.

KLIMAN, R. M., and J. HEY, 1992 DNA sequence variation at the *period* locus within and among species of the *Drosophila melanogaster* complex. Genetics **133:** 375–387.

PLUZHNIKOV, A., and P. DONELLY, 1996 Optimal sequencing strategies for surveying molecular genetic diversity. Genetics **144:** 1247–1262.

PRESS, W. H., S. A. TEUKOLSKY, W. T. VETTERLING and B. P. FLANNERY, 1992 *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge.

ROGERS, A. R., and H. HARPENDING, 1992 Populations growth makes waves in the distribution of pairwise differences. Mol. Biol. Evol. **9:** 552–569.

SATTA, Y., N. TAKAHATA, C. SCHÖNBACH, J. GUTKNECHT and J. KLEIN, 1991 Calibrating evolutionary rates at major histocompatibility complex loci, pp. 51–62 in *Molecular Evolution of the Histocompatibility Complex Loci,* edited by J. KLEIN and D. KLEIN. Springer-Verlag, Berlin.

SLATKIN, M., and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

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

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

TAJIMA, F., 1993 Measurement of DNA polymorphism, pp. 37–59 in *Mechanisms of Molecular Evolution,* edited by N. TAKAHATA and A. G. CLARK. Sinauer Associates, Sunderland, MA.

TAKAHATA, N., 1986 An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. Genet. Res. Camb. **48:** 187–190.

TAKAHATA, N., and M. NEI, 1985 Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110:** 325–344.

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.

WAKELEY, J., 1993 Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. **37:** 613–623.

WAKELEY, J., 1997 Using the variance of pairwise differences to estimate the recombination rate. Genet. Res. Camb. (in press).

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theoret. Pop. Biol. **7:** 256–276.

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

Communicating Editor: N. TAKAHATA