Inferring Coalescence Times From DNA Sequence Data

Simon Tavaré,* David J. Balding,[†,1] R. C. Griffiths[‡] and Peter Donnelly[§,2]

**Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, California 90089-1113,*
[†]*School of Mathematical Sciences, Queen Mary & Westfield College, University of London, London E1 4NS, United Kingdom,*
[‡]*Department of Mathematics, Monash University, Clayton VIC 3168, Australia and* [§]*Departments of Statistics, and Ecology & Evolution, University of Chicago, Chicago, Illinois 60637*

Manuscript received April 4, 1996
Accepted for publication October 30, 1996

ABSTRACT

The paper is concerned with methods for the estimation of the coalescence time (time since the most recent common ancestor) of a sample of intraspecies DNA sequences. The methods take advantage of prior knowledge of population demography, in addition to the molecular data. While some theoretical results are presented, a central focus is on computational methods. These methods are easy to implement, and, since explicit formulae tend to be either unavailable or unilluminating, they are also more useful and more informative in most applications. Extensions are presented that allow for the effects of uncertainty in our knowledge of population size and mutation rates, for variability in population sizes, for regions of different mutation rate, and for inference concerning the coalescence time of the entire population. The methods are illustrated using recent data from the human *Y* chromosome.

WITH the abundance of molecular genetic data that is now becoming available, and the interest in assessing current levels of human genetic diversity, attention has recently turned toward the question of what these data can tell us about human prehistory. More specifically, several recent papers have addressed the problem of inferring times since the most recent common ancestor of a sample of homologous DNA sequences drawn from a diverse range of contemporary humans.

Haploid sequences are convenient for such studies because the need to distinguish an individual's two haplotypes at an autosomal locus is then avoided. Studies have therefore tended to favor either the maternally inherited mitochondrial DNA (mtDNA) (*e.g.*, WALLACE 1995; WILLS 1995) or the male-specific part of the *Y* chromosome (HAMMER 1995; JOBLING and TYLER-SMITH 1995; WHITFIELD *et al.* 1995). These two sources of DNA represent, in effect, only two loci, one or both of which may be subject to selection, so that data from autosomal loci (AYALA 1995; HARDING *et al.* 1997) are also needed to obtain a reasonably good picture of recent human evolutionary history.

Extracting information about human history from these data requires careful modeling of the complex underlying processes such as mutation, demography and genealogical structure (DONNELLY 1996), by which we mean the ancestral relationships among the sequences. In recent years, a convenient mathematical framework has emerged for describing these processes, known as coalescent theory.

In Section 2, we give a brief outline of standard coalescent theory. This theory describes, in terms of probabilities, the ancestral relationships we would expect to find in a sample of sequences *before* any of the sequences are observed. Once the data have been examined, these probabilities can be revised in the light of the data. The correct method for revising the initial coalescent probabilities on the basis of the observed data is given by the usual rules of probability. The relevant calculations are, however, difficult, and a range of *ad hoc* alternative approaches have been developed in the literature. Some of these approaches are discussed in Section 4. Many of them make inefficient use of the data and some are fundamentally incorrect.

In this paper, we describe the correct approach for drawing inferences about coalescence times from sequence data, within the framework of infinitely-many-sites coalescent theory. Exact solutions to the inference problem in this setting have been obtained for samples of size two (TAJIMA 1983) (Section 3) and samples of any size that display no diversity (Section 5.2). In other cases, the theory and computational implementation of an exact solution are complex (GRIFFITHS and TAVARÉ 1994a). We describe in Section 5.1 a very simple and fast approximate simulation method based on replacing the full data set with only the number of segregating sites in the sample. Extensions are presented that allow for the effects of uncertainty in our knowledge of population size and mutation rates, for variability in population sizes, for regions of different mutation rate, and

Corresponding author: Simon Tavaré, Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, CA 90089-1113. E-mail: stavare@gnome.usc.edu

[1] *Present address:* Department of Applied Statistics, University of Reading, P.O. Box 240, Reading RG6 2FN, United Kingdom

[2] *Present address:* Department of Statistics, University of Oxford, Oxford, OX1 3TG, United Kingdom.

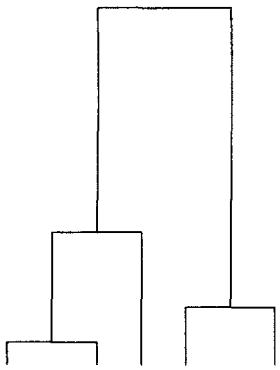

FIGURE 1.—Coalescent tree for a sample of five individuals.

for inference concerning the coalescence time of the entire population.

Our focus here is on methods that use population genetics modeling in inferring coalescence times from molecular genetic data. Another class of methods does not use population modeling (see Section 7.2 for an example). In principle, such approaches are inefficient, in not using all of the available information (see Section 3 for a comparison in the simplest case). They also suffer from various practical drawbacks. In particular, most implementations seriously underrepresent the uncertainty in the resulting estimates. For a critique of some of these difficulties in the context of human mtDNA data see TEMPLETON (1993).

2. The coalescent: Kingman's coalescent (KINGMAN 1982a) is a probability model for the genealogical tree of a random sample of n genes drawn from a large population. For recent reviews see HUDSON (1991), DONNELLY and TAVARÉ (1995). In the simplest formulation, the population size is a constant, N chromosomes, although this constraint is relaxed below. An example of a genealogical tree for a sample of size $n = 5$ is illustrated in Figure 1. Each branch tip represents a sequence in the current sample, and moving up the tree corresponds to going backward in time. Branches merge at a node when the associated sequences first share a common ancestor, so that the root of the tree occurs at the most recent common ancestor of the entire sample.

Time is measured continuously in the coalescent. In fact, the time W_j during which the sample has j distinct ancestors, $2 \le j \le n$, has the exponential distribution with parameter $j(j-1)/2$, the times for different j being independent. This description provides a close approximation to a range of population genetics models in which time is expressed in generations, provided that one unit of coalescent time is interpreted as N generations. An even larger class of models is approximated if a unit of coalescence time is interpreted as N/σ^2 generations, in which σ^2 is the variance in an individual's number of offspring (KINGMAN 1982a). Here, we shall assume $\sigma = 1$, but note that converting estimates of coalescence times into years requires a knowledge of σ.

There are two important quantities associated with a genealogical tree: the height of the tree, T_n, which is the time to the most recent common ancestor, and the length of the tree, L_n, which is the total of all the branch lengths. These are defined by

$$T_n = \sum_{j=2}^{n} W_j, \quad L_n = \sum_{j=2}^{n} jW_j. \tag{1}$$

The expectation of W_j is $\mathbb{E}(W_j) = 2/j(j-1)$, and so the expectations of T_n and L_n are given by

$$\mathbb{E}(T_n) = 2\left(1 - \frac{1}{n}\right), \quad \mathbb{E}(L_n) = 2\sum_{j=1}^{n-1} \frac{1}{j}. \tag{2}$$

Notice that as the sample size, n, gets large, $\mathbb{E}(T_n)$ approaches 2 units of coalescent time, equivalent to $2N$ generations, while $\mathbb{E}(L_n)$ increases without bound, growing like $2 \log(n)$. The variances of T_n and L_n are also readily obtained:

$$\text{Var}(T_n) = 8\sum_{j=2}^{n} \frac{1}{j^2} - 4\left(1 - \frac{1}{n}\right)^2,$$

$$\text{Var}(L_n) = 4\sum_{j=1}^{n-1} \frac{1}{j^2}. \tag{3}$$

For n large, Var (T_n) approaches $4\pi^2/3 - 12 \approx 1.16$, whereas Var (L_n) converges to $2\pi^2/3 \approx 6.58$. Notice that T_n, the height of the tree, has a high variance relative to its mean and that this ratio is not reduced by increasing the sample size. On the other hand, the length L_n has variance that becomes negligible relative to its mean as n increases. This is because L_n becomes dominated by the large number of very short branches that occur near the tips of the tree, whereas T_n is mainly affected by the two long branches that emanate from the root of the tree.

The times at which mutations occur are modeled in the coalescent by assuming that these times form a Poisson process of constant rate $\theta/2$. This means that if a branch of the tree has length w, then the number of mutations on that branch has a Poisson distribution with mean $w\theta/2$, independently of the mutations on the other branches. Here

$$\theta = 2N\mu, \tag{4}$$

in which μ is the mutation rate per gene per generation. If the data are DNA sequences, then μ is equal to the sequence length times the mutation rate per site per generation.

As well as describing the *locations* of mutations on the genealogical tree, a description of mutation *types* is also needed. A wide variety of different models for the mutation process can be incorporated into the coalescent. When the data are DNA sequences, the *infinitely-many-sites* model may be appropriate (WATTERSON 1975). In

this model, each gene is considered to be a sequence of completely linked sites, so that no recombination occurs within the sequence. Further, every mutation occurs at a site different from the sites of the previous mutations, so that a new segregating site arises.

It follows from the infinitely-many-sites assumption that, given the length of the tree L_n, the number S_n of segregating sites in the sample has a Poisson distribution with mean $\theta L_n/2$. Formally,

$$\mathbb{P}(S_n = k | L_n = l) = \mathrm{Po}(k, \theta l/2), \quad (5)$$

in which we introduce $\mathrm{Po}(k, \lambda)$ to denote the probabilities of the Poisson distribution,

$$\mathrm{Po}(k, \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad (6)$$

for $k = 0, 1, \ldots$, and $\lambda \geq 0$. The unconditional distribution of S_n is then obtained by integrating the conditional distribution (5) with respect to the distribution of L_n.

3. Samples of size two: It is helpful to consider initially this simplest case because both the principles underlying the correct approach to inference and the deficiencies of some alternative approaches are readily highlighted.

Under the infinitely-many-sites assumption, all of the information in the two sequences is captured in S_2, the number of segregating sites. Our goal, then, is to describe T_2, the time to the most recent common ancestor of the sample, or *coalescence* time, as fully as possible under the assumptions of the model and in the light of the data, which is the observed value of S_2.

One approach is to treat the realized value of T_2 as an unknown parameter that is then naturally estimated by $\tilde{T}_2 = S_2/\theta$, since $\mathbb{E}(S_2 | T_2) = \theta T_2$. Such an approach, however, does not use all of the available information. In particular, the information available about T_2 due to the effects of genealogy and demography are ignored.

Under the coalescent model, when $n = 2$ the coalescence time T_2 has an exponential distribution with mean 1 before the data are observed. As TAJIMA (1983) noted, it follows from Bayes Theorem that, after observing $S_2 = k$, the distribution of T_2 is gamma with parameters $1 + k$ and $1 + \theta$, which has probability density function (pdf)

$$f_{T_2}(t | S_2 = k) = \frac{(1+\theta)^{1+k}}{k!} t^k e^{-(1+\theta)t}, \quad t \geq 0. \quad (7)$$

In particular,

$$\mathbb{E}(T_2 | S_2 = k) = \frac{1+k}{1+\theta}, \quad (8)$$

$$\mathrm{Var}(T_2 | S_2 = k) = \frac{1+k}{(1+\theta)^2}. \quad (9)$$

The pdf (7) conveys all of the information available about T_2 in the light of both the data and the coalescent model. It represents a complete solution to the inference problem for coalescence time in the case $n = 2$. In some contexts it may be helpful to give interval estimates of T_2. For example, intervals that contain T_2 with probability 95% are readily obtained for particular data sets.

When, as here, the post-data distribution of T_2 is available, merely reporting a point estimate would usually be inappropriate. Nevertheless, if a point estimate were required, (8) suggests the choice $\hat{T}_2 = (1 + S_2)/(1 + \theta)$. Perhaps not surprisingly, the estimator $\hat{T}_2$, which is based on all of the available information, is superior to $\tilde{T}_2$ that ignores the pre-data information. For example, writing MSE for the mean square error of an estimator, straightforward calculations show that

$$\mathrm{MSE}(\hat{T}_2) = \frac{1}{1+\theta} < \frac{1}{\theta} = \mathrm{MSE}(\tilde{T}_2). \quad (10)$$

The difference in mean square errors could be substantial for small θ. In addition, the estimator $\tilde{T}_2$ is clearly inappropriate when $S_2 = 0$.

4. Previous approaches: In general terms, our problem is to describe the coalescence time T_n as fully as possible in the light both of the observed sequences and appropriate modeling assumptions. A complete solution is given by $f_{T_n}(t | D)$, $t > 0$, the pdf of T_n given the sequence data D. Within the infinitely-many-sites modeling framework, this solution is available in terms both of exact expressions and convenient simulation methodologies, described in Section 5. Here, we briefly discuss a number of alternative approaches that have been employed in the literature and point out some of their flaws and inefficiencies.

To distinguish clearly the distribution of T_n based on the genealogical model only and that based on both model and observed data, we will use the phrases "pre-data" and "post-data."

TEMPLETON (1993) considered the time since mitochondrial "Eve", that is the coalescence time of extant human mtDNA. For a particular reconstruction of the genealogical tree of the sampled sequences, he calculated the number of differences between each pair of sequences whose common ancestor is the root of the tree and then averaged this pairwise difference across all such pairs. He observed that the value, $\bar{k}$, of the average varied little over plausible reconstructed trees. He then substituted the value of $\bar{k}$ in place of k in (8) and (9), claiming that this represented the post-data mean and variance, respectively, of T, the coalescence time of the sampled sequences. He then referred to a result of KIMURA (1970) to the effect that the distribution of T is approximately gamma and obtained "confidence limits" for T using the gamma distribution with the mean and variance he had calculated.

There are several problems with this method. The results (8) and (9) give post-data moments for the

coalescence time of a sample of two randomly chosen sequences. Templeton chose pairs of sequences that were on different "sides" of the reconstructed genealogical tree. A pair of such sequences is *not* randomly chosen: they are chosen to have the longest coalescence time among all pairs of sequences in the tree and will thus tend to be more different than a typical pair. The discussion of the previous section, and in particular (8) and (9), do not apply for a pair of sequences chosen from different sides of a reconstructed tree. It is not clear what the correct expressions are for such sequences. Second, even if (8) and (9) had obtained for each of the pairs of sequences Templeton considered, it does not follow that they remain correct on substituting $\bar{k}$ for k.

DORIT *et al.* (1995) reported no variation in 38 human Y chromosome sequences of length 729 bp. They used a coalescent model of population genetics to infer confidence intervals for the coalescence time. Their analysis also contained serious errors. For a discussion see DONNELLY *et al.* (1996) and accompanying papers.

A different study of Y-chromosome variation was reported in HAMMER (1995). A 2.6-kb fragment of the male-specific portion of the Y chromosome was sequenced from 16 humans and four chimpanzees. The author presented estimates of the coalescence time for the human sequences. The largest value of k among all pairs was substituted into (8) and (9) to obtain an estimate of T and confidence intervals based on an approximating gamma distribution, as in (TEMPLETON 1993). This method is invalid for effectively the same reasons as Templeton's. In particular, (8) and (9) do not apply to a pair of sequences chosen because they are maximally different. In fact, DONNELLY and KURTZ (1997) have shown that the number of mutations between the two maximally different sequences in a sample of size n goes to infinity as n increases. This holds even though the coalescence time for the sample remains bounded. Use of the maximally different pair of sequences might thus be very misleading, particularly when the sample size is large.

5. Simulation methods for valid inferences: As stated above in Section 4, a complete solution to the problem of inferring T_n, the time to the most recent common ancestor of a random sample of n sequences, is given by $f_{T_n}(t|D)$, $t > 0$, the conditional pdf of T_n given the complete data D. In this section we describe methods for obtaining this solution within the infinitely-many-sites mutation model.

From the definition of conditional probability we have

$$f_{T_n}(t|D) = f_{T_n}(t)\frac{\mathbb{P}(D|T_n = t)}{\mathbb{P}(D)}. \qquad (11)$$

In words, the post-data pdf of T_n is given by the pre-data pdf times a factor that is the ratio of the probability of observing the data if $T_n = t$ to the unconditional probability of the data. Notice that (11) accords well with intuition in that, if the data are relatively more likely when $T_n = t$, then the post-data pdf will be larger than the pre-data pdf at that value of t, and vice versa.

Since $f_{T_n}(t|D)$ is a probability density function, there is no need to evaluate $\mathbb{P}(D)$ directly: it is a constant (with respect to t) and its value is determined by the constraint that $f_{T_n}(t|D)$ must integrate to one. In view of this simplification, we can write

$$f_{T_n}(t|D) \propto f_{T_n}(t)\,\mathbb{P}(D|T_n = t). \qquad (12)$$

Equation 12 is not directly useful in general, because an explicit expression for $\mathbb{P}(D|T_n = t)$ is only available in simple cases, such as $n = 2$ (Section 3) or $S_n = 0$ (Section 5.2). A computer-intensive approach is described in Section 5.6. First we describe a simple and fast approximate simulation method, based on replacing the complete data D with the summary statistic S_n.

5.1. Conditioning on S_n: An important simplification to (12) is obtained by replacing D with S_n, the number of segregating sites in the sample. The simplification arises because the distribution of S_n, described near (5), depends on the coalescent tree only through L_n, the length of the tree. The finer details of branch lengths and topological structure are irrelevant. Moreover, the distribution of L_n is completely specified in a form convenient for simulation by its definition (1) as a weighted sum of independent, exponential random variables.

Substituting $S_n = k$ for D in (12) and rearranging, we derive

$$\begin{aligned} f_{T_n}(t|S_n = k) &\propto f_{T_n}(t)\,\mathbb{P}(S_n = k|T_n = t) \\ &= \int_0^\infty f_{T_n L_n}(t, l) \\ &\qquad \times \mathbb{P}(S_n = k|T_n = t, L_n = l)\,dl \\ &= \int_0^\infty f_{T_n L_n}(t, l)\,\mathbb{P}(S_n = k|L_n = l)\,dl \\ &= \int_0^\infty f_{T_n L_n}(t, l)\,\mathrm{Po}(k, l\theta/2)\,dl, \qquad (13) \end{aligned}$$

in which $f_{T_n L_n}(t, l)$ denotes the joint pdf, under the coalescent model, of T_n and L_n. Evaluation of (13) by stochastic simulation is now straightforward, using the following algorithm:

Algorithm 1 *Rejection algorithm for $f_{T_n}(t|S_n = k)$.*

1. *simulate the W_j (independent, exponential random variables with parameter $j(j-1)/2$, $j = 2, \ldots, n$);*
2. *evaluate T_n and L_n according to the definition (1);*
3. *keep T_n with probability u, defined by*

$$u = \frac{\mathrm{Po}(k, L_n\theta/2)}{\mathrm{Po}(k, k)}, \qquad (14)$$

otherwise discard T_n and go to 1.

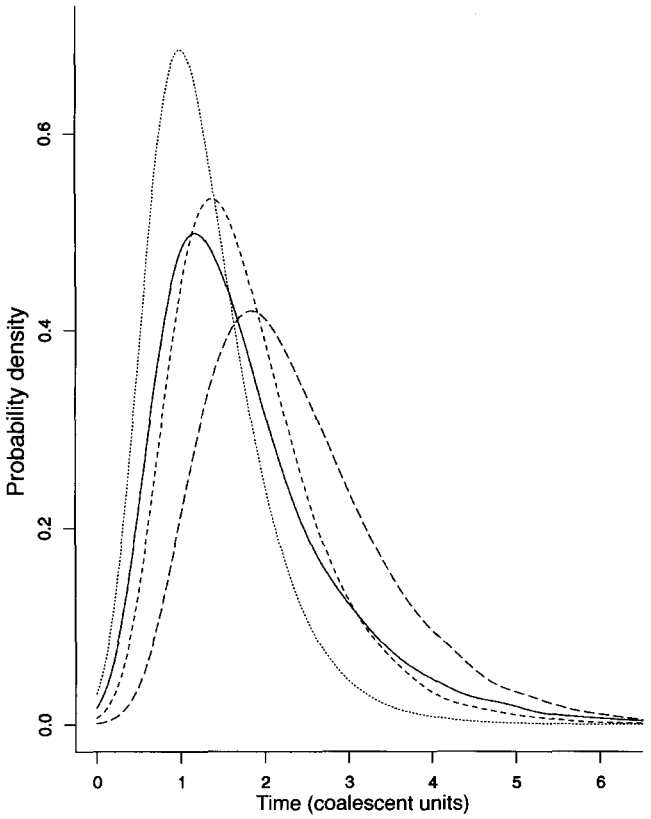

FIGURE 2.—Pre- and post-data density curves for T_{10} with $\theta = 1$. —, pre-data density; · · ·, $S_{10} = 1$; ---, $S_{10} = 3$; – – –, $S_{10} = 5$.

The resulting value of T_n is a sample of size one from the required distribution. The algorithm can be repeated arbitrarily often, to estimate quantities of interest as accurately as desired. Let $t_1, \ldots, t_m$ be the times returned from m iterations of Algorithm 1. The density $f_{T_n}(t|\, S_n = k)$ may be estimated from a histogram of the observations $t_1, \ldots, t_m$, or by using more sophisticated density estimation methods. Moments of the distribution can be estimated from the corresponding sample moments. For example, $(t_1 + \cdots + t_m)/m$ approximates the post-data mean of the coalescence time.

Notice that the denominator of (14) satisfies

$$\mathrm{Po}(k, k) = \max_{\lambda \geq 0} \mathrm{Po}(k, \lambda).$$

This constant can be replaced with any larger value, for example, 1. The resulting algorithm will be valid, but less efficient, since smaller values of u increase the chance that T_n is rejected. Algorithm 1 belongs to a class of simulation methods known as acceptance-rejection sampling (RIPLEY 1987). When the acceptance probability u is small, these algorithms can be time consuming. One might then use a Markov chain Monte Carlo method. A natural choice is the independence sampler with the pre-data distribution as the proposal (*cf.* GILKS *et al.* 1996, Chapter 1).

Figure 2 illustrates results from Algorithm 1 with $n = 10$ and $\theta = 1$. The solid curve indicates $f_{T_n}(t)$, the pre-data pdf of the coalescence time T_{10}. The other curves in the figure show the post-data density curves for T_{10} given three different observed values for S_{10}. Moments of each of the distributions for T_{10} are given in the first two columns of Table 1. From (2) and (3), the pre-data distribution of the length of the tree L_{10} has mean 5.66 and variance 6.16. Given the value of L_{10}, the number of segregating sites S_{10} has the Poisson distribution with mean $L_{10}/2$. It follows that S_{10} has mean 2.83 and variance 4.12. A feature of Algorithm 1 is that it is usually very fast: the simulations underlying Figure 2 and Table 1 require only a few seconds on a desktop workstation.

5.2. The case $S_n = 0$: A set of sequences displaying no variation was presented in DORIT *et al.* (1995) and discussed further in DONNELLY *et al.* (1996) (see also the accompanying discussion and authors' response). In this case, the data are fully summarized by the event $S_n = 0$. Although the simulation algorithms described earlier can be used, there is no need; exact results have been available for some time (TAVARÉ 1984). No segregating sites occur in the sample if and only if there are no mutations in the coalescent tree of the sample. Conditional on $D = \{S_n = 0\}$, it follows that the time $\tilde{W}_j$ during which the sample has j distinct ancestors has probability density proportional to

$$\tfrac{1}{2} j(j-1) \exp(-j(j-1)w/2) \cdot (\exp(-\theta w/2))^j.$$

Hence $\tilde{W}_j$ has an exponential distribution with parameter $j(j+\theta-1)/2$, and (since mutations are independent in different branches of the tree) the $\tilde{W}_j$ are independent random variables. Thus the post-data distribution of T_n is that of $\tilde{T}_n$ defined by

$$\tilde{T}_n = \tilde{W}_n + \cdots + \tilde{W}_2, \tag{15}$$

which has probability density function

$$f_{T_n}(t|\, S_n = 0) = \sum_{j=2}^{n} (-1)^j \times \frac{(2j+\theta-1)\, n_{[j]}\, (\theta+1)_{(j)}}{2(j-2)!\,(\theta+n)_{(j)}}\, e^{-j(\theta+j-1)t/2}, \tag{16}$$

where $x_{(j)} = x(x+1)\cdots(x+j-1)$ and $x_{[j]} = x(x-1)\cdots(x-j+1)$. This follows directly from equation 5.2 of TAVARÉ (1984).

It is clear that the mean time to the most recent common ancestor, given $S_n = 0$, is smaller than its pre-data mean, since

$$\mathbb{E}(T_n|\, D) = \sum_{j=2}^{n} \frac{2}{j(j+\theta-1)}.$$

Furthermore, given that the sample has no variability, the time to the most recent common ancestor of the sample is, as expected, stochastically smaller than the unconditional distribution.

5.3. The case $S_n = k > 0$: When θ is assumed known and the population size is assumed constant, some ex-

TABLE 1

Effect of uncertainty about θ

	$\theta = 1$		θ random, $\mathbb{E}(\theta) = 1$	
Data	Mean of T_{10}	Variance of T_{10}	Mean of T_{10}	Variance of T_{10}
None	1.80	1.16	1.80	1.16
$S_{10} = 1$	1.30	0.44	1.60	0.87
$S_{10} = 3$	1.79	0.75	1.78	1.04
$S_{10} = 5$	2.38	1.19	1.90	1.18

Pre-data and post-data moments of T_{10} for three values of S_{10}, for θ known and θ random. The pre-data values are exact, from (2) and (3), whereas the post-data values are estimated from 10,000 iterations of Algorithm 1 or Algorithm 2. For θ random, its value is $2N\mu$, where N is lognormal (9,1) and μ is gamma with shape parameter 2 and mean 1/26,719. It follows that θ has mean 1, median 0.48 and SD 1.75.

plicit results are available for the distributions of interest. In view of the independence structure, generating functions provide a useful tool. It follows from (5) and the independence of the waiting times W_j that

$$\begin{aligned}\phi_n(u, z) &\equiv \mathbb{E}(e^{-uT_n}z^{S_n}) \\ &= \mathbb{E}\mathbb{E}(e^{-uT_n}z^{S_n} \mid T_n, L_n) \\ &= \mathbb{E}(e^{-uT_n}e^{-\theta(1-z)L_n/2}) \\ &= \mathbb{E}(e^{-\sum_{j=2}^{n} W_j(u+\theta j(1-z)/2)}) \\ &= \prod_{j=2}^{n} \mathbb{E}e^{-(u+\theta(1-z)jW_j/2)} \\ &= \prod_{j=2}^{n} \frac{j(j-1)}{j(j-1+\theta(1-z))+2u}.\end{aligned}$$

The bivariate generating function $\phi_n(u, z)$ has the property that

$$\mathbb{E}(T_n z^{S_n}) = -\frac{\partial}{\partial u}\phi_n(u, z)\big|_{u=0}$$

(*cf.* FELLER 1968, Chapter XI). Differentiating and simplifying, we have

$$\mathbb{E}(T_n z^{S_n}) = \sum_{j=2}^{n} \frac{2}{j(j-1)} \frac{(j-1)}{j-1+\theta-\theta z}\mathbb{E}(z^{S_n}),$$

so that, by equating coefficients of z^k, we have

$$\mathbb{E}(T_n \mid S_n = k) = \frac{1}{\mathbb{P}(S_n = k)} \sum_{l=2}^{n} \frac{2}{l(l-1)} \times \sum_{m=0}^{k} \mathbb{P}(S_n = m)\left(\frac{\theta}{l+\theta-1}\right)^{k-m} \frac{l-1}{l+\theta-1}. \quad (17)$$

An explicit expression for $\mathbb{P}(S_n = m)$ appears in equation 9.5 of TAVARÉ (1984).

A similar analysis establishes that the conditional density $f_{T_n}(t \mid S_n = k)$ has the form

$$f_{T_n}(t \mid S_n = k) = \frac{1}{\mathbb{P}(S_n = k)}[z^k]\phi_n(0, z)g_n(t, \theta(1-z)), \quad (18)$$

where $g_n(t, \theta)$ is the density in (16), and $[z^k]f(z)$ is the coefficient of z^k in $f(z)$.

5.4. Incorporating uncertainty about N and μ: The analyses of Section 5.1 are conditional on the value of the scaled mutation parameter θ and thus are directly useful only when θ is known, which usually requires knowledge both of the (haploid) population size N and of the mutation rate μ. In practice, there will be substantial uncertainty about θ. Often, the value of μ can be estimated by comparisons with homologous sequences in related species, but there will always be some uncertainty in such estimates. Further, there is typically little information available about the (variance effective) population size.

The sequence data will be informative about both the quantity of direct interest, T_n, and the other unknowns such as N and μ. For example, an observation of little variation in the sample, compared with expectations under the model, is evidence for a relatively small value of T_n. However, such an observation may be due in part to the value of either N or μ, or both, being smaller than had been thought. An analysis that ignores uncertainty in N and μ may give misleading results since the effect of a small value of N or μ may wrongly be attributed to T_n.

Fortunately, Algorithm 1 is readily modified to allow for uncertainty about N and μ. We assume that, before observing the data, N and μ are mutually independent random quantities, and independent of T_n and L_n. Equation 13 can then be rewritten

$$f_{T_n}(t \mid S_n = k) \propto \int_0^\infty \int_0^\infty \int_0^\infty f_{T_n L_n}(t, l) \times \pi_N(u)\pi_\mu(v)\,\mathrm{Po}(k, luv)\,dl\,du\,dv, \quad (19)$$

in which we introduce π_N and π_μ for the pre-data densities of N and μ. Although N is, strictly speaking, a discrete variable, it is convenient, and leads in practice to negligible error, to describe it mathematically as a continuous variable.

The pre-data distributions π_N and π_μ should be chosen so as to summarize the information available about N and μ, for example from relevant genetic and anthro-

pological studies. Typically, such information will not uniquely specify π_N and π_μ. In practice, therefore, it is prudent to consider several different plausible choices and to investigate the sensitivity of conclusions to different formulations of the pre-data information.

Equation 19 suggests the following modification to Algorithm 1:

Algorithm 2 *Modified rejection algorithm for* $f_{T_n}(t|\, S_n = k)$.

1. *simulate* N *from* π_N;
2. *simulate* μ *from* π_μ;
3. *simulate the* W_j *(independent, exponential random variables with parameter* $j(j-1)/2$, $j = 2, \ldots, n$);
4. *evaluate* T_n *and* L_n *according to the definition (1);*
5. *keep* T_n, N, *and* μ *with probability* u, *defined by*

$$u = \frac{\mathrm{Po}(k, N\mu L_n)}{\mathrm{Po}(k, k)}, \tag{20}$$

otherwise discard them and go to 1.

Accepted values have the joint post-data distribution of (T_n, N, μ). These observations can be used as before to study properties of this post-data distribution. For example, it is often useful to give ancestral times in years rather than coalescent units. Denoting the time in years by T_n^y, we have

$$T_n^y = N \times G \times T_n,$$

where G denotes the number of years per generation, and we continue to assume that $\sigma = 1$. The post-data distribution of T_n^y can be found by returning values of NGT_n from each set of (T_n, N, μ) values accepted in step 5 of Algorithm 2. The post-data distribution of N and $\theta = 2N\mu$ can be studied similarly.

Some results from use of Algorithm 2 are displayed in the final two columns of Table 1. Notice that allowing for uncertainty in θ draws the expected value of T_{10} closer to the pre-data value, compared with the θ known case, while also increasing its post-data variance.

We remark that the rejection methods exploited here can be used to study a variety of other problems. For example, one might be interested in inferring times to the common ancestor of a sample given that the segregating sites arose in specified positions in the coalescent tree. All the method requires is that we be able to calculate the analogues of the rejection probabilities (14) and (20). It is also clear that the methods generate observations on the post-data distributions of the coalescence times W_j themselves; merely return the full vector $(W_2, \ldots, W_n)$ rather than the summary statistic T_n. Computer implementations of these algorithms are available from the authors.

5.5. Mean pairwise differences: Instead of simplifying the complete data set by reporting only the number of segregating sites in the sample, which leads to the inferential method described above, other approaches to data reduction sometimes employed are based on Π_n, the average of the nucleotide differences over all pairs of sequences in the sample. Unlike S_n, the distribution of Π_n does depend on the details of the genealogy and so a detailed simulation method, such as that described in Section 5.6, is required. However, since this approach is not simpler than the exact method, there seems no advantage in pursuing it.

5.6. Exact simulation methods: Under the infinitely-many-sites model, the full data D are equivalent to an unrooted tree (GRIFFITHS and TAVARÉ 1995). The probability distribution of such trees can be determined recursively (GRIFFITHS 1989), though direct recursive computation is feasible only for small sample sizes. In practice, Markov chain simulation techniques can be used to approximate any required probabilities to arbitrary accuracy. Further details and examples are given in GRIFFITHS and TAVARÉ (1994a, 1995). The same computer-intensive approach may be used to find the post-data distribution of T_n given D; see GRIFFITHS and TAVARÉ (1994a). In practice these methods can be quite time consuming, especially for large sample sizes n. The approximate methods based on replacing the full data D by the summary statistic S_n, and using Algorithms 1 and 2, are much quicker and hence allow a wider range of modeling assumptions to be investigated. An assessment of how this data summary affects the post-data distribution of T_n is given in GRIFFITHS and TAVARÉ (1996). We return to this issue in Section 8.

5.7. Variable population size: The effect of variable population size is to change the joint distribution of the times W_j (KINGMAN 1982b); see also GRIFFITHS and TAVARÉ (1994b), SLATKIN and HUDSON (1991), TAJIMA (1989), and DONNELLY and TAVARÉ (1995). In particular, these times are no longer independent. Suppose that the population size at the time of sampling is N, and measure time in units of N generations. We write $Np(t)$ for the population size a time t ago, and define $\lambda(t) = 1/p(t)$. Under a wide class of demographic models, the conditional distribution of the time W_j for which there are exactly j ancestors of the sample, given that the time in which there are more than j ancestors is s, is

$$\mathbb{P}(W_j > t|\, W_n + \cdots + W_{j+1} = s) = \exp\left(-\binom{j}{2}\int_s^{s+t} \lambda(u)\, du\right). \tag{21}$$

This provides a direct way to simulate times W_n, W_{n-1}, $\ldots$, W_2 having the required distribution; see GRIFFITHS and TAVARÉ (1994a), for example, for conditions under which (21) is valid.

A very useful way to think of the process $A^v(\cdot)$ that counts the number of distinct ancestors of a sample is to write

$$A^v(t) = A(\Lambda(t)), \quad t \geq 0, \tag{22}$$

where $A(\cdot)$ is the corresponding process for the constant population size case, and

$$\Lambda(t) = \int_0^t \lambda(s)\, ds.$$

Thus the variable population size model is just a deterministic time change of the constant size model. Despite this fact, it is often difficult to provide useful explicit results for many quantities of interest. We remark that Algorithms 1 and 2 may be employed directly as long as the W_j are simulated with the distribution determined by (21).

5.8. Other demographic scenarios: The coalescent approximation described in Section 2 applies to populations that are panmictic, and whose size is constant through time. The methods described in Sections 5.1–5.7 aim to use background knowledge of population genetics, together with the information in the data, to infer coalescence times. For these methods to be appropriate, it is necessary that the assumptions underlying the population genetics models apply (at least approximately) to the population from which the data are obtained.

The previous section explained how to relax the assumption of constant population size. Recall that the only alteration to the simulation Algorithms 1 and 2 was that the times W_j between coalescences were simulated from the distribution appropriate for the model with variable population sizes.

There are other situations in which the version of the coalescent described in Section 2 does not adequately describe the sample genealogy. In some such situations, the appropriate genealogical structure, and in particular the pre-data distribution of the times W_j, is known. These include models for populations with certain forms of geographical structure, genealogy at a neutral locus that is tightly linked to one at which a selective sweep has recently occurred, and a neutral locus that is tightly linked to a locus at which balancing selection is operating. For further details, see, for example, HUDSON (1991) and references therein.

Inferences about coalescence times on the basis of DNA sequence data can be undertaken for any of these situations, along the lines of the algorithms described above. The only change that is needed is that in step 1 of Algorithm 1 and step 3 of Algorithm 2, the times W_j between coalescences should be simulated from the version of the coalescent that is appropriate for the particular genetic or demographic scenario under consideration.

If limited information is available about the demographic scenario that actually pertained during the relevant period of the population's history, it would be prudent to undertake an analysis under several different assumptions to assess the sensitivity of the conclusions to the initial assumptions.

5.9. Variable mutation rates: These rejection methods can be employed directly to study the behavior of the infinitely-many-sites model that allows for several regions with different mutation rates. Suppose then that there are r regions, with mutation rates $\mu_1, \ldots, \mu_r$. The analysis also applies, for example, to r different types of mutations within a given region. We sample n individuals, and observe k_1 segregating sites in the first region, k_2 in the second, ..., and k_r in the rth. The problem is to find the conditional distribution of T_n, given the vector $(k_1, \ldots, k_r)$.

When N and the μ_i are assumed known, this can be handled by a modification of Algorithm 1. Conditional on L_n, the probability of $(k_1, \ldots, k_r)$ is

$$h(L_n) = \mathrm{Po}(k_1, L_n\theta_1/2) \times \cdots \times \mathrm{Po}(k_r, L_n\theta_r/2),$$

where $\theta_i = 2N\mu_i$, $i = 1, 2, \ldots, r$. It is easy to check that $h(L_n) \le h(k/\theta)$, where

$$k = k_1 + \cdots + k_r, \quad \theta = \theta_1 + \cdots + \theta_r.$$

Therefore in the rejection algorithm we may take $u = h(L_n)/h(k/\theta)$, which simplifies to

$$u = h(L_n)/h(k/\theta) = \frac{\mathrm{Po}(k, L_n\theta/2)}{\mathrm{Po}(k, k)}. \quad (23)$$

Equation 23 establishes the perhaps surprising fact that the conditional distribution of T_n given $(k_1, \ldots, k_r)$ and $(\theta_1, \ldots, \theta_r)$ depends on these values only through their respective totals: the total number of segregating sites k and the total mutation rate θ. Thus Algorithm 1 can be employed *directly* with the appropriate values of k and θ. This result justifies the common practice of analyzing segregating sites data through the total number of segregating sites, even though these sites may occur in regions of differing mutation rate.

If allowance is to be made for uncertainty about the μ_i, then this simplification no longer holds. However, Algorithm 2 can be employed with the rejection step (20) replaced by (24):

$$u = \frac{\mathrm{Po}(k_1, L_n\theta_1/2)}{\mathrm{Po}(k_1, k_1)} \cdots \frac{\mathrm{Po}(k_r, L_n\theta_r/2)}{\mathrm{Po}(k_r, k_r)}. \quad (24)$$

In this case, step 2 requires generation of a vector of rates $\mu = (\mu_1, \ldots, \mu_r)$ from the joint prior π_μ. Furthermore, the algorithm immediately extends to the case of variable population size.

6. The time to the MRCA of a population given data from a sample: In this section, we show how the rejection technique can be used to study the time T_m to the MRCA of a sample of m individuals, conditional on the number of segregating sites in a subsample of size n. In many applications of ancestral inference, the real interest is on the time to the MRCA of the *population*, given data on a *sample*. This can be obtained by setting $m = N$ below.

The quantities of interest here are A_m (the number of distinct ancestors of the sample), A_n (the number

of distinct ancestors of the subsample), and T_n (the time to the MRCA of the subsample). The results of SAUNDERS *et al.* (1984) justify the following algorithm.

Algorithm 3 *Rejection algorithm for* $f_{T_m}(t \mid S_n = k)$.

1. *Set* $A_m = m$, $A_n = n$, $T_n = 0$, $L_n = 0$.
2. *Generate W, exponential of rate* $A_m(A_m - 1)/2$. *Set* $T_n = T_n + W$, $L_n = L_n + A_n \cdot W$.
3. *Set* $p = A_n(A_n - 1)/A_m(A_m - 1)$. *Set* $A_m = A_m - 1$. *With probability p set* $A_n = A_n - 1$. *If* $A_n > 1$ *go to 2.*
4. *Set* $u = Po(k, \theta L_n/2)/Po(k, k)$. *Accept* (A_m, T_n) *with probability u, else go to 1.*
5. *If* $A_m = 1$, *set* $T_{nm} = 0$, *and return* $T_m = T_n$. *Else, generate independent exponentials* W_j *with parameter* $j(j-1)/2$, *for* $j = 2, 3, \ldots, A_m$, *and set* $T_{nm} = W_2 + \cdots + W_{A_m}$. *Return* $T_m = T_n + T_{nm}$.

Many aspects of the joint behavior of the sample and a subsample can be studied using this method. In particular, values of (A_m, T_n) accepted at step 5 have the joint conditional distribution of the number of ancestors of the sample at the time the subsample reaches its common ancestor and the time of the MRCA of the subsample, conditional on the number of segregating sites in the subsample. In addition, values of T_{nm} produced at step 5 have the conditional distribution of the time between the two most recent common ancestors. It is straightforward to modify the method to cover the case of variable population size, and the case where uncertainty in N and μ is modeled. With high probability, the sample and the subsample share a common ancestor and therefore a common time to the MRCA. However, if the two common ancestors differ, then the times to the MRCA can differ substantially. This is explored further in the examples below.

7. Examples: In this section we illustrate the methods described above by applying them to two recently published data sets on DNA variation on the human Y chromosome. One of the original papers aimed to use coalescent-based methods in estimating coalescence times, the other used methods that do not make use of population modeling. We concentrate on methods that assume a constant-sized, panmictic, population, and focus on the effects on the conclusions of various levels of uncertainty about the parameters involved. Use of such data in making inferences about early human evolution should involve a more detailed analysis of the extent to which the conclusions depend also on the underlying demographic assumptions. These can be investigated using the methods described in Sections 5.7 and 5.8.

7.1. Hammer data: HAMMER (1995) sequenced a 2.6-kb fragment containing a polymorphic Alu insertion for a sample of $n = 16$ human Y chromosomes. Having observed three segregating sites, he estimated the time to the common ancestral human Y chromosome to be 188,000 years, with a 95% confidence interval of 51,000–411,000 years.

The results of several reanalyses of the data are summarized in Table 2. The analyses differ in the aspects of the data they utilize and the amount of uncertainty associated with underlying parameters.

Line a of Table 2 summarizes the inference in HAMMER (1995). As mentioned above, there are several flaws in the methodology of the paper. Lines b and c present reanalyses of the data HAMMER used. In each case we have used the same values as HAMMER for the other parameters: a value of 4900 for N, the variance effective population size, and a value for μ_S, the rate per generation of point mutations in the region considered, of 9.88×10^{-5}, corresponding to $2600 \times 1.9 \times 10^{-9} = 4.94 \times 10^{-6}$ substitutions per sequence per year, with a generation time of 20 years.

The three polymorphic nucleotides in the data are consistent with the infinitely-many-sites assumption (GRIFFITHS and TAVARÉ 1994a). Analysis b is based on the number of segregating sites (here $S_{16} = 3$) as a summary of the data, and uses Algorithm 1. In contrast, analysis c uses the full data set and the method of GRIFFITHS and TAVARÉ (1994a). We note that, in this example, the analysis based on the full data gives similar results to that that uses only the summary statistic S_n, although the former has a smaller range of uncertainty. Both analyses also lead to similar conclusions to those originally reported by HAMMER.

The data also show presence or absence of a YAP element, which can have either a long or short poly(A) tail. This does not appear to have been utilized in HAMMER (1995) for estimating coalescence times. It seems plausible that the YAP element was inserted just once and we assume here that there was a single insertion event. Little information is available about the insertion rate, but it seems likely to be substantially smaller than the nucleotide substitution rate. We will not attempt to model the mutation mechanism that resulted in the lengthening or shortening of the YAP element, and so do not distinguish between the long and short version of the insert.

Analyses d and e use the YAP insertion in addition to the three polymorphic nucleotide sites. They use the values for N and μ_S given above, and in addition assume the rate of insertions, per generation, in the region examined, to be $\mu_I = 9.88 \times 10^{-8}$. Analysis d is based on segregating sites, three of which pertain to point mutations and one to the insertion. Since there are two different types of mutation we use the methods discussed in Section 5.9. Recall that the analysis based on segregating sites only depends on the total number (here four) of segregating sites and the sum of the mutation rates (here 9.89×10^{-5}). It follows that the analysis will not depend sensitively on the value of μ_I (about which little is known) provided this is small relative to μ_S. Analysis e is based on the full data for both the polymorphic nucleotides and the YAP insertion (although without distinguishing short from long

TABLE 2

Reanalyses of data of HAMMER

	Data	Model	Mean of T_{16} ($\times 10^3$)		95% Interval ($\times 10^3$)	
			Pre-data	Post-data	Pre-data	Post-data
(a)	$S_{16} = 3$	HAMMER (11)		188		51–411
(b)	$S_{16} = 3$	$N = 4900$ $\mu_S = 9.88 \times 10^{-5}$	184	173	56–460	62–377
(c)	Full data (not insert) $S_{16} = 4$	$N = 4900$ $\mu_S = 9.88 \times 10^{-5}$ $N = 4900$	184	172	56–460	65–341
(d)	(3 subs, 1 insert)	$\mu_S = 9.88 \times 10^{-5}$ $\mu_I = 9.88 \times 10^{-8}$ $N = 4900$	184	198	56–460	71–426
(e)	Full data	$\mu_S = 9.88 \times 10^{-5}$ $\mu_I = 9.88 \times 10^{-8}$ $N = 4900$	184	193	56–460	86–343
(f)	$S_{16} = 4$	μ_S gamma μ_I lognormal N gamma	184	210	56–460	72–485
(g)	$S_{16} = 4$	$\mu_S = 9.88 \times 10^{-5}$ $\mu_I = 9.88 \times 10^{-8}$ N gamma	186	200	34–560	65–455
(h)	$S_{16} = 4$	μ_S gamma μ_I lognormal N lognormal	186	239	34–560	61–606
(i)	$S_{16} = 4$	μ_S gamma μ_I lognormal	492	484	27–2,320	74–1,738

Line a gives the results reported by the author. Reanalyses are given both without a–c and with d–i taking the YAP insert into account. Mean and 95% interval are estimated from samples of size 10,000 generated by Algorithm 1 for b and d, and Algorithm 2 for f–i, while c and e make use of the exact simulation methods described in GRIFFITHS and TAVARÉ (1994a). Details of the gamma and lognormal distributions used to model uncertainty about N, μ_S and μ_I are given in the text.

inserts). It uses an extension of the algorithm in GRIFFITHS and TAVARÉ (1994a). The effect of including the YAP insert in the analysis is to increase both the mean of the post-data distribution of the coalescence time, and the uncertainty surrounding this time.

In practice, of course, we do not know the values of the parameters N, μ_S, and μ_I. There is uncertainty about each of these, perhaps rather more so for N and μ_I. Any analysis that treats these parameters as known will underrepresent the uncertainty in inferences about the coalescence time. The methods of Section 5.4, and in particular Algorithm 2, can be used to incorporate uncertainty about these parameters. Each of the analyses f–i is based on the number of segregating sites of each of the two types (three polymorphic sites, one insertion) in the data.

The analysis described at line f of Table 2 treats N as known (using the same value, 4900, used in analyses a–e) but incorporates uncertainty about the mutation rates μ_S and μ_I. We used a gamma distribution with mean 9.88×10^{-5} and shape parameter 2 to encapsulate uncertainty about the polymorphic site rate μ_S (Figure 3A). For the insertion rate μ_I we used a lognormal prior with parameters $(-17.5, 1.5)$. This distribution has mean 5.3×10^{-8}, fifth percentile 2.1×10^{-9} and 95th percentile 3.0×10^{-7}. There are no compelling reasons for the particular choices of gamma and lognormal here. These distributions both have the desirable properties of being smooth, unimodal and excluding negative values. In addition, the lognormal has heavier tails, reflecting greater uncertainty about the value of μ_I. Having adopted these functional forms, the parameter values were chosen to give desired means and variances.

Analysis g incorporates uncertainty about the value of N, while assuming that both μ_S and μ_I are known exactly. We modeled uncertainty about N in the form of a gamma random variable with mean 5000 and shape parameter 5 (Figure 3B, solid curve). This distribution has fifth percentile 1970; mode 4000; median 4671; mean 5000; and 95th percentile 9154. It thus concentrates largely on values between 0 and 10,000, and is "centred" around the value 4900 used by HAMMER, which, as he notes, is supported by other studies (TAKAHATA 1993; FULLERTON *et al.* 1994).

Analyses h and i incorporate the effects of uncertainty about each of N, μ_S and μ_I. Analysis i differs from h in examining the effect of a larger range of prior uncertainty about the value of N. Specifically, it incorporates a lognormal (9, 1) prior (Figure 3B, dashed curve). This distribution has fifth percentile 1564; mode 2981; median 8103; mean 13,360; and 95th percentile 41,976.

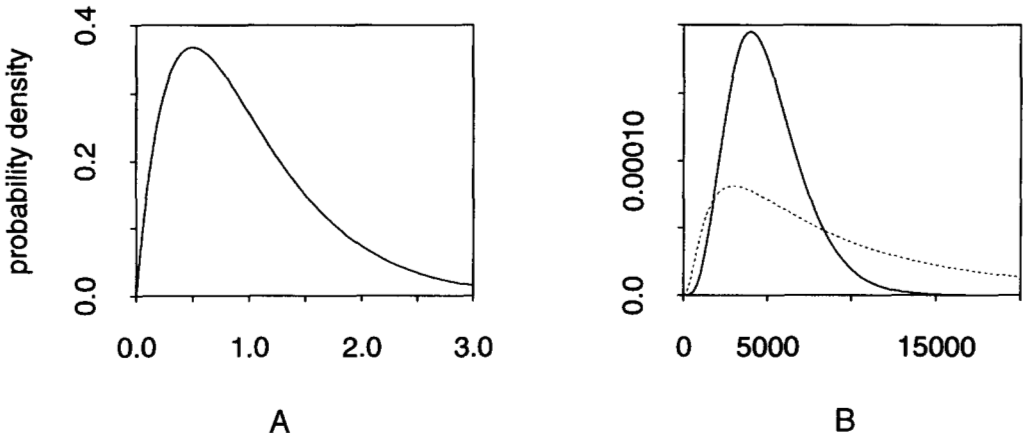

FIGURE 3.—(A) Probability density for the gamma distribution with mean 1 and shape parameter 2. This distribution, scaled so that the mean matched the original authors' point estimates, was used to model uncertainty about the polymorphic site mutation rate, μ_S. (B) Probability densities for the two distributions used to model uncertainty about the (variance effective) population size N (number of chromosomes). Solid curve, gamma distribution with mean 5000 and shape parameter 5; dashed curve, lognormal distribution with parameters 9 and 1.

As expected, each of the analyses f–i results in a wider range of uncertainty about the coalescence time T than the analyses that treat the parameters as known. (The natural comparison is with analysis d.) In addition, the post-data distribution of T, and in particular its mean, is shifted toward larger values in each case. Specifically, in comparing d with f and g, we see that the failure to allow for uncertainty about first the mutation rates and second the effective population size, may result in underestimation of T.

Direct comparison of analysis i with the other analyses in the table requires some caution. Recall that times in the coalescent analysis must be converted to generations through multiplication by N. Increasing the value of N will affect the coalescent analysis (by increasing θ), but it will also have a direct effect on the resulting times through this multiplication. [Even with data showing no variation, the direct effect of the multiplication by N can outweigh the indirect effect of shortening the coalescence time (DONNELLY *et al.* 1996), notwithstanding the comments of DORIT *et al.* (1996).]

Analysis i allows for the possibility that N may be substantially larger than 4900. It is thus not surprising that this results in rather larger post-data estimates of T. On the other hand, we do not have good prior information about N, and the beliefs encapsulated in this analysis, and the resulting conclusions, may not be unreasonable. At the very least, this analysis shows that inferences about T can depend sensitively on what is known about the relevant population size. Further, uncertainty about the value of N can result in considerable uncertainty about the coalescence time.

7.2. Whitfield data: WHITFIELD *et al.* (1995) describe another *Y*-chromosome data set that includes a sample of $n = 5$ humans. The 15,680-bp region has three polymorphic nucleotides that once again are consistent with the infinitely-many-sites model. WHITFIELD *et al.* estimated the coalescence time of the sample to be between 37,000 and 49,000 years. Again, we present several reanalyses, each of which is based on the number of segregating sites in the data. The results are summarized in Table 3 and illustrated in Figure 4.

In estimating the coalescence time, WHITFIELD *et al.* adopt a method that does not use population genetics modeling. While the method is not systematically biased, it may be inefficient to ignore pre-data information about plausible values of the coalescence time. In addition, the method substantially underrepresents the uncertainty associated with the estimates presented (see TEMPLETON 1993).

Here, we contrast the results of such a method with those of one that does incorporate background information. The conclusions of a coalescent analysis will depend on the assumptions about N, whereas the approach of WHITFIELD *et al.* (1995) does not involve this parameter. There is thus no particular value of N on which to base a direct comparison of the two methods. Pre-data beliefs about the variance effective male population size should not depend on the region of the Y chromosome being examined, and in particular should be the same for the analyses of the data of Sections 7.1 and 7.2. We use the same three assumptions about N that we employed in our analyses of HAMMER's data, for the reasons given earlier and to facilitate comparisons.

While it is natural to make the same assumptions about N in each analysis, the mutation rate per site may differ between the two regions investigated. We use the average figure of 1.123×10^{-9} substitutions per nucleotide position per year given in WHITFIELD *et al.* (1995), and a generation time of 20 years, to give $\mu = 15{,}680 \times 1.123 \times 10^{-9} \times 20 = 3.52 \times 10^{-4}$ substitutions per generation. For these parameter values, the post-data mean of T_5 is 87,000 years, which is much less than any of the post-data means of T_{16} for the HAMMER *et al.* data, but much greater than the upper estimate given by WHITFIELD *et al.* (1995).

TABLE 3

Renalyses of data of WHITFIELD *et al.*

	Model	Mean of T_5 ($\times 10^3$)		95% interval ($\times 10^3$)	
		Pre-data	Post-data	Pre-data	Post-data
(a)	WHITFIELD *et al.*				37–49
(b)	$N = 4900$ $\mu_S = 3.52 \times 10^{-4}$	157	87	31–429	30–184
(c)	$N = 4900$ μ_S gamma	157	125	31–429	32–321
(d)	N gamma $\mu_S = 3.52 \times 10^{-4}$	159	80	21–517	26–175
(e)	N gamma μ_S gamma	159	117	21–517	25–344
(f)	N lognormal μ_S gamma	428	149	19–2200	22–543

In each case the data are $S_5 = 3$. Line a gives the interval reported by the authors (but note that they assigned no probability to their interval). Mean and 95% interval are estimated from samples of size 10,000 generated by Algorithm 1 for b, and Algorithm 2 for c–f. The gamma and lognormal distributions are those of Figure 3; details are given in the text.

As noted in the previous section, the appropriate values of the parameters are not known. Analysis c incorporates uncertainty about μ, in the form of a gamma distribution with shape parameter 2 and mean 3.52×10^{-4}, while continuing to assume that N is known to be 4900. The effect is to greatly increase the post-data mean of T. Allowing N to be uncertain while μ_S is known has, on the other hand, the effect of slightly reducing the post-data estimates of T_5, compared with the case that N and μ_S are both known. This may be attributed to the data favoring values of N smaller than 4900.

Analyses e and f incorporate uncertainty about both N and μ_S. They use the same prior distributions as analyses g and i, respectively, of the previous section. Note that, as should be expected, the uncertainty about T is larger than when one or both of N and μ_S are assumed known exactly. The post-data distributions of T in e and f of Table 3 are shifted toward smaller values than are the respective lines of Table 2, although there is considerable overlap between the two distributions. (Because of the lack of recombination within the relevant region of the Y chromosome, the entire population must have

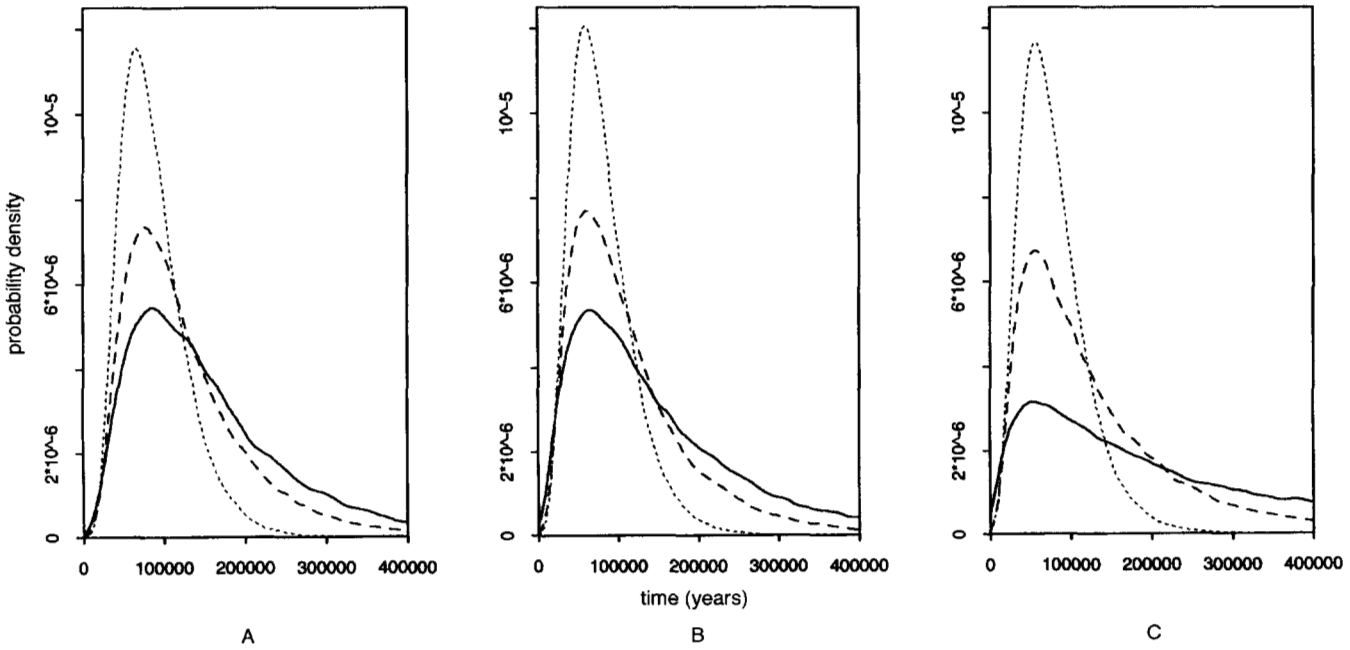

FIGURE 4.—Probability density curves for T_5 based on the data of WHITFIELD *et al.* (1995). In each panel the three curves correspond as follows: solid, pre-data; dashed, post-data assuming μ_S gamma; dotted, post-data assuming $\mu_S = 3.52 \times 10^{-4}$. The three panels correspond to (A) $N = 4{,}900$, (B) N gamma, (C) N lognormal. The gamma and lognormal distributions are those of Figure 3, details are given in the text.

the same coalescence time at the different regions considered by the two studies, but note that this need not be true of the samples examined in each study.)

WHITFIELD *et al.* (1995) point to their estimated coalescence time as being substantially shorter than those published for the human mitochondrial genome. In contrast, the ranges in each of our analyses b–e overlap with recent interval estimates (see for example TEMPLETON 1993) for the time since mitochondrial Eve. In addition, recall that the quantity T_5 being estimated in Table 3 is the coalescence time of the sample of five males sequenced in the study. As noted in Section 6, this time may be different from, and substantially shorter than, the coalescence time of *all* existing *Y* chromosomes. Under the assumption that $N = 4900$ and $\mu = 3.52 \times 10^{-4}$, Algorithm 3 can be used to show that the mean time to the common ancestor of the male *population*, given $S_5 = 3$, is 157,300 years, with a corresponding 95% interval of (58,900 – 409,800) years. These figures differ markedly from the corresponding values for the sample, given at line b of Table 3. It is the population values that are likely to be of primary interest.

8. Discussion: We have described a simple simulation approach based on the acceptance-rejection method to generate observations from the post-data distribution of TMRCA. This allows us to explore this distribution in settings in which analytical expressions are either intractable or uninformative. When the acceptance probability is very small the method can be slow; alternatives based on Markov chain Monte Carlo can then be exploited (*cf.* GILKS *et al.* 1996, Chapter 1).

The rejection method is particularly useful when the data are summarized by the number of segregating sites in the sample. It is important to know how inferences based on this data reduction compare to those based on the full data. In principle it is always better to use the full data if possible, and the differences can be marked (GRIFFITHS and TAVARÉ 1996). However, inferences using the method given here are often very close to those for the full data (as in the examples in Section 7), and it allows much more flexibility to explore the effects of different modeling assumptions.

We acknowledge helpful discussions with RICHARD NICHOLS and thank the communicating editor and referees for valuable suggestions. S. T. was supported in part by National Science Foundation (NSF) grant BIR 95-04393. D. J. B. was supported in part by the Science Research Fellowship scheme of the Nuffield Foundation and the UK EPSRC. R. C. G. was supported in part by an Australian Research Council Grant. P. D. was supported in part by NSF grant DMS 95-05129 and by the Block Fund of the University of Chicago.

Note added in proof: For results relating to Section 5.3, see also FU, Y-X (1996). Genetics **144:** 829–838.

LITERATURE CITED

AYALA, F. J., 1995 The myth of Eve: molecular biology and human origins. Science **270:** 1930–1936.

DONNELLY, P., 1996 Interpreting genetic variability: the effects of shared evolutionary history, pp 25–50 in *Variation in the Human Genome,* edited by K. WEISS. Wiley, Chichester, UK.

DONNELLY, P., and T. G. KURTZ, 1997 The asymptotic behaviour of an urn model arising in population genetics. Stoch. Proc. Appl. (in press).

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

DONNELLY, P., S. TAVARÉ, D. J. BALDING and R. C. GRIFFITHS, 1996 Estimating the age of the common ancestor of men from the ZFY intron. Science **272:** 1357–1359.

DORIT, R. L., H. AKASHI and W. GILBERT, 1995 Absence of polymorphism at the ZFY locus on the human Y chromosome. Science **268:** 1183–1185.

DORIT, R. L., H. AKASHI and W. GILBERT, 1996 Estimating the age of the common ancestor of men from the ZFY intron. Science **272:** 1361–1362.

FELLER, W., 1968 *An Introduction to Probability Theory and Its Applications,* Ed. 3. Wiley, New York.

FULLERTON, S. M., R. M. HARDING, A. J. BOYCE and J. B. CLEGG, 1994 Molecular and population genetic analysis of allelic sequence diversity at the human beta-globin locus. Proc. Natl. Acad. Sci. USA **91:** 1805–1809.

GILKS, W. R., S. RICHARDSON and D. J. SPIEGELHALTER, 1996 *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

GRIFFITHS, R. C., 1989 Genealogical-tree probabilities in the infinitely-many-site model. J. Math. Biol. **27:** 667–680.

GRIFFITHS, R. C., and S. TAVARÉ, 1994a Ancestral inference in population genetics. Statist. Sci. **9:** 307–319.

GRIFFITHS, R. C., and S. TAVARÉ, 1994b Sampling theory for neutral alleles in a varying environment. Phil. Trans. R. Soc. Lond. B **344:** 403–410.

GRIFFITHS, R. C., and S. TAVARÉ, 1995 Unrooted genealogical tree probabilities in the infinitely-many-sites model. Math. Biosci. **127:** 77–98.

GRIFFITHS, R. C., and S. TAVARÉ, 1996 Monte Carlo inference methods in population genetics. Math. Comput. Modelling **23:** 141–158.

HAMMER, M. F., 1995 A recent common ancestry for human Y chromosomes. Nature **378:** 376–378.

HARDING, R., S. M. FULLERTON, R. C. GRIFFITHS, J. BOND, M. J. COX *et al.,* 1997 African *and* Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. (in press).

HUDSON, R. R., 1991 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology,* edited by D. FUTUYMA and J. ANTONOVICS. Oxford University Press, Oxford.

JOBLING, M. A., and C. TYLER-SMITH, 1995 Fathers and sons: the Y chromosome and human evolution. Trends Genet. **11:** 449–456.

KIMURA, M., 1970 The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. Genet. Res. **15:** 131–133.

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

KINGMAN, J. F. C., 1982b Exchangeability and the evolution of large populations, pp. 97–112 in *Exchangeability in Probability and Statistics,* edited by G. KOCH and F. SPIZZICHINO. North-Holland Publishing Company, Amsterdam.

RIPLEY, B. D., 1987 *Stochastic Simulation.* Wiley, New York.

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.

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

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

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

TAKAHATA, N., 1993 Allelic genealogy and human evolution. Mol. Biol. Evol. **10:** 2–22.

TAVARÉ, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theoret. Popul. Biol. **46:** 119–164.

TEMPLETON, A. R., 1993 The "Eve" hypothesis: a genetic critique and reanalysis. Amer. Anthropologist **95:** 51–72.

WALLACE, D. C., 1995 1994 William Allan Award Address—Mitochondrial DNA variation in human evolution, degenerative disease, and aging. Am. J. Hum. Gen. **57:** 201–223.

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

WHITFIELD, L. S., J. E. SULSTON and P. N. GOODFELLOW, 1995 Sequence variation of the human Y chromosome. Nature **378:** 379–380.

WILLS, C., 1995 When did Eve live? An evolutionary detective story. Evolution **49:** 593–607.

Communicating editor: R. R. HUDSON