Genealogical Inference From Microsatellite Data

Ian J. Wilson*1 and David J. Balding2

*School of Biological Sciences, Queen Mary and Westfield College, University of London, London E1 4NS, England and 2Department of Applied Statistics, University of Reading, Reading RG6 6FN, England

ABSTRACT

Ease and accuracy of typing, together with high levels of polymorphism and widespread distribution in the genome, make microsatellite (or short tandem repeat) loci an attractive potential source of information about both population histories and evolutionary processes. However, microsatellite data are difficult to interpret, in particular because of the frequency of back-mutations. Stochastic models for the underlying genetic processes can be specified, but in the past they have been too complicated for direct analysis. Recent developments in stochastic simulation methodology now allow direct inference about both historical events, such as genealogical coalescence times, and evolutionary parameters, such as mutation rates. A feature of the Markov chain Monte Carlo (MCMC) algorithm that we propose here is that the likelihood computations are simplified by treating the (unknown) ancestral allelic states as auxiliary parameters. We illustrate the algorithm by analyzing microsatellite samples simulated under the model. Our results suggest that a single microsatellite usually does not provide enough information for useful inferences, but that several completely linked microsatellites can be informative about some aspects of genealogical history and evolutionary processes. We also reanalyze data from a previously published human Y chromosome microsatellite study, finding evidence for an effective population size for human Y chromosomes in the low thousands and a recent time since their most recent common ancestor: the 95% interval runs from \( \approx 15,000 \) to \( 130,000 \) years, with most likely values around \( 30,000 \) years.

PATTERNS of heritable genetic variation in contemporary samples can provide information both about parameters of evolutionary processes and about details of the genealogical history of the sample. Data from the male-specific part of the human Y chromosome, for example, can provide evidence both about mutation rates and about the number and reproductive behavior of human males. When combined with information from mitochondrial, autosomal, and X chromosome loci, additional insights about recent human evolutionary history may be obtained.

Extracting historical and evolutionary information from the genetic data is, however, difficult, due to the complex interaction of the underlying genetic processes. Traditionally, the interpretation of genetic samples has been based on summary statistics, such as heterozygosity or pairwise measures of identity (Nei 1987; Slatkin 1992). Such an approach can waste much of the information contained in the data (Felsenstein 1992). Intuitively, this is because pairwise measures of identity do not explicitly take account of the ancestral relationships underlying the data (Donnelly 1996). For microsatellite data, a network can be constructed (Cooper et al. 1996; Zerjal et al. 1997) that displays some of the structure of the data and suggests historical relationships, but does not provide a basis for assessing the uncertainty associated with inferences.

Recently, inferential methods that use the coalescent (Kingman 1982; Hudson 1991) to model explicitly the genealogical relationships underlying a genetic sample have become available (Griffiths and Tavare 1994; Kuhner et al. 1995). Tavare et al. (1997) present computational methods for genealogical inference under the assumptions of the coalescent model with infinite-sites mutation, so that back-mutation is assumed to not occur. Microsatellite loci present a particular challenge to genealogical inference because these loci form an important source of highly polymorphic molecular genetic data (Jarme and Lagoda 1996), but the mutation process is such that back-mutations cannot reasonably be ignored. Nielsen (1997) developed Griffiths and Tavare’s (1994) algorithm to obtain maximum likelihood estimates of the scaled mutation parameter \( \theta \) at microsatellite loci. The method was found to be computationally costly, even for a single locus, making accurate estimation difficult.

Here, we present a computationally tractable method for drawing inferences from microsatellite data, not only about \( \theta \) but also about population histories. Very briefly, the method is based on the coalescent model of genealogy together with a ladder (or stepwise) model of microsatellite mutation and is implemented via a Markov chain Monte Carlo (MCMC) simulation algorithm.

In the following section we start by outlining the coa-
lescent-with-ladder-mutation modeling framework and the process of drawing inferences from microsatellite data under this model. Following Tavaré et al. (1997), we adopt a fully probabilistic approach in which the uncertainty about an unknown parameter is expressed in terms of its probability distribution, given the data and the model. As well as making efficient use of all the available information, another important advantage of this approach is interpretability. For scalar parameters, either singly or in combination, inferences are naturally presented visually, in terms of probability density curves or surfaces. Even very complex unknowns, for example, the entire genealogy of the sample, can be described either in terms of probability density curves for important features, such as height or total branch length, or in terms of pictures of a sample of realizations from the probability distribution. A further advantage of a fully probabilistic analysis is flexibility. For example, inferences about the genealogical tree or about the effective population size, or both, can be obtained, according to the goals of the investigator.

In the recent past, fully probabilistic analyses of complex genetic processes were not computationally feasible. While computational cost remains an issue, advances in stochastic simulation methodology, such as MCMC algorithms, now allow problems of substantial size and complexity to be tackled. One important feature of the MCMC algorithm that we propose here is that the allelic type of the ancestral gene at each coalescence is assigned and successively updated according to its conditional probabilities. This simplifies the likelihood computations, which in turn allow flexibility in the choice of algorithms for stepping through the space of candidate trees.

The quality of genealogical inference that can be achieved under ideal circumstances is investigated using data simulated from the model. The method is then illustrated by reanalyzing the data of Cooper et al. (1996) for five microsatellite loci on the human Y chromosome. Because of the complexities of the genetic phenomena under study, we find that data from a single microsatellite locus do not suffice for accurate inferences, even when the modeling assumptions hold exactly. However, if data from a number of completely linked loci are available and the mutation process can be assumed to be the same at each locus, then much more precise inferences can be made.

**The Model**

**Genealogies and the coalescent:** Interpreting genetic data requires an understanding of the patterns of shared ancestry among the genes in the sample. Currently, the most successful mathematical description of the genealogical processes underlying these patterns is provided by the so-called coalescent model.

The simplest version of the coalescent, the standard coalescent, assumes neutrality, random mating, and a constant, large population size. These assumptions can each be weakened to some extent and at some computational cost. However, the novelty of this article is the introduction of a model for microsatellite mutation, and to simplify the presentation of this development we work primarily with the standard coalescent.

Time in the coalescent model is measured in units of $N$ generations, where $N$ is the (fixed and large) population size. Tracing backward in time the lineages of each gene in a sample of size $n$, the time $t_i$ until the first “coalescence” of two lineages at a common ancestor has the exponential distribution with mean $2/n(n-1)$. Continuing backward in time, the time $t_2$ between the first and the second coalescences has the exponential distribution with mean $2/(n-1)(n-2)$, and so forth, until the time $t_{n-1}$ between the final two coalescences (i.e., the time during which the sample has exactly two ancestors) has the exponential distribution with mean 1. Crucially, each of these times is independent of the others. Hence the joint probability density of the coalescence times $t_1, \ldots, t_{n-1}$ is

$$p(t_1, \ldots, t_{n-1}) = \prod_{i=1}^{n-1} \frac{n+1-i}{2} \exp \left( -\frac{n+1-i}{2} t_i \right).$$

Equation 1 pertains to the predata coalescent, in which the sample size $n$ is fixed but the allelic types are not yet observed. Once the allelic types are known, the coalescent probabilities are altered: evaluating the updated probabilities after observing a sample of microsatellite data is the primary goal of this article. A particular feature of the predata coalescent is that most of the lineages coalesce relatively quickly. (In other words, most branches are short.) On the other hand, the time period during which the tree has just two lineages is on average 1 coalescent unit, more than half the mean height of the coalescent tree. Another notable feature of the predata coalescent is the high variability in tree height: its standard deviation is about 60% of the mean height for typical values of $n$. See Donnelly and Tavaré (1995) and references therein for further details of the coalescent model.

**Microsatellite mutations:** Given the genealogy, mutations in the standard coalescent are assumed to occur independently and at constant rate $\theta/2$, where

$$\theta = 2N\mu,$$

and $\mu$ denotes the mutation rate per gene per generation. This means that the number of mutations in any section of the tree with total branch length $t$ has the Poisson distribution with parameter $t\theta/2$. 


Although additional variation can be distinguished in some cases, microsatellite alleles are usually characterized by the copy number of the repeat motif. For the data of Cooper et al. (1996) discussed below, the repeat motif is the four-base sequence GATA.

Mutations of microsatellite alleles are thought to be due predominantly to polymerase slippage (Levinson and Gutman 1987; Dover 1996), which produces mutant alleles close in length to the original; the mutant alleles differ by whole copies of the repeat motif. Direct studies of mutations using a large number of parent-offspring triplets (Weber and Wong 1993) for autosomal microsatellites, and using pedigrees over larger numbers of generations for Y chromosome microsatellites (Heyer et al. 1997), show only single gains or losses of the GATA motif for 11 observed mutations. The mechanisms for gains of repeats through slippage may well differ from those for losses. There may also be evidence of between-species differences (Rubinsztein et al. 1996).

For autosomal DNA, rare, large mutational steps are thought to occur (Di Rienzo et al. 1994), and there is evidence from somatic mutations in cancer patients of heterogeneity between loci (Di Rienzo et al. 1998). These may be due to unequal crossing over, and so it remains uncertain whether or not they occur on the nonrecombining portion of the Y chromosome.

Perhaps the simplest plausible model for the changes in repeat number at each mutation event is the stepwise, or ladder model (Ohta and Kimura 1973), under which the repeat number behaves like a simple random walk; i.e., it is equally likely to increase or decrease by 1 unit at each mutation, and changes of more than 1 unit do not occur. Although the ladder model may not describe fully the complexities of the microsatellite mutation process, it does incorporate “local” changes in allele length, while remaining tractable (Shriver et al. 1993; Valdes et al. 1993; Goldstein et al. 1996). More detailed models of microsatellite mutation, such as the extended models of Di Rienzo et al. (1994) and Slatkin (1995), can readily be incorporated into the inferential framework described here.

**Statistical Inference**

The direct probability paradigm: We have a sample, D, of genes at a particular microsatellite locus, and a collection of unknown parameters, N, μ, and the tree parameters—the coalescence time and the two descendant nodes of each internal node—which we collectively denote Y. We want to make valid and useful statements about N, μ, and Y, given D and the modeling assumptions. In the direct probability, or Bayesian, paradigm of statistical inference, such statements are based on the probability distribution of N, μ, and Y, conditional on D and the model. The required probability distribution is usually obtained by first specifying a distribution to model the background information before D is observed, then updating this prior distribution, via Bayes’ rule, to incorporate the information conveyed by D.

The coalescent model specifies a probability distribution for Y. This distribution can be thought of as a prior distribution for the genealogical tree, which should be updated in the light of the data D. Information about μ obtained from pedigree studies, such as those described above, can be summarized by a probability density curve that would usually be smooth and unimodal. Information about N is more difficult to specify, because N should be interpreted as an effective, rather than actual, population size. However, previous genetic studies, together with archaeological evidence, do give some idea of the effective population sizes for recent human evolution (Fullerton et al. 1994; Hammer 1995; Harding et al. 1997). Corresponding probability distributions would normally be very diffuse, reflecting the imprecise background information, but would again be smooth and unimodal.

Although the probability distributions chosen to represent knowledge about N and μ are not unique, in many cases the postdata inferences will be insensitive to reasonable specifications. If this is not the case, investigation of the sensitivity will indicate the information needed to produce more reliable inferences. An alternative approach sometimes adopted is to undertake analyses conditional on particular values for N and μ. As noted by Brooks (1997) and Tavaré et al. (1997), this approach can be seriously misleading, because information in the data that is informative about N or μ may be misinterpreted as informative about Y. Repeating the analysis for various values of N and μ cannot overcome the problem; the only satisfactory solution is to let the data speak simultaneously for all the parameters, N, μ, and Y.

**MCMC methods:** MCMC algorithms generate approximate random samples from a probability distribution II by constructing a Markov chain whose equilibrium distribution is II. Consecutive states of a Markov chain are usually correlated, but if the chain is run for a suitably long “burn-in” period, and then every ith state is recorded for some sufficiently large i, the resulting values will form an approximate random sample from II. Features of II can then be investigated by examining corresponding properties of this sample. For example the probability assigned by II to any region of the parameter space can be approximated by the proportion of the sample values that lie in this region. For a further discussion see Besag et al. (1995) and Brooks (1998).

It is not usually possible to prove that a Markov chain has converged to its equilibrium distribution. However, a number of diagnostic checks that allow many cases of nonconvergence to be detected have been proposed. The chains implemented below have been checked using the suite of diagnostic tools contained in the software package CODA (Best et al. 1995). In each case,
The likelihood is then the sum of these conditional
abilities for the data given the allelic state at the root. terminal (data) node, while
otes the prior probability density of
stein (1981). This algorithm proceeds recursively, start-
ing in more detail, we introduce some notation: for a node
and/or dif-
culty. If the parameter space becomes very large, 
calculated easily from
constructed in this way converges to \( p(\Theta | D) \), the proba-
bility distribution of the unknown parameters given the data, provided that \( q \) is such that the chain is aperiodic
irreducible
metropolis Hastings algorithm (Metropolis et al. 1953; Hastings 1970). Given a current location \( \Theta \) in parameter space, where \( \Theta \) stands for
the parameter vector \((N, \mu, Y)\), a new candidate location \( \Theta' \) is chosen from a proposed distribution \( q(\Theta' | \Theta) \). The new location is accepted according to the value of
\[
\begin{align*}
    u &= \frac{q(\Theta' | \Theta)}{q(\Theta | \Theta')} \ \frac{p(D | \Theta')}{p(D | \Theta)} \ \frac{\pi(\Theta')}{\pi(\Theta)},
\end{align*}
\] (2)
where \( p(D | \Theta) \) denotes the likelihood, the probability of the data given the parameter vector \( \Theta \), and \( \pi(\Theta) \) de-
notes the prior probability density of \( \Theta \).
If \( u > 1 \), the proposal \( \Theta' \) is accepted; otherwise it is
accepted with probability \( u \). If \( \Theta' \) is not accepted, the chain re-
ains in its current state, \( \Theta \). The Markov chain
constructed in this way converges to \( p(\Theta | D) \), the proba-
bility distribution of the unknown parameters given the data, provided that \( q \) is such that the chain is aperiodic
and/or dif-
cible approach to inference adopted here permits much
faster likelihood computations. The key idea is that the
likelihood would be relatively easy to compute if the allelic states at the internal nodes of the genealogical
tree were known. Then, the likelihood would be simply
a product of terms, one for each branch of the tree.
The term corresponding to a branch of length \( t \), linking
nodes whose allelic states differ by \( d \geq 0 \), is
\[
\psi_d(t, \theta) = e^{-t/2} \sum_{k=0}^{\infty} \frac{(t/4)^{k+d}}{k!(k+d)!} = e^{-t/2} I_d \left( \frac{t}{4} \right),
\] (3)
in which \( I_d(x) \) denotes the \( d \)-th order modified Bessel func-
tion of the first kind (Gradshteyn and Ryzhik 1980). Although \( \psi_d \) involves an infinite sum, in practice only
the first few terms are required for an accurate approxi-
mation. This is because the value of \( k \) corresponds
to the number of pairs of mutations in opposite directions,
which is usually very small. Fast algorithms for computing
\( I_d(x) \) are widely available; see, for example, Press et al. (1992).

Equation 3 specifies the likelihood that would apply
if the internal allelic states were known. Unfortunately,
they are unknown. However, the simple likelihood for-
mula based on (3) can nevertheless be exploited under
the direct probability paradigm, because the internal
allelic states can be regarded as additional parameters.
The parameter space is therefore augmented: in addition
to \( N, \mu, \) and \( Y \), there is an allelic state for each internal
node.
Increasing the dimension of the parameter space in this
way is impractical in traditional statistical approaches. With direct probability inference based on
an MCMC algorithm, however, there is no substantial
difficulty. If the parameter space becomes very large,
then convergence of the algorithm can become slow
and/or difficult to assess, but this did not arise for the
examples discussed below.
The augmented parameter space allows great flexibil-
ity in the choice of proposal distributions \( q \). We use
a very simple method for generating candidate trees. Basically, the method involves removing a branch from
the tree at random and adding it anywhere on the tree, but locations close to “similar” allelic types are preferen-
tially chosen. In this way large jumps in tree space are
possible, while acceptance rates remain sufficiently
high. Before describing the branch-swapping algorithm
in more detail, we introduce some notation: for a node
\( x \), we write \( t(x) \) for its coalescence time \( t(x) = \frac{1}{2} \theta \) if \( x \) is a
terminal (data) node, while \( \alpha(x) \) denotes the allelic
state at node \( x \).
The branch-swapping algorithm: Choose an internal node $x$ at random, except that the root may not be chosen. We then attempt to move the parent of $x$ to a new location in the tree. To this end, we choose a node $y$ above which to attach the parent of $x$. For this to be possible, either $y$ is the root or $t(z) > t(x)$, where $z$ denotes the parent of $y$. Choosing $y$ at random among nodes satisfying this condition is likely to be unsatisfactory: if $\alpha(y)$ is very different from $\alpha(x)$, the candidate tree will almost certainly be rejected. To avoid an excessive rejection rate, the probability of a node being selected is set to be a decreasing function of $|\alpha(y) - \alpha(x)|$. Specifically, we assign

$$P(y|x) \propto \frac{1}{1 + |\alpha(x) - \alpha(y)|} \quad (4)$$

For example, nodes whose allelic state differs from that of $x$ by one are half as likely to be chosen as nodes with the same allelic state. To simplify the computation, we set $P(y|x) = 0$ when $y$ is the parent of $x$. The distribution specified by (4) is somewhat arbitrary; there exist many other suitable distributions, but this choice seems to work well in practice.

Once $y$ has been chosen, if it is not the root then the parent of $x$ is inserted at a point chosen uniformly between $\max\{t(y), t(x)\}$ and $t(z)$. If $y$ is the root, the parent of $x$ is located at a time chosen from the standard exponential distribution above the root (and thus becomes the new root). Finally, a new allelic state for the parent of $x$ is chosen according to a discretized normal distribution, with mean $(\alpha(x) + \alpha(y))/2$ and standard deviation $(|\alpha(x) - \alpha(y)| + 1)/4$. Again, this choice is somewhat arbitrary but seems to lead to both reasonable acceptance rates and good mixing. The chain produced using this proposal distribution is clearly aperiodic and is irreducible because we can recreate any tree in, at most, $n - 1$ steps (where $n$ is the sample size) by successively moving terminal nodes, one at a time, to their position on the new tree, simultaneously changing the coalescence time and allelic state of the branch point.

Other updating algorithms: Although the branch swapping algorithm described above leads to acceptable convergence properties, we found that convergence rates could be improved by including between each branch-swapping step another updating algorithm that attempted to alter branch lengths only, not the tree topology. The two scalar parameters, $N$ and $\mu$, are updated using a uniform probability density on a logarithmic scale, centered on the current value, and with length tailored to optimize convergence.

### RESULTS

**Data simulated under the model:** Shown at the top of Figure 1 is a genealogical tree, labeled “true,” with a microsatellite copy number indicated at each of the $n = 10$ terminal nodes. This tree was simulated from the coalescent-with-ladder-mutation model with $\theta = 5$. The other four trees are simulated from the postdata distribution given the allelic data of the true tree. These trees are samples numbered 2000, 4000, 6000, and 8000 from the MCMC run corresponding to row 1 of Table 1.

The height of the tree, $T$, is 1.25 coalescent units, which is less than 1.54, the median height of the pre-data coalescent when $n = 10$, but is very close to the modal height. The value of $L$, the total branch length, is 4.82, which again is less than the median of 5.21 for the pre-data coalescent when $n = 10$, but very close to the modal value.

Note that, of the four genes with allelic type 6 in the true tree of Figure 1, only one pair has very recent shared ancestry. In fact, one of the other 6-alleles has no ancestry in common with this pair beyond the root of the tree, whereas its nearest relative in the sample is a 3-allele. Clearly, accurate reconstruction of the true genealogical tree from only the allelic-type data is unachievable here, although some information about key parameters, such as $\theta$, $T$, and $L$ is available.

Application of MCMC algorithm: What can be inferred from the allelic types shown on the true tree is suggested by the four other trees shown in Figure 1, which were simulated from the postdata coalescent. A uniform prior was assumed for $\theta$, and the value realized in each simula-
TABLE 1

Inferences for $\theta$, $T$, and $L$ from a single tree

<table>
<thead>
<tr>
<th>Sample size (n)</th>
<th>No. loci</th>
<th>$\theta$ Median Interval</th>
<th>$T$ Median Interval</th>
<th>$L$ Median Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>14.9 (2.9, 95)</td>
<td>1.32 (0.42, 4.0)</td>
<td>4.51 (1.8, 10)</td>
</tr>
<tr>
<td>40</td>
<td>1</td>
<td>13.0 (3.8, 38)</td>
<td>1.42 (0.55, 4.0)</td>
<td>7.33 (4.4, 13)</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>5.36 (2.3, 14)</td>
<td>1.33 (0.49, 3.3)</td>
<td>4.64 (2.0, 9)</td>
</tr>
<tr>
<td>40</td>
<td>5</td>
<td>5.67 (3.5, 9)</td>
<td>1.19 (0.58, 2.7)</td>
<td>6.59 (4.3, 11)</td>
</tr>
</tbody>
</table>

Median and 95% equal-tailed intervals of the posterior distributions for $\theta = 2N\mu$, tree height $T$, and total branch length $L$, based on samples of size $n = 10$, shown at the terminal nodes of the true tree of Figure 2, and $n = 40$ (not shown). The values of $T$ and $L$ are given in coalescent units; to obtain years, multiply by population size and generation time. The values used to generate the data were: $\theta = 5$, $T = 1.25$, $L = 4.82$ ($n = 10$), and $L = 7.15$ ($n = 40$). Table entries are estimated from 10,000 output values (corresponding to $2 \times 10^7$ attempts to update $N$ and $\mu$, and $1.6 \times 10^6$ branch-swapping steps); simulation error is $\sim 1\text{-}3\%$ of stated values.

Inferences for a single microsatellite, even when the modeling assumption interval (PIL) calculated from the MCMC run. For each combination of $\theta$, $n$, and $I$, the height of the bar gives an average of results from 140 datasets simulated from the coalescent-with-ladder-mutation model and part of the Y chromosome. The true tree of Figure 2 is the same as that of Figure 1 (N.B. different time scale), but in addition to the allelic data of Figure 1, a further four independent simulations of the ladder mutation process are given, each with $\theta = 5$. This simulation mimics data from five completely linked microsatellite loci with a common value of $\theta$. Once again, four trees are shown simulated from the coalescent based on the five-locus data, with a completely flat predata distribution for $\theta$.

As expected, the trees simulated from the postdata coalescent are, with information from $l = 5$ loci, more similar to the original tree than in the one-locus case. Nevertheless, none of the simulations comes close to reconstructing the original tree.

Summary statistics for the $n = 10$, $l = 5$, case are given in row 3 of Table 1. Even with five loci, the post data uncertainty about $T$ and $L$ remain large, although inference about $\theta$ is now much improved. Row 4 of Table 1 quantifies a further improvement when $n$ is increased to 40 (allelic data not shown).

Average performance over many trees: Each row of Table 1 corresponds to only one realization of a genealogical tree and allelic data. To obtain a better overall appreciation of the quality of inference achievable from microsatellite data, it is useful to assess average performance over many tree and mutation simulations. Care is needed to effectively summarize such a large quantity of simulation results, in part because the uncertainty in inference about $\theta$, $T$, and $L$ tends to increase with the magnitude of the true value.

For each of $\theta$, $T$, and $L$, Figure 3 shows both the mean absolute deviation (MAD) of the MCMC output values from the true value, and the length of the 95% probability interval (PIL) calculated from the MCMC run. For each combination of $\theta$, $n$, and $I$, the height of the bar gives an average of results from 140 datasets simulated from the coalescent-with-ladder-mutation model and

The effect of additional, linked loci: We have seen that there is only limited information about $\theta$, $T$, and $L$ at a single microsatellite, even when the modeling assumptions hold exactly. But is it perhaps possible to obtain good inferences from several completely linked loci? Such data arise, for example, from the nonrecombining
on averaging over many trees, especially for the latter case the MAD of \( u \) because the true value increases with \( n \).

In the pre-insert in our analyses, although it could readily have been incorporated by means of a further augmentation of the parameter space.

Figure 2.—The true tree (top) is the same as that of Figure 1, but the results of four additional, independent simulations of the mutation process are also shown, mimicking data from five completely linked loci, each having the same mutation mechanism and with \( \theta = 5 \). The other four trees are simulated from the postdata distribution given all five data sets. These trees are samples numbered 2000, 4000, 6000, and 8000 from the MCMC run corresponding to row 3 of Table 1.

**Human Y chromosome microsatellite data:** Human mitochondrial DNA sequences have been interpreted as supporting the theory—dubbed “Out of Africa”—that modern humans are descendants of a small group that lived in Africa perhaps about 200,000 years ago and subsequently spread throughout the world, eliminating most or all other extant human lineages. However, inferences about the time since the most recent common ancestor (TMRCA) of the sample generally underestimate the amount of variability (Tavare et al. 1997), and geographical location of the MRCA is problematic and contentious (Templeton 1997).

Patterns observed from autosomal DNA seem somewhat different. For example, \( \beta \)-globin data suggest a much longer TMRCA (Harding et al. 1997). These differing interpretations are not necessarily in conflict because autosomal and mitochondrial DNA reflect different aspects of human history, and the results may be affected by selection effects. Recombination of autosomal DNA sequences may also lead to some problems for inference.

A third potential source of evidence, reflecting a further aspect of human prehistory, comes from genetic variation on the human Y chromosome. Recently, a number of polymorphic microsatellites have become available for population surveys (Cooper et al. 1996; Deka et al. 1996; Ruiz Linares et al. 1996; Hammer et al. 1997; Zerjal et al. 1997).

A large effort has been concentrated on estimating the TMRCA of a sample of genes drawn from a locus—in this case the entire nonrecombining portion of the human Y chromosome. While the TMRCA may not be the most important time of human history (Brookfield 1997), it is central to interpreting genetic samples and has been investigated by several authors (Goldstein et al. 1996; Tavaré et al. 1997). Furthermore, the method proposed here allows simultaneous inferences about the TMRCA (the height of the tree) and, for example, the (effective) population size, \( N \).

Data: We consider the data of Cooper et al. (1996), which consist of the genotypes of 212 individuals at five Y chromosome microsatellite loci from East Anglia (UK), Sardinia, and Nigeria, together with a linked Alu insert. Since we are concerned here with inference from microsatellite haplotypes, we did not include the Alu insert in our analyses, although it could readily have been incorporated by means of a further augmentation of the parameter space.

Inferences become somewhat more precise as \( u \) increases and markedly better as \( n \) increases. Hence measures of uncertainty, expressed as a proportion of the true value. In ~5% of MCMC runs, the value of \( \theta \) lay outside the 95% probability interval, and similarly for \( T \) and \( L \), suggesting that the MCMC runs had inadequately converged.

The poor quality of inferences about \( \theta \) when \( l = 1 \), noted for the particular tree of Figure 1, remains evident on averaging over many trees, especially for \( n = 10 \). In the latter case the MAD of \( \theta \) is 3 to 5 times the true value and the PIL as much as 20 times the true value. Inferences become somewhat more precise as \( \theta \) increases and markedly better as \( n \) and \( l \) increase.

Increases in \( n \) and \( l \) are less effective in improving the precision of \( T \), with the improvement from worst to best cases only ~20% for both MAD and PIL when \( \theta = 1 \), rising to 30% for larger values of \( \theta \). The same patterns are shown as for \( \theta \), with precision increasing with \( \theta \), \( n \), and \( l \). Inferences about \( L \) are harder to interpret because the true value increases with \( n \). In the pre-data coalescent, the standard deviation of \( L \) decreases relative to its mean as \( n \) increases. Hence measures of
Two datasets were used: the complete set of Nigerian and Sardinian haplotypes, together with the initial sample of 22 East Anglians (dataset NSE), and all 174 East Anglian haplotypes (dataset EA). The first of these sets gives approximately equal weighting to the three regions; the second provides a larger sample from a single location. Although the coalescent-with-ladder-mutation model is unlikely to be exactly appropriate for these datasets, inferences based on this model can nevertheless be informative. It is of particular interest to see what aspects of the postdata distributions differ substantially from the corresponding predata distributions.

Priors: Under the standard coalescent, no information about the values $N$ and $\mu$ can be obtained from the allelic data except through their product, $\theta = 2N\mu$. Postdata inferences about $\theta$ are therefore more robust than inferences about either $N$ or $\mu$ separately. It is useful to distinguish the two because information about them can be obtained from other sources, particularly in the case of $\mu$. Heyer et al. (1997) used three observed mutations in 1491 meioses to obtain a point estimate of mutation rate of 0.2% per meiosis. Assuming a Poisson distribution for the number of mutations, and using a standard exponential pre-prior, the distribution of $\mu$...
such attempts between samples. After discarding the first 2000 samples (the burn-in), 10,000 samples were retained. Two such sets of samples were taken, with different starting trees, for each prior and dataset combination. The posterior distributions for \( m \), \( N \), \( T \), and \( L \) approximated from the two MCMC runs were checked and in each case found to be effectively indistinguishable. They were then combined to give a total of 20,000 samples.

Results are given in Figure 4 (probability density curves for dataset NSE; those for dataset EA are very similar and are not shown) and Table 2 (summary statistics for both datasets). For dataset NSE, a number of individual trees sampled from the MCMC output were examined in detail. Although there was some relation between geographic location of haplotype and tree structure, this was restricted to recent nodes. Clades of more than six haplotypes all from a single location were rare, and haplotypes from all locations were typically represented on both sides of the root node.

Inferences about \( \mu \): Figure 4 (top right) shows, for dataset NSE, the two postdata probability density curves for \( \mu = 2N\mu \), as well as the corresponding predata curves. The postdata curves are very similar, despite the differences in the two priors. For example, the postdata medians are both around 11, compared with prior medians of around 22 and 39, respectively, for the low- and high-variance priors (Table 2). Moreover, the two postdata densities are insensitive to this assumption: a uniform prior for \( \mu \) leads to very similar conclusions (results not shown).

Inferences about \( \theta \): Figure 4 (top right) shows, for dataset NSE, the two postdata probability density curves for \( \theta = 2N\mu \), as well as the corresponding predata curves. The postdata curves are very similar, despite the differences in the two priors. For example, the postdata medians are both around 11, compared with prior medians of around 22 and 39, respectively, for the low- and high-variance priors (Table 2). Moreover, the two postdata 95% probability intervals are practically indistinguishable: (7.7, 17.0) and (7.6, 16.4). For dataset EA, the postdata medians and upper 95% interval limit are both a little lower (Table 2).

As expected, the postdata distributions for the two components of \( \theta \), the mutation rate, \( \mu \), and the population size, \( N \), are negatively correlated, and each is more strongly affected by the prior than is the postdata distribution of \( \theta \). Figure 4 (top left) shows the two post-NSE-data density curves for \( \mu \), together with the predata curves. Both posterior curves are somewhat sharper than the prior, with diminished support for high values of \( \mu \). The postdata density curves for \( N \) (Figure 4, bottom left) are very similar, despite the substantial difference in the prior curves. The post-EA-data distributions are very similar to those for NSE. In all cases they reflect diminished support for high values of \( N \). The postdata medians are \( \sim 3000 \), with most likely values between \( \sim 1500 \) and 8000 for both datasets. Although the limitations of the modeling assumptions require that caution be attached to the interpretation of a particular analysis, the similarity of the postdata distributions provides some confidence for the conclusion that the Y chromosome effective population size during recent human history is a few thousands, consistent with the results of previous analyses.

Inferences about the TMRCA: An estimate for the number of generations since the MRCA of the sample can be made by multiplying together the postdata values for
TABLE 2
Summary of human Y chromosome analyses

<table>
<thead>
<tr>
<th></th>
<th>Low-variance prior</th>
<th></th>
<th>High-variance prior</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median Interval</td>
<td></td>
<td>Median Interval</td>
</tr>
<tr>
<td>$\theta$</td>
<td>Prior</td>
<td>22.0 (4.8, 75.9)</td>
<td>39.2 (4.0, 338)</td>
</tr>
<tr>
<td></td>
<td>NSE</td>
<td>11.4 (7.7, 17.0)</td>
<td>11.2 (7.6, 16.4)</td>
</tr>
<tr>
<td></td>
<td>EA</td>
<td>10.0 (7.4, 13.3)</td>
<td>9.8 (7.3, 13.1)</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Prior</td>
<td>2.5 (0.73, 5.9)</td>
<td>2.5 (0.73, 5.9)</td>
</tr>
<tr>
<td></td>
<td>NSE</td>
<td>1.7 (0.74, 3.7)</td>
<td>1.8 (0.59, 4.6)</td>
</tr>
<tr>
<td></td>
<td>EA</td>
<td>1.5 (0.67, 3.5)</td>
<td>1.8 (0.57, 4.3)</td>
</tr>
<tr>
<td>$N$</td>
<td>Prior</td>
<td>4.7 (1.6, 10.3)</td>
<td>8.2 (1.1, 56.4)</td>
</tr>
<tr>
<td>$(\times 10^3)$</td>
<td>NSE</td>
<td>3.5 (1.5, 7.4)</td>
<td>3.0 (1.1, 9.6)</td>
</tr>
<tr>
<td></td>
<td>EA</td>
<td>3.3 (1.4, 7.1)</td>
<td>2.7 (1.1, 8.6)</td>
</tr>
<tr>
<td>TMRCA</td>
<td>Prior (n = 60)</td>
<td>157 (39, 579)</td>
<td>281 (31, 2466)</td>
</tr>
<tr>
<td>$(\times 10^3 \text{yr})$</td>
<td>NSE</td>
<td>36 (13, 128)</td>
<td>33 (10, 138)</td>
</tr>
<tr>
<td></td>
<td>Prior (n = 174)</td>
<td>159 (39, 565)</td>
<td>289 (32, 2493)</td>
</tr>
<tr>
<td></td>
<td>EA</td>
<td>31 (11, 108)</td>
<td>27 (8.7, 113)</td>
</tr>
</tbody>
</table>

Median and 95% equal-tailed intervals of prior and posterior distributions for $\theta$, $\mu$, $N$, and TMRCA for the NSE sample (60 Y chromosome haplotyes, approximately equal numbers from Nigeria, Sardinia, and East Anglia), and for the EA sample (174 East Anglian haplotypes). Haplotypes consist of five microsatellite loci; data from Cooper et al. (1996). Prior distributions are: $\mu \sim \text{gamma}(4, 1492)$; $N \sim \text{gamma}(5, 1/1000)$ (low variance), and $N \sim \text{ln}(9, 1)$ (high variance). Table entries are based on 10,000 output values (corresponding to $4 \times 10^7$ branch-swapping steps).

N and T. Further multiplication by the generation time $G$ gives a posterior density curve for the number of years since the MRCA. Figure 4 (bottom right) shows both the pre-and post-NSE-data density curves, assuming $G = 20$. This value allows comparison with the results of Tavaré et al. (1997), but may be too low: alternative values can be implemented simply by proportional adjustment.

The two postdata curves are very similar and reflect a very marked shift of support toward smaller values compared with the predata distributions. For example, the postdata distributions are sharply peaked at values of $\sim 30 \text{ kyr}$, a value that has little a priori support. Most likely postdata values are between $\sim 10$ and $100 \text{ kyr}$, while values $>150 \text{ kyr}$ have probabilities of $\sim 1.5$ and $2\%$ for the low- and high-variance priors, respectively. For the much larger EA sample, drawn from a single geographic location, postdata distributions are shifted slightly downward compared with the post-NSE-data distributions (Table 2).

The posterior distributions for (scaled) tree height, $T$, have medians of $\sim 0.7$ in all cases compared with prior medians of $\sim 1.7$. The scaled lengths, $L$, are not reduced to the same extent. This may be evidence for “radial” type trees, suggesting some recent population growth. Nevertheless, the posterior values are also consistent with the standard coalescent model.

DISCUSSION

We have developed a methodology for carrying out fully probabilistic analyses of microsatellite samples, which opens up possibilities for inferences much more detailed than those previously possible. For example, the implications of the data for the scaled mutation parameter, $\theta$, and the height and shape of the genealogy can be assessed simultaneously. One key feature of our direct probability analysis is that likelihood calculations are greatly simplified by augmenting the parameter space to include the internal allelic states. This innovation permits great flexibility in algorithms for exploring the space of possible trees, as well as in the range of modeling assumptions that become practicable. Here, we have focused on perhaps the simplest, plausible modeling framework: the coalescent-with-ladder-mutation.

Results from simulation studies, in which the modeling assumptions are known to hold exactly, indicate that accurate inference about $\theta$ requires sampling several, tightly linked loci: a single locus provides little information, even when the sample size is large. With five loci, good quality inferences about $\theta$ are achievable, but those for other aspects of the tree, such as $T$ and $L$, remain far from precise.

Turning to analyses of published data, although our modeling assumptions are, inevitably, not fully realistic, our results provide support both for an effective population size of human Y chromosomes in the lower thousands and for relatively short times (point estimates around 30 kyr) since the most recent common ancestor. These conclusions in turn support the theory that extant human males have spread relatively recently from a small group. In addition, the relatively small value for effective population size may reflect high between-male variance
in reproductive success. The range of supported values for \( \theta \) is \( \sim 8 \) to \( 16 \). Improved predata estimates for the mutation rate \( \mu \) would enable more accurate inference about the population size \( N \) and the TMRCA. Inferences from the two datasets were very similar, despite the fact that one was geographically dispersed and the other geographically homogeneous and much larger. Additionally, there is little evidence of “clumping” of haplotypes from the same region, except in the very recent past from posterior trees.

Values of the TMRCA supported by our analyses are low compared both with times suggested by nongenic evidence and with published studies based on autosomal DNA and mitochondria (Templ et al. 1993; Harding et al. 1997). They are, however, broadly consistent with the analysis of Tavaré et al. (1997), based on \( Y \) chromosome sequence data and the coalescent-with-infinite-sites model. [Our 95% intervals are narrower than those of Tavaré et al. (1997), reflecting more information from five microsatellites than from >15 kb of sequence, despite the limitations imposed by recurrent mutations.] Wide variation between \( Y \) chromosome, mtDNA, and autosomal TMRCAs are plausible for purely stochastic reasons. Additional factors not accounted for in the model may also explain the difference: male generation time may be greater than female, and selective sweeps may play a large part in \( Y \) chromosome evolution.

Our analyses were based on males from three locations and may not represent all human \( Y \) chromosome history. Cooper et al. (1996) estimated the timing of population splits using a maximum divergence approach. This gives an estimate of \( \mu T \), where \( T \) is the TMRCA in generations. Their estimates of \( \mu T \) were 11.4 for the whole data set and 7.75 for EA. These give point estimates for the TMRCA of 110 kyr for the whole data set and 77 kyr for the EA dataset. Estimates of uncertainty are not available with this method. These values are toward the upper tails of our corresponding posterior distributions. Further, under our analyses the data suggest values for the TMRCA for the EA sample only slightly lower than those for the NSE sample. They also suggest little increase in inferential precision with increasing sample size, in contrast to the conclusions of the original authors.

Producing the first row of Table 1 required about 50 min on a desktop workstation—equivalent to 320,000 attempted tree rearrangements and 16,000 attempted changes to \( \theta \) per minute. Increasing the sample size and number of loci increases the time required. To perform the same number of steps on a tree with five loci and a sample size of 200 takes \( \sim 400 \) min. Computational resources should not provide a barrier to extending our analyses to incorporate more sophisticated modeling assumptions. These might include more detailed models for population growth and structure and for microsatellite mutation. Although such developments are well worth pursuing, it may turn out that the primary insights attainable from the data are apparent from the simpler analyses presented here.

We thank Mark Beaumont, Richard Nichols, and Bill Amos for helpful discussions and comments, and the latter also for drawing our attention to the dataset. This work was supported in part by the Stochastic Modeling in Science and Technology initiative of the United Kingdom Engineering and Physical Sciences Research Council (Grant no. K72599).

**LITERATURE CITED**


Communicating editor: R. R. Hudson