Frequentist Estimation of Coalescence Times From Nucleotide Sequence Data Using a Tree-Based Partition

Hua Tang,* David O. Siegmund,*† Peidong Shen,* Peter J. Oefner† and Marcus W. Feldman‡

*Department of Statistics, †Stanford Genome Technology Center and ‡Department of Biological Sciences,
Stanford University, Stanford, California 94305

Manuscript received July 19, 2001
Accepted for publication February 26, 2002

ABSTRACT

This article proposes a method of estimating the time to the most recent common ancestor (TMRCA) of a sample of DNA sequences. The method is based on the molecular clock hypothesis, but avoids assumptions about population structure. Simulations show that in a wide range of situations, the point estimate has small bias and the confidence interval has at least the nominal coverage probability. We discuss conditions that can lead to biased estimates. Performance of this estimator is compared with existing methods based on the coalescence theory. The method is applied to sequences of Y chromosomes and mtDNAs to estimate the coalescent times of human male and female populations.

METHOD

Interpretations of the estimator: Before reviewing the existing methods and introducing the new one, we wish to clarify the genetic and statistical interpretations of the quantity under investigation.

We restrict attention to the estimation of the sample TMRCA. Whether the TMRCA of a sample can be interpreted as an approximation of the TMRCA of a population depends on the history of the population as well as the sampling strategy. In the simplest case, for a random sample of size \( n \) from a panmictic population, the probability that the common ancestor of a sample coincides with the common ancestor of the population is \( \sim \frac{n-1}{n+1} \) (Saunders et al. 1984). In practice, the panmictic assumption is seldom satisfied, and the sampling scheme is rarely completely random, which lead to uncertainties in the interpretation of the sample TMRCA (Templeton et al. 2000).

Statistically speaking, because of the stochastic nature of the evolutionary process, TMRCA can be regarded
as the (unobserved) value of a random variable. More precisely, the DNA at each locus evolves as one realization of a stochastic process. Further, the TMRCA of one sample may differ from that of another sample. However, for a fixed sample and at each locus, there is an unknown true TMRCA, denoted by \( T \), at which all individuals in the sample share a common ancestor. We may imagine that an immortal person has kept a genealogy of each locus since the beginning of human history. If we had the genealogy, we would know \( T \). Since we do not have the genealogy, we estimate \( T \) using the genetic variation of the sample at that locus. In statistical terms, we estimate \( T \) conditional on its true value. Thus, in spite of our view that \( T \) is a random variable, in our analysis it becomes an unknown parameter that is estimated on the basis of a sample of current DNA.

**Review of existing methods:** The foundation of the majority of existing methods of estimating TMRCA is Kingman’s (1982) coalescent theory. It determines a probability distribution of TMRCA solely on the basis of the assumed population model, *i.e.*, in the absence of any sequence data. For example, under the standard model of a panmictic population of constant size \( N (N \rightarrow \infty) \), coalescence theory stipulates that the waiting times between consecutive coalescent events are independent exponential random variables with known parameters. When there is also mutation, an important parameter is \( \theta = 4N\mu \) (or \( \theta = 2N\mu \) for a haploid population), where \( \mu \) is the mutation rate per genetic locus. When time is measured in units of \( N \) generations, \( \theta \) is simply twice the mutation rate per unit time. Most existing methods share a common structure with three components:

1. Choose a population model and parameters, which may include population size, growth rate, and migration rate.
2. Formulate a prior distribution of TMRCA according to the coalescent.
3. Update the prior with a likelihood function or summary statistics derived from the data.

In practice, the selection of a population model involves somewhat arbitrary choices because of our ignorance of demographic history; it is also constrained by the computational tools currently available. Thus, these methods vary in their choice of population parameters and the method for evaluating the likelihood function.

One group of methods places prior distributions on one or more of the model parameters and reduces sequence data to one or a few summary statistics, such as the number of segregating sites, heterozygosity, or maximum pairwise number of nucleotide differences. With values of the parameters chosen from their prior distributions, proposal genealogies are simulated using the coalescent. The likelihood is the probability of observing data with identical (or similar) summary statistics to those observed in the sample. These methods often implement rejection sampling, where each proposed genealogy is accepted with a probability proportional to the likelihood. The TMRCA distribution of the accepted genealogies is taken as the posterior distribution. If the model parameters are chosen from prior distributions, the same procedure also produces posterior distributions of population parameters, such as population size, growth rate, etc. Examples of these methods can be found in Tavare et al. (1997) and Pritchard et al. (1999). We call these methods Bayesian approaches because they view \( T \) as a random variable with the prior distribution determined by the coalescent and summarize conclusions in terms of the posterior distribution of \( T \) given the data (or summary statistics of the data).

In contrast, a second group of methods estimates the model parameters on the basis of the likelihood of the complete data. The posterior distribution of TMRCA is computed at the estimated parameter values without regard to the variability due to estimation of unknown parameters. Evaluation of the full likelihood can be accomplished by methods such as importance sampling or Markov chain Monte Carlo. We refer to these methods as “empirical Bayes approaches,” to indicate that the posterior distributions of the TMRCA are computed at the maximum-likelihood estimate of population parameters. Examples of empirical Bayes methods include algorithms implemented in genetree (Griffiths and Tavare 1994; available from http://www.stats.ox.ac.uk/mathgen/software.htm) and the larmarc package (available from http://evolution.genetics.washington.edu/lamarc.html), which includes programs fluctuate (Kuhner et al. 1998), recombine (Kuhner et al. 2000), and migrate (Beerli and Felsenstein 1999). These programs have the advantage of producing maximum-likelihood estimates of model parameters as well as the posterior distribution of the sample TMRCA. At present, however, the convergence behavior of the Markov chain Monte Carlo (MCMC) algorithms is still not completely understood in the context of ancestral processes. We find that these MCMC procedures require considerable experience in the “art” of tuning the MCMC parameters and patience for the substantial computational costs incurred.

An estimator that is similar in certain respects to ours was suggested by Templeton (1993), who exploited expressions for the conditional moments of the coalescent time \( T \) between two randomly sampled sequences given \( k \), the observed number of nucleotide differences between those sequences. The expressions, derived by Tajima (1983), are

\[
E(T_2|k) = \frac{\theta(1 + k)}{2\mu(1 + \theta)} \tag{1}
\]

and

\[
\text{var}(T_2|k) = \frac{\theta^2(1 + k)}{4\mu^2(1 + \theta)^2}. \tag{2}
\]
where \( \theta = 2N\mu \) (haploid model as above). For a sample of \( n \) sequences, Templeton evaluated \( D \), the number of pairwise nucleotide differences averaged over all pairs whose most recent common ancestor is the root of a reconstructed genealogy. Substituting \( D \) for \( k \), he then applied (1) and (2). This method has been criticized on two grounds (Tavaré et al. 1997). First, because the derivation of (1) and (2) uses coalescence theory, Tajima’s results are applicable only for two randomly sampled sequences, while Templeton’s method uses pairs selected to be more different than average. Second, since \( D \) is derived from many dependent pairs of sequences, it has different statistical properties from \( k \), the number of nucleotide differences between two observed sequences.

It is important to recognize that all of these estimators of TMRCA assume knowledge in the form of prior distributions or precise values of population parameters such as effective population size, growth rates, or migration per site, whose most recent common ancestor is the root of a reconstructed genealogy. Substituting \( D \) for \( k \), he then applied (1) and (2). This method has been criticized on two grounds (Tavaré et al. 1997). First, because the derivation of (1) and (2) uses coalescence theory, Tajima’s results are applicable only for two randomly sampled sequences, while Templeton’s method uses pairs selected to be more different than average. Second, since \( D \) is derived from many dependent pairs of sequences, it has different statistical properties from \( k \), the number of nucleotide differences between two observed sequences.

Most of these assumptions are made in other methods as well. As explained in a later section, assumption 4 is not necessary, but is made to simplify the exposition. As long as 4 holds, 2 is not strictly required, provided the total mutation rate for the locus is known. Since we take this rate from phylogenetic estimates, where it is given as a per site rate, and since we want to apply our methods to mtDNA data, where hypothesis 4 is untenable, we are in practice required to make the assumption 2 in some applications.

Point estimate of TMRCA: Our method is motivated by the following observation. As a consequence of the constant molecular clock hypothesis and the infinitely many sites mutation model, the number of nucleotide differences \( d_{ij} \) between any two sequences \( i \) and \( j \) (not necessarily randomly sampled) follows a Poisson distribution of mean \( 2\mu\ell T_z \), where \( \mu \) is the mutation rate per site, \( \ell \) is the length of the sequenced region, and \( T_z = T_z(i, j) \) is the coalescent time of the sequences from their common ancestor measured in the same time units as the mutation rate. In particular, the expected number of nucleotide differences between pairs of sequences in a sample that diverged from a common ancestor is proportional to the TMRCA of the sample. This suggests a frequentist estimator:

\[
\hat{T}_z = \frac{d_{ij}}{2\mu\ell}.
\]

If \( T_z \) is the true coalescence time for a pair of sequences, by basic properties of the Poisson distribution, we have

\[
E(\hat{T}_z | T_z) = T_z
\]

and

\[
var(\hat{T}_z | T_z) = \frac{T_z}{2\mu\ell}.
\]

Thus, \( \hat{T}_z \) is an unbiased estimator of \( T_z \). A plug-in estimator for its variance is simply \( d_{ij}/(4\mu^2\ell^2) \).

To generalize \( \hat{T}_z \) to a sample of \( n \) sequences, we must solve two problems. First, not all pairs of sequences in the sample share the same TMRCA. Second, because the sequences are related through a genealogy, distances between pairs of sequences have complicated joint distributions.

To solve the first problem, we follow Templeton (1993) and partition the sample of sequences into two subsets, corresponding to the two clades that appear to have diverged from the common ancestor of the sample. Denote the two clades as \( \mathcal{L} \) and \( \mathcal{R} \); clade \( \mathcal{L} \) includes sequences \( l_1, l_2, \ldots, l_p \), and clade \( \mathcal{R} \) includes sequences \( r_1, r_2, \ldots, r_q \), where \( p + q = n \). Next, we compute the average pairwise nucleotide differences between \( \mathcal{L} \) and \( \mathcal{R} \). Let

\[
D = \frac{1}{pq} \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{R}} d_{ij}.
\]
The estimator for the TMRCA we propose is

$$
\hat{T} = \frac{D}{2\mu \ell}.
$$

(5)

Keeping in mind that $\mu$, in (1) and (2) equals $\mu \ell$ here, we find it instructive to rewrite Templeton’s estimator as

$$
\hat{T}_{\text{Temp}} = \frac{\theta \ell}{1 + \theta \ell} \left( \frac{1}{2\mu \ell} + \frac{D}{2\mu \ell} \right)
$$

$$
= \frac{\theta \ell}{1 + \theta \ell} \left( \frac{1}{2\mu \ell} + \frac{T}{\ell} \right),
$$

which illustrates that Templeton’s estimator is a linear transformation of our estimator. The shrinkage by a factor of $\theta \ell / (1 + \theta \ell)$ and the offset of $1 / (2\mu \ell)$ arise from Templeton’s use of the assumed constant population size model as a prior.

Partitioning algorithms: In practice, to select $L$ and $R$, one can use any tree-constructing algorithm and take the deepest branch (Felsenstein 1993). The root of the tree may be established by comparison with an outgroup, for example, using chimpanzee for human data. Although the topology of the entire genealogy is seldom known, and different criteria may lead to more than one plausible tree, the two clades that diverge most from the ancestor can often be constructed with less ambiguity. Templeton observed that all of the plausible genealogies in his example agreed on the branching pattern at the root. Our simulation results suggest the same unless the true genealogy is extremely star-shaped.

To analyze the simulated data, which do not include an outgroup sequence, we take advantage of two heuristic ideas to estimate the two clades directly. First, when the underlying genealogy is not star-shaped, the pair of sequences that differ most from each other are most likely to represent the two clades; each of the remaining sequences can be “classified” into one of the clades by its relative similarity to the existing sequences in each clade. Second, in the case of a perfectly star-shaped genealogy (that is, all sequences have evolved independently from the root of the genealogy), the above algorithm may lead to a completely incorrect partition and a biased point estimate. We prefer a random partition that produces a point estimate with less bias than would result from a deterministic partition. To accomplish this, we suggest the following “two-step iterative” partition algorithm, which seems to combine the best features of both these ideas.

1. Compute the Hamming distance matrix $K$ for the sample, where $K_{ij}$ is the number of nucleotide differences between sequences $i$ and $j$.
2. Find a pair $(i^*, j^*)$, such that $K_{i^*, j^*} \geq K_{i,j}$ for all pairs $(i, j)$. Assign $i^*$ and $j^*$ to the opposite clades.
3. Add each of the remaining sequences, in random order, to the clade closer to it, as measured by the mean pairwise Hamming distance of the candidate sequence to the current members of the clades.
4. Compute the average pairwise nucleotide difference, $D$, by Equation 4.
5. Compute the upper triangular matrix, $M_{ij} = \Lambda(K_{ij}, D)$ (i, j = 1, 2, ..., n, i ≥ j), where $\Lambda(x, \lambda)$ is the Poisson probability density function with mean parameter $\lambda$, evaluated at $x$.
6. Compute the likelihood ratio matrix, $L_{ij} = M_{ij} / \max_{i,j}(M_{ij})$ (i ≥ j). If $L_{ij}$ is less than a prespecified threshold (which we set at 0.2), set it to 0.
7. Choose a pair $(i, j)$ according to the multinomial probability proportional to $L_{ij}$ and assign $i$ and $j$ to the opposite clades. This pair of sequences, $i$ and $j$, takes the role of $i^*$ and $j^*$ in step 2.
8. Repeat step 3 to obtain a partition.

For our simulation, we used this algorithm to obtain 20 partitions and took the averaged point estimates and variance estimates over all the partitions.

Sampling variance of $\hat{T}$: The second problem in generalizing $\hat{T}$ in (3) is the correlation between pairwise distances, $d_{ij}$. Although the correlation does not affect the point estimate, it must be accounted for in the computation of the sampling variance of $\hat{T}$:

$$
\sigma^2 = \text{var}(\hat{T}|T) = \frac{\text{var}(D|T)}{4\mu \ell^2}.
$$

(6)

To estimate $\text{var}(D|T)$ from data, we ignore the variability due to the uncertainty of the clade partition. Then

$$
\text{var}(D|T) = \frac{1}{p^2 q^2} \left( \sum_{i \in L} \sum_{j \in R} \text{var}(d_{ij}|T) 
+ 2 \times \sum_{i \in L} \sum_{j \in R} \sum_{i \not\in R} \text{cov}(d_{ij}, d_{ij}|T)
+ 2 \times \sum_{i \in L} \sum_{j \in R} \sum_{j \not\in L} \text{cov}(d_{ij}, d_{ij}|T)
+ 2 \times \sum_{i \in L} \sum_{j \in R} \sum_{i \not\in L, j \not\in R} \text{cov}(d_{ij}, d_{ij}|T) \right).
$$

(7)

Since each $d_{ij}$ is a Poisson random variable, its variance is equal to its mean. Let $\overline{ik}$ be the most recent common ancestor of sequences $i$ and $k$, and let $T_{\overline{ik}}$ be the time between $i$ and $k$ (twice their coalescence time). Given $T_{\overline{ik}}$, the mutation processes from the two sequences, $i$ and $k$, to their most recent common ancestor, $\overline{ik}$, are independent; i.e., $\text{cov}(d_{ij}, d_{ij}, T_{\overline{ik}}, T) = 0$. Hence, for example, for $i, k \in L$, $j \in R$,

$$
\text{cov}(d_{ij}, d_{ij}|T) = \mathbb{E}[(\text{cov}(d_{ij}, d_{ij}|T)|T]
$$

$$
= \mathbb{E}(\text{cov}(d_{ij}|T)|T]
$$

$$
= \mathbb{E}(d_{ij}|T),
$$

where $\mathbb{E}$ denotes the expected value.
so

\[ \text{var}(d_i \mid T) = \mathbb{E}(d_i \mid T), \quad i \in \mathcal{L}, \quad j \in \mathcal{R} \quad (8) \]

\[ \text{cov}(d_i, d_j \mid T) = \mathbb{E}(d_i d_j \mid T), \quad i, k \in \mathcal{L}, \quad j \in \mathcal{R} \quad (9) \]

\[ \text{cov}(d_i, d_j \mid T) = \mathbb{E}(d_i d_j \mid T), \quad i, k \in \mathcal{L}, \quad j, l \in \mathcal{R}. \quad (10) \]

A plug-in estimator for (8) is \( d \). A plug-in estimator for (10) is \( d \), and this can be approximated by counting the number of bases that are the same in each \( i \) and \( k \) and in \( j \) and \( l \) but are different between, say, \( i \) and \( j \). Likewise, a plug-in estimator for (9) is the number of sequences from alternating clades. Thus, we can estimate \( \text{var}(D \mid T) \) by

\[
\hat{\text{var}}(D \mid T) = \frac{1}{p^2 q^2} \left( \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{R}} d_{ij} + 2 \times \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{R}} d_{ij}^2 
+ 2 \times \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{R}} d_{ij}^3 
+ 2 \times \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{R}} d_{ij}^4 \right). \quad (11)
\]

Substituting (11) into (6) gives the sampling variance of \( \hat{T} \). The amount of computation required for the variance calculation is on the order of \( p^2 q^2 \), which should not pose any difficulty unless the sample is very large and the tree is balanced (i.e., \( p = q = n/2 \)). Further, the variance can be calculated sequentially. We observe very fast convergence of the variance estimate by adding sequences from alternating clades.

**Confidence interval for \( T \):** Because \( \hat{T} \) is a weighted sum of Poisson random variables, its sampling distribution is skewed to the right. Our simulations indicate that because the distribution of \( \hat{T} \) is skewed, confidence intervals based on a normal approximation can have incorrect coverage probability in samples with a small number of segregating sites. To correct for the skewness, we utilize a square root transformation. We propose to approximate the sampling distribution of \( \sqrt{T} \) by the normal distribution; i.e.,

\[ \sqrt{T} \sim \mathcal{N}\left(\sqrt{T}, \frac{\sigma^2}{4T}\right). \]

where \( \sigma^2 \) is defined in (6) and estimated in (11). We observe that the MSE of \( \hat{T} \) can be reduced slightly by an adjustment based on the square root transformation. The modified estimator, \( \hat{T}' \), is

\[ \hat{T}' = \hat{T} - \frac{\sigma^2}{4T}, \quad (12) \]

where \( \sigma^2 \) is the sampling variance estimated by (11). Tables 1–4 show that the confidence intervals constructed this way have coverage probabilities close to or higher than the nominal levels, and the bias is often negligible.

**Correction for recurrent mutation:** The exposition so far has assumed the infinitely many sites mutation model. Our simulation program generates sequences according to a model involving two states per site. The pairwise distance, \( d \), is the number of nucleotide differences between sequences \( i \) and \( j \). To account for recurrent mutations, we replace all \( d \) with a corrected distance measure \( d' \), where

\[ d'_i = -\frac{\ell}{2 \log(1 - 2d_i/\ell)} \quad (13) \]

(\text{Nei} 1987). If a sample contains sites with three and more alleles, Kimura’s two parameter model can be used to correct for recurrent mutations (\text{Nei} 1987).

**Computational details:** For the numerical studies reported in the next section, sequence data are simulated on the basis of coalescence. The estimation procedure described above is implemented in MATLAB and is applied to both the simulated data and the real Y chromosome and mtDNA polymorphism. For genetree estimates of TMRCA in the simulations, we input the assumed values of \( \theta \) and use the mean TMRCA of 500,000 sample genealogies. The publicly available version of the program fluctuate does not estimate the sample TMRCA. However, by modifying the source code, we record the TMRCA of all the sample genealogies in the last long chain and use the mean of this distribution as an estimate. The parameters for the MCMC sampling used to estimate TMRCA for the Y and mtDNA sample are 10 short chains of 100,000 steps each and 5 long chains with \( 10^6 \) steps each.

**RESULTS FROM SIMULATED DATA**

**Simulations of a panmictic population:** Figures 1 and 2 display results from two simulation experiments. In the first experiment, 1000 genealogies from a sample...
of 20 haploid individuals are simulated under a model of a panmictic population whose effective size has remained constant at $N = 5000$. Mutation events are superimposed on each genealogy. The mutation model is quite simple: In a region of 15 kb, each site mutates between two states at a rate of $\mu = 2.5 \times 10^{-7}$/generation. These figures are comparable to what is known about mtDNA. We adopted a two-state mutation model rather than a four-state model for computational simplicity. Recurrent mutations do occur in these simulations and are corrected by (13).

The second experiment is done in a similar fashion: Each genealogy of 20 sequences is simulated assuming that the panmictic population has grown exponentially at a rate of $g = 1 \times 10^{-5}$/generation with a present effective population size of $N_0 = 4 \times 10^4$. The mutation rate and sequence length correspond to what is known about the Y chromosome: $\ell = 50$ kb and $\mu = 2.5 \times 10^{-5}$/generation (Thomson et al. 2000). The vertical bars in Figures 1 and 2 are the 95% confidence intervals.

In the same fashion, we have simulated genealogies under a panmictic population model with different values of the mutation rate and sequence length (Table 1), effective population size (Table 2), growth rate (Table 3), and the sample size (Table 4). Under each parameter set, we evaluate the performance of the estimator with two summary statistics: bias and coverage probability. For a measure of (relative) bias, we compute

$$\text{Bias} = \frac{1}{I} \sum_{i=1}^{I} \frac{\hat{T}_i - T_i}{T_i},$$

where $I$ is the number of genealogies simulated in an experiment. Coverage probability is defined as the proportion of runs in which the confidence interval covers the true $T$. Each of Tables 1–4 varies one aspect of the mutation model, the population model, or the number of samples, while keeping all other aspects fixed.

Several trends become evident from the simulation results. First, the bias of the estimate becomes nonnegligible in two situations. In one, the mutation rate is low or the sequenced region is short (Table 1). In the other, the present population has resulted from a small founding population that has gone through rapid growth (Table 3). In this case, the genealogy tends to be short and star-shaped, and nearly all branches evolve independently. In both situations, the resulting DNA samples

---

**TABLE 1**

<table>
<thead>
<tr>
<th>$\ell$</th>
<th>$\mu$</th>
<th>$N_0$</th>
<th>$g$</th>
<th>$T$</th>
<th>Coverage</th>
<th>Bias</th>
<th>$k$</th>
<th>$H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000</td>
<td>$2.5 \times 10^{-4}$</td>
<td>$4 \times 10^5$</td>
<td>0.001</td>
<td>6483</td>
<td>0.99</td>
<td>0.140</td>
<td>19</td>
<td>2.60</td>
</tr>
<tr>
<td>1,000</td>
<td>$2.5 \times 10^{-7}$</td>
<td>$4 \times 10^5$</td>
<td>0.001</td>
<td>6501</td>
<td>0.99</td>
<td>0.146</td>
<td>19</td>
<td>2.60</td>
</tr>
<tr>
<td>50,000</td>
<td>$2.5 \times 10^{-4}$</td>
<td>$4 \times 10^5$</td>
<td>0.001</td>
<td>6504</td>
<td>0.973</td>
<td>0.00163</td>
<td>95</td>
<td>12.85</td>
</tr>
<tr>
<td>5,000</td>
<td>$2.5 \times 10^{-7}$</td>
<td>$4 \times 10^5$</td>
<td>0.001</td>
<td>6512</td>
<td>0.977</td>
<td>0.00244</td>
<td>94</td>
<td>12.85</td>
</tr>
</tbody>
</table>

$\ell$ is the length of the sequenced region; $\mu$ is the per site mutation rate; $N_0$ is the present effective population size; $g$ is the growth rate per generation; $T$ is the mean simulated TMRCA; coverage is the estimated coverage probability of the 95% confidence interval; bias is defined in (14); $k$ is the number of segregating sites; and $H$ is the mean total heterozygosity among the segregating sites. The quantities $T$, bias, $H$, and $k$ are averaged over 1000 simulated genealogies for a given set of parameters. Twenty sequences are simulated in each genealogy.
TABLE 3
Coverage probability of the confidence intervals and average bias of the point estimates as functions of the population growth rate, $g$

<table>
<thead>
<tr>
<th>$g$</th>
<th>$T$</th>
<th>Coverage</th>
<th>Bias</th>
<th>$k$</th>
<th>$H$</th>
<th>$N_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7.66 $\times 10^4$</td>
<td>0.94</td>
<td>$-6.11 \times 10^{-4}$</td>
<td>3,392</td>
<td>908</td>
<td>$4 \times 10^4$</td>
</tr>
<tr>
<td>0.0005</td>
<td>11,583</td>
<td>0.948</td>
<td>$-0.0068$</td>
<td>158</td>
<td>22.64</td>
<td>1,402</td>
</tr>
<tr>
<td>0.001</td>
<td>6,504</td>
<td>0.973</td>
<td>0.00163</td>
<td>95</td>
<td>12.85</td>
<td>675</td>
</tr>
<tr>
<td>0.002</td>
<td>3,600</td>
<td>0.988</td>
<td>0.040</td>
<td>95</td>
<td>12.85</td>
<td>344</td>
</tr>
<tr>
<td>0.005</td>
<td>1,619</td>
<td>0.988</td>
<td>0.143</td>
<td>27</td>
<td>3.32</td>
<td>140</td>
</tr>
</tbody>
</table>

Other population parameters are fixed: $\ell = 50,000$ bp; $\mu = 2.5 \times 10^{-8}$; $N_e = 4 \times 10^5$. The quantities $T$, bias, $H$, and $k$ are averaged over 1000 simulated genealogies for a given set of parameters. The column labeled $N_f$ gives the extrapolated values of the population size at $T$. Twenty sequences are simulated in each genealogy.

TABLE 4
Coverage probability of the confidence intervals and average bias of the point estimates as functions of number of sequences included in the sample

<table>
<thead>
<tr>
<th>$n$</th>
<th>$T$</th>
<th>Coverage</th>
<th>Bias</th>
<th>$\sqrt{\text{MSE}}$</th>
<th>$w$</th>
<th>$k$</th>
<th>$H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>6423</td>
<td>0.972</td>
<td>0.00100</td>
<td>1043</td>
<td>4522</td>
<td>56</td>
<td>12.21</td>
</tr>
<tr>
<td>20</td>
<td>6504</td>
<td>0.973</td>
<td>0.00163</td>
<td>926</td>
<td>4105</td>
<td>95</td>
<td>12.85</td>
</tr>
<tr>
<td>50</td>
<td>6529</td>
<td>0.981</td>
<td>$7.97 \times 10^{-4}$</td>
<td>820</td>
<td>3832</td>
<td>185</td>
<td>13.51</td>
</tr>
<tr>
<td>100</td>
<td>6914</td>
<td>0.969</td>
<td>$-0.00159$</td>
<td>848</td>
<td>3783</td>
<td>298</td>
<td>13.47</td>
</tr>
</tbody>
</table>

$n$ is the number of sequences in the sample; $\sqrt{\text{MSE}}$ denotes the square root of the mean squared error averaged over all simulated genealogies; $w$ is the mean width of the 95% confidence intervals. Other population parameters are fixed: $\ell = 50,000$; $\mu = 2.5 \times 10^{-8}$; $N_e = 4 \times 10^5$; $g = 0.001$. The quantities $T$, bias, $H$, and $k$ are averaged over 1000 simulated genealogies for a given set of parameters.
only observe DNA sequences, but we also know the positions of each mutation relative to the coalescent events. A schematic genealogy is shown in Figure 4. The true times between two consecutive coalescence events are \( W_j \), indexed by the number of lineages existing during that time interval. We do not observe \( W_j \), but we do know the number of mutations that occurred during time interval \( W_j \). By coalescence theory under the constant population model, estimation of TMRCA amounts to estimating each \( W_j (j = 2, 3, \ldots, n - 1) \) independently, on the basis of \( S_j \), the number of mutations occurring in the corresponding time interval. Consider a Bayesian estimator and assume the population size, \( N \), is known without error. The prior distribution of \( W_j \), based on coalescent theory (Tavare’ et al. 1997), is

\[
\text{Prior}(W_j) \sim \exp\left(\frac{j(j-1)}{2}\right).
\]

Modeling the mutation as a Poisson process, we write the likelihood of observing \( S_j = s_j \) mutations during time \( W_j \) as

\[
P(S_j = s_j | W_j = w_j) = \frac{e^{-1/2\mu w}}{s_j!} ((1/2) \theta w)^s_j.
\]

The posterior distribution of \( W_j \) can be derived by Bayes theorem,

\[
P(W_j | S_j) = \frac{\text{Prior}(W_j) P(S_j | W_j)}{\int \text{Prior}(w) P(S_j | w) dw},
\]

and is a gamma distribution:

\[
\mathcal{L}(W_j | S_j) = \Gamma\left(1 + S_j, \frac{2}{j(j-1+\theta)}\right).
\]

Let \( \hat{W}_j \) be a Bayes estimator taken as the mean of the posterior distribution of \( W_j \). We have

\[
\hat{W}_j = \frac{2(1 + S_j)}{j(j-1+\theta)}.
\]
The MSE associated with $W_j$ is

$$\text{MSE}(W_j) = \int \sum_{w=0}^{\infty} (\tilde{w} - w)^2 P(s|w) \times \text{Prior}(w) dw$$

$$= \frac{4}{j^2(j - 1 + \theta)(j - 1)}, \quad (15)$$

Now we consider a frequentist estimator, $\hat{W}_j$, which is based only on the number of mutations in each time interval of duration $W_j$ and is defined by

$$W_j = \frac{S_j}{\theta/2j}.$$  

The MSE associated with this estimator is

$$\text{MSE}(\hat{W}_j) = \int \sum_{w=0}^{\infty} (\tilde{w} - w)^2 P(s|w) \times \text{Prior}(w) dw$$

$$= \frac{4}{j^3\theta(j - 1)}.$$

(16)

For the entire genealogy with $n$ sequences, let $T = \sum_{j=2}^{n} W_j$ and $\hat{T} = \sum_{j=2}^{n} \hat{W}_j$. Then

$$\text{MSE}(T) = \sum_{j=2}^{n} \text{MSE}(W_j),$$

$$\text{MSE}(\hat{T}) = \sum_{j=2}^{n} \text{MSE}(\hat{W}_j)$$

since all cross terms are 0.

Comparing (16) with (15), we see that $\text{MSE}(\hat{W}_j) < \text{MSE}(W_j)$ for each $j$ and all values of $\theta$. This is expected, as a Bayes estimator minimizes MSE under the assumed model. The difference between the two MSE terms, however, is

$$\Delta(j) = \frac{4}{(j - 1 + \theta)j^2} \rightarrow 0 \quad \text{for large } \theta. \quad (17)$$

The computation so far establishes lower bounds on the MSE of a Bayesian and a frequentist estimator, respectively. Both MSE terms converge to a finite limit as $n \rightarrow \infty$, while the difference converges to 0 as $\theta \rightarrow \infty$.

Our next task is to examine how close to the lower bound one gets by using the existing Bayesian or our frequentist estimator. We simulated genealogies and DNA sequences under a specific population model. For the Bayesian estimate, we used genetree with assumed values of $\theta$, and the estimate is based on 500,000 sampled genealogies. Table 5 indicates that both the Bayesian and frequentist estimators approach their respective theoretical lower bounds for large $\theta$. The discrepancy between the MSE of our estimator, $\hat{T}$, and the theoretical lower bound is due to our imperfect knowledge of the position of each mutation relative to the coalescence events. The discrepancy between the MSE of the gene tree estimator and the Bayesian theoretical lower bound is due to two factors: imperfect knowledge of the position of each mutation relative to the coalescence events and the sampling variance due to the finite number of trees sampled by the program. The simulation study in Stephens (2001) shows that millions of genealogies are required to reduce this latter source of variation. Finally, the value of $\theta = 2N\mu$ is seldom known. In fact, a primary goal of genetree is to estimate $\theta$ empirically. This additional source of uncertainty would increase the MSE associated with TMRCA from the genetree estimator.

### TABLE 5

<table>
<thead>
<tr>
<th></th>
<th>Simulation</th>
<th>Theoretical lower bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$</td>
<td>Bayesian</td>
<td>Frequentist</td>
</tr>
<tr>
<td>2.5</td>
<td>3309</td>
<td>4128</td>
</tr>
<tr>
<td>12.5</td>
<td>1830</td>
<td>1787</td>
</tr>
</tbody>
</table>

All genealogies are simulated under a constant population model. Simulated MSE for a Bayesian estimator is obtained from genetree with the shown values of $\theta$.

### APPLICATIONS TO Y-CHROMOSOMAL DNA AND mtDNA

As examples, we apply our method to the worldwide samples of Y chromosome and mtDNA we have recently sequenced.

Y chromosome TMRCA: The Y chromosome data are from $n = 108$ individuals sampled worldwide. A region of 69,000 bp has been screened, and 114 single nucleotide polymorphisms (SNPs) have been found. The assumed mutation rate is derived from the divergence between chimpanzees and humans in the corresponding genomic region:

$$\mu = \frac{2T_{\text{chimp}} \times \text{human}}{\text{substitutions between chimp and human}}.$$

Assuming a divergence time of 5 million years between chimpanzees and humans and a constant generation length of 25 years, the mutation rate is $\sim 3 \times 10^{-9}/\text{site} \times \text{generation}$ (Thomson et al. 2000). Under a model with constant population size and panmixia, the program fluctuates produces an estimate of 10,300 for the effective population size and 117,000 years for the worldwide TMRCA. If we assume an exponential growth model, however, the same program estimates the rate of growth to be $1.3 \times 10^{-3}$, the current effective population size to be 22,400, and the worldwide TMRCA to be 75,000 years (H. Tang, R. Thomson, L. L. Cavalli-Sforza, P. Shen, P. J. Oefner and M. W. Feldman, unpublished results). Clearly, a Bayesian approach depends heavily on the assumed population model. In the application of our approach, the clades are formed on the basis of the haplogroup definitions in Underhill et al. (2000). In particular, for the worldwide TMRCA, one clade is represented by the five San and Pygmy sam-
TABLE 6
Estimated Y chromosome TMRCA for each continent and for the world

<table>
<thead>
<tr>
<th>Region</th>
<th>Sample size</th>
<th>$k$</th>
<th>$H$</th>
<th>$\hat{T}_{25}$ (10$^3$ yr)</th>
<th>$\hat{T}_{20}$ (10$^3$ yr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>38</td>
<td>98</td>
<td>7.86</td>
<td>87 (57, 125)</td>
<td>104 (68, 150)</td>
</tr>
<tr>
<td>Europe</td>
<td>18</td>
<td>23</td>
<td>5.67</td>
<td>48 (26, 80)</td>
<td>58 (31, 96)</td>
</tr>
<tr>
<td>Asia</td>
<td>37</td>
<td>46</td>
<td>4.75</td>
<td>37 (21, 58)</td>
<td>44 (25, 70)</td>
</tr>
<tr>
<td>Oceania</td>
<td>8</td>
<td>17</td>
<td>4.69</td>
<td>39 (20, 68)</td>
<td>47 (24, 82)</td>
</tr>
<tr>
<td>America</td>
<td>7</td>
<td>11</td>
<td>2.86</td>
<td>34 (11, 53)</td>
<td>41 (13, 64)</td>
</tr>
<tr>
<td>World</td>
<td>108</td>
<td>114</td>
<td>7.49</td>
<td>91 (60, 130)</td>
<td>109 (72, 156)</td>
</tr>
</tbody>
</table>

Numbers in the column labeled $\hat{T}_{25}$ are the estimated TMRCA assuming a generation length of 25 years; those in the column labeled $\hat{T}_{20}$ are the estimated TMRCA assuming a generation length of 30 years. Parenthetical entries are 95% confidence intervals.

The coalescent time of the African $Y$ chromosomal sequences is quite similar to that of the whole world, but much older than that of other continents, providing support for the out of Africa hypothesis. That the values for the estimated TMRCA of all non-African continents are similar does not necessarily suggest simultaneous settlement on these continents; rather, it may be a consequence of population bottlenecks. In fact, the TMRCA represents only a lower bound of the founding date and can be substantially more recent than the time of the settlement.

mtDNA TMRCA: The mtDNA data are from 179 individuals sampled worldwide. The entire mitochondrial genome, except the hypermutating D-loop, has been sequenced, and 971 SNPs have been found. The mutation rate has been derived from the divergence between chimpanzees and humans in the corresponding genomic region. Because of the relatively high mutation rate, we used Kimura’s two-parameter model to correct for recurrent mutations (Nei 1987). The average mutation rate is $\sim 2.43 \times 10^{-7}/(\text{site} \times \text{generation})$. The sample heterozygosity is $\sim 39$. For the tree-partition step, we used dnapenny. The estimated TMRCA for individual continents and for the world are given in Table 7.

As with the estimates for the $Y$ chromosome TMRCA, the worldwide mitochondrial TMRCA is comparable to that of the African TMRCA, while mtDNA from populations on all other continents coalesces much more recently. It is also interesting to note that the TMRCA for European mtDNA populations is estimated to be smaller than that for Asian, Oceanic, and American populations, which may indicate a small female founding population in Europe, combined with a high population growth rate in that continent. More on this is in the DISCUSSION.

TABLE 7
Estimated mtDNA TMRCA for each continent and for the world

<table>
<thead>
<tr>
<th>Region</th>
<th>Sample size</th>
<th>$k$</th>
<th>$H$</th>
<th>$\hat{T}_{25}$ (10$^3$ yr)</th>
<th>$\hat{T}_{20}$ (10$^3$ yr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>52</td>
<td>463</td>
<td>55.65</td>
<td>238 (203, 276)</td>
<td>190 (162, 221)</td>
</tr>
<tr>
<td>Europe</td>
<td>55</td>
<td>250</td>
<td>17.70</td>
<td>69 (54, 87)</td>
<td>22 (43, 70)</td>
</tr>
<tr>
<td>Asia</td>
<td>37</td>
<td>316</td>
<td>26.47</td>
<td>103 (85, 128)</td>
<td>84 (68, 102)</td>
</tr>
<tr>
<td>Oceania</td>
<td>16</td>
<td>133</td>
<td>25.98</td>
<td>108 (87, 133)</td>
<td>86 (70, 106)</td>
</tr>
<tr>
<td>America</td>
<td>17</td>
<td>108</td>
<td>22.34</td>
<td>106 (81, 135)</td>
<td>85 (65, 108)</td>
</tr>
<tr>
<td>World</td>
<td>179</td>
<td>971</td>
<td>39.5</td>
<td>238 (200, 281)</td>
<td>190 (160, 225)</td>
</tr>
</tbody>
</table>

Numbers in the column labeled $\hat{T}_{25}$ are the estimated TMRCA assuming a generation length of 25 years; those in the column labeled $\hat{T}_{20}$ are the estimated TMRCA assuming a generation length of 20 years. Parenthetical entries are 95% confidence intervals.

DISCUSSION

In the construction of the variance of $\hat{T}$, we assume that the two clades are defined without error. In our simulation, we used a rather naive partitioning algorithm to reduce computation and to facilitate automation. One might have expected that the confidence intervals based on (7) would be too narrow due to the uncertainties in the clade-defining step. On the con-
trary, for our simulated data, the estimate and the confidence interval appear to be relatively robust. We reason that a branch that would be placed in the wrong clade is likely to have diverged from the rest of the sequences near the root; hence the point estimate and the confidence interval are not greatly affected. One can make a bias-variance tradeoff argument: A slightly mistaken but more balanced partition produces a small bias, but reduces the variability of the estimate. We do not pursue an optimal partitioning algorithm in this study.

Effects of population subdivision: So far, our simulation has dealt with panmictic populations only. A natural question is how our method performs when the sampled population is subdivided. We argue that, unless the generation length is subpopulation dependent, the constant molecular clock assumption is independent of population subdivision; therefore, this frequentist method should be valid in the presence of population subdivision. Our simulations in one-population (panmictic) situations indicate that the performance of our estimator depends on the shape of the underlying genealogy and the number of segregating sites, but does not depend on the specific mechanisms that generate such genealogies. In fact, since subdivided populations tend to generate genealogies with very long branches near the root, the sample partition tends to be more accurate. Hence, we hypothesize that the estimates of TMRCA should have smaller biases in these situations. We performed a few simulations under a model that incorporates population subdivision and migration. The goal of these simulations was not to prove the validity of our method in a subdivided population; rather, it is to provide some evidence supporting our intuitions.

A population model with subdivision and migration: We performed simulations in which sample genealogies are generated as follows: An ancient population of size $N_0(T_1 + T_2)$ began to grow $T_1 + T_2$ generations ago. At time $T_3$ generations ago, this population, now of size $N_3(T_3)$, was split into two subpopulations, of size $N_1(T_3) = f_1N_0(T_3)$ and $N_2(T_3) = f_2N_0(T_3)$. The two subpopulations each underwent exponential growth, at rates $g_1$ and $g_2$, respectively. At the end of each generation, a fraction $m_1$ of subpopulation 1 migrates to subpopulation 2, while a fraction $m_2$ of subpopulation 2 migrates to subpopulation 1. The ancestral population remains at the constant size, $N_0(T_1 + T_2)$. Figure 5 plots the true TMRCA vs. the estimated TMRCA. Specifically, $T_1 = 500$, $T_2 = 2000$, $m_1 = 0.0005$, $m_2 = 0.0001$, $g_1 = 0.005$, $g_2 = 0.001$, $f_1 = 0.003$, $f_2 = 0.75$, $f_3 = 0.25$, $N_0(T_1 + T_2) = 300$, $L = 15,000$ bp, and $\mu = 2.5 \times 10^{-7}$/site $\times$ generation. The mean simulated TMRCA is 2505 generations; on average each simulated data set contains 120 segregating sites. The bias defined in (14) over 1000 genealogies is 2% and the coverage probability is 96.5%. Thus, the method seems to perform well even in the presence of population structure.

The robustness of the estimator does break down in some situations. For a diagnostic test, we suggest using the sample heterozygosity. The coverage of the confidence interval is reasonably accurate when the sample heterozygosity is at least 15 and exceeds the nominal level otherwise. The relative bias of the point estimator is also reasonably small and becomes very small when the sample heterozygosity exceeds 5. Smaller values of $H$ indicate a lack of information in the data. In these cases the point estimator of the TMRCA is likely to be positively biased. To reduce bias, we recommend using a true tree-building algorithm instead of our naive algorithm to partition the sample into left and right clades.

At least for human populations, since rapid population growth is likely to be a recent phenomenon and since worldwide populations have experienced repeated isolation and admixture, we believe that the genealogy underlying a real sample is unlikely to have the star shape, where our method shows its poorest behavior, unless the sampled individuals are closely related.

Limitations of the frequentist estimator: As one may suspect, the chief weakness of this frequentist estimator is its sensitivity to misspecification of the mutation model. Unless the rate of recurrent mutation is very high, our frequentist estimate of TMRCA varies linearly with the estimate of the mutation rate. The uniform independent mutation model assumed here is clearly not realistic. It is possible to generalize our estimator to allow for different transition-transversion rates or to model a higher mutation rate at third codon positions. In estimating the divergence time of two species, even the validity of the constant molecular clock hypothesis can be challenged. The real problem is, however, that our current knowledge of the mutation process is limited. One could argue that, in cases where researchers have more faith in their knowledge of population histories
than they do in the estimates of mutation rate, Bayesian estimators can be more robust. Nonetheless, we believe that, with the availability of comparative genomes, our understanding of the mutation process will advance at a faster rate than our understanding of the demographic histories of the studied populations.

The simulation results in Tables 1 and 4 allow us to consider a question related to the sampling scheme. To acquire more information regarding a population TMRCA, we may elect to sequence a longer region for each individual already in the sample, or we may fix the size of the sequenced region but sample more individuals. Which scheme is more efficient? The answer seems to be the former. Equations 15 and 16 indicate that the MSE decreases as \( \theta \) increases, and \( \theta \) is proportional to the sequence length. Coalescent theory shows that if our sample is already of reasonable size, additional individuals are likely to be closely related to individuals already sequenced and thus would add little information regarding the TMRCA. These observations are also supported by the simulation results shown in Tables 1 and 4. However, in reality our methods estimate the sample TMRCA, not the population TMRCA [although in panmictic populations these are the same with probability close to one (Saunders et al. 1984)]. A small convenience sample may not fully represent the population diversity. When a population is subdivided, to increase the confidence that the sample TMRCA is close to the population TMRCA, we should sample additional sequences from underrepresented subpopulations.

Implications of the estimated TMRCA for human populations: Comparison between the estimates in Table 6 and those in Table 7 reveals that the Y chromosome TMRCA, of the world and on each continent, are much younger than their mtDNA counterparts. H. Tang, R. Thomson, L. L. Cavalli-Sforza, P. Shen, P. J. Oefner and M. W. Feldman (unpublished results) provide an explanation that is based on the disparity between the male and female effective population sizes. The estimates of TMRCA for the Y chromosome and mtDNA populations are puzzling in another way: How could the worldwide Y chromosome TMRCA be younger than that of the mtDNA of all continents: Africa, Asia, Oceanic, and America? At first sight, this seems to imply that the most recent mtDNA common ancestors for all continents (except for Europe) resided in Africa. One could certainly claim that all the confidence intervals overlap, and therefore the worldwide Y chromosome TMRCA could still be older than the non-African mtDNA TMRCA. Alternatively, as explained before, these estimates are sensitive to the errors in the estimated mutation rate. A small error in mutation rate will amplify to a much larger discrepancy in the estimated TMRCA. Thus, a slightly underestimated mtDNA mutation rate can account for the observation. However, we favor a simple explanation based on the unequal generation length between the human male and female populations. Because of the physiological constraints and social customs, it has been observed that the average generation length can be longer for the male than the female population (Cavalli-Sforza et al. 1964; Tremblay and Vezina 2000). In our analyses, we used a generation length of 25 years for both the male and the female populations. Our paradox can be reconciled if we assume a 25-year generation length for the females, but 30 years per generation for the males. As seen in Table 6, the adjusted worldwide Y population TMRCA then becomes 104,000 years; similar adjustment results in estimates of 40,000—60,000 years for the non-African continents.

Finally, it is tempting to associate the TMRCA of a sample collected on a single continent with the date of settlement of that continent. This can be misleading for two reasons. On the one hand, many of the founding human populations are likely to have been small and to have gone through expansion; hence the population TMRCA of a continent can be much more recent than the date of settlement. Second, the MRCA of the present population may be much more ancient than the founding population due to subsequent migration events. For example, in the analysis of the European Y chromosome samples, we detected two sequences that represent haplotypes that are found almost exclusively in Africa. These were not included in our analysis of the European sequences. The mtDNA of both of these individuals resemble those of other European mtDNA sequences. We suspect that there have been migrations from Africa in the male lineages of these individuals. Because of the lack of recombination on the Y chromosome, the whole African Y chromosome has been preserved. Inclusion of these two Y chromosome sequences would have increased the European Y chromosome TMRCA considerably. We detected these two sequences using the jackknife method, repeating the estimation procedure leaving one sequence out each time. However, when the sample is small, such as in the case of the American samples, it is often impossible to detect "outliers."

CONCLUSION

We have introduced a new method of estimating TMRCA on the basis of a sample of DNA sequences. Unlike many existing methods, this estimator does not require assumptions on the demographic model and associated parameters. Construction of point estimates and confidence intervals is simple and fast. Simulation studies have been performed under simple population models. The MSE calculation and (17) suggest that, with a sufficient amount of data, our estimator, which requires no knowledge of the population model, achieves similar predictive power to a Bayesian approach that uses exact knowledge of the demographic parameters.
We gratefully acknowledge Dr. Mary Kuhner’s advice on obtaining TMRCA estimates with fluctuate. We thank Luca Cavalli-Sforza, Neil Risch, Noah Rosenberg, and Russell Thomson for helpful discussions. Two anonymous reviewers provided useful comments and criticisms. Hua Tang is supported by a Howard Hughes predoctoral fellowship. Research was supported in part by the National Institutes of Health grants GM28016 and GM28428 to M.W.F. and by the National Science Foundation grant DMS-0072523 to D.O.S.

LITERATURE CITED


Communicating editor: J. B. Walsh