Estimation of Population Parameters and Recombination Rates From Single Nucleotide Polymorphisms

Rasmus Nielsen

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138

Manuscript received July 8, 1999
Accepted for publication October 14, 1999

ABSTRACT

Some general likelihood and Bayesian methods for analyzing single nucleotide polymorphisms (SNPs) are presented. First, an efficient method for estimating demographic parameters from SNPs in linkage equilibrium is derived. The method is applied in the estimation of growth rates of a human population based on 37 SNP loci. It is demonstrated how ascertainment biases, due to biased sampling of loci, can be avoided, at least in some cases, by appropriate conditioning when calculating the likelihood function. Second, a Markov chain Monte Carlo (MCMC) method for analyzing linked SNPs is developed. This method can be used for Bayesian and likelihood inference on linked SNPs. The utility of the method is illustrated by estimating recombination rates in a human data set containing 17 SNPs and 60 individuals. Both methods are based on assumptions of low mutation rates.

SINGLE nucleotide polymorphisms (SNPs) are single base changes in a DNA sequence. In the human genome, such polymorphisms are thought to exist in ~1 out of every 300–500 base positions. Much interest has centered on such genetic markers because of their potential use in gene mapping and in elucidating ancestral human demographic patterns. The recent advent of chip technology gives strength to the idea that human SNP data may soon become abundant. For example, Wang *et al.* (1998) constructed a human genetic map consisting of 2227 SNPs. They also reported the development of genotyping chips that allow simultaneous genotyping of 500 SNPs. However, the great promise of these new markers has not been followed by the development of statistical and population genetical methods for analyzing such data. This article attempts to correct this problem by suggesting new statistical methods for data analysis that take the special properties of SNPs into account.

An important characteristic of SNPs is that they are thought to have very low mutation rates, $\sim 10^{-8}$–10^{-9} in humans. The population genetical parameter $N_e\mu$ ($\mu =$ mutation rate per generation, $N_e =$ effective population size) was estimated as 10^{-4} by Wang *et al.* (1998). This implies that the probability of two mutations occurring in the same locus is very low and consequently, the data are essentially diallelic. Another important property of SNPs is that, per definition, only variable markers are included in a data set. Often little or no information is available regarding the identity of base positions located between the SNPs in a particular population. In some cases the SNPs have originally been identified by sequencing. In such cases it may be advantageous to include information regarding the invariable sites in any statistical analysis. However, in other cases, information regarding invariable sites may not be available or was never obtained. This may occur, for example, if the SNPs were obtained by screening databases for expressed sequence tags (ESTs). In these cases, standard methods for analyzing DNA sequences are not appropriate in the analysis of SNPs. Instead, these types of data must be analyzed by conditioning on each locus being variable.

Two general methods for analyzing SNPs that take these properties into account are developed in this article. The common feature of these approaches is that the sampling probability is calculated conditional on variability in each locus. Because only variable loci are included in the analysis, the mutation rate may in itself be of little interest. The mutation rate is therefore treated as a nuisance parameter and is eliminated by considering the limit of $\mu \rightarrow 0$.

First, a likelihood approach based on markers in linkage equilibrium for use in population genetical and demographic studies is presented. In addition, a likelihood/Bayesian approach to linked SNP markers based on a Markov chain Monte Carlo (MCMC) method is presented. Both approaches are illustrated by applications to real data sets.

SNPs IN LINKAGE EQUILIBRIUM

Considered first are SNPs in linkage equilibrium (*i.e.*, it is assumed that the recombination rate between the markers is so high that they can be treated as independent loci). This assumption is reasonable when the SNPs

Address for correspondence: Department of Organismic and Evolutionary Biology, Harvard University, 288 Biol. Labs., 16 Divinity Ave., Cambridge, MA 02138 E-mail: rnielsen@oeb.harvard.edu

are obtained at random positions in the genome. The data (X) for k loci can then be represented as a collection of k diallelic data patterns, *e.g.*, $X = \{X_1, X_2, \ldots, X_k\} = \{(x_{11}, x_{12}), (x_{21}, x_{22}), \ldots, (x_{k1}, x_{k2})\}$, where the x_{i1}'s and x_{i2}'s are unordered. The fact that all data patterns are diallelic is a consequence of the method used for scoring the data and of the low mutation rates. The likelihood function for a vector of parameters Θ is then given by

$$L(\Theta|X) = \prod_{i=1}^{k} L(\Theta|X_i) \tag{1}$$

under the assumption of linkage equilibrium.

We first consider the case in which the isolation of variable loci and the estimation of population parameters are performed using the same population sample. However, it should be noted that most schemes for obtaining SNPs are more complicated than this and that the definition of the likelihood function depends on the ascertainment scheme. Assuming this simple ascertainment scheme, we can calculate the contribution to the likelihood function from each locus as

$$L(\Theta|X_i) = \Pr(X_i|\Theta, S_i > 0), \tag{2}$$

where S_i is the number of mutations in the ith locus. This conditioning is necessary to take account of the fact that only variable loci are included in the analysis.

It is assumed that mutations occur according to a Poisson process on the edges of an ancestral genealogy with rate $\theta/2$ and that Θ, therefore, can be divided into parameters (Ω) that are independent of the mutation process conditional on the genealogy (such as demographic parameters) and θ. Conditioning on the underlying gene genealogy (G), the sampling probability can be rewritten as

$$\Pr(X_i|\Theta, S_i > 0) = \frac{1}{\Pr(S_i > 0|\Theta)} \int \Pr(X_i|\theta, G)\, dF(G|\Omega). \tag{3}$$

A genealogy consists of $2n - 1$ edges, where n is the sample size. Let the jth edge in the ith genealogy be denoted by b_{ij} and let the length of such an edge be denoted by T_{ij} (Figure 1). The total tree length in the gene genealogy associated with the ith locus (T_i) is given by $T_i = \Sigma_j T_{ij} = \Sigma_{j=2}^{n} j\tau_{ji}$, where τ_{ji} is the time in the genealogy associated with the ith locus in which there exist j genes ancestral to the sample. Let B_i be the set of edges in the genealogy in which a single mutation could have caused data pattern i, if that was the only mutation occurring in the genealogy. For example, for the genealogy depicted in Figure 1, $B_i = \{b_{i3}, b_{i5}\}$. If a mutation happened on edge b_{i3} and no other mutations occurred in the genealogy, there would be three gene copies with the mutant type and two gene copies with the ancestral type. Likewise, if a mutation happened on edge b_{i5} and no other mutations occurred in the genealogy, there would be two gene copies with the mutant type and three gene copies with the ancestral type. In both cases we would observe the data pattern $X_i = \{3, 2\}$. Let t_i be the sum of the length of all edges in the ancestral gene genealogy in which a mutation could have caused the observed configuration (X_i), *i.e.*, $t_i = \Sigma_j T_{ij} I_{(b_{ij} \in B_i)}$. For example, in the genealogy depicted

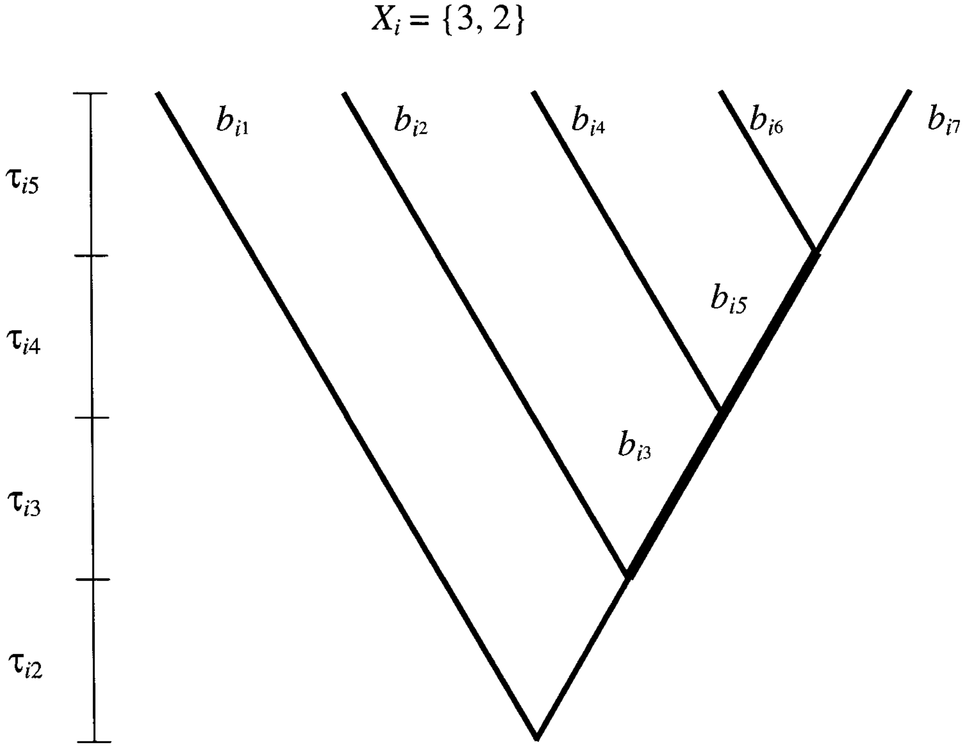

Figure 1.—An example of a coalescence genealogy. The edges of the genealogy, in which a single mutation would have caused the observed data pattern (X_i), are shown in bold.

in Figure 1, the edges in bold are the ones in which a mutation would have caused the observed configuration $\{3, 2\}$ and $t_i = T_{i3} + T_{i5} = \tau_{i4} + \tau_{i3}$. Assuming that mutations occur according to a Poisson process along the edges of the genealogy and assuming that the mutation rates are so low that we can ignore the possibility of back mutation, we realize that $\Pr(X_i|\theta, G) = \Sigma_{j:b_{ij}\in B_i} (1 - e^{-\theta T_{ij}/2})e^{-\theta(T_i - T_{ij})/2}$, the sum over all edges in which a single mutation could cause the observed site pattern, of the probability that at least one mutation happens in that edge multiplied by the probability that no other mutations happen in any of the other edges of the genealogy. Therefore, the sampling probability may be written as

$$\Pr(X_i|\Theta, S_i > 0) = \frac{\int \sum_{j:b_{ij}\in B_i} (1 - e^{-\theta T_{ij}/2})e^{-\theta(T_i - T_{ij})/2} dF(G|\Omega)}{\int (1 - e^{-\theta T_i/2}) dF(G|\Omega)}. \tag{4}$$

We now use the assumption that the mutation rate is low ($\theta \longrightarrow 0$) to eliminate the nuisance parameter θ.

$$\begin{aligned} L(\Omega|X_i) &= \lim_{\theta\to 0} \Pr(X_i\,|\Omega,\theta,S_i > 0) \\ &= \lim_{\theta\to 0} \frac{\int (\theta/2)^{-1} \sum_{j:b_{ij}\in B_i} (1 - e^{-\theta T_{ij}/2})e^{-\theta(T_i - T_{ij})/2}\, dF(G|\Omega)}{\int (\theta/2)^{-1}(1 - e^{-\theta T_i/2}) dF(G|\Omega)} \\ &= \frac{\int t_i dF(G|\Omega)}{\int T_i dF(G|\Omega)} = \frac{\mathrm{E}(t_i\mid\Omega)}{\mathrm{E}(T_i\,|\Omega)}. \end{aligned} \tag{5}$$

The interchange of limit and integral in both denominator and numerator is justified by the assumption that $\mathrm{E}[T_i] < \infty$, an assumption that will be valid for the relevant biological models. A similar result was previously obtained by Griffiths and Tavaré (1998), using arguments based on the infinite-sites model.

Note that the only other assumptions made when deriving Equation 5 are the existence of a well-behaved ancestral genealogy, that the mutational process is a Poisson process along the ancestral genealogy, and the mutation rate is low ($\theta \longrightarrow 0$). The above result is therefore quite general and should be applicable to a wide variety of models. Using Equations 5 and 1 directly, the likelihood function can be evaluated efficiently using analytical methods or simulations for a wide variety of models.

If it is assumed that all gene copies in the population are exchangeable (*e.g.*, a random population sample of neutral genes from a randomly mating population), some further progress can be made. Divide the graph representing the genealogy for the ith locus into $n(n + 1)/2 - 1$ edges, by inserting a node in all edges at the time of a coalescence event. Let the jth edge occurring in the kth coalescence interval be b_{ijk}. Then, because the tree topology is independent of the coalescence times,

$$\begin{aligned} E(t_i|\Omega) &= \sum_{k=2}^{n}\left(E(\tau_{ik}|\Omega)\sum_{j=1}^{k}\Pr(b_{ijk} \in B_i)\right) \\ &= \sum_{k=2}^{n}\left(E(\tau_{ik}|\Omega)k\frac{\binom{x_{i1}-1}{k-2} + (1-\delta_{x_{i1},x_{i2}})\binom{x_{i2}-1}{k-2}}{\binom{x_{i1}+x_{i2}-1}{k-1}}\right), \end{aligned} \tag{6}$$

where δ_{ij} is the Kronecker delta function. The latter expression follows from the fact that all configurations are equally likely when the genes are exchangeable (Kingman 1982). Because $E(T_i|\Omega) = \Sigma_{j=2}^{n} jE(\tau_{ij}|\Omega)$, the likelihood function can be expressed simply in terms of expected coalescence times for any model of exchangeable alleles. These expectations can usually be obtained quite easily analytically or by simulation. For a given data set, the expectations can be evaluated just once, and the sampling probability can thereafter be evaluated for many loci. For the standard neutral coalescence models of a single population of constant size, the expression (Equation 5) reduces to the well-known form of the conditional Ewens sampling formula (Ewens 1972). This is no surprise because the number of alleles is a sufficient statistic for θ in this model.

Estimating growth rates: In the following, the utility of this approach is illustrated by estimating the growth rate of the American Caucasian population for a data set published by Picoult-Newberg *et al.* (1999). They presented a new method for extracting SNPs from publicly available EST databases. They further confirmed the existence of some of these by a method coined genetic bit analysis (GBA) and estimated gene frequencies in the Caucasian-, African-, and Hispanic-American populations. A subset of the data containing 37 polymorphic loci, with an average of 16 haplotypes, from the American Caucasian population was provided by L. Picoult-Newberg and is used here for illustrating the utility of the new method (Equation 6).

The model chosen here to describe population growth is a model of constant exponential growth of a single panmictic population. In this model, r is the exponential growth rate defined by $N(t) = N_0e^{-rt}$, where $N(t)$ is the population size t generations in the past and N_0 is the present population size. Using Equations 5 and 6, we can estimate the growth rate if the expected coalescence times can be evaluated. There exists no simple analytical method for calculating the expected coalescence times in this model, but Slatkin and Hudson (1991) provided a simple method for simulating coalescence times under such a model. Letting t be scaled by $1/r$, the time in which there are i lineages can be generated by

$$\tau_i = \ln\left[1 + \alpha e^{-t}\frac{-2}{i(i-1)}\ln(U)\right], \tag{7}$$

where $\alpha = N_0r$, U is a random deviate drawn from a

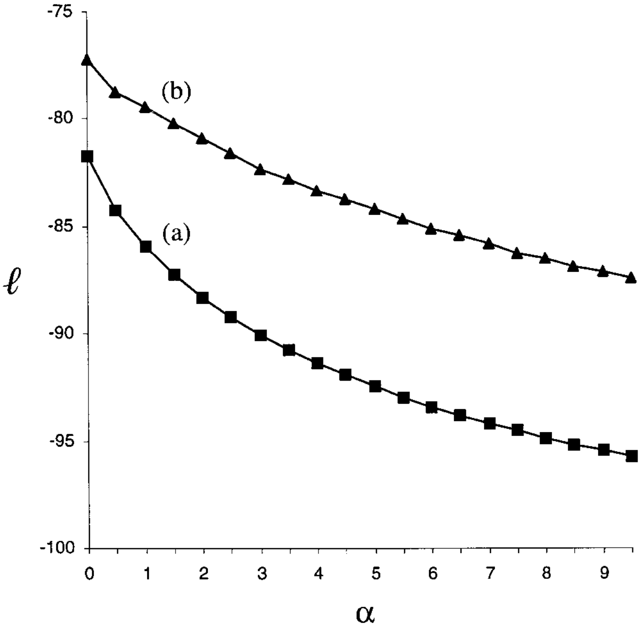

Figure 2.—The log-likelihood function for α conditioned on (a) variability in the sample and (b) variability in the first two sampled gene copies. The data analyzed consist of 37 variable SNP loci published by Picoult-Newberg *et al.* (1999).

uniform (0, 1) density, and t is the time where $i + 1$ genes coalesced into i genes [this corrects a trivial typo in Slatkin and Hudson (1991)]. $E[\tau_i|\alpha]$ can then be estimated by repeated simulations and the likelihood function for α can be evaluated using Equations 5 and 6.

The estimate of the likelihood function on a grid of 20 values of α was obtained by using 100,000 simulations to evaluate $E[\tau_i|\alpha]$ for each gridpoint. This took <1 min on a 450-MHz Pentium II machine; the computational time would not increase significantly as more loci are included in the analyses. The computer program is available from the author upon request.

The results of the analysis are depicted in Figure 2a. Note that the likelihood function is a strictly decreasing function of α, and a maximum-likelihood estimate of $\alpha = 0$ is obtained. There is no evidence in the data for population growth based on SNP loci. This observation contrasts with the pattern found in mitochondrial DNA in which there are strong deviations from the equilibrium model in the direction expected under population growth (*e.g.*, Excoffier 1990). A similar discrepancy between nuclear and mitochondrial data was first described by Hey (1997). It was suggested that the difference could be due to natural selection at the molecular level and/or demographic factors that have not been taken into account, such as population subdivision.

Taking account of ascertainment biases: A possibility that may also be considered for the SNP data is that loci with high frequency alleles have preferentially been chosen. Population growth will lead to an excess of loci with rare alleles. If loci with rare alleles tend to not be included in the sample, much of the evidence for population growth may be lost. This might occur if loci originally were chosen because variability was detected between only two or a few copies. For example, the loci extracted by Picoult-Newberg *et al.* (1999) were identified initially by the screening of published ESTs. This implies that variability was first detected by comparing only a few gene copies. A simple way of taking this screening procedure into account is by conditioning on variability in the first analyzed ESTs (a subset of the sample). The protocols used for isolating SNPs may vary and most protocols may be more complex than this; however, conditioning on variability in the first analyzed ESTs provides for a mathematically tractable way of correcting for the biases arising from preferential selection of loci with alleles of intermediate frequency. Considering the extreme case of only two ESTs, we can calculate the likelihood function as Pr(X| variability in the first two copies sampled) = Pr(variability in the first two copies sampled |X) Pr(X)/Pr(variability in the first two copies sampled). Noting that Pr(variability in the first two copies sampled |X) = $2(x_{i1}x_{i2})/(n(n-1))$ and using the same arguments as in the derivation of Equations 3–5, we find that this likelihood function can be expressed as

$$L_2(\Omega|X_i) := \frac{x_{i1}x_{i2}E(t_{iG}|\Omega)}{n(n-1)E(\tau_2|\Omega)}, \tag{8}$$

where $E(\tau_2)$ is the expected coalescence time in a sample size of two.

The likelihood function for α was recalculated using Equation 8. Note that again, a strictly decreasing likelihood surface is obtained, although the likelihood surface is not quite as steep as before (Figure 2b). This suggests that the apparent pattern of no population growth is not an artifact but may reflect a real biological property of the data. Presumably there are some biological factors that the model does not take into account such as population subdivision or selection.

Because the likelihood function can be written as a product of the likelihood in independent loci (Equation 1), the usual large sample approximations from statistical theory should be applicable as the number of loci becomes large. For example, by inspection of the likelihood function depicted in Figure 2, we can obtain an ~95% upper bound for α of ~{α : α < 1.0} using $L_2(\alpha|X_i)$.

SNPs IN LINKAGE DISEQUILIBRIUM

The analysis of SNPs in linkage disequilibrium is in many ways much more complicated because the sampling probability cannot be expressed as a simple product of the marginal sampling probability of each locus. However, linked loci are in many ways more interesting data than independent loci. They may contain more information about the parameters of interest and they may be used for linkage disequilibrium mapping. Recently, several new methods have emerged for analyzing population samples of linked loci. The approach by Griffiths and Marjoram (1996), based on the infinite-sites model, is a derivative of the general Monte Carlo recursion methods of Griffiths and Tavaré (1994a,b). The method of Kuhner (1999) is based on MCMC. In the following, we present a method applicable to SNPs similar to the Kuhner (1999) method. The two methods are similar in that they are both based on Metropolis-Hastings (Metropolis *et al.* 1953; Hastings 1970) MCMC, but they differ on several important points. For example, our method uses a Bayesian approach to the problem of parameter estimation, whereas the method of Kuhner uses importance sampling to estimate the likelihood surface for the relevant parameters(s). Also, calculations of sampling probabilities conditional on an ancestral graph are greatly simplified under the model of SNP evolution considered here. The present method should therefore be much faster than the method of Kuhner (1999).

The ancestral recombination graph: To describe the genealogical process governing the evolution of the SNPs, we use the familiar coalescence process with recombination (*e.g.*, Hudson 1983; Griffiths and Marjoram 1996). We make the standard assumptions associated with the coalescence process of a single panmictic population of constant size. The entire ancestral process is described by an ancestral graph (A) and a set of marginal genealogies. A contains information regarding the ancestral linkage of the different genes so the marginal genealogies can be deduced from A, whereas A cannot be deduced from the marginal genealogies. A is generated by the following stochastic process: at time zero, there exist n edges in the ancestral graph. Each edge contains genetic material from the k loci. Let the distances between the k loci, in number of base pairs, be described by a vector $\mathbf{d} = (d_1, d_2, \ldots, d_{k-1})$ and the per base pair rate of recombination be $R = \rho/(2N)$. Then, looking back in time, each edge initially recombines at rate $\rho\Sigma_{i=1}^{k-1}d_i$ when time is scaled in units of $1/(2N_e)$. If an edge recombines, a breakpoint δ is chosen uniformly in the interval $(0, \Sigma_{i=1}^{k-1}d_i)$ and two new edges are formed, containing the ancestral genetic material from the original edge in the interval $(0, \delta)$ and $(\delta, \Sigma_{i=1}^{k-1}d_i)$, respectively. In general, if the distance between the two most distant ancestral sites in edge j is denoted by D_j, edge j will recombine at rate ρD_j.

Each pair of edges also coalesce with each other at rate 1 so the total rate of coalescence events is $j(j-1)/2$ when there are j active edges in the ancestral graph. When two edges coalesce, the new edge contains the genetic material from both daughter edges. For example, if two edges containing sites (0, 1, 2, 3, 4) and (2, 6, 7) coalesced, the resulting edge would contain the ancestral genetic material of sites (0, 1, 2, 3, 4, 6, 7). The stochastic process describing the number of edges in the ancestral graph is therefore given by a birth-and-death process in which deaths occur at rate $j(j-1)/2$ and births occur at rate $\rho\Sigma_{i=1}^{j}D_i$. The process stops when a common ancestor is reached, *i.e.*, when only one edge containing ancestral genetic material is left.

Data from linked SNP loci can be represented as a set of ordered site patterns X and the associated vector of distances between sites $\mathbf{d}$. For example, a data set consisting of three SNPs from four individuals could be represented as

$$X = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix},$$

where the two allelic types in an SNP are represented as 1's and 0's, respectively. This representation of the data is similar to the representation used for sequences under the infinite-sites model. However, the models differ because in the infinite-sites model, the number of variable loci is considered a random variable. Here we condition on the number of variable loci and consider the limit of $\mu \rightarrow 0$. The likelihood function can then easily be derived using a multilocus extension of Equation 5. Using the exact same arguments as in the derivation of Equation 5, we obtain

$$L(\Omega|X) = \lim_{\theta_i \to 0} \Pr(X|\Omega,\mathbf{d},\theta_i,S_i > 0,\ i = 1\ldots k)$$
$$= \frac{E\left(\Pi_{i=1}^{k}t_i|\Omega\right)}{E\left(\Pi_{i=1}^{k}T_i|\Omega\right)}, \tag{9}$$

where now T_i refers to the total tree length of the ith marginal genealogy and t_i is the sum of the length of edges in the ith marginal genealogy in which one mutation could have caused the ordered site pattern i. Again, in the derivation we must assume $E(\Pi_{i=1}^{k}T_i|\Omega) < \infty$ to justify the interchange of limit and integral. Although this condition may be difficult to prove, we conjecture that it is true in the case of the standard neutral coalescence process with recombination, because

$$E\left(\prod_{i=1}^{k}T_i|\rho = 0\right) = k!\sum_{i=2}^{n}\frac{(n-1)!(2/(i-1))^{k+1}}{2\Pi_{j=2}^{i-1}(j-i)\ \Pi_{j=i+1}^{n}(j-i)}$$

(appendix) for this model and $E(\Pi_{i=1}^{k}T_i|\rho)$ appears to be a strictly decreasing function of ρ.

The above representation assumes that the map distances of the markers ($\mathbf{d}$) are known. This will usually be the case for SNPs because of genomic sequencing.

If the genealogy is not consistent with the observed site pattern, $t_i = 0$. For most data sets, under any reasonable genealogical model, the vast majority of all possible ancestral graphs will contain at least one marginal site genealogy that is not consistent with the observed site pattern. $E(\Pi_{i=1}^{k}t_i|\Omega)$, therefore, cannot be efficiently evaluated by simple simulations of the prior distribution as was the case for SNPs in linkage equilibrium. In contrast, $E(\Pi_{i=1}^{k}T_i|\Omega)$, does not depend on the data and it can be evaluated relatively easily by simulation. In the following, a MCMC method to estimate $L(\Omega|X)$ in this model is devised. This method allows Bayesian or likelihood estimation of the relevant parameters regarding both the genealogical and the mutational process. We illustrate the method in terms of Bayesian estimation, but the method could be used as well in a likelihood framework. Our main motivation for choosing a Bayesian approach is that the large sample approximations usually applied in likelihood analysis may not be justified for linked loci. Adopting a Bayesian view may therefore simplify the interpretation of the results.

A MCMC method: In the following, a MCMC method based on Metropolis-Hastings sampling (Metropolis *et al.* 1953; Hastings 1970) for approximating $f(\Omega|X)$ is described. Previous application of Metropolis-Hastings sampling in population genetics that the reader may be familiar with include the methods by Kuhner *et al.* (1995), Wilson and Balding (1998), and Beerli and Felsenstein (1999).

First, note that the posterior density, being proportional to the product of the prior times the likelihood function, can be written as

$$f(\Omega|X) = \frac{cf(\Omega)}{E(\Pi_{i=1}^{k}T_i|\Omega))}\int\prod_{i=1}^{k}t_i dF(A|\Omega), \qquad (10)$$

where c is an unknown constant. This representation suggests the following method for estimating $f(\Omega|X)$. The first step is to evaluate $E(\Pi_{i=1}^{k}T_i|\Omega)$, which does not depend on the data, directly by simulation (see below). We then run a Markov chain on (A, Ω) and use the Metropolis-Hastings method to ensure that the chain has stationary distribution proportional to

$$h(\Omega, A) = \frac{f(A|\Omega)f(\Omega)\ \Pi_{i=1}^{k}t_i}{E(\Pi_{i=1}^{k}T_i|\Omega)}.$$

By sampling values of Ω from this chain at equilibrium, we can approximate $f(\Omega|X)$. If the current state of the chain is (Ω_0, A_0) an update to another state (Ω_1, A_1) is proposed according to the proposal density $q[(\Omega_0, A_0), (\Omega_1, A_1)]$. As is usual in Metropolis-Hastings sampling, a proposed update to the current state is accepted with probability

$$\alpha[(\Omega_0, A_0), (\Omega_1, A_1)] = \min\{w_{01}, 1\},$$

$$w_{01} = \frac{h(\Omega_1, A_1)q[(\Omega_1, A_1),(\Omega_0, A_0)]}{h(\Omega_0, A_0)q[(\Omega_0, A_0),(\Omega_1, A_1)]}.$$

Under general conditions, such as the existence of a unique stationary distribution, this chain will converge if the proposal density is constructed such that all states of the chain eventually can be reached from all other possible states (Ripley 1987). An implementation of this method is described in the appendix.

Evaluation of the method: Using the Markov chain described in the appendix, the posterior distribution of parameters of interest can be evaluated. In the following, the method is evaluated in terms of its properties as a Bayesian estimator of ρ, but many other applications of the method are possible. For example, it is obvious to use the method for linkage disequilibrium mapping, although this application is not pursued in this article.

We assume a uniform prior distribution of ρ. The posterior distribution is therefore proportional to the likelihood function and the results can be directly interpreted in a likelihood framework in addition to a Bayesian framework.

To evaluate the MCMC method, multiple independent runs of the Markov chain were performed for the simulated data set discussed in the appendix, containing 50 chromosomes and nine SNPs. In these runs, initial ancestral graphs were generated by simulating marginal genealogies for each site separately, conditional on the genealogies to the 5′ end of the site. The simulation algorithm would start with the site closest to the 5′ end and stop when the 3′ end was reached. If the genealogy generated for a particular site is not consistent with the site pattern in that site, the genealogy is abandoned and a new genealogy is simulated. This algorithm thereby runs along the sequence, generating a random ancestral graph consistent with the data. In some cases, the algorithm may take a very long time to find a marginal genealogy consistent with the data. In such cases, recombination and coalescence events are forced on the genealogy, guaranteeing that an appropriate genealogy will be found. This approach for obtaining an initial ancestral graph was chosen to minimize correlation between independent runs.

$E(\Pi_{i=1}^{k}T_i|\rho)$ was estimated independently in each run on a grid containing only two points, each based on 100,000 simulations. Each run of the Markov chain consisted of 45% proposed changes of type 1, 5% of type 2, 45% of type 3, and 5% of type 4 (see the appendix). This mixture appeared to provide a reasonable rate of convergence upon inspection of individual chains. Each run consisted of 1,000,000 steps in the chain and a burn-intime of 200,000 steps was chosen. The entire estimation procedure took <10 min on a 450 MHz Pentium II machine.

The first property of the method examined here is the degree of autocorrelation in the likelihood along the chain. The likelihood averaged over 1000 steps for four different runs is plotted in Figure 3. Note that there appears to be little long-range autocorrelation in the likelihood along the Markov chain. This is a good sign and may indicate that the Markov chain converges relatively fast. However, there appear to be some trends in the likelihood over tens of thousands of replicates.

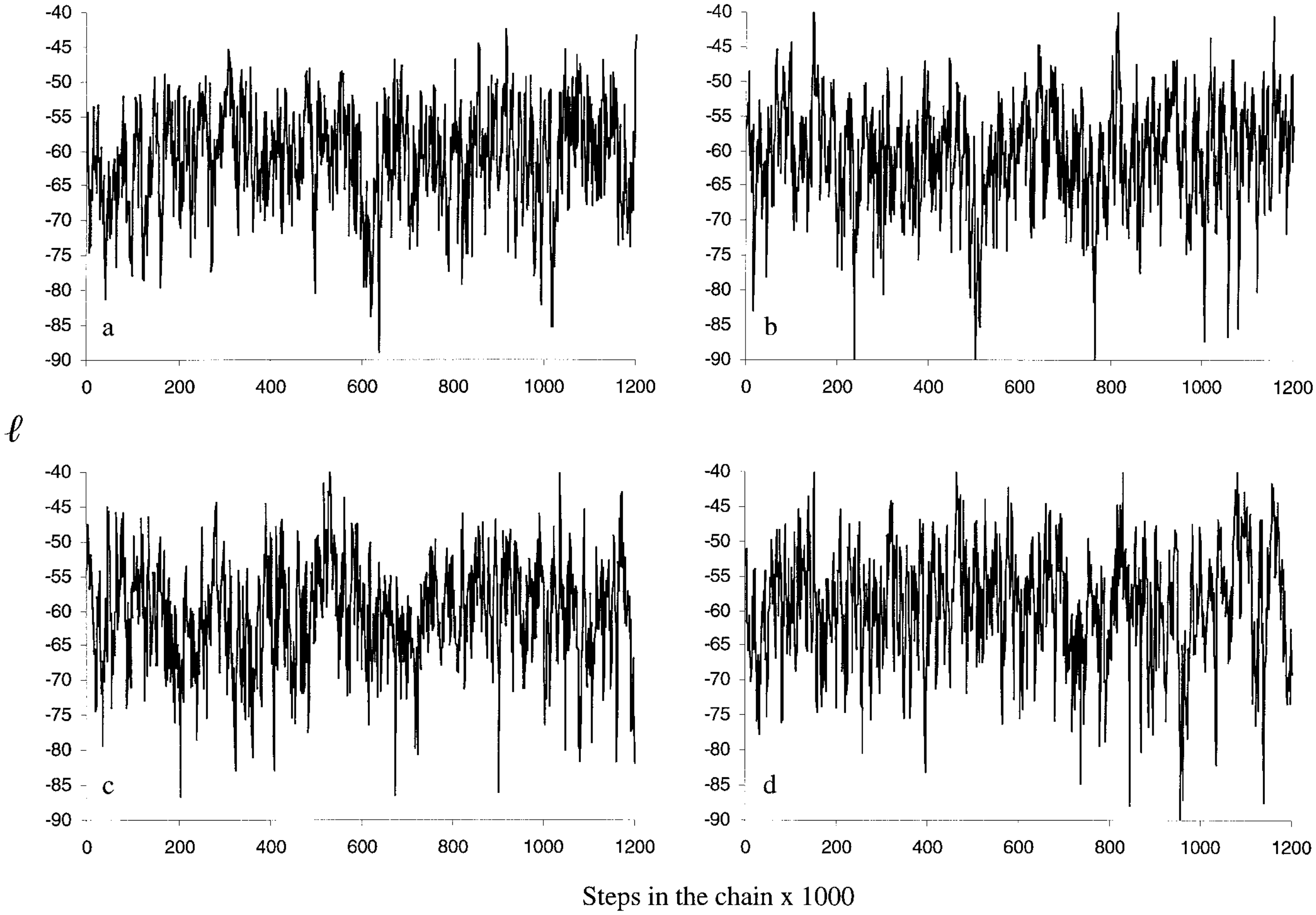

Figure 3.—The log-likelihood as a function of the number of steps in the Markov chain for four independent runs of the chain, based on simulated data containing 50 chromosomes and nine SNPs. The points are averages over 1000 steps in the chain.

This suggests that millions and not thousands of steps in the Markov chain are required for convergence.

The posterior distributions for ρ, obtained from the same four independent runs, are depicted in Figure 4. The posterior distributions obtained in these four runs are almost identical, suggesting that the chain does in fact converge in 1,000,000 steps. Gelman and Rubin's (1992) convergence statistic was calculated for ρ using CODA (Best *et al.* 1995). The 50 and 97.5% quantile of the sampling distribution of the shrink factor were 1.01 and 1.03, respectively, suggesting that convergence may have been achieved (see Gelman and Rubin 1992). Some runs involving 100,000 steps in the chain were also performed (not shown). The posterior distribution could vary significantly among such runs, again suggesting that a large number of steps in the chain (*i.e.*, millions, not thousands) are necessary.

Combining the distributions from the four runs gives an estimate of $\rho = 0.0019$, using the mode of the posterior distribution as an estimator, corresponding to the maximum-likelihood estimator. Alternatively, the mean of the posterior distribution could be used as a point estimator of ρ. Griffiths and Marjoram (1996) obtained maximum-likelihood estimates of approximately $\rho = 0.0015$ and $\rho = 0.002$ in two different runs for this simulated data set. It appears that there is good agreement between the estimates obtained using the present method and the estimates obtained using the method of Griffiths and Marjoram (1996), despite the differences in the models used to analyze the data. Griffiths and Marjoram (1996) assume that the number of variable loci is a random variable and they estimate $N_e\mu$ simultaneously with ρ.

Data analysis: To illustrate the utility of the method, we analyze a data set published by Fullerton *et al.* (1994) of 60 human DNA sequences of length 3007 bp containing 17 SNPs. The SNPs are spaced at distances of {157, 10, 15, 59, 129, 24, 374, 452, 58, 7, 585, 546, 80, 2, 156, 153} bp. This data set was previously analyzed as part of an illustration of the method of Hey and Wakeley (1997) for estimating recombination rates from DNA sequence data. The aligned sequences were provided by J. Wakeley. To analyze the data, two independent runs were performed. In each run, 500,000 simulations were performed for each of two gridpoints in the estimation of $E(\Pi_{i=1}^{k} T_i|\rho)$. A burn-in time of

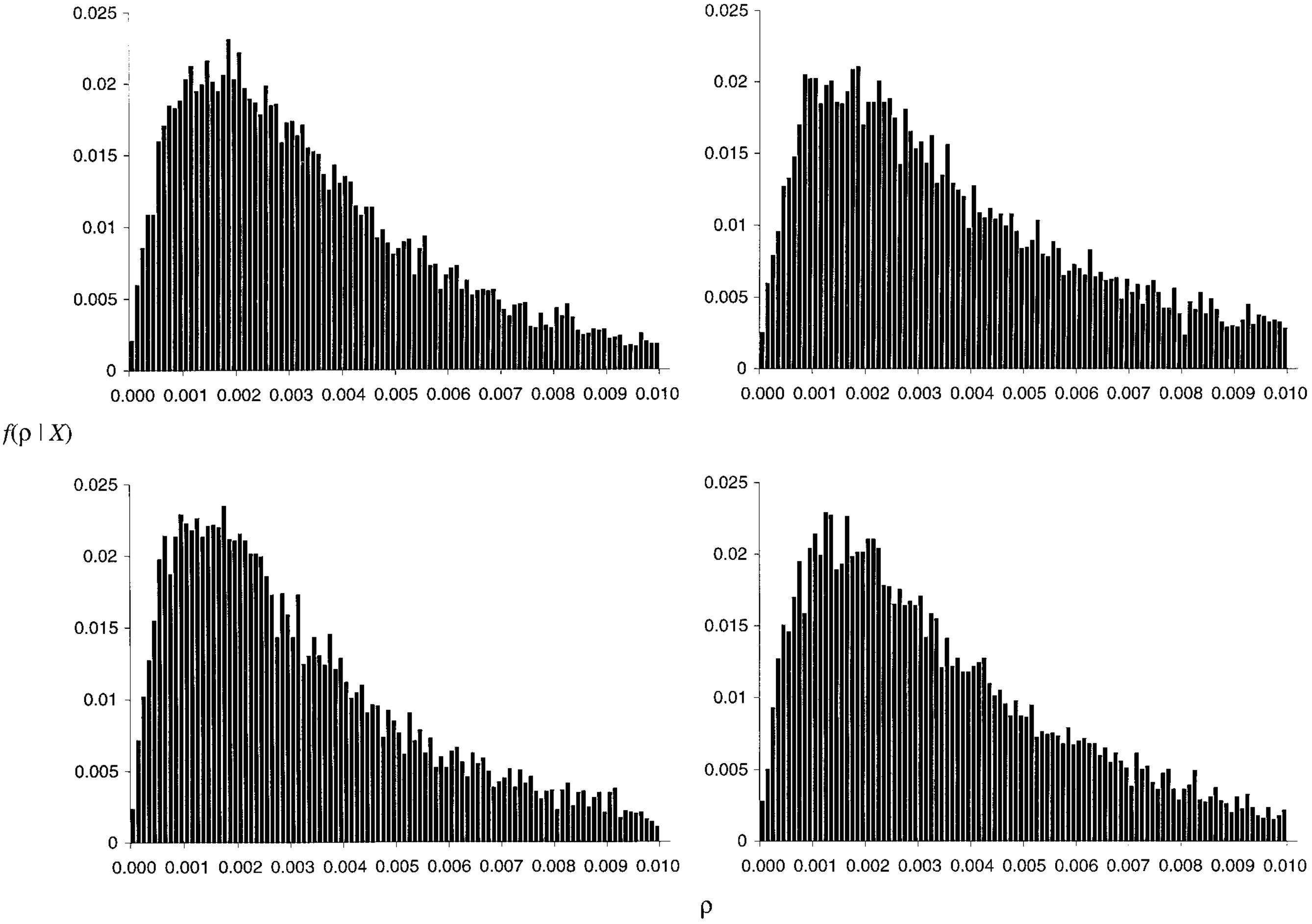

Figure 4.—The discrete approximation to the posterior distribution of ρ obtained in the four independent runs of the Markov chain shown in Figure 3.

500,000 steps of the chain was chosen and 10,000,000 steps were thereafter performed to evaluate the posterior distribution of ρ. The remaining parameters are the same as in the example described above. The entire estimation procedure took ~2 hr.

The posterior distribution of ρ for these data is depicted in Figure 5. An estimate of $\rho = 0.0009$ was obtained using the mode of the posterior distribution as the estimator, corresponding to the maximum-likelihood estimate. An ~95% Bayesian credibility interval is obtained as $C_r(\rho) = \{\rho : 0.0004 < \rho < 0.0023\}$. Hey and Wakeley (1997) obtained an estimate of $\rho = 0.00085$ using an estimator based on multiple subsets consisting of four sequences. The high correspondence between the maximum-likelihood estimate and the estimate obtained by Hey and Wakeley (1997) may indicate that the latter successfully approximates the maximum-likelihood method.

DISCUSSION

SNP loci in linkage equilibrium can be analyzed under reasonable assumptions regarding the sampling process used when typing such loci. The fact that most of the currently available SNP loci are not initially discovered by analyzing large random samples should not discourage population geneticists from using such loci in the analysis of demographic or evolutionary models. In this article, some likelihood methods for analyzing SNP loci in linkage equilibrium were developed that take account of the special methods used in the initial identification of SNP loci. These methods allow fast and

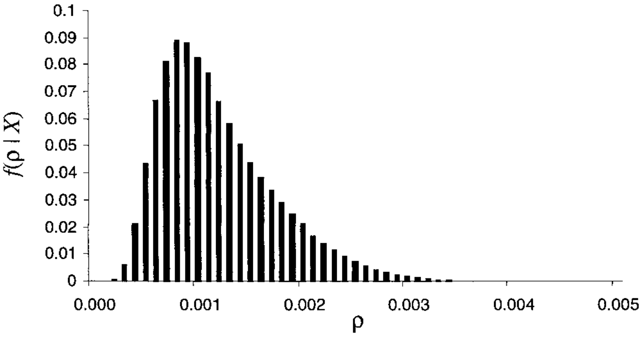

Figure 5.—The discrete approximation to the posterior distribution of ρ for a data set containing 60 DNA sequences and 17 polymorphic sites published by Fullerton *et al.* (1994).

efficient analyses of even very large data sets. Given that several thousand humans SNPs have already been identified, methods such as the one described here should be useful for elucidating the evolution and diversification of human populations.

However, the assumptions regarding the ascertainment schemes were somewhat simplified in this study. In many cases, some initial sorting of the SNP loci is done. In other cases, the SNP loci are initially identified in one population, and subsequently, population samples are obtained from another population. In such cases, correct statistical inference would require the modeling of this complex isolation protocol if the loci are to be used in the estimation of population parameters. This in return requires that the exact protocols used when isolating SNPs are made publicly available. If such information is not available, or if the resulting models are mathematically intractable, it may be necessary to settle for simpler models such as those discussed in this article.

In this analysis it was found that there was no evidence for population growth in a data set containing 37 human SNPs. This result is in accordance with previous observations based on nuclear sequence data (Hey 1997) but is obviously in stark contrast to the large amounts of direct demographic data showing strong population growth in human populations the last 10,000–100,000 years. Several explanations for this discrepancy can be given. Balancing selection is an obvious explanation, although this explanation would require that most randomly selected loci are under strong selection, an assumption that most population geneticists would be unwilling to accept. The explanation for the apparent lack of evidence for population growth is most likely that the assumed demographic model does not take population subdivision into account. One could imagine several demographic scenarios in which any evidence for population growth would be offset by the effects of population subdivision (Wakeley 1999). Other factors that may be of importance in explaining the discrepancy between nuclear and mitochondrial DNA are the difference in effective population size between the two types of markers, selection in the mtDNA, and the fact that analyses based on mtDNA are based on a single random realization of a stochastic process.

Linked SNPs can be analyzed using MCMC. It was demonstrated that such an analysis is feasible for realistic-sized data sets. Because of the simplicity of the mutational model, millions of steps in the Markov chain can be performed. It appears that this many steps are necessary to ensure convergence of the chain. The main limitation of the method is that it will become very slow as the recombination rate increases. The reason for this is that the number of edges in the ancestral graph grows quite rapidly when the recombination rate increases. Therefore, it does not seem possible to develop a full likelihood/Bayesian approach applicable to large genomic regions.

The method can be improved in several ways from its current form. For example, the entire ancestral graph is represented in the computer memory in the current implementation. Computational time could be saved by storing only the part of the ancestral graph required for calculation of the likelihood. Also, considerable computational time is spent estimating the function $E(\Pi_{i=1}^{k} T_i|\rho)$ by simulation. Analytical results facilitating a numerical evaluation of this function could therefore greatly reduce the computational time.

However, even in its current implementation, the method allows relatively fast likelihood and Bayesian inference on linked SNPs. A Bayesian approach to the problem of estimation was chosen here. One of the reasons for this choice is that the large sample approximations usually applied in the likelihood framework may not be applicable in the case of a single population sample. However, more theoretical work is needed to examine this problem in the context of moderate recombination.

The posterior density was approximated by sampling values of ρ from a Markov chain at stationarity. An alternative method is used by Kuhner *et al.* (1995). They use importance sampling to evaluate the likelihood function for multiple values of the relevant parameter on a grid (see Kuhner *et al.* 1995 for details). A Markov chain is run similarly to the present case, using a single fixed value of the parameter, say Θ_0. The likelihood function for the parameter (Θ) is then evaluated for multiple values of Θ, using importance sampling.

A similar approach was also implemented for the current method. The Markov chain was run using a single value of ρ (ρ_0) and the likelihood was evaluated using importance sampling for multiple values of ρ. However, it was found that the Monte Carlo variance was very large for values of ρ just slightly larger or smaller than ρ_0. Some reasons why a large Monte Carlo variance may be expected are provided by Stephens (1999). This method was therefore abandoned. The method used by Kuhner *et al.* (1995) involves running multiple chains to find the mode of the likelihood function, which may alleviate some of the problems encountered in the current case, at least in the context of point estimation.

I thank L. Picoult-Newberg and J. Wakeley for providing data analyzed in this publication. This project benefited greatly from advice and discussion from M. Slatkin, J. Wakeley, and J. P. Huelsenbeck and from comments from the two anonymous reviewers and the associate editor S. Tavaré. This study is supported by a fellowship to the author from the Danish Research Council and National Science Foundation grant 9815367 to J. Wakeley.

LITERATURE CITED

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

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

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

Excoffier, L., 1990 Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations in equilibrium. J. Mol. Evol. **30:** 125–139.

Fullerton, S. M., R. M. Harding, A. J. Boyce and J. B. Clegg, 1994 Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus. Proc. Natl. Acad. Sci. USA **91:** 1805–1809.

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

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

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

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

Griffiths, R. C., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Stoch. Mod. **14:** 271–295.

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

Hey, J., 1997 Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol. Biol. Evol. **14:** 166–172.

Hey, J., and J. Wakeley, 1997 A coalescent estimator of the population recombination rate. Genetics **145:** 833–846.

Hudson, R. R., 1983 Properties of the neutral allele model with intergenic recombination. Theor. Popul. Biol. **23:** 183–201.

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

Kuhner, M., 1999 Recombine. Computer program available from http://evolution.genetics.washington.edu/lamarc/recombine.html.

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

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:** 1087–1091.

Picoult-Newberg, L., T. E. Ideker, M. G. Pohl, S. L. Taylor, M. A. Donaldson *et al.*, 1999 Mining SNPs from EST databases. Genome Res. **9:** 167–174.

Ripley, B., 1987 Stochastic simulation. Wiley, New York.

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

Stephens, M., 1999 Problems with computational methods in population genetics. Contribution to the 52nd session of the International Statistical Institute, August 1999. Available from http://www.stats.ox.ac.uk/~stephens/group/publications.html.

Wakeley, J., 1999 Non-equilibrium migration in human evolution. Genetics **153:** 1863–1871.

Wang, D. G., J. B. Fan, C. J. Siao, A. Berno, P. Young *et al.*, 1998 Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science **280:** 1077–1082.

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

Communicating editor: S. Tavaré

APPENDIX

This appendix describes the details of the MCMC method used to evaluate $f(\rho|X)$. In this discussion, "up" in the ancestral graph implies closer to the present and "down" means further back in the past. An edge is connected up to one "daughter" edge if it "originated" in a recombination event or it is connected up to two daughter edges if it originated in a coalescence event. Likewise, an edge is connected down to one "parental" edge if it "ends" in a coalescence event or down to two parental edges if it ends in a recombination event.

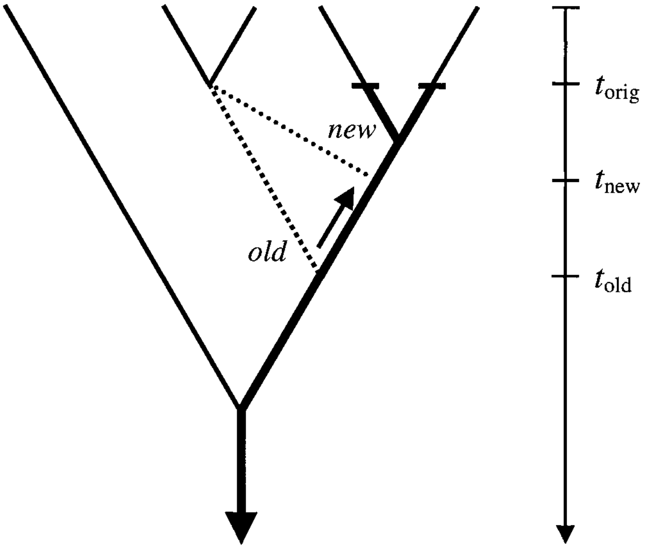

Figure A1.—An illustration of the method used to propose changes of coalescence events in the ancestral graph. The part of the genealogy in bold is the part to which the end of the edge can move in a single update.

It is assumed that the only parameter of interest in Ω is ρ and that the prior distribution of this parameter is uniformly distributed. The neutral equilibrium model is adopted as the prior distribution of *A*, facilitating fast computation of $f(A|\rho)$. Four different types of updates to *A* and ρ are proposed: (1) moving a coalescence event, (2) moving a recombination event, (3) adding or removing a recombination event, and (4) updating ρ. The proposal distribution of the Markov chain consists of a mixture of these four types of changes.

Moving a coalescence event: The first type of update to *A* proposed is the moving of a coalescent according to the following scheme: an edge ending in a coalescence event is chosen uniformly among all edges in the ancestral graph ending in a coalescence event. The end of the edge is moved randomly to a new time t_{new} while the origination of the edge does not move. Denoting the time of the original end of the edge by t_{old}, we let the time $\Delta t = t_{\text{old}} - t_{\text{new}}$ be normally distributed with mean 0 and variance σ^2 (Figure A1). In the cases described in this article, a value of $\sigma^2 = 0.5$ was chosen. If t_{new} is less than the time of the origination of the edge (t_{orig}), we set $t_{\text{new}} = 2t_{\text{orig}} - (\Delta t + t_{\text{old}})$, thereby reflecting t_{new} around t_{orig}. This ensures reversibility of the chain. The edge is moved by sliding it up or down in the ancestral graph (Figure A1). If $t_{\text{new}} < t_{\text{old}}$, the end of the edge is moved upward in the graph. When a coalescence event is encountered, the edge will follow each of the two daughter edges with probability 0.5. Likewise, if $t_{\text{new}} > t_{\text{old}}$, the end of the edge is moved downward in the graph. When a recombination event is encountered, each of the two parental edges in the ancestral graph is followed with probability 0.5. After moving the edge, all other edges in the genealogy are updated accordingly. This algorithm for proposing changes to the ancestral graph was chosen because it has the desirable consequence that the probability that an edge will be involved in a change in the topology of the graph depends on the length of the edge. Presumably, short edges tend to be edges that are less supported by the data. The algorithm should therefore tend preferentially to change the topology of the graph in regions where edges are poorly supported by the data.

Weighting: If this type of change changes the ancestral graph from A^0 to A^1 and t_i^0 to t_i^1, $i = 1, 2, \ldots,$ k, then the weight associated with such a change is

$$w_{01} = \frac{\Pi_{i=1}^{k} t_i^1 f(A^1|\rho)}{\Pi_{i=1}^{k} t_i^0 f(A^0|\rho)} 2^{(\beta-\gamma)}$$

if the edge was moved upward in the genealogy and

$$w_{01} = \frac{\Pi_{i=1}^{k} t_i^1 f(A^1|\rho)}{\Pi_{i=1}^{k} t_i^0 f(A^0|\rho)} 2^{(\gamma-\beta)}$$

if the edge was moved downward in the genealogy. β is the number of recombination events and γ is the number of coalescence events encountered while moving the edge.

Moving a recombination event: An existing recombination event may be moved. In that case, an edge originating in a recombination event is chosen uniformly among all edges originating in a recombination event. The time of the new recombination event is bounded upward by the time of the origination of the daughter edge. It is bounded downward by the minimum of the time of the end of the edge and the time of the end of the other daughter edge of the parental edge. The time of the new recombination event is chosen uniformly in this interval.

Weighting: If this type of change alters the ancestral graph from A^0 to A^1 and t_i^0 to t_i^1, $i = 1, 2, \ldots, k$, then the weight associated with such a change is

$$w_{01} = \frac{\Pi_{i=1}^{k} t_i^1 f(A^1|\rho)}{\Pi_{i=1}^{k} t_i^0 f(A^0|\rho)}.$$

Adding and removing a recombination event: Recombination events are added to the chain with probability 0.5 by choosing an edge uniformly among all edges. A recombination event occurs on this edge at a time uniformly chosen along the length of the edge, and the breakpoint δ is chosen uniformly in the interval between the two most distant sites in the edge. The recombination event results in two new edges: one edge following the path of the original edge and a new edge. With probability 0.5, the new edge will contain the ancestral genetic material of the original edge in the region (0, δ) and with probability 0.5 the new edge will contain the ancestral genetic material of the original edge in sites numbered larger than δ. The new edge is chosen to coalesce with another edge uniformly chosen among all edges. The time of coalescence is chosen uniformly along the length of the new edge.

Elimination of recombination events is proposed with probability 0.5 by choosing an edge to be eliminated uniformly among all edges in the ancestral graph. After adding or removing a recombination event, all other edges in the graph are updated accordingly. However, no additional recombination events are allowed.

Weighting: When adding a recombination event, it may easily occur that the receiving edge ends at a time before the recombination event. In such cases, the recombination event is not possible and the proposed change is given weight 0. Also, if adding the recombination event eliminates any other edges in the graph, the change is given weight 0. Elimination of an edge occurs when the edge contains no SNP sites. In all other cases the weight associated with adding a recombination event, changing the ancestral graph from state A^0 to state A^1, is given by

$$w_{01} = \frac{\Pi_{i=1}^{k} t_i^1 f(A^1|\rho)(j+3) t_{\mathrm{don}} t_{\mathrm{rec}} D}{\Pi_{i=1}^{k} t_i^0 f(A^0|\rho) j^2},$$

where t_{don} is the length of the donating edge in which the recombination event occurs, t_{rec} is the length of the receiving edge in which the new edge ends, j is the number of edges in the genealogy, and D is the distance between the two most distant ancestral SNP sites in lineage j. The factor of $j^2/(j+3)$ arises because adding a recombination event introduces three new edges in the genealogy.

The weight associated with removing a recombination event is 0 if the chosen edge does not originate as a recombination event or if removing the edge eliminates another edge in the graph. Otherwise, the weight associated with this type of change is

$$w_{01} = \frac{\Pi_{i=1}^{k} t_i^1 f(A^1|\rho)(j-3)^2}{\Pi_{i=1}^{k} t_i^0 f(A^0|\rho) j t_{\mathrm{don}} t_{\mathrm{rec}} D},$$

where j is the number of edges in the graph before the recombination has been removed and t_{don}, t_{rec}, and D

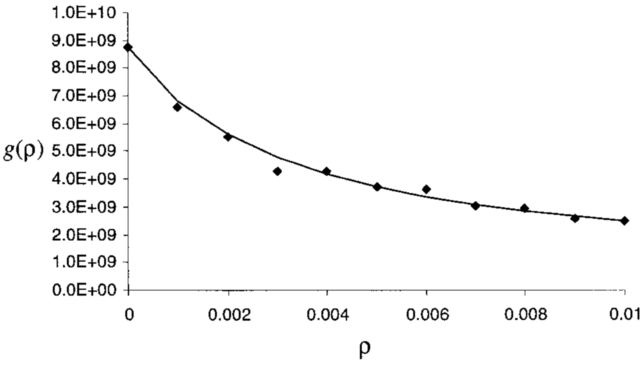

Figure A2.—The fit of the functiong$(\rho) = E(\Pi_{i=1}^{k} T_i|\rho)$ in the case of the simulated data set described in the text.

refer to lengths and distances after the recombination event has been removed.

Changing ρ: As mentioned above, a uniform distribution is assumed for the prior of ρ. ρ is updated using a sliding window technique. If the current state of the chain is ρ_0, new values of $\rho(\rho_1)$ are chosen uniformly from the interval $(\rho_0 - \Delta\rho, \rho_0 + \Delta\rho)$, where $\Delta\rho$ is some specified value. If $\rho_0 - \Delta\rho < 0$, we set $\rho_1 = \Delta\rho - \rho_0$. This ensures reversibility of the chain.

Weighting: The weights associated with this type of change are simply given by

$$w_{01} = \frac{f(A|\rho^1)E(\Pi_{i=1}^{k}T_i|\rho^0)}{f(A|\rho^0)E(\Pi_{i=1}^{k}T_i|\rho^1)}.$$

Estimating $E(\Pi_{i=1}^{k}T_i|\rho)$: To run the Markov chain it is necessary first to calculate $E(\Pi_{i=1}^{k}T_i|\rho)$. This can be easily done analytically in the case of no recombination ($\rho = 0$) and in the case of free recombination ($\rho \rightarrow \infty$). $E(\Pi_{i=1}^{k}T_i|\rho\rightarrow\infty) = E(T)^k = (2\Sigma_{i=1}^{n-1}\ 1/i)^k$, where T now is total tree length of the common genealogy shared by all SNP sites. $E(\Pi_{i=1}^{k}T_i|\rho = 0)$is given by the kth moment of a marginal genealogy. The moment-generating function for the total tree length in a marginal genealogy is

$$\prod_{i=2}^{n}\frac{i(i-1)/2}{i(i-1)/2 - si} = \prod_{i=1}^{n-1}\frac{i}{i-2s}.$$

Upon differentiation we find

$$E\left(\prod_{i=1}^{k}T_i|\rho = 0\right) = k!\sum_{i=2}^{n}\frac{(n-1)!(2/(i-1))^{k+1}}{2\Pi_{j=2}^{i-1}(j-i)\ \Pi_{j=i+1}^{n}(j-i)}. \tag{A1}$$

Unfortunately, it does not appear possible to find similar expressions for intermediate values of ρ. Instead, $E(\Pi_{i=1}^{k}T_i|\rho)$can be evaluated on a grid for arbitrary values of ρ by simulations. To get a smooth surface, a function must be fit to the simulated values. In this article, the functional form chosen was

$$\frac{c-d}{1+a\rho^b} + d, \tag{A2}$$

where $c = E(\Pi_{i=1}^{k}T_i|\rho = 0)$, $d = E(\Pi_{i=1}^{k}T_i|\rho\rightarrow\infty)$, and a and b are constants to be estimated using simulations. This function appeared to provide a reasonable fit in all examined cases.

An example of the fit of Equation A2 is given in Figure A2. The example is based on simulated data shown in Table 4 of Griffiths and Marjoram (1996). This data set was chosen to allow easy comparison with the method developed by Griffiths and Marjoram (1996). It contains 50 sequences and nine polymorphic sites. The vector of distances between polymorphic sites is {9, 26, 25, 8, 1, 2, 10, 7}. It was assumed that the values of ρ of interest were in the interval [0, 0.01], corresponding to a total rate of recombination between the two most distant sites in the interval $[0, 1.74/N_e]$. A total of 100,000 simulations were performed on two gridpoints ($\rho = 0.005$ and $\rho = 0.01$) and the function (Equation 12) was fitted to the simulation results. Subsequently, estimates of the function for $\rho = 0.001$, $\rho = 0.002$, $\rho = 0.003$, $\rho = 0.004$, $\rho = 0.006$, $\rho = 0.007$, $\rho = 0.008$, and $\rho = 0.009$ were obtained, again using 100,000 simulations. Note that the function appears to provide a reasonable fit, considering the Monte Carlo error.