The Age of a Unique Event Polymorphism

Lada Markovtsova,* Paul Marjoram† and Simon Tavaré*,†,‡

**Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113, †Biostatistics Division, Department of Preventive Medicine, University of Southern California, Los Angeles, California 90033 and ‡Program in Molecular Biology, Department of Biological Sciences, University of Southern California, Los Angeles, California 90089-1340*

Manuscript received August 2, 1999
Accepted for publication May 19, 2000

ABSTRACT

We develop a Markov chain Monte Carlo approach for estimating the distribution of the age of a mutation that is assumed to have arisen just once in the history of the population of interest. We assume that in addition to the presence or absence of this mutation in a sample of chromosomes, we have DNA sequence data from a region completely linked to the mutant site. We apply our method to a mitochondrial data set in which the DNA sequence data come from hypervariable region I and the mutation of interest is the 9-bp region V deletion.

ESTIMATION and inference for population quantities such as mutation rates and demographic history are often based on molecular data sampled from populations. Underlying any such data is a genealogy that describes the way in which the sampled chromosomes are related. The behavior of such a genealogy is often approximated by a stochastic process called the *coalescent*, introduced by Kingman (1982). There are now several approaches to estimation and inference for coalescent-based models, among them the Markov chain approach of Griffiths and Tavaré (1994a,b), and the Markov chain Monte Carlo (MCMC) approach initiated by Kuhner *et al.* (1995, 1998). Bayesian approaches to inference in the coalescent are described by Tavaré *et al.* (1997) and Wilson and Balding (1998).

Our data come from a random sample of n chromosomes, from each of which we have the DNA sequence of a region of interest. We denote the collection of sequences by **D**. In addition to the sequence data, we have information on the presence or absence of a unique event polymorphism (UEP) mutation, a neutral mutation that is assumed to have arisen just once in the population of interest. In this article, we develop an MCMC approach for studying the age of a UEP mutation that is segregating in the sample, under the assumption of no recombination in **D** or between the mutant site and **D**. Our MCMC approach generates observations from the joint conditional distribution of a number of quantities of interest, including the time to the most recent common ancestor (MRCA) of the sample, the time to the MRCA of the chromosomes carrying the UEP mutation, the length of the branch on which the UEP mutation arises, the age of the UEP, and the mutation rates, given **D**.

We illustrate our method by finding the conditional distribution of the age of the 9-bp deletion in mitochondrial region V in a sample of Native Americans from the Yakima tribe (Shields *et al.* 1993). This deletion, arising in the intergenic region between the cytochrome oxidase II gene and the lysine tRNA gene, has been used to trace the history of human migrations (*cf.* Shields *et al.* 1993, Figure 7; Lorenz and Smith 1994; Redd *et al.* 1995; Soodyall *et al.* 1996; and Watkins *et al.* 1999). The age of the deletion may provide information about the timing of such migrations. In a sample of $n = 42$ individuals, **D** comprises the sequence of 360 bp from region I of the control region, and $b = 26$ of the individuals carried the deletion. Under reasonable demographic assumptions, we find a mean conditional age for the deletion of ~4100 years. The 2.5th percentile of the distribution is 1700 years, and the 97.5th percentile is 8200 years; thus the data support values as small as 1700 years and as large as 8200 years.

The coalescent: The simplest version of the coalescent assumes the population is random mating and of a large constant size N, from which we sample n sequences from the present-day population. In the coalescent, time runs backward and is recorded in units of N/σ^2 generations, where N is the (effective) population size and σ^2 is the variance of the distribution for the number of offspring produced by a parent in a single generation. In common with most authors, we assume $\sigma^2 = 1$. Thus a coalescence time of 1 translates into N generations ago, and so on. The time T_j during which the sample has j distinct ancestors has an exponential distribution with parameter $j(j-1)/2$, and coalescent events occur at random among the ancestors of the sample. The process of coalescences terminates when a single line of ancestry remains.

The genealogy can be viewed as consisting of two

Corresponding author: Simon Tavaré, Program in Molecular Biology, Department of Biological Sciences, SHS 172, University of Southern California, Los Angeles, CA 90089-1340.
E-mail: stavare@gnome.usc.edu

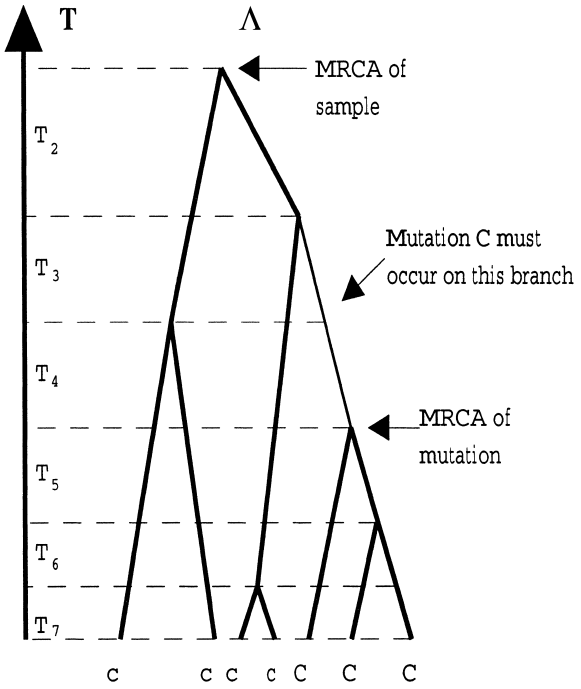

FIGURE 1.—Coalescent tree with UEP.

components: the topology of the tree structure and the times between coalescent events. We use Λ to denote the topology of the tree, and $T = \{T_n, T_{n-1}, \ldots, T_2\}$ to denote the set of coalescence times in the sample. An example of a genealogy for seven sequences is given in Figure 1. Accessible reviews of the coalescent are given by HUDSON (1991) and DONNELLY and TAVARÉ (1995), for example.

Theory for the age of a UEP: Because the UEP has arisen by mutation just once in the ancestry of the sampled population, the individuals in the sample can be divided into two groups, those that carry the mutation corresponding to the UEP and those that do not. Further, it must be the case that the sequences carrying the mutation have coalesced with each other before sharing a common ancestor with any sequence not carrying the mutation. Figure 1 shows an example with $n = 7$ sequences, of which $b = 3$ carry the UEP mutation C and four carry the ancestral allele c.

We suppose that the scaled mutation parameter of the UEP mutation is μ; that is, $\mu = 2Nv$, where v is the probability of the mutation occurring in a given sequence per generation. Potential mutation events occur on the branches of the coalescent tree according to independent Poisson processes of rate $\mu/2$. Several theoretical results are known about the age of a UEP. In the limiting case $\mu \to 0$, GRIFFITHS and MARJORAM (1996) derive a formula for the mean of the age ξ_{nb} of a mutation observed b times in a sample of size n, for $0 < b < n$, and the mean time to the MRCA. The probability density of ξ_{nb} and the time to the MRCA, under both constant and variable population size scenarios, are given in GRIFFITHS and TAVARÉ (1998, Equations 5.3 and 6.3). The case $\mu > 0$ can be treated directly by the approach of GRIFFITHS and TAVARÉ (1994b); some related results appear in STEPHENS (2000).

WIUF and DONNELLY (1999) derived a number of results about the topology of a conditional coalescent tree having the property

$$E \equiv \{\text{A particular set of } b \text{ sequences coalesces before any of the remaining } n - b \text{ sequences share a common ancestor with any of the } b\}. \quad (1)$$

They show that until the sequences carrying the mutation have found a common ancestor, the transition probabilities are as follows: If there are currently $j \geq 2$ lines of ancestry carrying the mutation and l lines not doing so, the probability that the next coalescent event involves two lines carrying the mutation is

$$(j + 1)/(l + j), \quad (2)$$

while the probability that it involves two lines not carrying the mutation is given by $(l - 1)/(l + j)$. Once an MRCA has been attained for the sequences carrying the mutation, coalescence occurs as normal, randomly between any two existing lines of ancestry. Since the conditioning involves just the topology of the tree, the event times in the conditional coalescent have the same distribution as in the unconditional case.

SLATKIN and RANNALA (1997) use a different approach to estimate the age of an allele when **D** consists of (an estimate of) the number of mutations that have arisen in a region close to that allele *in just those chromosomes carrying the given allele*. In their approach and that of THOMPSON and NEEL (1997), the age of an allele is treated as a parameter to be estimated from the data, together with an appropriate confidence interval. In our approach, the age of a UEP is an unobservable random variable, and what is reported is then the conditional distribution of the age of the UEP given **D**. For more on the Slatkin and Rannala model in the coalescent setting, contact R. C. GRIFFITHS and S. TAVARÉ (unpublished results).

Mutation model for sequences: The variation observed in the DNA sequences **D** is a consequence of mutations occurring in the coalescent tree of the sample. We model the evolution of **D** using a finite-sites model due to Felsenstein, described in detail in THORNE *et al.* (1992). In this model there are two parameters: g, the general substitution rate, and w, the within-group substitution rate. We parameterize the model by setting $w = \kappa g$, where κ is the transition/transversion parameter. In our implementation of this model, we assume that the base frequencies are known (and given by their observed frequencies in the sample). The unknown parameters in this part of the model are then κ and g. It is conventional to report the parameter θ, where $\theta/2$ is

the mean number of mutations per unit time that *change* a base along a given branch. We have

$$\theta = 2g\left\{\left(1 - \sum_{i=1}^{4}\pi_i^2\right) + 2\kappa\left(\frac{\pi_A\pi_G}{\pi_A + \pi_G} + \frac{\pi_C\pi_T}{\pi_C + \pi_T}\right)\right\}. \quad (3)$$

Note that θ can be calculated from κ, g, and the base frequencies.

MATERIALS AND METHODS

Let U denote the single event that causes the UEP mutation. This corresponds to a single (rate $\mu/2$) mutation arising on the branch indicated in Figure 1 and no other mutations on the rest of the coalescent tree. Let A denote the age of the UEP, and denote the mutation parameters by $M = (g, \kappa, \mu)$. In what follows, we assume a prior distribution for M, and develop an MCMC method for generating observations from the conditional density $f(A, G \mid \mathbf{D}, E, U)$ of A and $G = (\Lambda, T, M)$ given $\mathbf{D}$, E, and U, where E is the event defined in (1). To do this, we express the required conditional density as a product of simpler terms and describe how each can be calculated. First we note that

$$f(A, G \mid \mathbf{D}, E, U) = f(A \mid G, \mathbf{D}, E, U)f(G \mid \mathbf{D}, E, U). \quad (4)$$

The first term on the right of (4) can be evaluated by considering Figure 1 once more. Given that a single mutation occurs on the indicated branch, the Poisson nature of the mutation process for the UEP means that the location of the mutation is uniformly distributed over that branch. Thus we can simulate observations from the conditional distribution of A by simulating from the second term on the right of (4), reading off the length of the branch on which the UEP mutation occurs, and adding a uniformly distributed fraction of that length to the height of the subtree containing all the chromosomes carrying the UEP. Our task is therefore reduced to simulating from the second term on the right of (4).

Let $p_1(\Lambda \mid E)$ denote the conditional distribution of the coalescent tree Λ given E, $p_2(T)$ the density of the coalescence times T, and $p_3(M)$ the prior for the mutation rates $M = (g, \kappa, \mu)$. We can then write

$$f(G \mid \mathbf{D}, E, U) = \mathbb{P}(\mathbf{D}, U \mid G, E)p_1(\Lambda \mid E)p_2(T)p_3(M)/\mathbb{P}(\mathbf{D}, E, U). \quad (5)$$

The term $\mathbb{P}(\mathbf{D}, U \mid G, E)$ is the product of two terms,

$$\mathbb{P}(\mathbf{D}, U \mid G, E) = \mathbb{P}(\mathbf{D} \mid G, E)\mathbb{P}(U \mid G, E).$$

The first of these, the likelihood of $\mathbf{D}$, can be computed using a peeling algorithm (*cf.* Felsenstein 1981) and the mutation model described above, while the second is

$$\frac{\mu S}{2}e^{-\mu S/2} \times e^{-\mu(L-S)/2} = \frac{\mu S}{2}e^{-\mu L/2}, \quad (6)$$

where S is the length of the branch on which the single UEP mutation must occur, and $L = \sum_{i=2}^{n} iT_i$ is the total length of the tree. The normalizing constant $\mathbb{P}(\mathbf{D}, E, U)$ is unknown, and hard to compute. As a consequence, we use a version of the Metropolis-Hastings algorithm, due originally to Metropolis *et al.* (1953) and Hastings (1970), to simulate from the required conditional distribution.

Markov chain Monte Carlo method: The algorithm produces correlated samples from a distribution π of interest, in our case

$$\pi(G) \equiv f(G \mid \mathbf{D}, E, U).$$

It starts with an arbitrary choice of T and M, and a Λ consistent with the conditioning event E. New realizations of G are then proposed and accepted, or rejected, according to the following scheme:

Basic Metropolis-Hastings Algorithm:

1. Denote the current state by $G = (\Lambda, T, M)$.
2. Output the current value of G.
3. Propose $G' = (\Lambda', T', M')$ according to a proposal kernel $Q(G \to G')$.
4. Compute the Hastings ratio

$$h = \min\left\{1, \frac{\pi(G')Q(G' \to G)}{\pi(G)Q(G \to G')}\right\}. \quad (7)$$

5. Accept the new state G' with probability h, otherwise stay at G.
6. Return to step 1.

Let $X(t)$ denote the state of this chain after t iterations. Once $X(\cdot)$ has reached stationarity, its values represent samples from the distribution $\pi(G)$. Note that consecutive outputs are often highly correlated. If we wish to simulate approximately independent samples from the posterior distribution, we commonly use output from every mth iteration for a suitable choice of m.

We have some freedom in choosing the updating kernel $Q(\cdot, \cdot)$. Ideally $Q(\cdot, \cdot)$ should be relatively easy to calculate, since the scheme above may need to be iterated many times to converge to stationarity. Furthermore, the chain $X(\cdot)$ must be irreducible (so that all states can be reached from any other) and positive recurrent to ensure that the limiting distribution is indeed the required $\pi(G)$.

An updating mechanism: The updating kernel Q defines a Markov process on the state space of trees, times, and mutation rates, $G = (\Lambda, T, M)$. Some samplers that might be adapted to our problem are given in Kuhner *et al.* (1995), Wilson and Balding (1998), Larget and Simon (1999) and Mau *et al.* (1999). We have chosen to make local changes to the genealogy in a somewhat different way.

We define level l of the genealogy to be the first point at which there are l distinct ancestors of the sample. The bottom of a genealogy of n individuals is referred to as level n. The topmost level is referred to as level 1 (this is the most recent common ancestor of the sample) and T_l is the time between levels l and $l - 1$. The sampler proposes a new graph (Λ', T') to which we might move. We consider this in two parts: proposing a new tree topology Λ', and proposing new times for the coalescence events therein, T'. We describe updates to M later. We begin by specifying the scheme for proposing a new tree topology Λ', ignoring the effects of conditioning on E. Our approach changes the structure of two adjacent levels of the genealogy. For a genealogy with n individuals, we begin by picking a level l ($l = n, n - 1, \ldots, 3$) according to an arbitrary distribution F; in practice, we generally used a uniform distribution. Once we have chosen l, we observe the pattern of coalescence at levels l and $l - 1$. This pattern falls into two cases, according to whether the coalescence at level $l - 1$ involves the line that results from the coalescence at level l. These two cases are illustrated in Figure 2. In case A, our kernel randomly generates a new topology involving the same three lines of ancestry; this new topology is also case A. These are illustrated in Figure 3. In case B, we change the order of the two coalescence events, resulting in another case B topology. For the example illustrated above, we move to the state shown in Figure 4.

We make a minor modification to this algorithm to ensure that new trees are also consistent with the event E. If, when

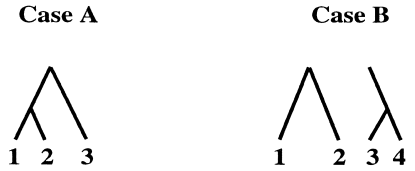

FIGURE 2.—Two possible coalescence patterns.

we pick a level, we find we are in case A, and exactly two of the lines carry the UEP, then we cannot change the order in which the two coalescences occur, since such a change would produce a new tree topology that is inconsistent with E. In such a situation, we leave the topology unchanged.

Having constructed a new tree topology, which may be the same as the existing topology, we now generate a new set of times, T'. We generate new times T'_l and T'_{l-1} according to an arbitrary distribution and leave other times unchanged. Thus, we only alter the times corresponding to the levels at which the topology has been changed. This ensures that (Λ', T') is similar to (Λ, T) and will therefore have a reasonable probability of being accepted. We found that a kernel that proposes new values of T'_l and T'_{l-1} having the predata coalescent distribution worked well on the data sets described later in the article. We also found that proposals that are Normally distributed with mean equal to the currently accepted value worked well. We chose to truncate the Normal distribution to ensure that negative times were not proposed. The variances of the Normal distributions are parameters that can be tuned to get good mixing properties. This choice effects the efficiency of the method, but makes no difference to the numerical results.

Finally, we update $M = (g, \kappa, \mu)$, where g and κ are the rate parameters for the sequence and μ is the rate parameter for the UEP. Parameters g and κ were updated every 10th iteration, and μ was updated on each iteration for which g was not updated. These were updated using truncated Normals, as in the last paragraph; for example, we generate a new value g' according to a Normal distribution with mean g. The variances of these distributions required some tuning to ensure well-behaved, *i.e.*, uncorrelated, output. We have explored a number of modifications to this basic approach, some of which are described further in MARKOVTSOVA *et al.* (2000).

The Hastings ratio: Writing $G = (\Lambda, T, M)$, the kernel Q can be expressed as the product of three terms:

$$Q(G' \to G) = Q_1(\Lambda' \to \Lambda)\, Q_2(T' \to T \mid \Lambda' \to \Lambda)\, Q_3(M' \to M).$$

Consequently, using (4), (5), and (6), the Hastings ratio, the probability with which we accept the new state, can be written in the form

$$h = \min\left\{1, \frac{\mathbb{P}(\mathbf{D} \mid G', E)}{\mathbb{P}(\mathbf{D} \mid G, E)} \frac{\mathbb{P}(U \mid G', E)}{\mathbb{P}(U \mid G, E)} \frac{p_1(\Lambda' \mid E)}{p_1(\Lambda \mid E)} \frac{p_2(T')}{p_2(T)} \frac{p_3(M')}{p_3(M)} \times \frac{Q_1(\Lambda' \to \Lambda)}{Q_1(\Lambda \to \Lambda')} \frac{Q_2(T' \to T \mid \Lambda' \to \Lambda)}{Q_2(T \to T' \mid \Lambda \to \Lambda')} \frac{Q_3(M' \to M)}{Q_3(M \to M')}\right\},$$

the unknown term $\mathbb{P}(\mathbf{D}, E, U)$ canceling. For our choice of transition kernel Q, (2) can be used to show that $p_1(\Lambda' \mid E) = p_1(\Lambda \mid E)$. We also have $Q_1(\Lambda \to \Lambda') = Q_1(\Lambda' \to \Lambda)$, and we note that Q changes only two of the times associated with T or T'. Hence h reduces to

$$h = \min\left\{1, \frac{\mathbb{P}(\mathbf{D} \mid G', E)}{\mathbb{P}(\mathbf{D} \mid G, E)} \frac{\mathbb{P}(U \mid G', E)}{\mathbb{P}(U \mid G, E)} \frac{p_2(T')p_3(M')}{p_2(T)p_3(M)} \times \frac{f_l(t_l)f_{l-1}(t_{l-1})}{f_l(t'_l)f_{l-1}(t'_{l-1})} \frac{Q_3(M' \to M)}{Q_3(M \to M')}\right\}, \quad (8)$$

where $f_l(\cdot)$ and $f_{l-1}(\cdot)$ are the densities of the time updating mechanism given that changes occur to the tree Λ at levels l and $l-1$.

Practical considerations: A key feature of the Metropolis-Hastings algorithm is that one wishes to observe the process $X(\cdot)$ *once stationarity has been reached*, so that the process has come to a steady state under which the distribution of $X(t)$ is the required π. There are many heuristic tests one might employ to assess whether $X(\cdot)$ is stationary. For a critique of these, see GILKS *et al.* (1996). We chose to look at functions of the statistics of interest such as autocorrelations and moving averages. Another useful diagnostic is to run the chain starting from widely dispersed starting points to see whether the long-term behavior is the same in each case. We also used the tests contained in the software package CODA (BEST *et al.* 1995). The output discussed later in the article performed well on such tests.

Significant time can be saved by starting the process from a genealogy (Λ, T) for which $P(\Lambda, T \mid \mathbf{D}, E, U)$ is relatively high. For example, one might use a UPGMA tree generated from the sequence data $\mathbf{D}$, as described in KUHNER *et al.* (1995); the resulting tree should satisfy the constraints required by E.

For the analyses discussed in the next section, the output typically appeared to be nonstationary for at least 200,000 iterations of the algorithm. In a bid to be conservative, we generally discarded the first 25 million iterations. After this, we sampled every 5000th iteration. Our output is typically based on 5000 samples from our stationary process. The acceptance rate was generally ~70%. For runs in which, for example, we needed to "tune" the variance parameter, the burn-in length varied, but the estimated parameter values were unchanged for the different variances we tried.

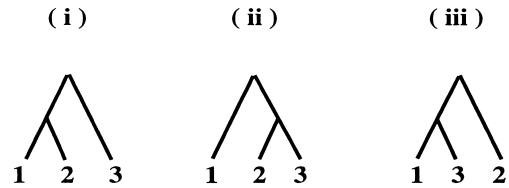

FIGURE 3.—Possible moves in case A.

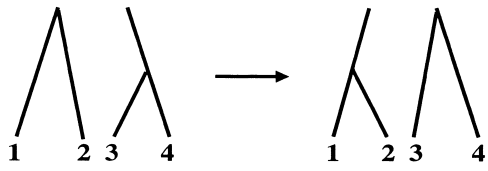

FIGURE 4.—Possible moves in case B.

RESULTS

We applied our method to find the conditional distribution of the age of the mitochondrial region V deletion in a sample of Yakima described by SHIELDS *et al.* (1993). The sample comprises $n = 42$ individuals, of whom $b = 26$ have the deletion. The data $\mathbf{D}$ comprise 360 bp from hypervariable region I of the control region, sequenced for all 42 individuals. The observed base frequencies are $(\pi_A, \pi_G, \pi_C, \pi_T) = (0.328, 0.113, 0.342, 0.217)$. We note that all individuals having a given control region sequence had the same deletion status,

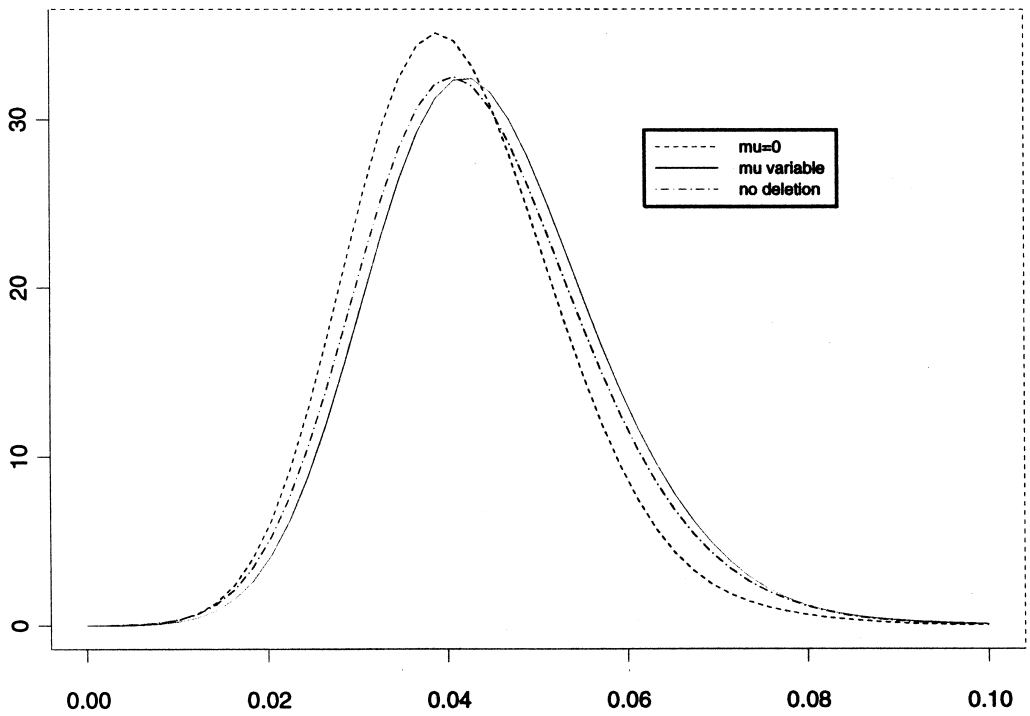

FIGURE 5.—Posterior density of mutation rate θ.

as might be expected if the deletion arose once quite recently.

Preliminary analysis of the sequence data (without regard to presence or absence of the deletion) was performed using the approach outlined in MARKOVTSOVA *et al.* (2000). For the present mutation model, we took uninformative priors (in the form of uniform densities having wide but finite support) for the mutation rates g and w and examined the posterior distribution of $\kappa = w/g$ (data not shown). The posterior median was 65.9, the distribution having 25th percentile of 34.0 and 75th percentile of 160.2. The data are certainly consistent with the value of $\kappa = 100$ assumed by KUHNER *et al.* (1995) in their analysis of the same region in a sequence of Nuu Chah Nulth sequences of WARD *et al.* (1991). We therefore chose to take $\kappa = 100$ as fixed in the subsequent analyses; from (3) we find that $\theta = 88.17g$.

We repeated the analysis with an uninformative prior, uniform on (0, 0.1), for the single parameter g. This resulted in the posterior density for θ given in Figure 5. Summary statistics are shown in Table 1. Our approach also provides a way to find the maximum-likelihood estimator (MLE) of θ, since with a flat prior the posterior is proportional to the likelihood. From a kernel density estimate we obtained an MLE of $\hat{\theta} = 0.039$ with an estimated standard error of 0.010. This is consistent with the estimate of $\hat{\theta} = 0.040$ found for the Nuu Chah Nulth data by KUHNER *et al.* (1995). Since the base frequencies in both data sets are similar and the mutation rates are likely to be the same, we conclude that the effective sizes of the two populations are also approximately equal. The effective population size of the Nuu Chah Nulth was estimated (from anthropological data) by WARD *et al.* (1991) to be $N \sim 600$, a number we take for the Yakima as well.

TABLE 1

Summary statistics for θ

θ	No deletion	μ variable	$\mu = 0$
Mean	0.044	0.045	0.041
Median	0.042	0.043	0.040
25th percentile	0.036	0.037	0.034
75th percentile	0.050	0.051	0.047

In the absence of data, the mean time to the MRCA of a sample of $n = 42$ is $2(1 - 1/42) = 1.95$. With an effective size of $N = 600$ and a 20-year generation time, this is ~23,500 years. The posterior density of the time to the MRCA given the control region data **D** is shown in Figure 6. The posterior mean is 0.72, or ~8600 years. Summary statistics are given in Table 2. The posterior distribution of the total tree length $L = \Sigma_{j=2}^{42} jT_j$ has mean 5.68.

Including the deletion: We turn now to the deletion data. We ran our MCMC algorithm using a uniform (0, 10) prior for μ and a uniform (0, 0.1) prior for g. The posterior density of θ is shown in Figure 5. Summary statistics are presented in Table 1. The distribution is qualitatively the same as that obtained by ignoring the deletion data. The posterior density of the deletion parameter μ is shown in Figure 7. The posterior mean is 0.75, the median is 0.61, the 25th percentile is 0.34, and the 75th percentile is 0.99.

The posterior density of the time to the MRCA of the group carrying the deletion is shown in Figure 8. The summary statistics are found in Table 3. The deletion

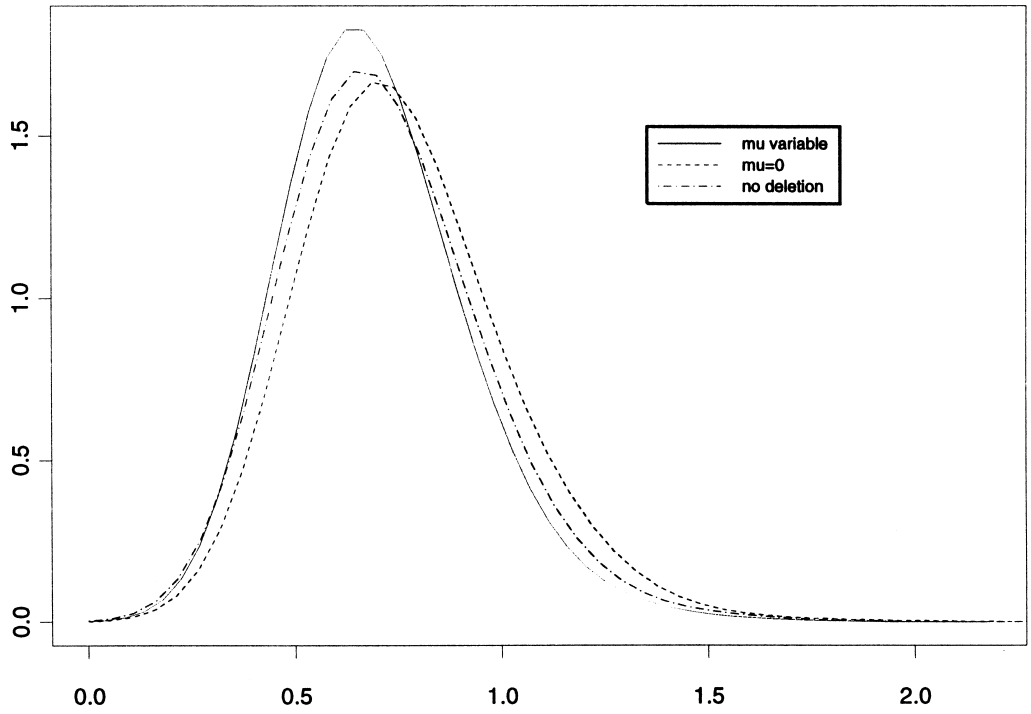

FIGURE 6.—Posterior density of TMRCA.

arises uniformly on the branch indicated in Figure 1, so that the age of the mutation is the time to the MRCA of the deletion group plus a uniform fraction of the mutation branch length. The posterior distribution of the age is given in Figure 9, and summary statistics are in Table 4. We also looked at the time to the MRCA of the entire sample when the deletion status of each sequence is included. The posterior density of this time is shown in Figure 6, with summary statistics given in Table 2. For these data the inclusion of deletion status has little effect on the posterior distribution.

The output from the MCMC runs can be used to assess whether the UEP assumption is reasonable. We first generated 5000 observations of the tree length L conditional on the data **D**; as noted above, the sample mean is 5.68. The modal posterior value of μ is 0.30, a value that we treat as a point estimate of μ. The expected number of deletions arising on the coalescent tree is then $0.30\ \mathbb{E}(L|\mathbf{D})/2$, which we estimate from the posterior mean tree length as $0.30 \times 5.68/2 = 0.85$. We can also use this value of μ and the simulated values of L to estimate the probability that exactly one mutation would occur on such a tree; we obtained an estimate of 0.36. Similarly, we estimated the probability of at least one mutation occurring as 0.57, so that the conditional probability that the mutation occurred once, given it occurred at least once, is estimated to be 0.63. Thus it is not unreasonable to assume that the deletion arose just once.

The case $\mu = 0$: In the Introduction, we pointed to a number of theoretical results concerning the age of a UEP given its frequency in the sample in the limiting case $\mu \to 0$. To compare these results with those obtained by including the sequence information, we modified our algorithm to allow $\mu = 0$. The mutation parameter M is now one-dimensional: $M = (g)$. The other change occurs to the conditional probability in (6), since now $\mathbb{P}(U \mid G, E) \propto S$, the length of the branch on which the UEP mutation must occur. This change appears in the Hastings ratio (8), where

$$\frac{\mathbb{P}(U \mid G', E)}{\mathbb{P}(U \mid \mathrm{G}, E)} = \frac{S'}{S}.$$

The posterior density of θ is also shown in Figure 5,

TABLE 2

Summary statistics for time to MRCA of the sample

Time to MRCA	No deletion	μ variable	$\mu = 0$
Mean	0.72 (8,600 yr)	0.70 (8,400 yr)	0.76 (9,200 yr)
Median	0.69 (8,300 yr)	0.67 (8,000 yr)	0.73 (8,800 yr)
25th percentile	0.57 (6,800 yr)	0.56 (6,700 yr)	0.61 (7,300 yr)
75th percentile	0.84 (10,100 yr)	0.81 (9,700 yr)	0.88 (10,600 yr)

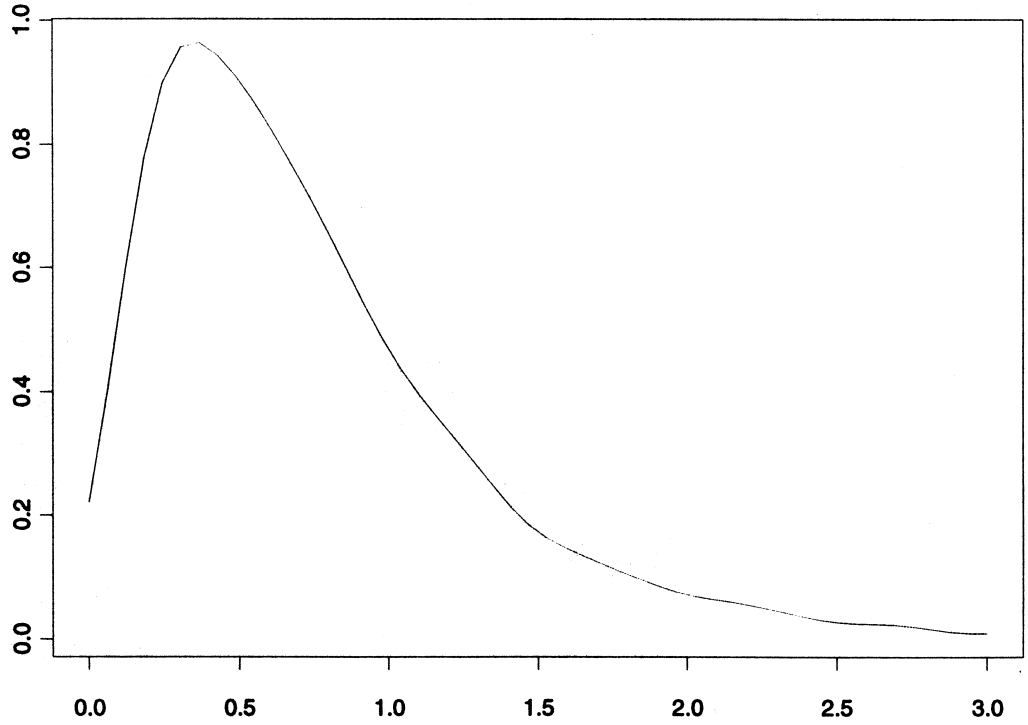

FIGURE 7.—Posterior density of mutation parameter μ.

with summary statistics given in Table 1; there is little difference from the case where μ is allowed to vary. The posterior density of the time to the MRCA is given in Figure 6, with summary statistics in Table 2. The mean time of 0.76 (or ~9100 years) stands in marked contrast to the value of 2.68 (~32,200 years) obtained from GRIFFITHS and MARJORAM (1996).

The summary statistics for the posterior distribution of the time to the MRCA of the group carrying the deletion are given in Table 3. The results are qualitatively the same as the case of variable μ. The posterior density of the age of the deletion appears in Figure 9, with summary statistics shown in Table 4. The posterior mean is 0.36 (or ~4400 years), compared to the value

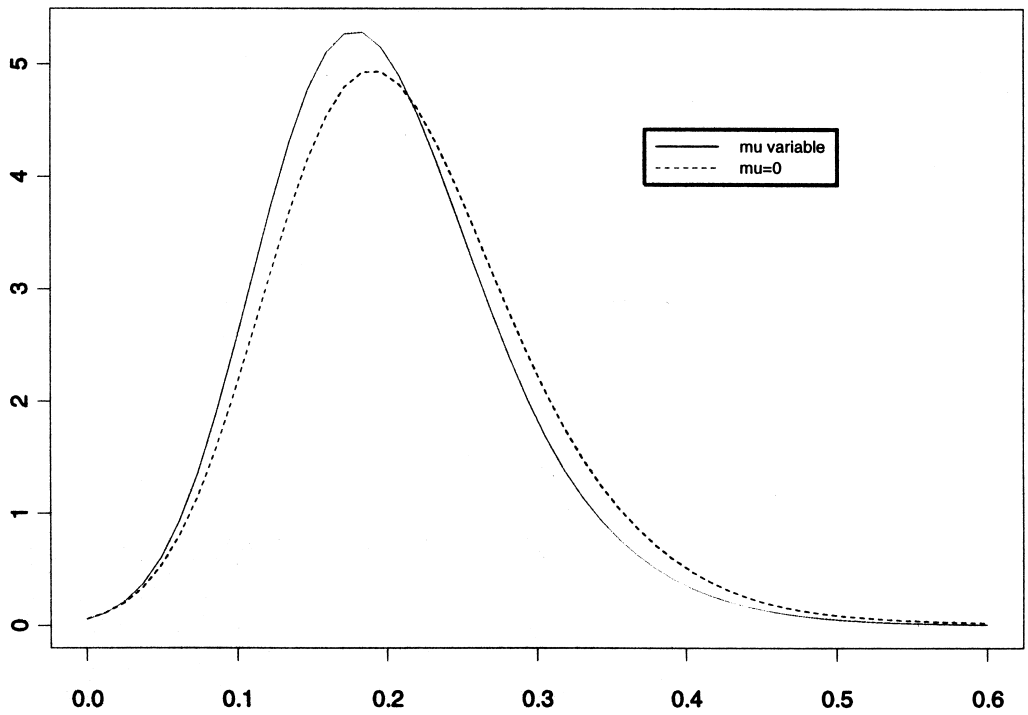

FIGURE 8.—Posterior density of TMRCA of deletion.

TABLE 3

Summary statistics for the time to MRCA of the group carrying the deletion

Time to MRCA	μ variable	μ = 0
Mean	0.20 (2400 yr)	0.21 (2600 yr)
Median	0.19 (2300 yr)	0.20 (2400 yr)
25th percentile	0.15 (1800 yr)	0.16 (1900 yr)
75th percentile	0.24 (2900 yr)	0.25 (3100 yr)

TABLE 4

Summary statistics for age of the deletion

Age of deletion	μ variable	μ = 0
Mean	0.34 (4100 yr)	0.36 (4400 yr)
Median	0.31 (3700 yr)	0.33 (4000 yr)
25th percentile	0.23 (2800 yr)	0.25 (3000 yr)
75th percentile	0.41 (5000 yr)	0.44 (5300 yr)

of $\mathbb{E}\xi_{42,26} = 1.54$ (or ~18,500 years) when the sequence data are ignored. As expected, the mean age is higher than it is when μ is nonzero.

DISCUSSION

We have described a Markov chain Monte Carlo method for finding the conditional distribution of the age of a mutation that is assumed to have arisen once in the history of the population under study, when further data in the form of completely linked DNA sequences are found for the individuals in the sample. There are several comments that should be made. In our analysis of the region V mitochondrial DNA deletion we assumed a constant population size (*cf.* Shields *et al.* 1993). We have also implemented a version of our algorithm that allows for deterministic population size fluctuations.

Several other variants on the theme are also readily implemented. For example, if several mutations are required to occur at the locus of interest (in our case, just one mutation corresponded to the deletion), then only terms of the form $\mathbb{P}(U \mid G, E)$ need be modified. Slatkin and Rannala (1997) and R. C. Griffiths and S. Tavaré (unpublished results) considered the case where further molecular data are obtained *only* for those individuals carrying the UEP mutation. We have implemented this ascertainment scheme when the extra molecular data come in the form of DNA sequences. We have also implemented a version of the algorithm that allows both the mutation rates g and w to vary; this is equivalent to allowing the transition/transversion parameter κ to vary. Code that implements the methods described in this article is available in the form of C++ source code and executables from the authors, and at http://hto-e.usc.edu.

We thank the referees for helpful comments on an earlier version of this article. The authors were supported in part by National Science Foundation grant BIR 95-04393 and National Institutes of Health grant GM 58897.

LITERATURE CITED

Best, N. G., M. K. Cowles and S. K. Vines, 1995 *CODA Manual Version 0.30.* MRC Biostatistics Unit, Cambridge, UK.

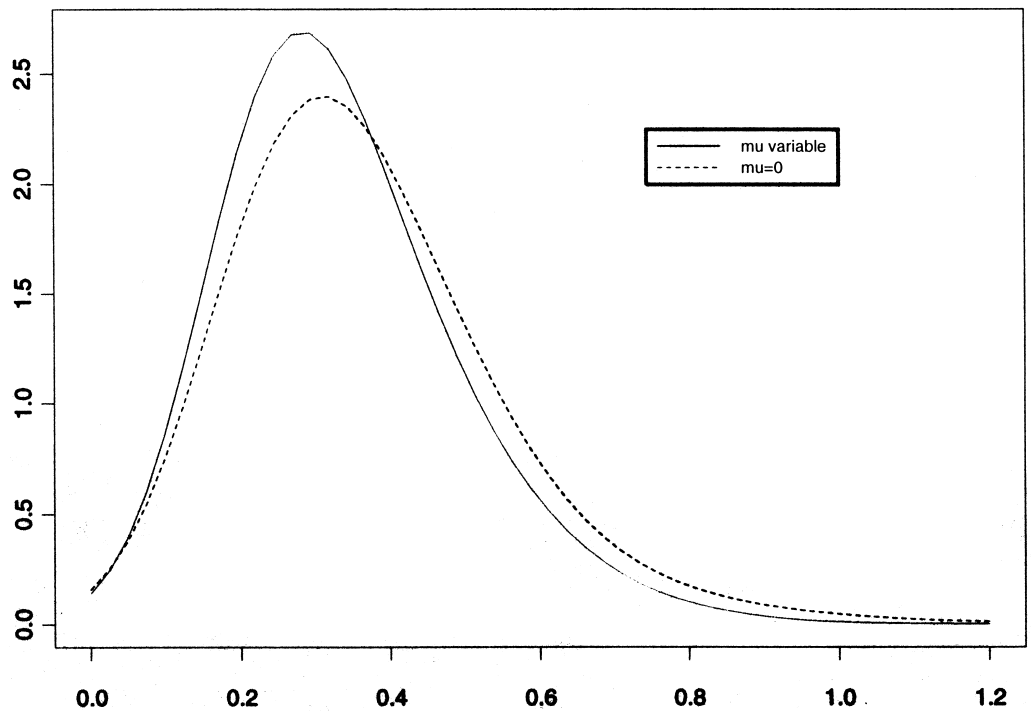

Figure 9.—Posterior density of age of deletion.

Donnelly, P., and S. Tavaré, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401–421.

Felsenstein, J., 1981 Evolutionary trees from DNA sequence data: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter (editors), 1996 *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London/New York.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comp. Biol. **3:** 479–502.

Griffiths, R. C., and S. Tavaré, 1994a Simulating probability distributions in the coalescent. Theor. Popul. Biol. **46:** 131–159.

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

Griffiths, R. C., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Stoch. Models **14:** 273–295.

Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97–109.

Hudson, R. R., 1991 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. Futuyma and J. Antonovics. Oxford University Press, Oxford.

Kingman, J. F. C., 1982 On the genealogy of large populations. J. Appl. Prob. **19A:** 27–43.

Kuhner, M. K., J. Yamato and J. Felsenstein, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

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

Larget, B., and D. L. Simon, 1999 Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. **16:** 750–759.

Lorenz, J. G., and D. G. Smith, 1994 Distribution of the 9-bp mitochondrial DNA region V deletion among North American Indians. Hum. Biol. **66:** 777–788.

Markovtsova, L., P. Marjoram and S. Tavaré, 2000 The effects of rate variation on ancestral inference in the coalescent. Genetics **156:** (in press).

Mau, R., M. A. Newton and B. Larget, 1999 Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics **55:** 1–12.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equations of state calculations by fast computing machine. J. Chem. Phys. **21:** 1087–1091.

Redd, A. J., N. Takezaki, S. T. Sherry, S. T. McGarvey, A. S. Sofro *et al.*, 1995 Evolutionary history of the COII/tRNA^{Lys} intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. Mol. Biol. Evol. **12:** 604–615.

Shields, G. F., A. M. Schmeichen, B. L. Frazier, A. Redd, M. I. Vovoeda *et al.*, 1993 mtDNA sequences suggest a recent evolutionary divergence for Beringian and Northern North American populations. Am. J. Hum. Genet. **53:** 549–562.

Slatkin, M., and B. Rannala, 1997 Estimating the age of alleles by use of intraallelic variability. Am. J. Hum. Genet. **60:** 447–458.

Soodyall, H., L. Vigilant, A. V. Hill, M. Stoneking and T. Jenkins, 1996 mtDNA control-region sequence variation suggests multiple independent origins of an "Asian-specific" 9-bp deletion in sub-Saharan Africans. Am. J. Hum. Genet. **58:** 595–608.

Stephens, M., 2000 Times on trees and the age of an allele. Theor. Popul. Biol. **57:** 109–119.

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

Thompson, E. A., and J. V. Neel, 1997 Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. Am. J. Hum. Genet. **60:** 197–204.

Thorne, J. L., H. Kishino and J. Felsenstein, 1992 Inching towards reality: an improved likelihood model of sequence evolution. J. Mol. Evol. **34:** 3–16.

Ward, R. H., B. L. Frazier, K. Dew and S. Pääbo, 1991 Extensive mitochondrial diversity within a single Amerindian tribe. Proc. Natl. Acad. Sci. USA **88:** 8720–8724.

Watkins, W. S., M. Bamshad, M. E. Dixon, B. Bhaskara Rao, J. M. Naidu *et al.*, 1999 Multiple origins of the mtDNA 9-bp deletion in populations of South India. Am. J. Phys. Anthropol. **109:** 147–158.

Wilson, I. J., and D. J. Balding, 1998 Genealogical inference from microsatellite data. Genetics **150:** 499–510.

Wiuf, C., and P. Donnelly, 1999 Conditional genealogies and the age of a neutral mutant. Theor. Popul. Biol. **56:** 183–201.

Communicating editor: G. A. Churchill