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

Hua Tang,* David O. Siegmund,*,[1] 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.

FOR a sample of DNA sequences that exhibits variation, a quantity of interest is the time until all of the sequences coalesce. This quantity is termed time to the most recent common ancestor (TMRCA). The age of the common ancestor indicates the degree to which the sample sequences relate to one another. Strictly speaking, we aim to estimate the TMRCA of the sample, which may not coincide with that of the underlying population. This issue is discussed further in the next section. Because unlinked loci have different genealogies, this ancestor is defined for a specific genetic locus. In studies of human populations, two loci have received great attention because of their implications for the demographic histories of the male and female human populations: the nonrecombining region of the Y chromosome and mitochondrial DNA (mtDNA), respectively. In addition, we may imagine a situation in which the sample consists of sequences sharing some unique event polymorphism (UEP). In this case, the TMRCA represents a lower bound for the age of the mutation that defines this polymorphism. Other applications of TMRCA are illustrated in Takahata *et al.* (2001) and Ruvolo (1996).

It is, therefore, desirable to have an estimator of TMRCA whose statistical properties are well understood. In the next section, we review some of the methods that have been developed recently. Most of these methods make assumptions about the population structure, *e.g.*, population size and rate of growth, amount of immigration, etc., as well as the mutation process. The robustness of these methods against model misspecification has not been characterized. A recent article by Stumpf and Goldstein (2001) advocates a model-free approach in population studies. Unfortunately, their "model-free" approach actually assumes a very special model, namely a star-shaped genealogy. Our goal in this article is to find an estimator of TMRCA that does not assume that the population parameters are known and whose statistical properties can be assessed easily. We propose an inferential approach to TMRCA that explicitly models the mutation process but not the population history. Simulations demonstrate that the estimate is relatively unaffected by some important components of population structure.

[1]*Corresponding author:* Department of Statistics, Sequoia Hall, 390 Serra Mall, Stanford University, Stanford, CA 94305-4065. E-mail: dos@stat.standford.edu

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 $\sim(n-1)/(n+1)$ (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) coalescence 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 $\theta = 4N\mu_1$ (or $\theta = 2N\mu_1$ for a haploid population), where μ_1 is the mutation rate per genetic locus. When time is measured in units of N generations, θ 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 Tavaré *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 Tavaré 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_2 between two *randomly* sampled sequences given k, the observed number of nucleotide differences between those sequences. The expressions, derived by Tajima (1983), are

$$\mathbf{E}(T_2|k) = \frac{\theta(1 + k)}{2\mu_1(1 + \theta)} \tag{1}$$

and

$$\mathbf{var}(T_2|k) = \frac{\theta^2(1 + k)}{4\mu_1^2(1 + \theta)^2}, \tag{2}$$

where $\theta = 2N\mu_1$ (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 rates, which are difficult to estimate accurately. Taking into consideration the uncertainty of the estimated parameter values will increase the sampling variance of the TMRCA estimators. This is demonstrated by the Y chromosome microsatellite analysis of Pritchard *et al.* (1999), which concludes that the TMRCA for the worldwide Y population may range between 16,000 and 126,000 years.

Worse yet, most populations have gone through periods of expansion and bottleneck. The existing parametric models are quite restrictive and may fail to describe the population history adequately. In particular, genetree, fluctuate, migrate, and micsat employ only a few greatly simplified models. For example, both genetree and fluctuate assume a constant and deterministic rate of exponential growth; migrate assumes constant population size and constant migration rate with respect to time. A large number of simulations would be required to investigate the bias and variance associated with those estimators of TMRCA when the underlying population model deviates from those allowed in these programs. We are not aware of any such study.

We now describe a new method that is explicit about the mutation model but not the population model.

Model and assumptions: We make the following assumptions on the mutation process:

1. Mutations occur as a Poisson process with constant rate along all branches (*i.e.*, constant molecular clock).
2. The mutation rate is the same known constant at all sites.
3. Mutation is independent among all sites.
4. The mutation rate is sufficiently low that the infinitely many sites model is a reasonable approximation (*i.e.*, there are no recurrent mutations).

In addition, we assume the following:

5. The sample consists of fully sequenced regions of DNA. There is no recombination.
6. Generations are nonoverlapping and of constant length.

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 with mean $2\mu\ell T_2$, where μ is the mutation rate per site, ℓ is the length of the sequenced region, and $T_2 = T_2(i, j)$ 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:

$$\hat{T}_2 = \frac{d_{ij}}{2\mu\ell}. \tag{3}$$

If T_2 is the true coalescence time for a pair of sequences, by basic properties of the Poisson distribution, we have

$$\mathbf{E}(\hat{T}_2 \mid T_2) = T_2$$

and

$$\mathbf{var}(\hat{T}_2 \mid T_2) = \frac{T_2}{2\mu\ell}.$$

Thus, $\hat{T}_2$ is an unbiased estimator of T_2. A plug-in estimator for its variance is simply $d_{ij}/(4\mu^2\ell^2)$.

To generalize $\hat{T}_2$ 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

$$D = \frac{1}{pq} \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{R}} d_{ij}. \tag{4}$$

The estimator for the TMRCA we propose is

$$\hat{T} = \frac{D}{2\mu\ell}. \tag{5}$$

Keeping in mind that μ_1 in (1) and (2) equals $\mu\ell$ here, we find it instructive to rewrite Templeton's estimator as

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

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 $\mathcal{L}$ and $\mathcal{R}$, 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^0, by Equation 4.
5. Compute the upper triangular matrix, $M_{ij} = \Lambda(K_{ij}, D^0)$ $(i, j = 1, 2, \ldots, n, i \geq j)$, where $\Lambda(x, \lambda)$ is the Poisson probability density function with mean parameter λ, evaluated at x.
6. Compute the likelihood ratio matrix, $L_{ij} = M_{ij}/\max_{k,l}(M_{kl})$ $(i \geq j)$. 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 T: The second problem in generalizing $\hat{T}_2$ 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 $\hat{T}$:

$$\sigma^2 = \mathbf{var}(\hat{T}|T) = \frac{\mathbf{var}(D|T)}{4\mu^2\ell^2}. \tag{6}$$

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

$$\begin{aligned}\mathbf{var}(D|T) = \frac{1}{p^2q^2}\Bigg(&\sum_{i\in\mathcal{L}}\sum_{j\in\mathcal{R}}\mathbf{var}(d_{ij}|T)\\ &+ 2\times\sum_{i,k\in\mathcal{L},i\neq k}\sum_{j\in\mathcal{R}}\mathbf{cov}(d_{ij}, d_{kj}|T)\\ &+ 2\times\sum_{i\in\mathcal{L}}\sum_{j,l\in\mathcal{R},j\neq l}\mathbf{cov}(d_{ij}, d_{il}|T)\\ &+ 2\times\sum_{i,k\in\mathcal{L},i\neq k}\sum_{j,l\in\mathcal{R},j\neq l}\mathbf{cov}(d_{ij}, d_{kl}|T)\Bigg).\end{aligned} \tag{7}$$

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_{ik} be the time between i and k (twice their coalescence time). Given T_{ik}, the mutation processes from the two sequences, i and k, to their most recent common ancestor, $\overline{ik}$ are independent; *i.e.*, $\mathbf{cov}(d_{\overline{ik},i}, d_{\overline{ik},k}|T_{ik}, T) = 0$. Hence, for example, for $i, k \in \mathcal{L}, j \in \mathcal{R}$

$$\begin{aligned}\mathbf{cov}(d_{ij}, d_{kj}|T) &= \mathbf{E}[\mathbf{cov}(d_{ij}, d_{kj}|T_{ik}, T)|T]\\ &= \mathbf{E}[\mathbf{E}(_{\overline{ik},j}|T_{ik}, T)|T]\\ &= \mathbf{E}(d_{\overline{ik},j}|T),\end{aligned}$$

so

$$\mathbf{var}(d_{ij}|T) = \mathbf{E}(d_{ij}|T), \quad i \in \mathcal{L}, \quad j \in \mathcal{R} \tag{8}$$

$$\mathbf{cov}(d_{ij}, d_{kj}|T) = \mathbf{E}(d_{\bar{ik},j}|T), \quad i, k \in \mathcal{L}, \quad j \in \mathcal{R} \tag{9}$$

$$\mathbf{cov}(d_{ij}, d_{kl}|T) = \mathbf{E}(d_{\bar{ik},\bar{jl}}|T), \quad i, k \in \mathcal{L}, \quad j, l \in \mathcal{R}. \tag{10}$$

A plug-in estimator for (8) is d_{ij}. A plug-in estimator for (10) is $d_{\bar{ik},\bar{jl}}$, 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 sites that are the same in sequences i and k, but are different in sequence j. Thus, we can estimate $\mathbf{var}(D|T)$ by

$$\widehat{\mathbf{var}}(D|T) = \frac{1}{p^2q^2}\left(\sum_{i\in\mathcal{L}}\sum_{j\in\mathcal{R}} d_{ij} + 2\times\sum_{i\neq k,i,k\in\mathcal{L}}\sum_{j\in\mathcal{R}} d_{\bar{ik},j} + 2\times\sum_{i\in\mathcal{L}}\sum_{j\neq l,j,l\in\mathcal{R}} d_{i,\bar{jl}} + 2\times\sum_{i\neq k,i,k\in\mathcal{L}j\neq l,j,l\in\mathcal{R}}\sum d_{\bar{ik},\bar{jl}}\right). \tag{11}$$

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^2q^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 $\sqrt{\hat{T}}$ by the normal distribution; *i.e.*,

$$\sqrt{\hat{T}} \sim \mathcal{N}\left(\sqrt{T}, \frac{\sigma^2}{4T}\right),$$

where σ^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, $\hat{T}^b$, is

$$\hat{T}^b = \hat{T} - \frac{\hat{\sigma}^2}{4\hat{T}}, \tag{12}$$

where $\hat{\sigma}^2$ 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_{ij}, is the number of nucleotide differences between sequences i and j. To account for recurrent mutations, we replace all d_{ij} with a corrected distance measure d'_{ij}, where

$$d'_{ij} = -\frac{\ell}{2\log(1 - 2d_{ij}/\ell)} \tag{13}$$

(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 (Nei 1987).

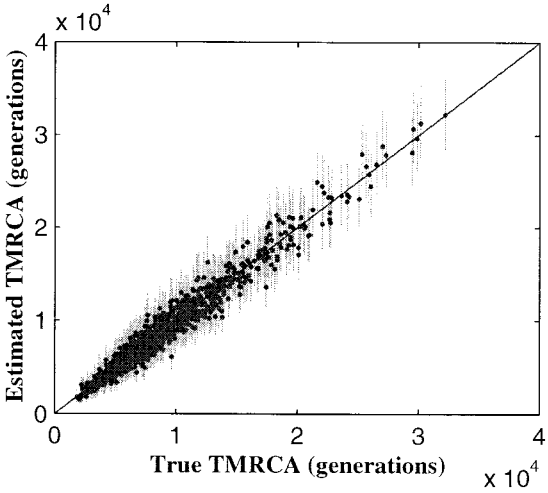

Figure 1.—Scatter plot of true TMRCA *vs.* $\hat{T}$ (dots) under a constant, panmictic population model, with 95% confidence intervals calculated using (6) and (11) (vertical bars). Of 1000 simulated genealogies, 953 true TMRCAs are covered by the 95% confidence interval.

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 θ 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

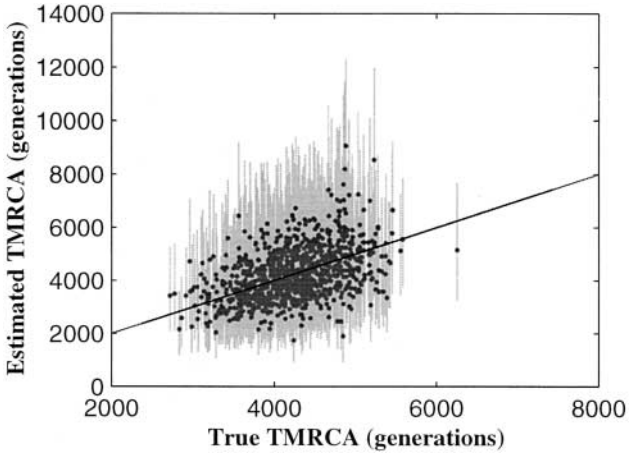

FIGURE 2.—Scatter plot of true TMRCA *vs.* $\hat{T}$ (dots) under an exponential growth, panmictic population model, with 95% confidence intervals calculated using (6) and (11) (vertical bars). Of 1000 simulated genealogies, 960 true TMRCAs are covered by the 95% confidence interval.

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 $g = 1 \times 10^{-3}$/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 $\mu = 2.5 \times 10^{-8}$/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

$$\text{Bias} = \frac{1}{I}\sum_{i=1}^{I}\frac{\hat{T}_i^b - T_i}{T_i}, \tag{14}$$

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 2

Coverage probability of the confidence intervals and average bias of the point estimates as functions of the present effective population size, N_0

N_0	T	Coverage	Bias	k	H
10^4	2858	0.967	−0.00523	25	4.81
4×10^4	4235	0.960	0.00415	47	7.74
2×10^5	5807	0.969	0.00223	80	11.29
4×10^5	6504	0.973	0.00163	95	12.85

Other population parameters are fixed: $\ell = 50{,}000$ bp; $\mu = 2.5 \times 10^{-8}$; $g = 0.001$. 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 1

Coverage probability of the confidence intervals and average bias of the point estimates as functions of mutation rate and sequence length

ℓ	μ	N_0	g	T	Coverage	Bias	k	H
10,000	2.5×10^{-8}	4×10^5	0.001	6483	0.99	0.140	19	2.60
1,000	2.5×10^{-7}	4×10^5	0.001	6501	0.99	0.146	19	2.60
50,000	2.5×10^{-8}	4×10^5	0.001	6504	0.973	0.00163	95	12.85
5,000	2.5×10^{-7}	4×10^5	0.001	6512	0.977	0.00244	94	12.85

ℓ is the length of the sequenced region; μ 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

g	T	Coverage	Bias	k	H	$N_{\hat{T}}$
0	7.66×10^5	0.94	-6.11×10^{-4}	3,392	908	4×10^5
0.0005	11,583	0.948	−0.00683	158	22.64	1,402
0.001	6,504	0.973	0.00163	95	12.85	675
0.002	3,600	0.988	0.040	95	12.85	344
0.005	1,619	0.988	0.143	27	3.32	140

Other population parameters are fixed: $\ell = 50{,}000$ bp; $\mu = 2.5 \times 10^{-8}$; $N_0 = 4 \times 10^5$. The quantities T, bias, H, and k are averaged over 1000 simulated genealogies for a given set of parameters. The column labeled $N_{\hat{T}}$ gives the extrapolated values of the population size at $\hat{T}$. Twenty sequences are simulated in each genealogy.

contain only a few segregating sites for most of which there is a single representation of the rare type. Thus, the observed genetic variation contains little information regarding the relationship among the sequences. Consequently, distance-based partition algorithms often introduce bias. The two-step algorithm described above reduces the bias by averaging over 20 partitions. Parsimony- or likelihood-based algorithms are less likely to introduce bias in these situations. Second, as seen in Table 4, a larger sample size will reduce the mean square error (MSE) and the width of the confidence interval, but the improvement tends to be negligible, especially when the sample size is already moderately large. Finally, the total heterozygosity (summed over all sites), H, increases with the mutation rate, the length of the sequenced region (Table 1), and the population size near the root of the genealogy (Table 2); it decreases as the growth rate increases (Table 3), but is roughly constant with changing sample size (Table 4). Hence, H can be used to predict the performance of the estimator on a particular data set. Roughly speaking, an H value <5 provides a warning sign that the data do not contain enough information, in which case special care should be taken in the partition step by using non-distance-based tree-building algorithms or by considering more than one partition.

Next, we compare our estimator, $\hat{T}^b$, to two existing methods: the Bayesian method based on summary statistics and the empirical Bayes approach as implemented by genetree, using complete sequence data.

Comparison with methods based on summary statistics: As an example, we compare our approach with the rejection method of Tavaré *et al.* (1997). A data set is simulated under a chosen population model, and its TMRCA is estimated by (5). For the rejection method, we assume perfect knowledge of both the population model and parameter values. Each proposed genealogy is accepted with **Prob**(k segregating sites|the length of the proposal genealogy). Figure 3 shows that the 95% confidence interval constructed by our frequentist approach is narrower than the 95% credibility interval from the rejection method. This is hardly surprising, since our frequentist estimator uses the complete sequence information, while the rejection method is based on summary statistics.

Comparison with methods based on the complete sequence data: It is important to compare the relative efficiencies of our estimator and a coalescent sampling estimator that uses complete data instead of summary statistics. We compare the MSE of the two approaches for the model of constant population size. Recall that the mutation rate parameter for the entire sequenced region is $\theta = 2N\mu\ell$.

First, consider an idealized situation, in which we not

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

n	T	Coverage	Bias	$\sqrt{\text{MSE}}$	w	k	H
10	6423	0.972	0.00100	1043	4522	56	12.21
20	6504	0.973	0.00163	926	4105	95	12.85
50	6529	0.981	7.97×10^{-4}	820	3832	185	13.31
100	6914	0.969	−0.00159	848	3783	298	13.47

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: $\ell = 50{,}000$; $\mu = 2.5 \times 10^{-8}$; $N_0 = 4 \times 10^5$; $g = 0.001$. The quantities T, bias, H, and k are averaged over 1000 simulated genealogies for a given set of parameters.

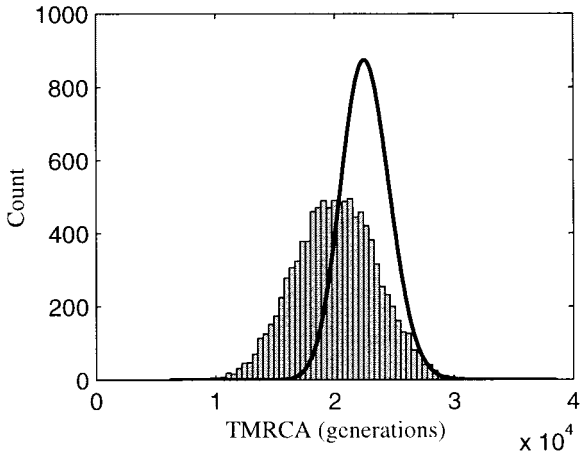

FIGURE 3.—Comparison of our approach to the posterior distribution obtained by rejection methods. The histogram is the posterior distribution obtained by the rejection method. The curve is the probability density function of the square of a $N(\sqrt{\bar{t}}, \widehat{\text{var}}(\sqrt{\bar{t}}))$ variate. The data are simulated under a constant size model, $N = 5000$, $g = 0$, $\mu = 5 \times 10^{-8}$, $\ell = 50$ kb.

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 (TAVARÉ *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

$$P(S_j = s_j | W_j = w_j) = \frac{e^{(-1/2)\theta j w_j}\,((1/2)\theta j w_j)^{s_j}}{s_j!}.$$

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

$$P(W_j|S_j) = \frac{\text{Prior}(W_j)\,P(S_j|W_j)}{\int_w \text{Prior}(w)\,P(S_j|w)\,dw},$$

and is a gamma distribution:

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

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

$$\check{W}_j = \frac{2(1+S_j)}{j(j-1+\theta)}.$$

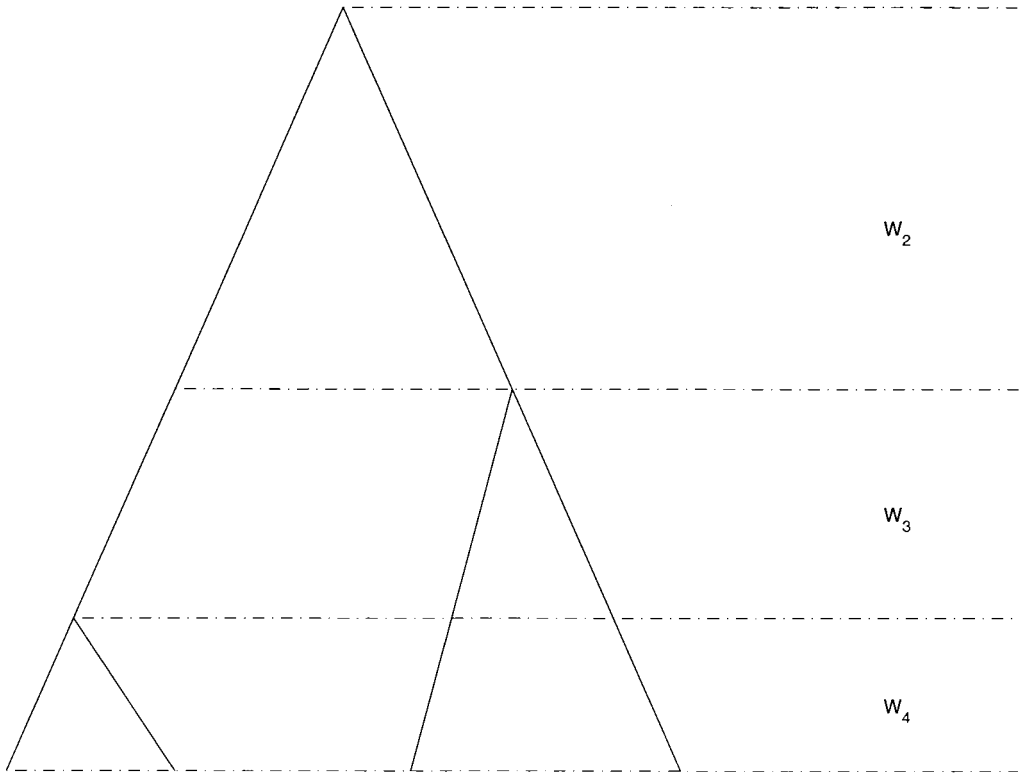

FIGURE 4.—A schematic genealogy with four samples.

The MSE associated with $\check{W}_j$ is

$$\mathrm{MSE}(\check{W}_j) = \int_w \sum_{s=0}^{\infty} (\check{w} - w)^2 P(s|w) \times \mathrm{Prior}(w)\, dw$$

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

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

$$\tilde{W}_j = \frac{S_j}{\theta/2j}.$$

The MSE associated with this estimator is

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

$$= \frac{4}{j^2\theta(j-1)}. \quad (16)$$

For the entire genealogy with n sequences, let $\check{T} = \Sigma_{j=2}^{n} \check{W}_j$ and $\tilde{T} = \Sigma_{j=2}^{n} \tilde{W}_j$. Then

$$\mathrm{MSE}(\check{T}) = \sum_{j=2}^{n} \mathrm{MSE}(\check{W}_j),$$

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

since all cross terms are 0.

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

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

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 $n \to \infty$, while the difference converges to 0 as $\theta \to \infty$.

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 θ, 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 θ. 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 $\theta = 2N\mu\ell$ is seldom known. In fact, a primary goal of genetree is to estimate θ empirically. This additional source of uncertainty would increase the MSE associated with TMRCA from the genetree estimator.

TABLE 5

$\sqrt{\mathrm{MSE}}$ associated with Bayesian and frequentist estimators of TMRCA

	Simulation		Theoretical lower bound	
θ	Bayesian	Frequentist	Bayesian	Frequentist
2.5	3309	4128	3016	3762
12.5	1830	1787	1588	1682

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

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:

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

Assuming a divergence time of 5 million years between chimpanzees and humans and a constant generation length of 25 years, the mutation rate is $\sim 3 \times 10^{-8}$/(site $\times$ generation) (Thomson *et al.* 2000). Under a model with constant population size and panmixia, the program fluctuate 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 1.3×10^{-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

Region	Sample size	k	H	$\hat{T}_{25}$ (10^3 yr)	$\hat{T}_{30}$ (10^3 yr)
Africa	38	98	7.86	87 (57, 125)	104 (68, 150)
Europe	18	23	5.67	48 (26, 80)	58 (31, 96)
Asia	37	46	4.75	37 (21, 58)	44 (25, 70)
Oceania	8	17	4.69	39 (20, 68)	47 (24, 82)
America	7	11	2.86	34 (11, 53)	41 (13, 64)
World	108	114	7.49	91 (60, 130)	109 (72, 156)

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

ples in haplogroup I, while haplogroups II–X form the other clade. The sample heterozygosity is ~7.49. TMRCA estimates for the world and for the five continents are given in Table 6.

Although the sample heterozygosity is relatively small, the simulations reported above suggest that the confidence interval is still reasonable. In view of the care with which the tree has been constructed (Underhill *et al.* 2000), we believe the bias due to the partition choice in the point estimators is relatively insignificant.

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 $\sim 2.43 \times 10^{-7}$/(site $\times$ generation). The sample heterozygosity is ~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 coalesce 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.

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-

TABLE 7

Estimated mtDNA TMRCA for each continent and for the world

Region	Sample size	k	H	$\hat{T}_{25}$ (10^3 yr)	$\hat{T}_{20}$ (10^3 yr)
Africa	52	463	55.65	238 (203, 276)	190 (162, 221)
Europe	55	250	17.70	69 (54, 87)	22 (43, 70)
Asia	37	316	26.47	105 (85, 128)	84 (68, 102)
Oceania	16	133	25.98	108 (87, 133)	86 (70, 106)
America	17	108	22.34	106 (81, 135)	85 (65, 108)
World	179	971	39.5	238 (200, 281)	190 (160, 225)

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.

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_2 generations ago, this population, now of size $N_0(T_2)$, was split into two subpopulations, of size $N_1(T_2) = f_1 N_0(T_2)$ and $N_2(T_2) = f_2 N_0(T_2)$. 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, $T_1 = 500$, $T_2 = 2000$, $m_1 = 0.0005$, $m_2 = 0.0001$, $g_0 = 0.005$, $g_1 = 0.001$, $g_2 = 0.003$, $f_1 = 0.75$, $f_2 = 0.25$, $N_0(T_1 + T_2) = 300$, $L = 15{,}000$ 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.

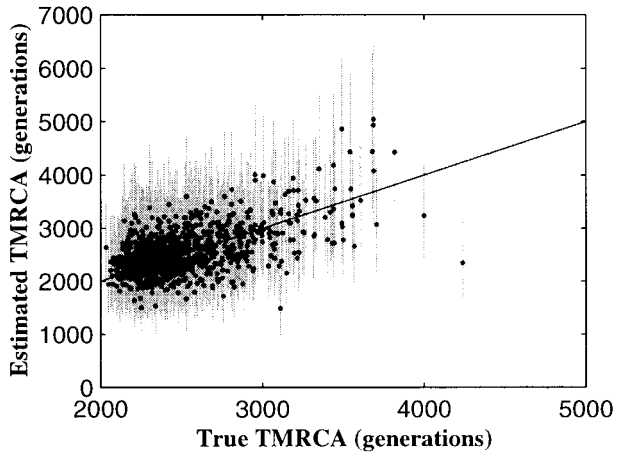

FIGURE 5.—Scatter plot of true TMRCA *vs.* $\hat{T}$ (dots) under a subdivided population model, with 95% confidence intervals calculated using (6) and (11) (vertical bars). Of 1000 simulated genealogies, 965 true TMRCAs are covered by the 95% confidence interval.

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 θ increases, and θ 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–$\sim$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

Beerli, P., and J. Felsenstein, 1999 Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152:** 763–773.

Cavalli-Sforza, L. L., I. Barrai and A. W. F. Edwards, 1964 Analysis of human evolution under random genetic drift. Cold Spring Harbor Symp. Quant. Biol. **29:** 7–20.

Felsenstein, J., 1993 *PHYLIP (Phylogeny Inference Package) Version 3.5c.* Distributed by the author. Department of Genetics, University of Washington, Seattle.

Griffiths, R. C., and S. Tavaré, 1994 Ancestral inference in population genetics. Stat. Sci. **9:** 307–319.

Kingman, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Kuhner, M. K., J. Yamato and J. Felsenstein, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. Genetics **149:** 429–434.

Kuhner, M. K., J. Yamato and J. Felsenstein, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393–1401.

Nei, M., 1987 *Molecular Evolutionary Genetics.* Columbia University Press, New York.

Pritchard, J. K., M. T. Seielstad, A. Pérez-Lezaun and M. W. Feldman, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16:** 1791–1798.

Ruvolo, M., 1996 A new approach to studying modern human origins: hypothesis testing with coalescence time distributions. Mol. Phylogenet. Evol. **5:** 202–219.

Saunders, I. W., S. Tavaré and G. A. Watterson, 1984 On the genealogy of nested subsamples from a haploid population. Adv. Appl. Prob. **16:** 471–491.

Stephens, M., 2001 Inference under the coalescent, pp. 213–238 in *Handbook in Statistical Genetics*, edited by D. J. Balding, C. Cannings and M. Bishop. Wiley, Chichester, UK.

Stumpf, M., and D. B. Goldstein, 2001 Genealogical and evolutionary inference with the human Y chromosome. Science **291:** 1738–1742.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

Takahata, N., S.-H. Lee and Y. Satta, 2001 Testing multiregionality of modern human origins. Mol. Biol. Evol. **18:** 172–183.

Tavaré, S., D. J. Balding, R. C. Griffiths and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145:** 505–518.

Templeton, A. A., 1993 The "Eve" hypothesis: a genetic critique and reanalysis. Am. Anthropol. **95:** 51–72.

Templeton, A. R., A. G. Clark, K. M. Weiss, D. A. Nickerson, E. Boerwinkle *et al.*, 2000 Cladistic structure within the human lipoprotein lipase gene. Genetics **156:** 1259–1275.

Thomson, R., J. K. Pritchard, P. Shen, P. J. Oefner and M. W. Feldman, 2000 Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc. Natl. Acad. Sci. USA **97**(13): 7360–7365.

Tremblay, M., and H. Vezina, 2000 New estimates of intergenerational time intervals for the calculation of age and origins of mutations. Am. J. Hum. Genet. **66:** 651–658.

Underhill, P. A., P. Shen, A. A. Lin, L. Jin, G. Passarino *et al.*, 2000 Y chromosome sequence variation and the history of human populations. Nat. Genet. **26:** 358–361.

Communicating editor: J. B. Walsh