Detecting Population Expansion and Decline Using Microsatellites

Mark A. Beaumont

Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom

Manuscript received January 4, 1999
Accepted for publication August 30, 1999

ABSTRACT

This article considers a demographic model where a population varies in size either linearly or exponentially. The genealogical history of microsatellite data sampled from this population can be described using coalescent theory. A method is presented whereby the posterior probability distribution of the genealogical and demographic parameters can be estimated using Markov chain Monte Carlo simulations. The likelihood surface for the demographic parameters is complicated and its general features are described. The method is then applied to published microsatellite data from two populations. Data from the northern hairy-nosed wombat show strong evidence of decline. Data from European humans show weak evidence of expansion.

IT is widely believed that the results from genetic surveys may be used to infer the demographic history of populations (Avise 1994). This has stimulated many studies using a wide variety of markers and statistical techniques in, for example, conservation biology and biological anthropology (Cavalli-Sforza *et al.* 1994; Roy *et al.* 1994). Before strong conclusions are drawn, and, in the case of conservation biology, before management decisions are taken, there is a need to better understand both the limitations and potentials of genetic data analysis.

Traditionally most analyses in population genetics have used methods whereby statistics calculated from genetic data, such as heterozygosity, are equated with their theoretical expectations under some demographic and mutational model, allowing parameters of the model to be inferred. Since the advent of genealogical modeling (Hudson 1991; Donnelly and Tavaré 1995), this moment-matching approach has been improved by the use of Monte Carlo simulations of the coalescent process to compare statistics calculated from a data sample, such as the average number of pairwise differences between sequences, with the distribution of the statistics in simulated samples (Rogers 1995; Weiss and von Haeseler 1998). This extension is sometimes called a likelihood analysis because the probability of obtaining the statistic within a given range is estimated as a function of the parameters in the model. However, as pointed out by Felsenstein (1992), statistics calculated from the genetic data do not capture all the information present, and are therefore less efficient in comparison with methods that estimate the probability of obtaining the sample configuration itself, the true likelihood. A general method for estimating this likelihood was described by Griffiths and Tavaré (1994a,b,c). Other methods have been described by Lundstrom *et al.* (1992), Kuhner *et al.* (1995), and Wilson and Balding (1998; see also Felsenstein *et al.* 1999 for a description of the relationships between some of these methods).

An interesting problem on which to apply these methods is the detection of population growth or decline. Such studies have primarily been concerned with the human demographic expansion. For example, moment-matching approaches have been taken by Slatkin and Hudson (1991), Rogers and Harpending (1992), Rogers (1995), and Reich and Goldstein (1998). Population decline has tended to be analyzed as a separate phenomenon from population growth (Cornuet and Luikart 1996). However, Weiss and von Haeseler (1998) have studied a model where decline and growth are described within the same framework. Likelihood-based approaches to the detection of past changes in population size have been described by Griffiths and Tavaré (1994b) and Kuhner *et al.* (1998).

These studies have primarily used models of exponential growth. An alternative that has been neglected is gradual linear growth or decline. While it is reasonable to suppose that on a short timescale the magnitude of fluctuations in population size is proportional to the population size, and hence more accurately modeled by an exponential function, average changes in population size over long periods are more likely to be functions of environmental and evolutionary factors and may be more linear. One aim of this article is to see how inferences differ depending on which model is used.

A method is described here whereby it is possible to draw random samples from the Bayesian posterior distribution of demographic and mutational parameters, using Markov chain Monte Carlo (MCMC). In

Address for correspondence: School of Animal and Microbial Sciences, University of Reading, Whiteknights, P.O. Box 228, Reading RG6 6AJ, United Kingdom. E-mail: m.a.beaumont@reading.ac.uk

MCMC, a Markov chain is simulated whose equilibrium distribution is the required probability distribution. The methods is applied to microsatellite data assumed to be evolving by a stepwise mutation model sampled from a population that has varied in size. The performance of the method on small test data sets is described, where the results can be compared with likelihoods estimated by Monte Carlo (MC) integration. The method is then applied to larger simulated data sets to see how well the original parameters can be inferred. Finally data sets from two natural populations are analyzed. In one case, that of the northern hairy-nosed wombat (Taylor *et al.* 1994), the population is believed to have declined rapidly over the last 100 years. The other data set consists of a subset from the survey of 60 tetranucleotide microsatellite loci among 15 human populations described in Jorde *et al.* (1997).

THEORY

Background: A sample of size n_0 chromosomes is taken from a closed panmictic population. The genealogical history of the sample can be considered as a sequence of mutation and coalescent events going back in time until a coalescent event occurs that is the most recent common ancestor (MRCA) of the sample (Griffiths and Tavaré 1994a). There are e events in the genealogical history including the last coalescent event. The events occur at times $T_1, T_2. . . T_e$ going back into the past, relative to the time when the sample was taken. Following Griffiths and Tavaré (1994b), the T_i are scaled in units of the current population size. At each event the state of the sample configuration changes, either because two lineages coalesce or because the mutational state of a lineage changes. For example the state of the sample configuration at the ith event could be represented as $S_i = \{n_i, (l_1, l_2, . . . , l_{n_i})\}$, where n_i is the number of lineages at the ith event, and the l_j are a set of labels for each lineage recording their mutational state and genealogical relationships, commonly represented as a tree. When each event occurs, the state changes from S_i to S_{i+1}. The observed sample can be given state S_0. Thus we can represent the genealogical history as a sequence

$$G = (\{S_1, T_1\}, \{S_2, T_2\} . . . \{S_e, T_e\}).$$

The shape of the tree, and the position and types of mutations on it, represented by the sequence G, depend on parameters describing the mutation rate, mutation model, and how the population size varies in time. These parameters are denoted here by a list, Φ, and the specific parameters used in this article are introduced later.

From standard coalescent theory, it is possible to calculate a probability distribution, $p(\{S_0, G\}|n_0, \Phi)$. In the following, "density" and "distribution" are used interchangeably, and $p(x|y)$ indicates a density (distribution) of x conditional on a particular y. The density is assumed to be marginal to (integrated over) any other variables not appearing within the brackets. A related distribution, $p(\{S_0, G_n\}|n_0, \Phi)$, has been described by Wilson and Balding (1998), where G_n is a genealogical history consisting of a sequence of coalescent events marginal to all possible mutation events, where the mutational states at each coalescent node are recorded. This latter density is different from the former in that it explicitly requires the branch length information recorded in the S_i, whereas that for G does not.

The probability distribution of S_0, given n_0 and Φ, $p(S_0|n_0, \Phi)$, is the marginal distribution over G (or G_n). Given this distribution, which need only be estimated up to a multiplying constant, k, likelihoods for Φ could be obtained. A straightforward method for estimating $p(S_0|n_0, \Phi)$ might be to simulate from $p(\{S_0, G\}|n_0, \Phi)$ and count the proportion of times that a target S_0* was observed. This can be regarded as a MC integration over G. For small n_0 this can be quite effective, as demonstrated later in this article. However, for most n_0, S_0 would never be observed in a practicable number of simulations.

One approach to overcome this problem is to note that for purposes of inference we are often more interested in the posterior distribution, $p(\Phi|n_0, S_0)$, which is proportional to the product of the likelihood, $p(S_0, n_0, \Phi)$, and the prior, $\Pr(\Phi)$. Using Metropolis-Hasting simulation (Metropolis *et al.* 1953; Hastings 1970), a commonly used MCMC method, it is possible to draw samples from $p(G, \Phi|n_0, S_0)$ for fixed S_0, knowing only $p(\{S_0, G\}|n_0, \Phi)$ and $\Pr(\Phi)$. The (marginal) posterior distribution, $p(\Phi|n_0, S_0)$, can be estimated from these simulated samples using standard density estimation techniques. It is also straightforward to estimate the marginal posterior densities for components of Φ and G. In the approach of Wilson and Balding (1998), samples are taken from $p(G_n, \Phi|n_0, S_0)$, while in this article, samples are taken from $p(G, \Phi|n_0, S_0)$.

Demographic model: The modeling approach follows that taken by Griffiths and Tavaré (1994b). The population is of size N_0 chromosomes when the sample is taken. Time is measured in units of N_0 generations. The population is sampled at time $t = 0$ with time increasing into the past. The size of the population and breeding structure is assumed to be such that the genealogical history is well approximated by standard coalescent theory (see Donnelly and Tavaré 1995). Looking backward in time, the population size, $N(t)$, changes deterministically to an ancestral size N_1 at time t_f and then remains constant at N_1 for $t > t_f$. Two models are considered: a linear change in population size with time; and an exponential change. To make coalescent modeling easier, population size in the demographic models is expressed in units of N_0 generations with $v(t) = N(t)/N_0$. In addition, it is useful to define the quantity $r = N_0/N_1$. If $r < 1$, the population has declined; if $r = 1$, the population has remained stable; and if $r > 1$, the population has expanded. In the case of the linear model,

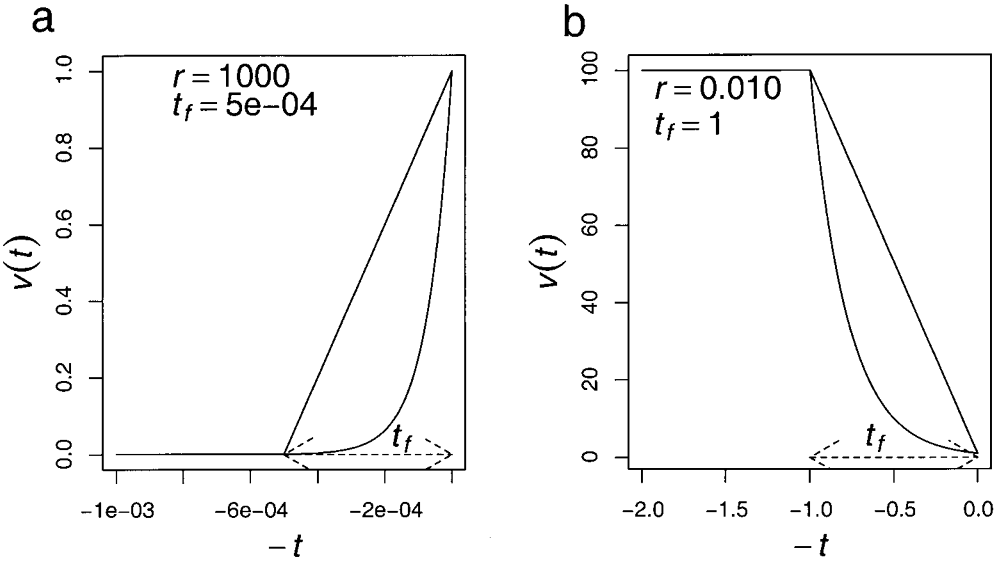

Figure 1.—Both graphs show the scaled population size, $v(t)$, against scaled time, t. Time has been reversed to make it flow in the standard direction on the graphs. Trajectories for exponential change (curved lines) and linear change (straight lines) are shown in each graph for a growing (a) and declining (b) population.

$$v(t) = \begin{cases} \dfrac{rt_{\mathrm{f}} + (1 - r)t}{rt_{\mathrm{f}}} & (0 \leq t \leq t_{\mathrm{f}}) \\ \dfrac{1}{r} & (t \geq t_{\mathrm{f}}). \end{cases} \tag{1}$$

In the case of the exponential model,

$$v(t) = \begin{cases} r^{-t/t_{\mathrm{f}}} & (0 \leq t \leq t_{\mathrm{f}}) \\ \dfrac{1}{r} & (t \geq t_{\mathrm{f}}). \end{cases} \tag{2}$$

(as in Weiss and von Haeseler 1998).

Growth trajectories for the two different models in growing and declining populations are shown in Figure 1. The two demographic parameters, r and t_{f}, are components of Φ in the coalescent model described below. Later in this article a reparameterization of the model is considered, where the interval over which the population size varies is measured in units of generations ($t_{\mathrm{a}} = N_0 t_{\mathrm{f}}$).

Calculation of likelihoods: The aim of this section is to show how $p(\{S_0, G\}|n_0, \Phi)$ can be obtained for the demographic model using the theory in Griffiths and Tavaré (1994b). As described earlier, the genealogical history of a sample can be considered as a sequence of events occurring at times $T_1, T_2, \ldots T_e$. The joint density of these times can be calculated from coalescent theory. At each time there is a set of possible outcomes, and the joint probability of the observed sequence of events can be calculated. The distribution $p(\{S_0, G\}|n_0, \Phi)$ is the product of the joint density of times and the joint distribution of events.

For convenience, $\lambda(t)$ is defined to be the reciprocal of the scaled population size at time t, $\lambda(t) = 1/v(t)$, where $v(t)$ is given by (1, 2) above. Events occur at rate

$$\gamma(t) = \binom{n_i}{2}\lambda(t) + \frac{n_i\theta}{2}, \tag{3}$$

where $\theta = 2N_0\mu$ with mutation rate μ. When an event occurs, the probability that it is a coalescence is

$$\frac{\binom{n_i}{2}\lambda(t)}{\gamma(t)}$$

and the probability that it is a mutation is

$$\frac{n_i\theta/2}{\gamma(t)}.$$

The genealogical model is a function of the three parameters $\Phi = \{\theta, r, t_{\mathrm{f}}\}$.

For an event occurring at time T_{i+1}, $\Pr(T_{i+1} > t_{i+1}| T_i = t_i)$ is

$$C(t_i, t_{i+1}) = \exp(-\int_{t_i}^{t_{i+1}} \gamma(u)du) \tag{4}$$

(Griffiths and Tavaré 1994b). Expressions for $C(t_i, t_{i+1})$ for the two demographic models considered in this article are given in appendix a.

The conditional density of T_{i+1} given T_i is

$$\begin{aligned} p(t_{i+1}|t_i) &= \frac{d}{dt_{i+1}}(1 - C(t_i, t_{i+1})) \\ &= \gamma(t_{i+1})C(t_i, t_{i+1}). \end{aligned} \tag{5}$$

Each event is either a coalescence or a mutation, involving either two or one lineages, respectively. Let $d(t_i,t_{i+1})$ denote the conditional probability density of obtaining the observed event at T_{i+1} given T_i. If the event is a coalescence involving a particular pair of lineages,

$$\begin{aligned} d(t_i, t_{i+1}) &= \frac{\binom{n_i}{2}\lambda(t_{i+1})}{\gamma(t_{i+1})} \frac{2}{n_i(n_i - 1)} p(t_{i+1}|t_i) \\ &= \lambda(t_{i+1})C(t_i, t_{i+1}). \end{aligned}$$

The equivalent expression for a mutation occurring at a particular lineage, with a probability 0.5 that the length is longer/shorter by one unit, is

$$d(t_i, t_{i+1}) = \frac{\theta}{4}C(t_i, t_{i+1}).$$

Summarizing,

$$d(t_i, t_{i+1}) = \begin{cases} \lambda(t_{i+1})C(t_i, t_{i+1}) & \text{if event is coalescence} \\ \theta C(t_i, t_{i+1})/4 & \text{if event is mutation.} \end{cases} \tag{6}$$

Because the events are conditionally independent of each other, the joint density for $\{S_0, G\}$ is the product of the $d(t_i, t_{i+1})$ over all e events in the genealogical history. When the MRCA is reached, the joint density must be multiplied by the probability that the MRCA has the observed mutational state, $p(S_e)$ (*i.e.*, the number of repeats in microsatellites), giving

$$p(\{S_0, G\}|n_0, \Phi) = p(S_e)\prod_{i=0}^{i=e-1} d(t_i, t_{i+1}). \tag{7}$$

It is assumed in this article that $p(S_e)$ is a constant independent of length.

METHOD

In the Metropolis-Hastings simulations described here, a starting point is taken within the domain of the density $L_1 = p(\{S_0, G\}|n_0, \Phi)\Pr(\Phi)$, where S_0 is fixed. New candidate values $\{G, \Phi\} \rightarrow \{G', \Phi'\}$ are chosen, while keeping S_0 unchanged. The values are chosen from some known distribution (the proposal distribution) so that relative frequency of choosing (new values) from (old values), P_f, and the relative frequency of choosing (old values) from (new values), P_r, are both known. The new density is calculated as $L_2 = p(\{S_0, G'\}|n_0, \Phi')\Pr(\Phi')$ and then

$$P = \frac{L_2}{L_1} \times \frac{P_r}{P_f}. \tag{8}$$

If $P \geq 1$ the new value is accepted; otherwise it is accepted only with probability P. If it is not accepted, the original is retained. This process is repeated many times. Provided certain conditions are met (see appendix c), the equilibrium distribution of this Markov chain is the density $p(G, \Phi|n_0, S_0)$. Thus Metropolis-Hastings simulation serves two purposes: it allows us to convert a density from $p(\{S_0, G\}|n_0, \Phi)\Pr(\Phi)$ to $p(G, \Phi|n_{0,} S_0)$, and it allows us to estimate $p(\Phi|n_0, S_0)$ (information on G is ignored in the MCMC output).

Definition of initial state: For each independent MCMC simulation the trees were constructed by coalescing lineages and adding mutations to produce genealogical histories consistent with data, but differing between simulations. The initial tree topology (*i.e.*, initial value of G) strongly influences the early dynamics of $\Phi = \{\theta, r, t_f\}$ in the simulations, and therefore the choice of starting values of Φ has little effect on the initial trajectory taken by Φ. In some simulations the initial demographic and mutational parameters were the same among independent runs and were different in others. The trajectories of Φ tended to diverge rapidly among independent runs in both cases, and no differences in the rate of convergence was observed among either case.

Updating the parameters: The parameters that need to be updated are the sequence of events in the genealogical history, the times of the events, T_i, and the components of Φ: θ, r, and t_f. The general scheme is outlined first, followed by the details. The components of Φ are parameterized on the $\log_{10}$ scale. The parameters are updated in a random order with (conditional) probabilities given in parentheses below. This scheme was devised by trial and error to obtain good rates of convergence.

(0.95) Update sequence of events.
(0.05) Update some of T_i, $\log_{10}(\theta)$, $\log_{10}(r)$, $\log_{10}(t_f)$.

(0.5) T_i only.
(0.5) Some of T_i, $\log_{10}(\theta)$, $\log_{10}(r)$, $\log_{10}(t_f)$.

(1/12) $\log_{10}(\theta)$; (1/12) $\log_{10}(\theta)$, T_i.
(1/12) $\log_{10}(r)$; (1/12) $\log_{10}(r)$, T_i.
(1/6) $\log_{10}(t_f)$, T_i.
(1/12) $\log_{10}(\theta)$, $\log_{10}(r)$; (1/12) $\log_{10}(\theta)$, $\log_{10}(r)$, T_i.
(1/6) $\log_{10}(\theta)$, $\log_{10}(t_f)$, T_i.
(1/6) $\log_{10}(\theta)$, $\log_{10}(r)$, $\log_{10}(t_f)$, T_i.

Updating the sequence of events: There are many possible proposal distributions for updating the sequence of events in the genealogical history, G. The scheme described below was devised from two considerations. First, the Metropolis-Hastings acceptance step depends on $L_2/L_1 \times P_r/P_f$. Although the L_2/L_1 term is straightforward to calculate, it is important to have proposal distributions for which P_r/P_f can be easily calculated. Second, it is necessary to have an updating scheme that can be demonstrated to explore all possible sequences of events. As shown in appendix c, both these requirements are met by the following scheme.

1. Addition of two mutations within a lineage. Two mutations are added at points chosen uniformly randomly along a lineage. Because a stepwise mutation model is assumed here, these are canceling pairs that shorten (-1 mutations) and lengthen ($+1$ mutations) the microsatellite by one unit.
2. Removal of two mutations within a lineage. The reverse of (1).
3. Addition of three mutations around a coalescent node. A mutation of the same sign [either ($+1$) or (-1)] is added to each of the two descendent lin-

eages, and a mutation of opposite sign is added to the ancestral lineage.

4. Removal of three mutations around a coalescent node. The reverse of (3).
5. Interchange of lineages. Two temporally adjacent events occurring on separate lineages are chosen and the ancestral portions of the lineages are swapped.
6. Nearest-neighbor interchange of lineages. See appendix b.
7. Swapping the order of events. Two temporally adjacent events are chosen and the order of occurrence is swapped.

While classes (1) and (3) are always available irrespective of the state of the system, there can exist states of the system in which the other classes of update cannot be applied. For example, class (2) cannot be applied when there are no lineages with pairs of $+1/-1$ mutations. Probabilities $(R_1, \ldots, R_7) = (0.1, 0.1, 0.2, 0.2, 0.1, 0.2, 0.1)$ give the chance that each class of update is chosen in the ideal case that all classes of update are possible. A set of weights E_j is then defined such that $E_j = R_j$ if the transformation can be applied to at least one transformable element of that class, $E_j = 0$ otherwise. Then, with probability $E_j/\Sigma E_j$ the jth class is chosen at each update step. The details of the proposal distributions associated with these classes of updates, and the calculation of P_r/P_f, are given in appendix b.

Updating T_i and Φ: Candidate values of T_i were obtained by generating a uniform random variable, substituting this for $C(t_i, t_{i+1})$ in (4), and solving for t_{i+1}, given t_i. P_r/P_f can be calculated from (5). In the case of populations that are not changing in size the updates are always accepted because $L_2/L_1 \times P_r/P_f = 1$.

Values of $\log_{10}(\theta)$, $\log_{10}(r)$, or $\log_{10}(t_f)$ were updated by adding normal random deviates with mean 0 and standard deviation 0.5. As shown later, in some regions of parameter space, $\log_{10}(\theta)$ and $\log_{10}(r)$ can be positively correlated, whereas $\log_{10}(\theta)$ and $\log_{10}(t_f)$ can be negatively correlated. Therefore, when any of $\log_{10}(\theta)$, $\log_{10}(r)$, or $\log_{10}(t_f)$ were updated jointly, as shown in the general scheme given earlier, the same random deviate was used for each updated parameter, but of differing sign for $\log_{10}(t_f)$. For all these cases $P_r/P_f = 1$.

Determination of run length and assessing output from MCMC simulations: It is important to determine whether the simulations have been run for a sufficient number of iterations to give an adequate estimate of $P(G, \Phi | n_0, S_0)$. The two most widely used approaches are either to run one long chain (Raftery and Lewis 1996) or several shorter chains with widely dispersed starting points (Gelman *et al.* 1995). The former method gives the number of iterations required to estimate quantiles from the posterior distribution of a monitored parameter to a specified degree of precision. The latter method estimates the quantity $\sqrt{(V_w + V_b)/V_w}$, where V_w is the variance of the parameter within a chain and V_b is the variance of the means among chains (Gelman and Rubin 1992). Gelman *et al.* (1995) suggest that values <1.1 (*i.e.*, where V_b is $\sim$5% of V_w) are adequate.

In the analysis of simulated data sets, unless otherwise stated (see description below), five independent chains were run, with different starting genealogical histories. The Gelman-Rubin statistic was monitored for $\log_{10}(\theta)$, $\log_{10}(r)$, and $\log_{10}(t_f)$, and the run length was determined by the need to keep the statistic <1.1. On the basis of these simulations, when real data sets were analyzed, as described in results, the number of simulations was reduced to a single long run for some data sets, and the Raftery-Lewis statistic was monitored for the 0.025 quantile.

In general the first 1% of sampled points for each run were discarded to ensure that the distributions were not influenced by unrepresentative initial values. A total of 10,000–50,000 points were collected from each run. Approximate densities have been calculated from the sampled parameters using the program Locfit (Loader 1996) implemented in R (http://stat.auckland.ac.nz/r/r.html), and contours corresponding to the 0.1, 0.5, and 0.9 highest posterior density (HPD) limits are plotted. The HPD limits are values of the variate that have the same density and define a region within which the probability is some critical value (*e.g.*, 0.9).

SIMULATION STUDIES

To check that the MCMC method performs correctly, a number of tests were carried out using simulated data sets. Samples were simulated from growing, declining, and stable populations. The method of simulating genetic samples is based on that of Hudson (1991). The intervals between coalescent events are first simulated by inverting (4) as described earlier, with $\theta = 0$ (*i.e.*, with no mutation). Lineages have equal probabilities of coalescing. The number of mutations, m_j, occurring down the jth lineage of length t_j is simulated as a Poisson random variable with parameter μt_j. The ancestral length is taken to be 0. m_j additions of $+1$ or -1 are successively applied down each lineage. The resulting sample is then centered by subtracting the length of the shortest chromosome from all chromosomes.

A sample of size 108 chromosomes was generated with parameter set MOD1 = $\{\theta = 2000.0, r = 1000.0, t_f = 0.0005\}$ (growing, Figure 1a) from the linear model. This was then subdivided into two samples of sizes 100 and 8. The smaller sample was then used to make comparisons between likelihood surfaces calculated by direct simulation with those obtained by the MCMC method. In addition a sample of 100 was generated from MOD2 = $\{\theta = 0.2, r = 0.01, t_f = 1.0\}$ (declining, Figure 1b) using the exponential model. Also a sample was simulated from a stable population MOD3 = $\{\theta = 10.0\}$. These were then analyzed with the MCMC method to assess how accurately the parameters were estimated, as described below. The frequency distribution of centered lengths in order of increasing size in each of these models was

MOD1   8 (1, 4, 3)
MOD1 100 (3, 23, 41, 28, 4, 1)
MOD2 100 (3, 0, 0, 0, 0, 33, 46, 10, 8)
MOD3 100 (4, 26, 9, 7, 6, 28, 18, 2).

Another test of the MCMC method is to examine the joint posterior distributions when the method is applied to data from a star genealogy, where all lineages have equal length, radiating from the MRCA. This approximates the genealogy expected in a population growing so rapidly that all coalescent events after the first occur within an interval negligibly small in comparison to the time over which the population has been growing. In this case, a sample size of 100 was simulated by drawing 100 Poisson random variables, x_i with mean 2. x_i random increments of $+1$ and -1 were added to an initial length of 0 to generate a distribution of lengths: (1, 3, 10, 27, 27, 21, 7, 4). This corresponds to a star genealogy with $\theta t_{\mathrm{f}} = 2\mu t_{\mathrm{a}} = 4$.

Comparison of conditional posterior distributions with likelihoods estimated from MC integration: For these studies the sample of size 8 from MOD1 was used. Posterior distributions were estimated separately for $\log_{10}(\theta)$, $\log_{10}(r)$, and t_{f} (in each case, the other parameters were fixed at the values used to simulate the sample). Five independent MCMC simulations were made for each parameter and the results are illustrated in Figure 2, a–c. The general updating scheme given earlier was modified to take into account that only one parameter was allowed to vary. The value of t_{f} was updated using lognormal deviates. Flat (improper) priors were used for $\log_{10}(\theta)$ and t_{f}. In the case of $\log_{10}(r)$, because the likelihood function asymptotes, a rectangular prior with limits (0–7) was used in the MCMC simulations. Each chain was run for 10^7 iterations. The curves were obtained by joining the midpoints in histograms constructed from the MCMC output. These were normalized so that the area under the histograms between the two endpoints sums to 1.

These curves can be compared with likelihoods estimated by MC integration. In this method a large number of data sets are simulated as described in the previous section. The likelihood is estimated as the proportion of simulated samples whose distribution of lengths is the same as that of the target sample. The process is repeated for different parameter values to obtain a likelihood surface. The likelihoods were estimated using 10^7 simulations. These were then normalized so that the resulting curves could be compared with the MCMC histograms. There is a good fit between the two approaches.

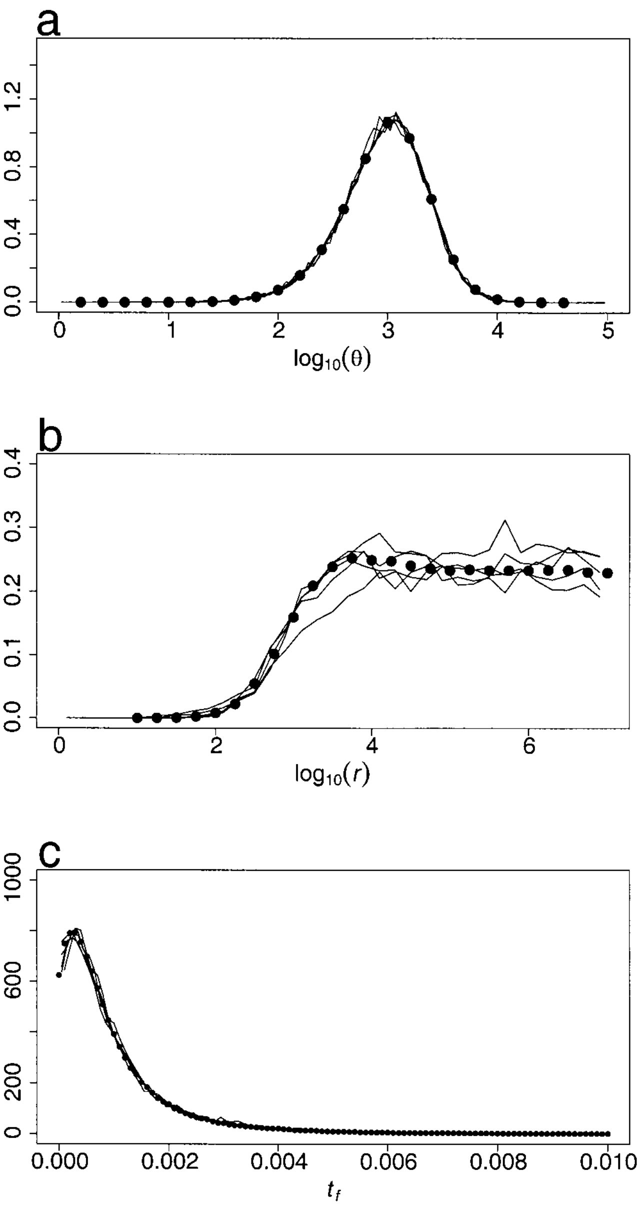

Figure 2.—(a) Estimated posterior density of $\log_{10}(\theta)$. The solid lines are estimated by the MCMC method. The solid circles are the relative likelihoods estimated by the direct method, scaled to enclose a unit area. Plots for $\log_{10}(r)$ and t_{f} are given in b and c. The densitites for each parameter are conditional on the following values: $\theta = 2000$, $r = 1000$, and $t_{\mathrm{f}} = 0.0005$ (1, 4, 3).

Comparison of joint posterior distributions with simulated likelihoods: The tests described above only consider conditional univariate distributions. To compare MCMC likelihoods with simulated likelihoods over the three parameters jointly, the MCMC simulation was performed using a rectangular prior on the log scale. The limits for $\log_{10}(r)$ and $\log_{10}(\theta)$ were (-5–5) and for $\log_{10}(t_{\mathrm{f}})$ were (-5–1). This volume was binned into $20 \times 20 \times 20$ cells. Five replicate runs were performed. Based on examination of the output of the MCMC simulation, 48 bins were selected for comparison with the likelihoods estimated directly from simulations. The bins were chosen both to have a wide spread of likelihoods and to sample the parameter space widely. Each

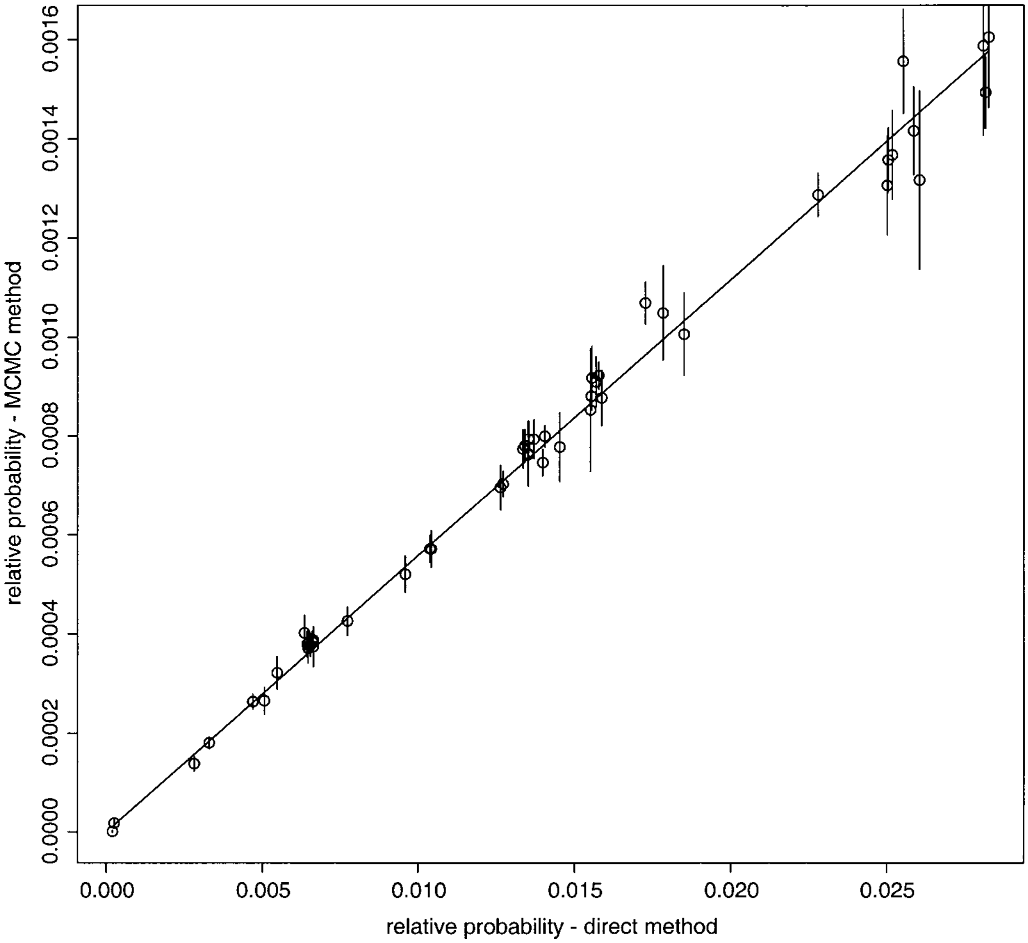

Figure 3.—A plot of the estimated frequencies in 48 bins from the joint posterior distribution of $\log_{10}(\theta)$, $\log_{10}(r)$, and $\log_{10}(t_f)$ against the likelihood estimated by the direct method. The likelihood is averaged within the bin, as described in the text. The vertical lines indicate 2 standard errors estimated from the 5 replicate MCMC simulations.

bin was subdivided into 10 × 10 smaller bins, and likelihoods were estimated from 10^4 simulations within each small bin and summed over all small bins to give a value proportional to the volume of the larger bins. Figure 3 shows the volume (proportion of points observed) of each chosen bin from the MCMC simulations, averaged over the five runs, plotted against the volumes estimated from direct simulation. It can be seen that there is a good correspondence between the two methods.

The joint posterior distribution for data simulated from a star genealogy: The MCMC simulation was run for 10^8 steps with the sample simulated from a star genealogy, using the same rectangular prior as discussed earlier. The 0.025, 0.5, and 0.975 quantiles of the distribution of mutations were 65, 143, and 228, respectively, compatible with the expected number of 200. Figure 4, a–c, shows the three bivariate marginals for the demographic parameters. The 0.1, 0.5, and 0.9 HPD limits are plotted. Also plotted is a sample of 10,000 points (2000 from the 50,000 recorded in each independent chain). Using statistics summarizing the shape of the genealogy as a guide (results not shown), a number of different regions can be identified in these figures.

Region A corresponds to genealogies that arise when t_f is sufficiently short that the shape of the genealogy depends on N_1 only—*i.e.*, a nonvarying population with mutation parameter θ/r, independent of t_f. Thus in Figure 4a the points lie around the line $\log_{10}(\theta) - \log_{10}(r) = \log_{10}(2N_1\mu) = k$.

Region B corresponds to the case when t_f is sufficiently long that the shape of the genealogy depends on N_0 only, independent of t_f and r. In this example, region B does not extend substantially below $\log_{10}(r) = 0$ in Figure 4a, but this is not generally the case.

Region C corresponds to star genealogies. Within region C, the likelihood becomes independent of $\log_{10}(r)$ and is strongly ridged along a line $\log_{10}(t_f) + \log_{10}(\theta) = \log_{10}(2\mu t_a) = k$. In this case, because the data were simulated from a star genealogy and $2\mu t_a$ is twice the expected number of mutations within a lineage, we may expect k to be close to $\log_{10}(4)$. The existence of a mode within the body of region C in Figure 4b, rather than at the edges as in Figure 4, a and c, is a consequence of this being the marginal distribution over r, which restricts t_f: if $\log_{10}(r)$ is extended beyond 5, the ridge in Figure 4b extends toward the lower corner of the graph (results not shown).

It is useful to assess how accurately the expected number of mutations within a lineage can be recovered from the analysis. A posterior distribution for $\log_{10}(2\mu t_a)$ (conditional on $r > 1$) can be obtained from the distribution of the sum of $\log_{10}(\theta)$ and $\log_{10}(t_f)$ evaluated at each sample point. In this case, however, the prior is no longer uniform because it is the distribution of the sum of two uniform variates. It is a trapezoid distribution with bounds (-10, 6). It is possible to correct for this by dividing the fitted density by the density of the trapezoid. This is most easily done in Locfit by weighting each point by the reciprocal of the density of the trapezoid, and is a special case of the important sampling method described in Tanner (1993, p. 34) for altering the output of MCMC simulations to reflect different priors. It should be noted, however, that, despite this correction, the posterior distribution, especially in its tails, still depends on the original priors for the three parameters. In the particular case here, the posterior distribution is hardly affected by the correction. For these data the estimated mode is 0.62 with 0.9 HPD limits 0.23–0.84, which agrees well with the expected number of $\log_{10}(4) \approx 0.602$.

The lower limit of the 0.9 HPD region for $\log_{10}(r)$ is 1.82. The proportion of sample points with $\log_{10}(r) > 0$ is 0.96. This corresponds to a Bayes factor of 26.8 favoring growth *vs.* contraction where the Bayes factor is the posterior odds divided by the prior odds (O'Hagan 1994). Thus in this artificially extreme example a single microsatellite locus can provide strong support for growth.

The effect of the demographic model: The joint posterior distributions of the demographic and mutational parameters were estimated using the samples of size 100 generated for MOD1, MOD2, and MOD3. The samples from MOD1 and MOD2 were analyzed assuming both linear and exponential population change. The sample from a stable population, MOD3, was analyzed assuming a model of linear population change. In each case the MCMC simulation was run for 10^8 updates. This was replicated with different starting trees five times for samples from MOD1 and MOD2, and twice for the samples from MOD3. The same rectangular prior as described earlier was used for the growing and stable populations. In the case of the contracting population the bounds on $\log_{10}(t_f)$ were (-3–3). The main aim of this test was a qualitative comparison of the two models. A detailed separation of the effects of sampling error in the MCMC output from small differences between the two models has not been performed. The parameters analyzed here are $\log_{10}(r)$, $\log_{10}(t_f)$, and the reparameterized quantity, $\log_{10}(2\ \mu t_a) = \log_{10}(t_f) + \log_{10}(\theta)$, introduced in the previous section.

Figure 5, a–d, shows the joint distribution of $\log_{10}(r)$ and $\log_{10}(t_f)$, marginal over θ for the two different parameter sets, analyzed using MOD1 and MOD2. The estimated 0.1, 0.5, and 0.9 highest posterior density re-

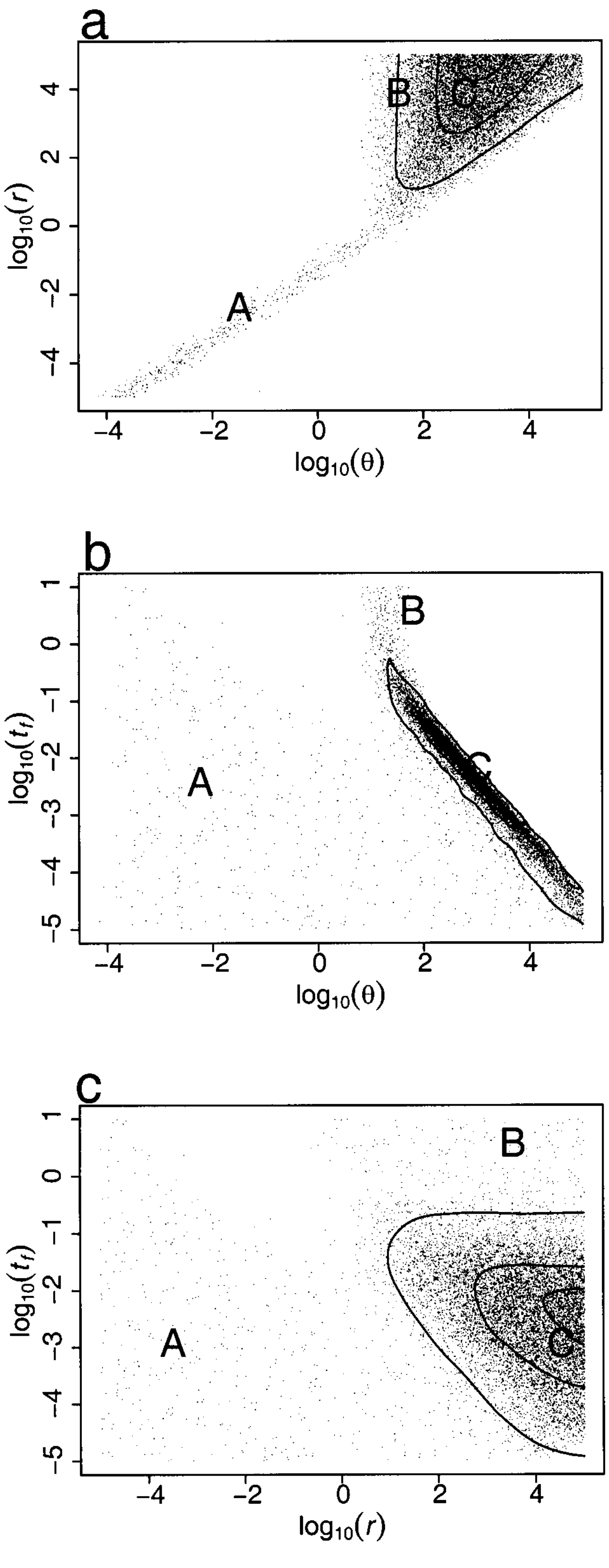

Figure 4.—Plots of 10,000 simulated points from the joint distribution of $\log_{10}(\theta)$, $\log_{10}(r)$, and $\log_{10}(t_f)$ estimated by the linear model using data simulated from a star genealogy. The three bivariate marginal distributions are shown. The labeled regions A, B, and C are described in the text. The contours correspond to estimated 0.9, 0.5, and 0.1 HPD intervals.

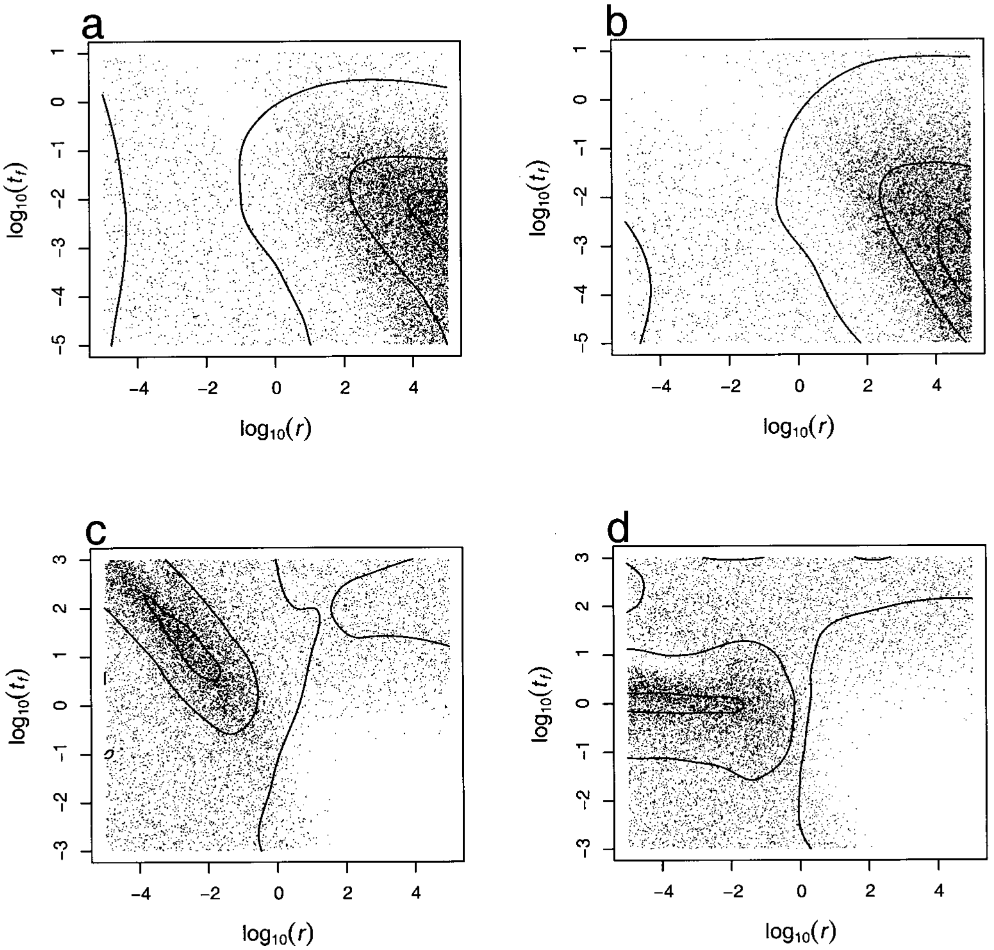

Figure 5.—(a and b) 10,000 simulated points from the marginal posterior distributions of $\log_{10}(r)$ and $\log_{10}(t_f)$ estimated from data simulated from a linearly growing population (MOD1). (c and d) Distributions estimated from data simulated from an exponentially declining population (MOD2). a and c assume a linearly varying population. b and d assume an exponentially varying population. The contours correspond to estimated 0.9, 0.5, and 0.1 HPD intervals.

gions are indicated. A sample of 10,000 points is also plotted. As can be seen from the distribution of points, the density is very flat outside the modes, and estimation of the 0.9 HPD limit in two dimensions is imprecise.

The sample simulated from a linearly growing population (Figure 5, a and b) seems to contain good evidence that it is from a growing population. The 0.9 HPD limit on $\log_{10}(r)$ is greater than 0 (Table 1). The posterior odds for a growing *vs.* declining population are 9.2 when analyzed under the linear model and 9.8 under the exponential model (this difference is commensurate with sampling error). The sample simulated from a declining population (Figure 5, c and d) does not provide equally strong evidence of decline, with an HPD limit on the marginal distribution of r that is substantially >0. The posterior odds for a declining *vs.* growing population are 4.4 in the exponential model and 4.8 in the linear model.

With data from a growing population the posterior distributions are very similar when analyzed assuming linear (Figure 5a) or exponential (Figure 5b) growth. The modes and 0.9 HPD limits are similar under the two models (Table 1). In the case of data simulated from a contracting population it can be seen from Figure 5, c and d, that the joint distribution of $\log_{10}(r)$, $\log_{10}(t_f)$ is clearly different between the two models. In the exponential model, the 0.1 HPD region lies around $\log_{10}(t_f) = 0$, independent of $\log_{10}(r)$, whereas in the linear model it appears to lie around a line $\log_{10}(r) + \log_{10}(t_f) = \log_{10}(t_a/N_1) = k$, *i.e.*, where time is scaled by ancestral population size.

The results from the data simulated from a stable population (MOD3) are summarized in Table 1. Plots of the joint marginal distributions for the three parameters are very similar to those given in Figure 4 except that the region C has very low density. The Bayes factor for a growing *vs.* declining population is 0.61. The distributions for t_f and $2\mu t_a$ are strongly bimodal. Because the HPD region is split around the two modes, both values are reported. The bimodality arises because region C is absent. A qualitative explanation for the shape is that the mode at smaller times arises from density in region A, and the mode at larger times arises from density in region B.

EXAMPLES

Northern hairy-nosed wombat: The data are from 16 microsatellite dinucleotide repeat loci scored from 28 northern hairy-nosed (NHN) wombats by Taylor *et*

TABLE 1

Summary statistics for simulated data analyzed using different models

		Linear model			Exponential model		
Growing							
$\log_{10}(r)$	(3.0)	0.082	(5)	(5)	0.257	(5)	(5)
$\log_{10}(t_f)+$	(−3.3)	−4.64	−1.97	−0.58	−4.74	−2.24	−0.49
$\log_{10}(2\mu t_a)+$	(0.0)	−0.51	0.27	0.66	−0.42	0.38	1.10
Stable							
$\log_{10}(r)$	(0.0)	(−5)	0.12	2.58			
$\log_{10}(t_f)+a$		(−5)	(−5)	−2.06			
$\log_{10}(t_f)+b$		(−0.84)	(1)	(1)			
$\log_{10}(2\mu t_a)+a$		−2.41	−0.36	0.22			
$\log_{10}(2\mu t_a)+b$		0.44	1.29	2.09			
Declining							
$\log_{10}(r)$	(−2.0)	−5.00	−2.41	1.70	−5.0	−1.88	1.53
$\log_{10}(t_f)-$	(0.0)	−1.68	1.74	(3)	−2.45	0.07	1.86
$\log_{10}(2\mu t_a)-$	(−0.70)	−5.59	−0.6	0.8	−5.86	−2.72	3.10

The lower 0.9 HPD limit, mode, and upper 0.9 HPD limit for three parameters, analyzed using the linear model and exponential model. The data have been simulated from growing, stable, or declining populations, as described in the text. The symbols + and − indicate whether the summary statistics have been calculated by conditioning on positive or negative $\log_{10}(r)$. Disjunct HPD regions are indicated by *a* and *b*. Parameter values with which the samples were simulated are given in the first column.

al. (1994), from which the following details have been obtained. The range of these wombats is restricted to a single locality (Epping Forest National Park, Queensland, Australia) from which the individuals were sampled. The species is believed to have suffered an extreme population decline in the past 120 years from a population size of many thousands to 20–30 individuals in 1981. The population size was around 70 when the individuals were genotyped. Seven of the loci are monomorphic. One locus has repeat length variation among the NHN wombats that is not a consistent multiple of 2, and was not used in this analysis. Taylor *et al.* (1994) also present data for the southern hairy-nosed (SHN) wombat, which has not suffered the same decline, and for some museum specimens of the NHN wombats collected around 1884, but these are not used in this analysis. However, these data show that the loci that are monomorphic in the NHN wombats are either polymorphic or (in the case of one locus) have a fixed difference between the two species. The generation time is believed to be 10 years.

In the analysis, rectangular priors on the $\log_{10}$ parameter values have been used. The limits for N_0, N_1, μ, and t_a, broadly reflecting the imformation given above, were taken to be (200–400), (2000–200,000), (10^{-2}–10^{-7}), and (5–100), respectively. From this, taking the most extreme combination of parameter values and rounding outward to the nearest whole number on a $\log_{10}$ scale the limits for $\log_{10}(\theta)$ are (−6–1), for $\log_{10}(r)$ are (−4–0), and for $\log_{10}(t_f)$ are (−2–1). The prior limits for μ were considered reasonable even for the monomorphic loci, based on the knowledge that these are polymorphic within the NHN/SHN clade.

The frequencies for the polymorphic loci are

L1 = (28, 0, 0, 0, 0, 0, 0, 20)

L2 = (11, 0, 0, 29)

L3 = (12, 14, 6)

L4 = (50, 0, 6)

L5 = (42, 14)

L6 = (33, 2, 0, 0, 0, 0, 0, 21)

L7 = (16, 0, 18, 22)

L8 = (10, 0, 0, 0, 46).

Each polymorphic set and a monomorphic set were analyzed separately. Based on the results for simulated data sets, one single run of length 6×10^7 updates was simulated for each set. Because the decline was believed to be recent, an exponential model was considered appropriate. The Raftery-Lewis method indicated that the 0.025 quantile was estimated to ±0.0125 for the three parameters in all loci, and to ±0.005 for all parameters in six loci.

To combine information across loci it is assumed that r and t_f are the same for all loci but θ differs among loci. A combined joint posterior distribution for r and t_f (marginal over θ) can be obtained by estimating the joint density (proportional to the likelihood over the interval) for each locus separately and then multiplying. The 0.9, 0.5, and 0.1 HPD regions for the combined density from the 8 polymorphic loci are given in Figure 6. Also shown is the combined density for all 15 loci. It can be seen that there would be an appreciable bias had only the results from the polymorphic loci been reported.

The effect of adding the monomorphic loci appears counterintuitive in that they support a shorter timescale

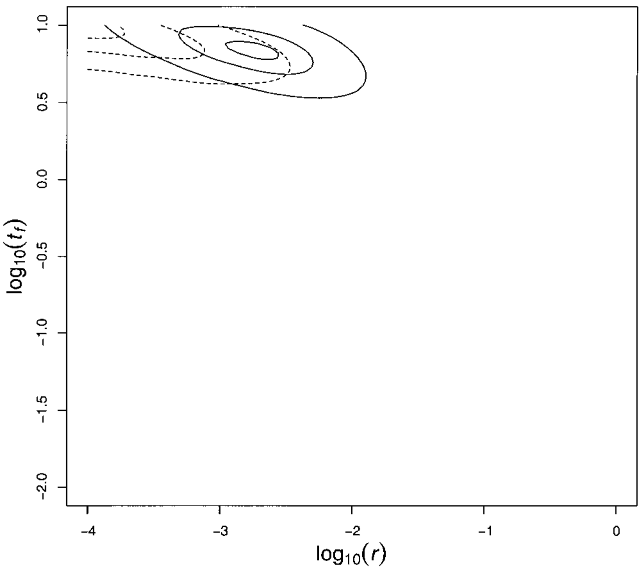

Figure 6.—Plot of the marginal posterior distribution of $\log_{10}(r)$ and $\log_{10}(t_f)$ for the northern hairy-nosed wombat data described in the text. The dotted lines give the 0.9, 0.5, and 0.1 HPD limits estimated from the polymorphic loci. The solid lines give the same HPD limits for both polymorphic and monomorphic loci.

over which the contraction has happened and a less severe contraction. The explanation is that, over the range of $\log_{10}(t_f)$ considered here, the number of mutations within the genealogy is determined by $\theta/r = 2N_1\mu$, small values of which have higher posterior density. Thus, on average, larger values of r will have higher density. Because the joint density for $\log_{10}(r)$ and $\log_{10}(t_f)$ is ridged, smaller values of $\log_{10}(t_f)$ also have higher density. In fact larger values of $\log_{10}(t_f)$ appear to have higher density than might be expected from this argument. This can be explained by observing that, when $\theta/r = 2N_1\mu$ is large, the bulk of mutations occur in the genealogy for $t > t_f$. However, a larger value of t_f makes the genealogy length $T_e < t_f$, and hence long t_f has higher density.

In conclusion, the analysis of the wombat data suggests that the population has been declining over a longer period and has suffered a sharper decline than the historical data imply. The wombat data also highlight the biases inherent in reporting data from only polymorphic loci. Inferences about the decline have to be treated cautiously, however, because the simulation results in the previous section suggest that parameter estimates are strongly affected by model assumptions. Taylor *et al.* (1994) assumed that the ancestral population had heterozygosity levels comparable with present-day levels in the SHN wombat. On this basis, they calculated that N_e would need to be <10 over the entire 120-year period. It is possible that if a step model of decline were used, the results may be consistent with their conclusions. An additional caveat is that the single-step mutation model is unlikely to be correct.

European humans: In this study a subset of 10 tetranucleotide loci was taken from a much wider set of 60 loci scored from 15 populations described in Jorde *et al.* (1995, 1997). The Northern European and French samples were combined. The 10 tetranucleotide loci were chosen uniformly randomly from the 60 kindly provided by Dr. Jorde, and have the following length distributions:

L1 = (1, 0, 0, 2, 7, 16, 16, 16, 18, 48, 27, 15, 9, 5)
L2 = (5, 19, 7, 12, 41, 20, 29, 33, 12, 2)
L3 = (1, 34, 24, 55, 53, 7)
L4 = (65, 115)
L5 = (2, 4, 1, 65, 26, 49, 13, 2)
L6 = (6, 12, 107, 5, 3, 12, 6, 14, 3)
L7 = (8, 12, 31, 56, 39, 19, 7)
L8 = (3, 1, 127, 39, 2)
L9 = (1, 25, 108, 33, 3)
L10 = (7, 0, 7, 61, 38, 24, 12, 14, 10, 1).

The following limits for N_0, N_1, μ, and t_a were considered: (10^5–10^9), (10^3–10^8), (10^{-2}–10^{-7}), and (25,000–10^6), respectively. Taking the same approach as for the wombats, the limits for $\log_{10}(\theta)$ are (−2–7), for $\log_{10}(r)$ are (−3–6), and for $\log_{10}(t_f)$ are (−6–0).

Run lengths and number of independent runs were determined individually for each of the 10 loci, using the Gelman-Rubin statistic as a guide. For most loci three independent runs were carried out. The run lengths varied from 10^8 to 2×10^8 updates. A single run of 10^8 updates for the locus 2 sample takes ~6 hr on a standard 333 Mhz Pentium II PC under Linux.

To combine information across loci it is assumed, as with the example above, that in the joint likelihood funtion for the 10 loci, the demographic parameters are the same for all loci but θ differs among loci. Thus to obtain a marginal posterior density for t_f and r, the joint marginals for t_f and r are estimated for each locus and then multiplied together. The marginals for $\log_{10}(t_f)$ and $\log_{10}(r)$ are obtained by integrating over this function. Note that, in general, this will give a different result from the case of multiplying the marginals for, *e.g.*, $\log_{10}(t_f)$ across loci, which assumes that both $\log_{10}(r)$ and $\log_{10}(\theta)$ differ among loci, but $\log_{10}(t_f)$ is the same across loci.

The mean numbers of mutations within the genealogies of the 10 loci were estimated as 270, 118, 38, 2, 40, 38, 187, 12, 51, and 83, respectively. Two out of the 10 loci (numbers 7 and 9) show strong evidence of having been drawn from a growing population with lower 90% HPD limits for the marginal distribution of $\log_{10}(r)$ of 0.98 and 1.95, respectively. The proportions of sampled points with $\log_{10}(r) > 0$ were 0.93 and 0.97 for the two loci, giving Bayes factors in favor of growth *vs.* decline of 6.6 and 16.2, respectively. However, combined over loci, the marginal posterior distribution of $\log_{10}(r)$ has one mode at 0.45 with 90% HPD limits of −2.2–3.27. This mode is relatively sharp, but with very thick tails,

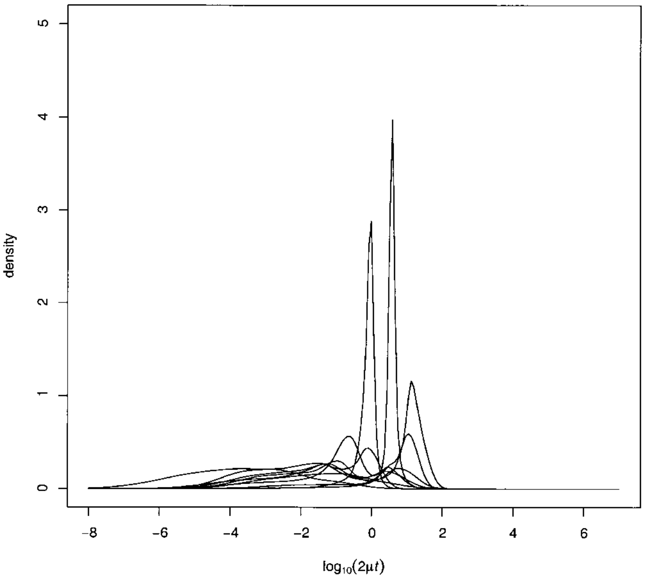

Figure 7.—The marginal posterior distributions of $\log_{10}(2\mu t_a)$ estimated from the 10 loci sampled from European humans.

hence the wide HPD limits. Thus there is little overall evidence to suggest that the data are drawn from a population that has been growing. The proportion of this combined density with $\log_{10}(r) > 0$ is 0.8, giving a Bayes factor of 2 in favor of population growth. Conditioning on $\log_{10}(r)$ being positive, the marginal distribution of $\log_{10}(t_f)$ can be combined over loci. This has a taller mode at −0.12 and a smaller mode at the −6 limit. The HPD region covers two intervals: −6–−4.1 and −2.8–0. Thus the information on t_f is very diffuse.

It is also possible to consider the reparameterization, $\log_{10}(2\mu t_a) = \log_{10}(\theta) + \log_{10}(t_f)$, conditional on $\log_{10}(r) > 0$, as discussed earlier. The results are illustrated in Figure 7. Many of the loci exhibit the bimodal distribution noted in the data simulated from a stable population. As noted earlier, the tails of these distributions depend on the priors for the three parameters jointly. The two tall peaks come from the two loci showing strong evidence of population growth. The distribution of modes is broad, with the taller modes centered around −1–1. This variation probably reflects different mutation rates at different loci. The leftmost mode corresponds to the locus with two alleles. Because a common mutation rate cannot be assumed, the posterior distributions (proportional to likelihoods over the interval) cannot be multiplied to give an overall posterior distribution. These could be integrated over some distribution of mutation rates to obtain distributions of t_a, which could then be combined, but, because alternative approaches are preferable (see discussion), this has not been attempted here.

It is possible to summarize the shape of genealogies by calculating the sum of the lengths of all branches leading from each data point to its first coalescent node as a proportion of the sum of all branch lengths within the tree. A value close to 1 indicates a star genealogy, while a value close to 0 indicates an "etiolated" genealogy. Figure 8 shows the distribution of this statistic for the MOD1, MOD2, and MOD3 data, each analyzed with the linear model (with, however, the prior limits for the examples in Figure 5, which are less biased toward star genealogies). Plotted alongside is the distribution for the 10 human loci. It can be seen that 5 of the loci have distributions very similar to that of MOD3 and to each other. The 2 loci showing high values of this "distal branch index" are loci 7 and 9 identified earlier. The reason for the higher density near 1 in comparison with the simulated example probably reflects the different prior limits used with the latter.

In conclusion, the 10 loci so far studied for the human European data show conflicting evidence of past population growth. Two loci show strong evidence of growth with the lower 90% HPD limit greater than $\log_{10}(r) > 0$. Yet the mode estimated from all loci is only 0.45, with wide bounds. The taller peaks for the posterior density of $\log_{10}(2\mu t_a)$ (conditional on $\log_{10}(r) > 0$) appear centered around −1–1. Assuming a point estimate of mutation rate of 5×10^{-4} and generation time of 20 years, this would imply that the population has been growing over the last 2000–200,000 years. The best estimate of the amount that it has grown is around threefold. Such calculations cannot be taken seriously at this stage until a more complete analysis of the human data is performed, using the natural parameters and more informative priors in the manner of Tavaré *et al.* (1997) and Wilson and Balding (1998), and more loci.

There are three likely causes for the apparently discrepant results between loci. First, it may be due to sampling, and there is no discrepancy. This would imply that there is little evidence of star genealogies in the European data. Second, the discrepancies may arise from making the strong assumption that t_a, N_0, and N_1 are the same for all loci. While strictly correct, selection on nearby loci may tend to shrink or expand the genealogies (Nordborg 1997). Thus it may make inference more robust if a hierarchical Bayesian model were used in which the variance in demographic parameters between loci could be separately estimated. Third, as discussed with the wombat example, the assumption of a strict single-step mutation model may also contribute to discrepancies between loci.

DISCUSSION

The introduction of general likelihood-based methods of inference by Griffiths and Tavaré (1994a) and Kuhner *et al.* (1995) promises to revolutionize the analysis of genetic data. However, the results of currently described methods (including this one) have to be treated cautiously. The essential difficulty is that (1) implementation is complicated and there is always the

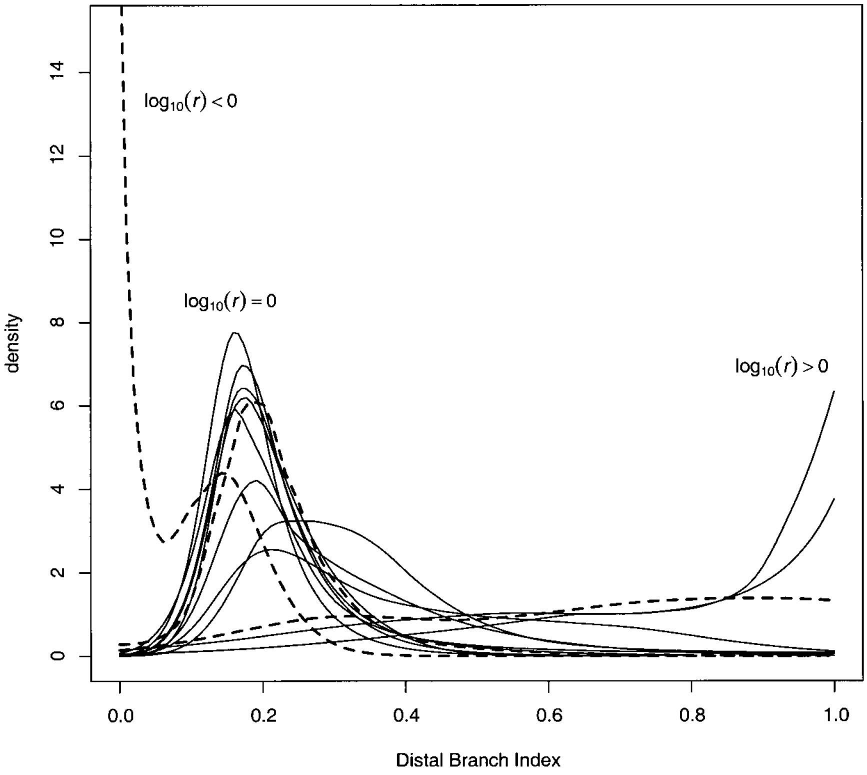

Figure 8.—The distribution of the distal branch index, defined in the text, for the 10 loci sampled from European humans. Plotted alongside are distributions of the distal branch index taken from the analysis of the simulated data for declining, stable, and expanding populations, assuming linear population growth (*i.e.*, corresponding to the first column of Table 1).

possibility of programmer error, and (2) all methods depend on sampling. Although there are theorems that state that the required result will be obtained if the sample is sufficiently large, there is no good way of knowing, in practice, how large this should be.

In this article, the method has been applied to small data sets where the likelihoods can be compared with those obtained from MC integration. Likelihoods estimated using MC integration should be unbiased, whereas those estimated from MCMC may not be if the chain has not yet converged. Although small, a sample of size 8 is not trivial, there being 1.6×10^6 distinguishable tree topologies. The posterior density estimates appear to be broadly accurate, especially in the tails of the distribution (Figure 2). This is also shown in Figure 3, where the absolute error appears to be proportional to the likelihood. Checks such as these are useful in the implementation and validation of MCMC methods.

The method has then been applied to larger data sets simulated with known parameter values. Although the data simulated assuming a star genealogy had no demographic model, the MCMC analysis made reasonable inferences about the number of mutations within the genealogy and the branch length multiplied by mutation rate (equivalent to a notional μt_a, assuming a star genealogy). In addition, when applied to data simulated from demographic models with known parameters, the posterior distributions of these parameters were compatible with the known values.

It is clear that the likelihood surface for growing populations is complex and, with a flat prior distribution, the posterior distribution for Φ has infinite volume. Under various regimes the dimensionality of the likelihood function is reduced—*i.e.*, the model becomes underdetermined. When t_a is very short or very long the likelihood is a function of $2N_0\mu$ or $2N_1\mu$. When r is close to 1, the likelihood is a function of $2N_0\mu = 2N_1\mu$. When r is very large it becomes a function of μt_a and $2N_1\mu$ (an assumption used by Rogers 1995).

It would appear from the parameterization used here that the most suitable summary statistic for detecting population growth or decline is $\log_{10}(r)$. In the case of population decline, the joint distribution of $\log_{10}(r)$ and $\log_{10}(t_f)$ appears to be informative and useful as illustrated in the wombat example. In the case of growth, a statistic that may be useful is the ratio of the posterior odds for growth/decline against the prior odds. This can be regarded as a Bayes factor for testing whether the data support a model of growth or decline. A criticism of Bayes factors is that they are very sensitive to the priors (O'Hagan 1994; Gelman *et al.* 1995), especially when they are quite vague as in the examples here. A sharply peaked distribution of $\log_{10}(2\mu t_a)$ appears indicative of growth. However, unless it is sharply peaked, the shape is likely to depend on the priors chosen, and broad, bimodal distributions are compatible with stable or declining populations. From the posterior distribution of genealogical histories, there are a large number of possible summary statistics that can be monitored. For example, the distal branch index appears to be a useful summary of tree shape.

Taking a fully Bayesian approach, it is probably more sensible to use the natural parameters, N_0, N_1, μ, and t_a directly as described in Tavaré *et al.* (1997) and Wilson and Balding (1998). In this case informative priors are specified for all parameters. Thus, for example, the human data require estimation of t_a. This could be obtained by integrating the likelihoods for $\log_{10}(2\mu t_a)$ (Figure 7) over some prior distribution of mutation rates among loci. However, by using the natural parameters and specifying priors for μ beforehand this could be carried out as part of the MCMC simulation.

In conclusion it would appear that the MCMC approach to modeling genetic data provides useful results for specific data sets and also gives additional insights into the way genealogical history influences gene frequency distributions. It has not been the purpose of this article to make direct comparisons with other methods. It should be noted, however, that the direct modeling of individual mutations in the genealogical history does not appear to produce a substantial overhead in terms of the computer time taken to obtain useful results with data sets of a reasonable size. This is very much the early stages of a revolution in population genetics, and a theoretical and empirical comparative study of the advantages and disadvantages of the currently published methods would be beneficial.

I am grateful to Simon Tavaré for his encouragement and advice on earlier drafts of the manuscript. I thank David Balding, Mike Bruford, Lounès Chikhi, Kevin Dawson, Bill Jordan, and two anonymous referees for their helpful comments.

LITERATURE CITED

Avise, J. C., 1994 *Molecular Markers, Natural History and Evolution*. Chapman & Hall, London.

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza, 1994 *History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.

Cornuet, J. M., and G. Luikart, 1996 Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics **144:** 2001–2014.

Donnelly, P., and S. Tavaré, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 410–421.

Felsenstein, J., 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59:** 139–147.

Felsenstein, J., M. K. Kuhner, J. Yamato and P. Beerli, 1999 Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data, in *Proceedings of the AMS Summer Workshop on Statistics in Molecular Biology*, edited by F. Seillier-Moiseiwitsch. American Mathematical Society, Providence, RI (in press).

Gelman, A., and D. B. Rubin, 1992 Inference from iterative simulation using multiple sequences. Stat. Sci. **7:** 457–472.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 1995 *Bayesian Data Analysis*. Chapman & Hall, London.

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

Griffiths, R. C., and S. Tavaré, 1994b Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B Biol. Sci. **344:** 403–410.

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

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

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

Jorde, L. B., M. J. Bamshad, W. S. Watkins, R. Zenger, A. E. Fraley *et al.*, 1995 Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am. J. Hum. Genet. **57:** 523–538.

Jorde, L. B., A. R. Rogers, W. W. Watkins, P. Krakowiak, S. Sung *et al.*, 1997 Microsatellite diversity and the demographic history of modern humans. Proc. Natl. Acad. Sci. USA **94:** 3100–3103.

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

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

Loader, C. R., 1996 Local likelihood density estimation. Ann. Stat. **24:** 1602–1618.

Lundstrom, R., S. Tavaré and R. H. Ward, 1992 Estimating substitution rates from molecular data using the coalescent. Proc. Natl. Acad. Sci. USA **89:** 5961–5965.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1092.

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics **146:** 1501–1514.

O'Hagan, A., 1994 *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*. Arnold, London.

Raftery, A. E., and S. M. Lewis, 1996 Implementing MCMC, pp. 115–130 in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter. Chapman & Hall, London.

Reich, D. E., and D. B. Goldstein, 1998 Genetic evidence for a palaeolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA **95:** 8119–8123.

Rogers, A. R., 1995 Genetic evidence for a Pleistocene population explosion. Evolution **49:** 608–615.

Rogers, A. R., and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552–569.

Roy, M. S., E. Geffen, D. Smith, E. A. Ostrander and R. K. Wayne, 1994 Patterns of differentiation and hybridization in North American wolflike canids, revealed by analysis of microsatellite loci. Mol. Biol. Evol. **11:** 553–570.

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

Tanner, M. A., 1993 *Tools for Statistical Inference*. Springer-Verlag, New York.

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

Taylor, A. C., W. B. Sherwin and R. K. Wayne, 1994 Genetic variation of microsatellite loci in a bottlenecked species: the northern hairy-nosed wombat *Lasiorhinus krefftii*. Mol. Ecol. **3:** 277–290.

Tierney, L., 1996 Introduction to general state-space Markov chain theory, pp. 59–74 in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter. Chapman & Hall, London.

Weiss, G., and A. von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539–1546.

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

Yang, Z., and B. Rannala, 1997 Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. **14:** 717–724.

Communicating editor: S. Tavaré

APPENDIX A: LIKELIHOODS FOR LINEAR AND EXPONENTIAL MODELS

Below I give the expressions for $\gamma(t)$ and $C(t_i, t_{i+1})$ for the linear and exponential models of population change. These can be substituted into Equation 5 and 6 to calculate the densities and likelihoods used in this article.

Linear case: In this case

$$\lambda(t) = \begin{cases} \dfrac{rt_{\mathrm{f}}}{rt_{\mathrm{f}} + (1 - r)t}, & t \leq t_{\mathrm{f}} \\ r, & t \geq t_{\mathrm{f}} \end{cases};$$

hence

$$\gamma(t) = \begin{cases} \dbinom{n_i}{2}\dfrac{rt_{\mathrm{f}}}{rt_{\mathrm{f}} + (1 - r)t} + \dfrac{n_i\theta}{2}, & t \leq t_{\mathrm{f}} \\ \dbinom{n_i}{2}r + \dfrac{n_i\theta}{2}, & t \geq t_{\mathrm{f}} \end{cases},$$

giving for t_i, $t_{i+1} \leq t_{\mathrm{f}}$

$$C(t_i, t_{i+1}) = \exp\left(-\frac{n_i}{2}\left(\theta(t_{i+1} - t_i) + \frac{(n_i - 1)rt_{\mathrm{f}}}{1 - r}\log\left(\frac{rt_{\mathrm{f}} + t_{i+1}(1 - r)}{rt_{\mathrm{f}} + t_i(1 - r)}\right)\right)\right), \quad (9)$$

for $t_i \leq t_{\mathrm{f}}$ and $t_{i+1} > t_{\mathrm{f}}$

$$C(t_i, t_{i+1}) = \exp(-\frac{n_i}{2}(\theta(t_{\mathrm{f}} - t_i) + \frac{(n_i - 1)rt_{\mathrm{f}}}{1 - r}\log\left(\frac{t_{\mathrm{f}}}{rt_{\mathrm{f}} + t_i(1 - r)}\right) + (\theta + r(n_i - 1))(t_{i+1} - t_{\mathrm{f}}))), \quad (10)$$

and for t_i, $t_{i+1} > t_{\mathrm{f}}$

$$C(t_i, t_{i+1}) = \exp\left(-\frac{n_i}{2}(\theta + r(n_i - 1))(t_{i+1} - t_i)\right). \quad (11)$$

Exponential case: In this case

$$\lambda(t) = \begin{cases} r^{t/t_{\mathrm{f}}}, & t \leq t_{\mathrm{f}} \\ r, & t > t_{\mathrm{f}} \end{cases}$$

and

$$\gamma(t) = \begin{cases} \dbinom{n_i}{2}r^{t/t_{\mathrm{f}}} + \dfrac{n_i\theta}{2}, & t \leq t_{\mathrm{f}} \\ \dbinom{n_i}{2}r + \dfrac{n_i\theta}{2}, & t > t_{\mathrm{f}} \end{cases}$$

giving for t_i, $t_{i+1} \leq t_{\mathrm{f}}$

$$C(t_i, t_{i+1}) = \exp(-\frac{n_i}{2}(\theta(t_{i+1} - t_i) + \frac{(n_i - 1)t_{\mathrm{f}}}{\log(r)}(r^{t_{i+1}/t_{\mathrm{f}}} - r^{t_i/t_{\mathrm{f}}}))), \quad (12)$$

for $t_i \leq t_{\mathrm{f}}$ and $t_{i+1} > t_{\mathrm{f}}$

$$C(t_i, t_{i+1}) = \exp(-\frac{n_i}{2}(\theta(t_{\mathrm{f}} - t_i) + \frac{(n_i - 1)t_{\mathrm{f}}}{\log(r)}(r - r^{t_i/t_{\mathrm{f}}}) + (\theta + (n_i - 1)r)(t_{i+1} - t_{\mathrm{f}}))), \quad (13)$$

and for t_i, $t_{i+1} > t_{\mathrm{f}}$, $C(t_i, t_{i+1})$ is given by (11) above.

Note that for both models, although (9, 10, 12, and 13) → (11) as $r \rightarrow 1$, the equations cannot be evaluated at $r = 1$. Therefore, for computational convenience, if $|r - 1| < 0.0001$ (11) is used with $r = 1$.

APPENDIX B: UPDATING THE GENEALOGY

Addition and deletion of two mutations within lineage: One set of reversible updates that can be made is to add or delete a pair of +1 and −1 mutations in a lineage. A +1 mutation means that the length immediately ancestral to the mutation event is one unit longer. Hereafter, a prime (′) is used to denote variables whose values may differ in candidate genealogies. Within a genealogy there are n_{c} coalescent nodes ($= n_0 - 1$) and $2n_{\mathrm{c}}$ lineages connecting either a sample node or a coalescent to another coalescent node.

In the case of the addition of a pair of mutations, lineage i is chosen with probability $w_i/\Sigma w_j$, where w_i are weights. If the lineages were given equal weight this would correspond to $1/2n_{\mathrm{c}}$. However, better convergence is obtained by weighting the choice of lineages by their squared length, $(\delta t_i)^2$, where δt_i is the difference in time (measured in the same units as for which the likelihoods are calculated) between the ancestral and descendent node. The mutations are added independently and uniformly randomly along the lineage. The joint density of two points along the lineage is therefore $1/(\delta t_i)^2$. This gives

$$P_{\mathrm{f}} = \frac{E_1}{\Sigma E_j \Sigma w_j}.$$

Once the pair of mutations have been added, the lengths of the microsatellite at intervening mutations along the lineage are updated. Thus, for example, if a +1 mutation is added at the bottom of a lineage, +1 is added to all the lengths up to the −1 mutation. The variables describing the candidate genealogy are then updated, and the transition probabilities for the reverse operation, deletion of the mutations, are obtained. In this case a lineage is chosen with probability $1/n'_{\mathrm{d}}$, where n'_{d} is the number of descendent lineages from which at least one pair of +1 and −1 mutations can be deleted in the candidate genealogy. Within a lineage there are n'_{p} distinguishable pairs of +1 and −1 mutations without regard to order [*i.e.*, (the number of +1 mutations in lineage) × (the number of −1 mutations in lineage)]. Thus

$$P_r = \frac{E_2{}'}{\Sigma E_j' n_d' n_p'}.$$

For deletion of two mutations, the procedure is the reverse of that above. The expressions remain the same, except that P_f and P_r and primed and nonprimed variables are interchanged.

Addition and deletion of three mutations around the coalescent node: Three mutations are added/deleted around the coalescent node. With probability 1/2 a +1 mutation is added to the upper lineage, and two −1 mutations are added to the lower lineages, and with probability 1/2 the alternative is carried out. Node *i* is chosen with probability $w_i/\Sigma w_j$, where $w_i = \delta t_{il}\delta t_{ir}\delta t_{iu}$ (the indices refer to left, right, and upper lineages). This gives

$$P_f = \frac{E_3}{\Sigma E_j 2\Sigma w_j}$$

and for the reverse process a node is chosen with probability $1/n_t'$, where n_t' is the number of nodes from which at least one triplet of mutations can be deleted. A particular triplet is deleted with probability $1/n_r'$, where n_r' is the number of distinguishable triplets [(number of +1, upper) × (number of −1, left) × (number of −1, right) + (number of −1, upper) × (number of +1, left) × (number of +1, right)]. This gives

$$P_r = \frac{E_4'}{\Sigma E_j' n_t' n_r'}.$$

As in the previous section, deletion of three mutations is the reverse of the above.

For the MRCA node, the same equations apply, but with variables describing the upper lineage removed. Thus only pairs of either +1 or −1 mutations are added to the two descendent lineages, and the value of the node changes accordingly.

Interchange of lineages: Two lineages can be interchanged, altering the branching structure of the genealogy, by choosing two succeeding events (mutation or coalescent) according to the following criteria: the first event must not be a sample node; the two events must have the same value; the succeeding event must not be the ancestor of the preceding event. The lineage descending from and including the first event is then attached to the ancestor of the second event and similarly for the second event.

Thus a first event is chosen with probability $1/n_l$, where n_l is the number of events that satisfy the criteria given above, giving

$$P_f = \frac{E_5}{\Sigma E_j(i) n_l(i)}$$

and

$$P_r = \frac{E_5'}{\Sigma E_j' n_l'}$$

This is reversible and $E_5 = E_5'$.

Nearest-neighbor interchange of lineages: A subset of interior nodes (those that are not sample nodes) can be chosen to be interchanged. A candidate node, N_1, must be interior. It has two descendent lineages. The node (N_2) at the end of the shorter descendent lineage (L_2) must be interior. The longer branch, L_1, must have no mutations between the time of N_1 and the time of N_2. The value of N_1 and N_2 must be the same. One of the descendent lineages from N_2 is chosen with probability 0.5 and swapped with L_1.

A node is chosen with probability $1/n_n$, where n_n is the number of nodes that satisfy the criteria given above. The two descendent lineages of N_2 have equal probability of being swapped, giving

$$P_f = \frac{E_6}{\Sigma E_j 2n_n}$$

and

$$P_r = \frac{E_6'}{\Sigma E_j' 2n_n'}.$$

This is reversible and $E_6 = E_6'$. Nearest-neighbor interchange of lineages has been used for updating tree topologies in MCMC by Yang and Rannala (1997).

Interchanging order of events: Two temporally adjacent coalescent or mutation events can be interchanged provided the succeeding event is not the ancestor of the preceding event, or, if it is ancestral, both events are mutation events.

The first event is chosen with probability $1/n_e$, where n_e is the number of nodes that satisfy the criteria above:

$$P_f = \frac{E_7}{\Sigma E_j n_e}$$

and

$$P_r = \frac{E_7'}{\Sigma \mathrm{E}_j' n_e'}.$$

This transformation is reversible and $E_7 = E_7'$.

APPENDIX C: STATIONARY DISTRIBUTION

An important consideration is whether the procedures outlined here will yield serially correlated samples from $p(\Phi, G|n_0, S_0)$ as desired. The conditions under which a Metropolis-Hastings simulation will converge to the required density have been well studied (see Tierney 1996). Essentially it is necessary to demonstrate that (1) the required density is proper (has finite volume); (2) the Markov chain is reversible; and (3) the Markov chain is irreducible.

Condition (1) is satisfied by using proper prior distributions, which ensures that the posterior distribution is also proper.

Because Φ has been updated using standard methods it is reasonable to assume that conditions (2) and (3) hold for Φ. The proposal distributions for G are described above, and are demonstrably reversible, satisfying (2). It is necessary to demonstrate that the Markov chain for G is irreducible (*i.e.*, it is necessary to show that starting at any G_i conditional on $\{n_0, S_0\}$ any other G_j can be reached in a finite number of update steps). This is satisfied if the following hold:

1. The interior coalescent nodes (including the MRCA) should be able to take any value.
2. In a single-step mutation model the minimum number of mutations in the lineage between two interior coalescent nodes is given by the absolute difference in length between two nodes. Otherwise there can be an infinite number of pairs of $+1/-1$ mutations.
3. It should be possible for any pair of available lineages to be joined at each coalescent event.

Three of the update classes described earlier are sufficient to ensure that these requirements are satisfied. Condition 1 is satisfied by the addition/deletion of three mutations around a node, which change its value in steps of ±1. Condition 2 is satisfied by the addition/deletion of a pair or $+1$ and -1 mutations within a lineage. Condition 3 is satisfied by the lineage swapping update class: if condition 2 is satisfied it follows that there is a finite probability that any pair of lineages will have temporally adjacent mutation events with the same value allowing them to be swapped, thereby allowing any pair of lineages to be joined. The remaining update classes, nearest neighbor interchange and swapping the order of events, are therefore unnecessary but may improve convergence.