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 descendents, 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 descendents 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 descendents, 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 descendents 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 one single descendendent 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

Genetics 145: 847-855 (March, 1997)
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 and a dearth of middle-frequency polymorphic sites.

We adopt general and easily interpreted versions of the isolation and size-change models. Before generation in the past, there was a single, panmictic population of size . Exactly at , 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 and , and no restrictions are put on the relative sizes of , , and . All populations conform to the commonly used Wright–Fisher model (Fisher 1930; Wright 1931). Generations are nonoverlapping and , and remain constant over time except at , 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 per sequence per generation. Four parameters, then, describe the isolation model: , , , and . The size-change model is characterized by three parameters: , , and .

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 and . 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 . Fourth, there are sites showing fixed differences between the two populations, the number of which are referred to as .

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 and . Finally, we can simply count the total number of polymorphic sites in the entire sample, and the number of these is referred to as . These different categories of sites are related in the following way:

\[
S_1 = S_{11} + S_{12}, \quad S_2 = S_{21} + S_{22}, \quad S = S_{11} + S_{12} + S_{21} + S_{22}.
\]

**Single-population expectations:** To calculate the expectations of , , and , 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 . Let the numbers of sequences sampled be and . Watterson (1975) showed that

\[
E(S) = \theta \sum_{i=1}^{n} \frac{1}{i},
\]

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

\[
E(S_1) = \theta \sum_{i=1}^{n_1} \frac{1}{i} \quad \text{and} \quad E(S_2) = \theta \sum_{i=1}^{n_2} \frac{1}{i}.
\]
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 $S_j$ 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 $\theta\left(1 + \frac{1}{n_1} + \frac{1}{n_2}\right)$, where $\theta$ is 2 if $n_1 = n_2$ and 1 otherwise (Tajima 1989b; Fu and Li 1993; Fu 1995). The chance that the $n_1$ and $n_2$ bases are distributed among the two subsamples as a fixed difference is related to the hypergeometric distribution and is just $\delta(\frac{1}{n_1} + \frac{1}{n_2})$. Thus,

$$E(S_j) = \frac{\theta\left(1 + \frac{1}{n_1} + \frac{1}{n_2}\right)}{n}$$

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

Then, using (1) - (3),

$$E(S_{X1}) = \theta \left[ \sum_{i=1}^{n_1-1} \frac{1}{i} - \sum_{i=1}^{n_2-1} \frac{1}{i} - \frac{1}{n_1} \right]$$

$$E(S_{X2}) = \theta \left[ \sum_{i=1}^{n_1-1} \frac{1}{i} - \sum_{i=1}^{n_2-1} \frac{1}{i} - \frac{1}{n_2} \right]$$

$$E(S_{s}) = \theta \left[ \sum_{i=1}^{n_1-1} \frac{1}{i} + \sum_{i=1}^{n_2-1} \frac{1}{i} - \frac{1}{n_1} \right]$$

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 the presently sampled $n_1$ sequences:

$$P_{n_1'}(T_1) = 1 - \sum_{i=2}^{n_1} \frac{1}{\prod_{j=2,j\neq i}^{n_1} \left(1 - \frac{i(i-1)}{j(j-1)}\right)}$$

if $n_1' = 1$ (10)

$$P_{n_1'}(T_1) = \frac{1}{2} \sum_{i=1}^{n_1} \frac{1}{\prod_{j=1,j\neq i}^{n_1} \left(1 - \frac{i(i-1)}{j(j-1)}\right)}$$

if $2 \leq n_1' \leq n_1$. (11)

Note that in (14) when $n_1' = 1$ the middle term in the brackets is defined to be equal to zero. Again, the equation that applies to $S_{X1}$ 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. $E(S_j \text{ before } t)$ is calculated similarly to (12) and (14) above,

$$E(S_j \text{ before } t) = \theta \sum_{n_1'=2}^{n_1} \sum_{n_2'=1}^{n_2} P_{n_1'}(\tau/\theta_1) P_{n_2'}(\tau/\theta_2)$$

$$\times \left[ \sum_{i=1}^{n_1'-1} \frac{1}{i} + \sum_{i=1}^{n_2'-1} \frac{1}{i} - \frac{1}{n_1'} // \frac{1}{n_2'} \right]$$

(12)

In words, $E(S_j \text{ before } t)$ 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 $t$:

$$E(S_{X1} \text{ before } t) = \theta \sum_{n_1'=2}^{n_1} \sum_{n_2'=1}^{n_2} P_{n_1'}(\tau/\theta_1) P_{n_2'}(\tau/\theta_2)$$

$$\times \left[ \sum_{i=1}^{n_1'-1} \frac{1}{i} + \sum_{i=1}^{n_2'-1} \frac{1}{i} - \frac{1}{n_1'} // \frac{1}{n_2'} \right]$$

(13)

Note that in (13) when $n_1' = 1$ the middle term in the brackets is defined to be equal to zero. Again, the equation that applies to $S_{X1}$ 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.
In addition, if there was only one 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

\[ E(S_i| \text{after } t) = \frac{\theta_1}{2} \sum_{i=2}^{n_1} \left( \frac{i}{2} \right) \prod_{j=2}^{n_2} \frac{1 - e^{-\left(\frac{i(i-1)}{j(j-1)}\right)}}{1 - \left(\frac{i(i-1)}{j(j-1)}\right)} \]

and, again, \( E(S_i) = E(S_i| \text{after } t) + E(S_i| \text{after } t) \).

**Site frequencies:** Let \( z_{i,j} \) 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, \( z_{i,j} \) represents mutations of size \( j \) 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, \( E(z_{i,j}| \text{before } t) \) is given by

\[ E(z_{i,j}| \text{before } t) = \sum_{i=2}^{n_1} P_{n_1} \left( \frac{\theta_1}{2} \right) \sum_{i=2}^{n_1} P(k \rightarrow i|n_1, n_1') E(z_{i,k}), \quad (17) \]

where \( P(k \rightarrow i|n_1, n_1') \) is the probability that a mutation of size \( k \) in the sample \( n_1' \) grows to size \( i \) in the sample \( n_1 \) and \( E(z_{i,k}) \) 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 \( z_{i,j} \) is equal to \( \theta_1/j \) (Tajima 1989b; Fu 1995). \( P(k \rightarrow i|n_1, n_1') \) is represented by the Polya-Eggenberger distribution; for example, see Johnson and Kotz (1977), section 4.2. In words, \( P(k \rightarrow i|n_1, n_1') \) 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

\[ P(k \rightarrow i|n_1, n_1') = \binom{n_1 - n_1'}{i-k} \frac{(n_1' - k)^{n_1' - i-k}}{i-k}, \quad (18) \]

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

The expectation of \( z_{i,j} \) after \( t \) is calculated similarly to (15):

\[ E(z_{i,j}| \text{after } t) = \frac{\theta_1}{i} \sum_{i=2}^{n_1} P_{n_1} \left( \frac{\theta_1}{2} \right) \sum_{i=2}^{n_1} P(k \rightarrow i|n_1, n_1') \frac{1}{k}, \quad (19) \]

and overall, i.e., \( E(z_{i,j}) = E(z_{i,j}| \text{before } t) + E(z_{i,j}| \text{after } t) \).

Again, the expression for \( E(z_{i,j}) \) is gotten simply by changing subscripts. Equation 20 is the decomposition of \( E(S) \) into site frequencies; \( E(z_{i,j}) \) is taken without regard to polymorphism in population 2. Thus, \( (20) \), when summed over all possible frequencies, \( i = 1 \) to \( i = n_1 - 1 \), 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 \( \eta_i \) be the number of polymorphic sites with frequency \( i/n_1 \) in the sample, where now \( i = n_1/2 \).

Then

\[ E(\eta_i) = \frac{E(z_{i,j}) + E(z_{i,n_1-i})}{\delta}, \quad (21) \]

where \( \delta \) is two if \( i = n_1 - i \) and one otherwise (Fu 1995).

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

\[ E(z_{i,j}) = \theta_1 \sum_{n_1=2}^{n_1} \sum_{n_1'=2}^{n_1} P_{n_1} \left( \frac{\theta_1}{2} \right) P_{n_1'} \left( \frac{\theta_1}{2} \right) \]

\[ \times \sum_{s_{k-1} \leq k-1} \sum_{s_{k-1} \leq k-1} P(k_1 \rightarrow i|n_1, n_1') \times P(k_2 \rightarrow j|n_2, n_2') \]

\[ \times P(k_1, k_2|n_1, n_2) \frac{1}{k_1 + k_2}, \quad (22) \]

where

\[ P(k_1, k_2|n_1, n_2) = \frac{n_1' \left( \frac{n_1'}{k_1} \right) \left( \frac{n_2'}{k_2} \right)}{n_1 + n_2} \]

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, \( E(S_{11}), E(S_{10}), E(S_{01}), \) and \( E(S_{00}) \), are given by (12) to (16), and, although complicated, are simply functions of the four parameters, \( \theta_1, \theta_2, \theta_3, \) and \( \tau \). By equating observed values of \( S_{11}, S_{10}, S_{01}, \) and \( S_{00} \), we can solve numerically to find the values of \( \theta_1, \theta_2, \theta_3, \) and \( \tau \) 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: \( \theta_1, \theta_2, \) and \( \tau \). 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 \( \eta_1, \eta_2, \) and \( \eta_3 \) are given by (21) and 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 1989; 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 \( S_n \) and \( S_r \), or of \( \eta_1 \), \( \eta_2 \), and \( \eta_3 \) tend to be closer to their expected values. Thus recombination should improve the quality of parameter estimates (Tajima 1989; Pluzhnikov and Donelly 1996; Wakeley 1997). If samples from 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 \( S_2 \) and \( S_3 \)) 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 within and between loci. However, the human mitochondrial DNA data to which the size-change model fits 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 descendent until generation \( t \), in the past, when the remaining sequences were united in the ancestral population. One set of values of \( \theta_1 \), \( \theta_2 \), and \( \theta_3 \) was chosen to illustrate the estimation procedure, and simulations were done over a range of the scaled time parameter, \( \tau \). Five thousand replicate data sets were generated for each set of parameter values. For each data set, \( S_{n1} \), \( S_{n2} \), \( S_n \), and \( S_r \) 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 \( \theta_1 \), \( \theta_2 \), \( \theta_3 \), and \( \tau \) 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, \( \theta_1 \) and \( \theta_2 \), are estimated with a fairly high degree of accuracy. To illustrate, for the case of \( \tau = 40 \) per locus, the standard error of \( \theta_1 \) is only \( \sim 10\% \) larger than when Watterson's (1975) estimator is used to estimate a single-population \( \theta \) from identical data, i.e., 20 sequences from each of 10 loci with \( \theta = 20 \) per locus. In this case, the standard error of \( \theta_2 \) is only \( \sim 5\% \) 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 \( \theta_3 \) is estimated with somewhat less accuracy than \( \theta_1 \) and \( \theta_2 \). 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 \( \theta_3 \) 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, \( \tau \), is estimated quite accurately, except when the time of separation is long. In this case, \( \tau \) tends to be underestimated, and, correspondingly, \( \theta_3 \) tends to be overestimated.

Just one set of parameter values was chosen to illustrate the effectiveness of using site frequencies to estimate \( \theta_1 \), \( \theta_3 \), and \( \tau \) in the size-change model. These were \( \theta_1 = 21.70 \), \( \theta_3 = 0.00 \), 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 \( \pm 1 \) SE of the estimates of the parameters over 10,000 simulation replicates were \( \hat{\theta}_1 = 20.4 \pm 0.5 \), \( \hat{\theta}_3 = 0.25 \pm 0.64 \), 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...
estimates cluster mainly around the true parameter values. Figure 3b shows the extreme L-shape of the distribution of $\widehat{\theta}_2$. While the mean quoted above indicates bias in estimating $\theta_A$, fully 75% of the estimates were smaller than $10^{-3}$.

APPLICATION TO DNA DATA

We used some previously reported DNA sequence data from two closely related species of Drosophila to illustrate the estimation of $\theta_1$, $\theta_2$, $\theta_A$, and $\tau$ in the isolation model. Species 1 was D. simulans and species 2 was D. mauritiana. These two species separated only $\sim 770,000$ 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 $\theta_1$, $\theta_2$, and $\theta_A$ will be correspondingly lower. Kliman and Hey (1992) sequenced 1878 bp of the period locus and Hey and Kliman (1992) sequenced 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, $S_1 = 56$, $S_2 = 47$, $S_0 = 11$, 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 2](image2.png)

**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 $\tau$ 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 $\tau$ equal to 10, 20, 30, and 40 per locus, respectively. Bars, 1 SE of the estimates over replicates.

![Figure 3](image3.png)

**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.
A dataset of human mitochondrial DNA serves to illustrate the use of site frequencies in estimating current vs. historical population sizes. In estimating \( \theta_i \) and \( \theta_A \), 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 \( \theta_1 > \theta_i \). They estimated that \( \theta_i = 17.55, \theta_A = 0.76, \) and \( \tau = 3.99 \) for Sardinia and \( \theta_i = 3117.40, \theta_A = 0.00, \) and \( \tau = 7.34 \) for the Middle East (Rogers and Harpending 1992). Fitting expectations, given by Equation 21 above, to the averages of \( \eta_1, \eta_2, \) and \( \eta_3 \) over 100,000 random subsamples of seven sequences, we obtain \( \theta_i = 21.70, \theta_A = 0.00, \) and \( \tau = 4.77 \) for Sardinia and \( \theta_i = 21.94, \theta_A = 0.00, \) 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

### Table 1

<table>
<thead>
<tr>
<th>Parameter estimates: total</th>
<th>Corresponding expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta_i = 30.2 ) (0.0076)</td>
<td>( E(S_{i1}) = 56.0 )</td>
</tr>
<tr>
<td>( \theta_2 = 23.0 ) (0.0058)</td>
<td>( E(S_{i2}) = 47.0 )</td>
</tr>
<tr>
<td>( \theta_A = 28.6 ) (0.0072)</td>
<td>( E(S彳) = 11.0 )</td>
</tr>
<tr>
<td>( \tau = 18.0 ) (0.0045)</td>
<td>( E(S productService) = 6.0</td>
</tr>
</tbody>
</table>

Values in parentheses are number per site.
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,
showed in Figure 4, b and d. Our estimates based on
site frequencies do not reproduce the distribution of
shown and different situation is found for the Middle East,
wise differences predict
expected, the estimates made here provide a nearly per-
dividuals. This intro-
observed. However, this is not possible if
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 in-
troduces a strong negative correlation between $S_S$ and $S_T$
in the isolation model, which decreases only when
there is recombination in the sequences or when mul-
tiple loci are studied.

The Drosophila data analyzed above show ample evi-
dence of recombination. Applying the “four gamete”
test of Hudson and Kaplan (1985) to the period locus
sequences, a minimum of seven and nine recombin-
ation 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 cor-
rrelations 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 $\theta_A$ and $\tau$ in the iso-
lation 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 distinctly related
enough that shared polymorphisms are rare, Takaha-
ta’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 prefera-
tible to use the method developed here (provided that
samples of more than one sequence are available) be-
cause it extracts the information contained in ancestral
polymorphisms.

When only shared polymorphisms are observed, as
will often be the case for very recently diverged popu-
lations or species, especially when there is no recombi-
nation, the following method could be used to extract
information about the common ancestor, independent
of the descendents. In this case, the minimum interpo-
lar launch pairwise differences, i.e., the smallest $k_i$, where
$k_i$ is the number of differences between sequence $i$
from
population 1 and sequence $j$ from population 2, will
give a reasonable estimate of $\tau$ (Takahata and Nei
1985). The estimate will, of course, be somewhat larger
than the true value of $\tau$, but the magnitude of this bias
is small when there are two or more ancestral lineages,
as required to observe shared polymorphisms. The aver-
age number of interpopulation pairwise differences,

\begin{equation}
\text{d}_{12} = \frac{1}{n_1 n_2 \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} k_{ij}}
\end{equation}

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

The reciprocal disagreement, shown in Figure 4, be-
tween the Middle East data and the expectations for
the size-change model from parameter estimates made
here using site frequencies and by Rogers and Har-
pending (1992) using pairwise differences, is interest-
ing 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)
might explain the discrepancies. Figure 5 shows the 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 $\theta_1 = 24.46, \theta_2 = 0.00$, 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-7745-01 to J.W. and National Science Foundation grant DEB-9306625 to J.H.

LITERATURE CITED


Communicating Editor: N. Takahata