Estimating Recombination Rates From Population Genetic Data

Paul Fearnhead and Peter Donnelly

Department of Statistics, University of Oxford, Oxford, OX1 3TG, United Kingdom

Manuscript received October 18, 2000
Accepted for publication August 13, 2001

ABSTRACT

We introduce a new method for estimating recombination rates from population genetic data. The method uses a computationally intensive statistical procedure (importance sampling) to calculate the likelihood under a coalescent-based model. Detailed comparisons of the new algorithm with two existing methods (the importance sampling method of Griffiths and Marjoram and the MCMC method of Kuhner and colleagues) show it to be substantially more efficient. (The improvement over the existing importance sampling scheme is typically by four orders of magnitude.) The existing approaches not infrequently led to misleading results on the problems we investigated. We also performed a simulation study to look at the properties of the maximum-likelihood estimator of the recombination rate and its robustness to misspecification of the demographic model.

ESTIMATION of recombination fractions using pedigree data is impracticable for very fine scales (<0.1 cM), because thousands of meioses are needed per recombination event. As there are a large number of meioses in the history of a sample of population data, such data could be used for estimating recombination rates over fine scales. However, estimation of the recombination rate from population data presents a difficult challenge.

In this article, we take a full-likelihood-based approach to this problem. We consider the case of a constant-sized panmictic population evolving under neutrality. Our aim is to approximate the joint-likelihood surface for the recombination and mutation rates, on the basis of all the information contained in the data. This likelihood surface involves a sum over all possible genealogies consistent with the data. Evaluating this sum exactly is impossible, so we develop an importance sampling method that approximates it. By approximating the optimal proposal density for the importance sampling, we obtain an algorithm that appears to be substantially more efficient than existing importance sampling and Markov chain Monte Carlo (MCMC) approaches to this problem.

Historically, inference about the recombination and mutation rates has been achieved using summary statistics. For example, estimation of the mutation rate for a set of sequence data, in the absence of recombination, could be based on the number of alleles, the number of segregating sites, or the number of pairwise differences (see, *e.g.*, Donnelly and Tavaré 1995). For estimating the recombination rate, estimators based on the number of pairwise differences (Hudson 1987; Wakeley 1997) or the minimum number of recombinations required (Hudson and Kaplan 1985; Wall 2000) have been proposed.

However, inference based on a summary statistic is inefficient, as it involves ignoring some of the information contained in the data. Due to strong correlations between the sampled chromosomes, the amount of extra information that can be obtained by increasing the sample size is small. As a result, it is imperative to try to use all the information contained in the sample. This involves calculating likelihood surfaces on the basis of the whole data for any unknown parameters of interest.

To calculate the likelihood surface for the data, a population genetics model must be assumed. We work within the framework of the coalescent (Kingman 1982a) and its extensions to include recombination (Hudson 1983; Griffiths and Marjoram 1996b). These processes model directly the genealogical history of a sample of chromosomes from the population. They provide good approximations to the genealogies that arise for a wide range of the classical models for population demography that are usually specified forward in time and in particular can be thought of as describing the genealogy of a large population, evolving according to the Wright-Fisher model (Kingman 1982b). The extension of the coalescent to incorporate recombination is called the ancestral recombination graph (ARG).

No explicit formulas are known for the likelihoods of interest. Recent developments in computationally intensive statistical methods have provided techniques for approximating likelihoods in complex problems such as this. For models with no recombination both importance sampling (Griffiths and Tavaré 1994a,b,c) and MCMC (Kuhner *et al.* 1995; Wilson and Balding 1998; Beaumont 1999) methods have been proposed.

Corresponding author: Peter Donnelly, Department of Statistics, University of Oxford, 1 S. Parks Rd., Oxford, OX1 3TG, England. E-mail: donnelly@stats.ox.ac.uk

More recently, Stephens and Donnelly (2000) suggested an improved importance sampling algorithm, which can be orders of magnitude more efficient than previous importance sampling schemes. They characterize the optimal proposal density for importance sampling. While this density is intractable, they show how it can be approximated by approximating one-dimensional sampling densities. This approximation can then be used as the proposal density in the importance sampling scheme.

Less work has been done for estimating likelihood surfaces in the presence of recombination. However, Griffiths and Marjoram (1996a) extended the importance sampling approach of Griffiths and Tavaré (1994a,b,c) and Kuhner *et al.* (2000) developed an MCMC scheme for this problem. Nielsen (2000) considered the related problem of estimating recombination rates from single nucleotide polymorphism data and developed an MCMC algorithm for estimating the likelihood curve in this case.

A comparison of the maximum-likelihood estimates of the recombination rate based on the likelihood surfaces produced by these methods, and estimates of the recombination rate based on summary statistics, can be found in Wall (2000). While basing estimation of the recombination rate on the likelihood surface of the whole data, as opposed to just some summary statistic of the data, is optimal (in the sense that it uses all the information in the data), it is unclear as to whether the existing methods are able to approximate the likelihood surface accurately enough to give a good approximation of the maximum-likelihood estimate.

Here we use a similar idea to that of Stephens and Donnelly (2000) to obtain a more efficient importance sampling algorithm for this class of problems. The reader is referred to that article for background on the different (MCMC and importance sampling) approaches to inference that are possible for this kind of problem. Here, as in Stephens and Donnelly (2000), the optimal proposal density can be characterized and related to one-dimensional sampling densities. These densities are approximated by considering a simpler model, which still contains many of the features of the ARG, but for which the sampling distribution is tractable. The approximations to the sampling densities can then be used to obtain an approximation to the optimal proposal density, which we then use as our proposal density. In choosing our proposal density we also take account of the ease of sampling from it.

The contents of the article are as follows. We give an informal introduction to the problem and our approach and then we introduce our model and notation. We next describe the idea of using importance sampling for approximating the likelihood surface for a given value of the recombination and mutation rates. In particular, we derive the optimal proposal density and our approximation to it. We consider estimating the likelihood surface over a grid of possible recombination and mutation rates. We then compare our method with those of Kuhner *et al.* (2000) and Griffiths and Marjoram (1996a) and demonstrate that the new algorithm gives more accurate estimates, typically substantially so, of the likelihood surface for fixed computing time. Finally, properties of the maximum-likelihood estimator (MLE), and its robustness to the demographic assumptions, are studied. This is based on a simulation study.

The *Model and notation*, *Importance sampling*, and *Implementation* sections describe the population genetics and statistical background and provide the details of the new method. Some level of technicality in the treatment seems unavoidable. We provide an informal description of the problem and of our approach in the next section. Readers interested primarily in the application of the method should be able to skip the three sections listed above and move directly from the next section to the implementation of the new method and the comparisons with other approaches, described in *Comparisons with existing methods*.

Informal description of the new method

Suppose we have a sample of chromosomes taken from a population, from which we wish to estimate the recombination rate or perhaps jointly estimate the recombination and mutation rates. In fact, if r and μ denote, respectively, the probability of recombination and the probability of mutation in the region of interest per chromosome per generation, it is possible from such data only to estimate $\rho = 4Nr$ and $\theta = 4N\mu$, where N is the effective population size. We focus on approximating the joint likelihood for ρ and θ, that is, the probability, as a function of ρ and θ, of obtaining the data actually observed.

Now imagine that in addition to observing the sampled chromosomes we were also told their complete ancestral history: the details of their genealogical tree at each locus and of the recombination and mutation events in that history. With this extra information, calculating the likelihood for ρ and θ is straightforward, and, for example, estimation of ρ and θ would typically just involve counting the number of recombination and mutation events and dividing these by the total time over which they could have occurred. Sadly, we do not observe this additional information, and the likelihood we actually want involves an average (or integral) over the (uncountable) number of possible histories consistent with the data. The dimension of the space of unobservable histories is so large that standard (naive) simulation or Monte Carlo methods for approximating the likelihood are impracticable. (For the examples we consider, of the order of 10^{50} simulations are required before we would simulate even one history that is consistent with the data.) Several more sophisticated, but nonetheless computationally intensive, methods for approximating

the likelihood are available. Two broad classes of methods are MCMC and importance sampling. Informally, each involves trickery to sample preferentially from genealogical histories that are relatively likely, *given the observed data*. Kuhner *et al.* (2000), and on a related but different problem, Nielsen (2000), adopt an MCMC approach. Griffiths and Marjoram (1996a) and this article use importance sampling.

Importance sampling involves repeatedly sampling independent ancestral histories from a proposal distribution. In our implementation, all sampled histories are consistent with the data, and, for each one, an *importance weight* is calculated. Loosely, the importance weight measures the probability of the sampled history under the coalescent model, relative to its probability under the proposal distribution. To estimate the value of the likelihood at a particular (ρ, θ) value, one simply averages the importance weights associated with the sampled histories. (Various methods are available to extend this to an estimate of the entire likelihood surface; see *Implementation*.)

The key to a good importance sampling scheme is in making a good choice of the proposal distribution. It turns out that for this problem we can characterize the optimal choice of proposal distribution, but that like the likelihood itself, this optimal proposal distribution is inaccessible. Nonetheless, and this is the key to our approach, the characterization, and an understanding of the structure of the underlying stochastic models, can be used to suggest proposal distributions that should provide good approximations to the optimal proposal distribution and hence lead to an efficient importance sampling scheme. The proposal distribution we actually used is developed in the following sections.

The proposal distribution that we choose is an approximation to the optimal proposal distribution for approximating the likelihood for one (ρ, θ) value (which is called the driving value). However, the ancestral histories that are sampled from it are used to approximate the likelihood over a grid of (ρ, θ) values. For (ρ, θ) values that are not close to the driving value, the resulting approximation of the likelihood can be poor (this is also a problem for the existing importance sampling method and the MCMC method of Kuhner *et al.* 2000). An additional novelty here is that our importance sampling method actually uses a set of driving values, which allows the likelihood to be estimated accurately over a much larger grid.

Model and notation

Assume we have data from a segment of a chromosome. Each locus within this segment is assigned a position, x, which lies in the interval $[0, 1]$. The positions 0 and 1 refer to the extremes of the segment of interest, and a locus at position x will be that fraction of the genetic distance along the segment.

We allow two distinct models for the segment. The first is a finite-sites model. We allow a finite number of loci, each at a discrete position. For example, if our data come from three equally spaced microsatellite loci, then we use a three-locus model, with loci at positions 0, 0.5, and 1. At each locus we allow a general K-allele mutation model. We assume the mutation transition matrix, P, is the same for each locus and is known. (The elements of the matrix P, $P_{\alpha\beta}$, are, for each α and β, the probability that when a mutation occurs to an allele of type α it mutates to an allele of type β.) Furthermore, we assume that the mutation rate is the same for each locus, but this is easily generalized.

Second, we consider an infinite-sites model. This model allows for an infinite number of loci at a continuum of positions. If a mutation occurs within our segment, then we assume it is equally likely to have occurred at any position in $[0, 1]$ (formally we simulate a position for the mutation from a continuous uniform distribution on $[0, 1]$). This is a suitable model for sequence data where the probability of repeat mutations is negligible. The data are summarized by the positions of the sites where mutations have occurred (the segregating sites) and which of the two types (arbitrarily labeled) each chromosome is at each segregating site.

Recombination within our segment causes different loci to have different genealogies. These genealogies can be compactly represented by a single graph, called the ARG (Griffiths and Marjoram 1996b). An example for a three-locus model is given in Figure 1. (Note that the difference between the ARG as we define it here and the recombinant genealogy of Kuhner *et al.* 2000 is the inclusion of mutations in the genealogy.)

Because different loci have different genealogies, they may also have different most recent common ancestors (MRCAs). If the ARG of a sample is known back to the oldest MRCA (OMRCA), then there is a simple analytic expression for the distribution of the types in the sample. If the types of the CA at each locus and the types of the mutations are known, then this specifies precisely the type of the sample (see Figure 1). (The ARG of Griffiths and Marjoram goes back further than the OMRCA; however, the knowledge of the ARG beyond the OMRCA contains no information about the sample.)

We define a branch in the ARG to represent the lineage of a single chromosome. Initially the number of branches in the ARG is equal to the sample size. At a coalescence event the number of branches decreases by one, and at a recombination event the number of branches increases by one. Some branches will appear in the genealogy of only a subset of the loci of interest. For each locus, if a branch appears in the genealogy for that locus then it is "ancestral" at that locus; otherwise the branch is "nonancestral" at that locus. In Figure 1, immediately after the first recombination (going back in time) there are five branches: Three are ancestral at

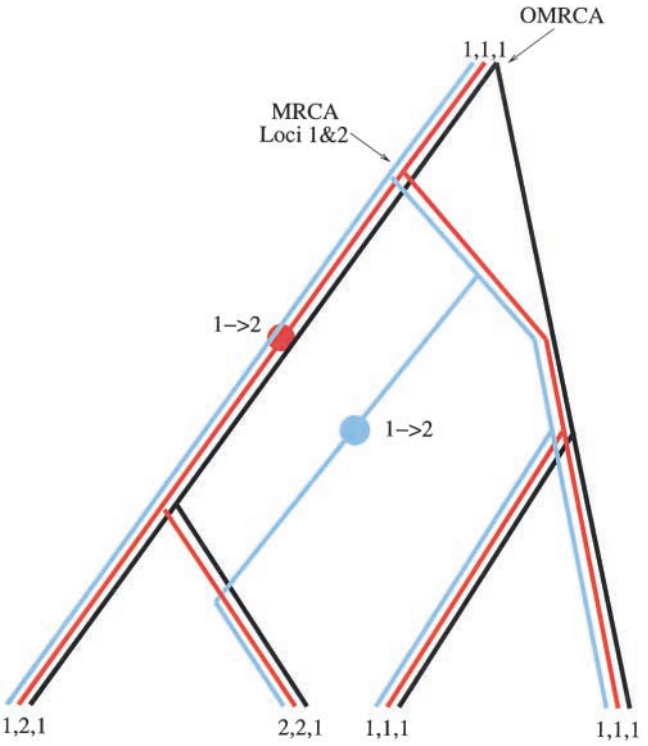

FIGURE 1.—An ARG: a graphical description of the genealogies at three linked loci for a sample of size 4. The ARG has been drawn back to the OMRCA (see text). Mutations are depicted by circles. Both the types of the loci at the top of the graph and the types of the mutations have been specified. This uniquely determines the type of the sample.

all three loci; one is ancestral for the leftmost locus and one is ancestral for the two rightmost loci.

The distribution of the ARG of a sample depends on the specific population genetic model. However, coalescent theory gives us a good approximation to this distribution for a large class of population genetic models (which includes the Wright-Fisher model). This approximate distribution is parameterized by population-scaled mutation and recombination rates, θ and ρ. These parameters depend on the effective population size, N, and the probabilities of mutation and recombination per segment per meioses, u and r. For a Wright-Fisher model and a diploid population, $\theta = 4Nu$ and $\rho = 4Nr$.

The distribution also depends on the demographic model, and we assume a constant population size and random mating (though violations of either of these assumptions can be incorporated within a coalescent framework). For the specific details of the coalescent-based approximation to the distribution of ARGs see HUDSON (1983) and KAPLAN and HUDSON (1985). Here we just note that this distribution can be characterized as a continuous-time Markov chain if viewed backward in time. Simulation of ARGs is straightforward, and they provide a computationally efficient method for simulating samples from population genetic models with recombination.

Note that by defining the position of a locus in terms of genetic distance, we are implicitly assuming that recombination is uniform across the [0, 1] interval. Other assumptions, such as uniform mutation rates and (for the finite-sites case) the assumption of the same mutation model at each locus, are purely for notational simplicity.

Importance sampling

While sampling from the distribution of ARGs is straightforward, a more difficult problem is to draw inference about the graph, and parameters of it, from a sample of chromosomes at the tips. We perform full-likelihood inference for the mutation and recombination parameters of the model by using importance sampling to estimate the likelihood of the data for a given value of θ and of ρ. See STEPHENS and DONNELLY (2000) for background on the use of importance sampling for likelihood inference in population genetic models.

Denote the types of the n sampled chromosomes by A_n. We define the ancestral history of a sample to be the ARG including mutations and their types (but without interevent times). If we let G be an ancestral history, then the likelihood, $L(\rho, \theta)$, can be written as

$$L(\rho, \theta) = p(A_n|\rho, \theta) = \int p(A_n|G, \rho, \theta)p(G|\rho, \theta)\mathrm{d}G,$$

with the integral being taken over all ancestral histories with n tips.

By our definition of G in the previous paragraph, $p(A_n|G, \rho, \theta)$ is an indicator function, taking the value 1 if G is consistent with the data and 0 otherwise. Let $\mathcal{G}$ be the set of all ancestral histories that are consistent with the data. Furthermore, let $q(G)$ define a density whose support contains $\mathcal{G}$. Then

$$\begin{aligned} L(\rho, \theta) &= \int_{\mathcal{G}} \frac{p(G|\rho, \theta)}{q(G)} q(G)\mathrm{d}G \\ &\approx \frac{1}{M}\sum_{i=1}^{M} p(G_i|\rho, \theta)/q(G_i), \end{aligned} \qquad (1)$$

where $\{G_1, \ldots, G_M\}$ are an independent sample from the density $q(G)$. Equation 1 is an importance sampling approximation of the likelihood, and $q(G)$ is called the importance sampling proposal density. Each term in the sum (1), $p(G_i|\rho, \theta)/q(G_i)$, is called an importance sampling weight. The approximation given by (1) is unbiased and consistent. Its accuracy for finite M depends on the variance of the importance sampling weights, which in turn depends on the choice of proposal density $q(G)$.

We now consider how to make a sensible choice of proposal density. Our approach is to calculate the optimal proposal density for a given value of ρ and θ. While sampling from this density is not possible, it can be approximated. It is this approximation that is our pro-

posal density in the importance sampling scheme. For notational simplicity we omit the conditioning on ρ and θ in the rest of this section.

Optimal proposal density: We consider the class of proposal densities that generate events back in time in a Markov fashion. The proposal density for the next event back in time will depend on the current state of the ARG: which loci are ancestral along which branches and the types of the chromosomes at ancestral loci. Thus a realization from the proposal density will be an ARG that is consistent with our sample, generated back to the OMRCA. Equivalently, it will consist of a series of states $H_0(=A_n), \ldots, H_\tau$, where H_τ is the state of the ARG when the OMRCA is reached. The optimal proposal density is in this class of Markov proposal densities and has transition probabilities

$$q(H_{i+1}|H_i) = p(H_i|H_{i+1})\pi(H_{i+1})/\pi(H_i), \qquad (2)$$

where $p(H_i|H_{i+1})$ are the (known) forward-transition probabilities of the ARG, and $\pi(H_i)$ is the probability that a sample from the population is of the same type as H_i at the ancestral loci in H_i. (This can be proven in an analogous way to Theorem 1 of Stephens and Donnelly 2000.)

We can simplify the "transition probabilities" (2) of the proposal density, but before doing so we need to make the following definition.

Definition 1: Let $\pi(\cdot|H)$ be the conditional distribution of the last chromosome in a sample, given that the other chromosomes are of type H,

$$\pi(\alpha|H) = \pi(\{H, \alpha\})/\pi(H).$$

The ratio $\pi(H_{i+1})/\pi(H_i)$ can be simplified because, for all possible transitions, the majority of chromosomes are unaffected. For example, at a coalescent event of two chromosomes of type α, H_{i+1} is equal to H_i, but with one less chromosome of type α, which we denote by $H_i - \alpha$. In this case

$$\pi(H_{i+1})/\pi(H_i) = 1/\pi(\alpha|H_i - \alpha).$$

Similar equations can be derived for the other possible transitions.

Thus we can rewrite (2) in terms of one-dimensional densities of the form $\pi(\alpha|H)$ and the known forward transition probabilities $p(H_i|H_{i+1})$. While the densities are unknown, and thus we cannot directly sample from the optimal proposal density, they can be approximated. Substituting these approximations into our formula for the optimal proposal density will give us a sensible proposal density for the importance sampling. We now consider how to approximate $\pi(\alpha|H)$.

Approximating $\pi(\alpha|H)$ when $\rho = 0$: To consider a suitable approximation for $\pi(\alpha|H)$ for the ARG, we first review the case where there is no recombination. In their work, Stephens and Donnelly (2000) approximated this conditional density by considering a related, but simpler, process.

When H contains j chromosomes, the new type α is obtained by choosing a chromosome from H at random and then mutating it a geometric number of times. If j_β is the number of chromosomes of type β in H, then the Stephens and Donnelly (2000) approximation to $\pi(\alpha|H)$ is

$$\hat{\pi}(\alpha|H) = \sum_\beta \frac{j_\beta}{j} \sum_{k=0}^{\infty} \frac{j}{j+\theta}\left(\frac{\theta}{j+\theta}\right)^k P^k_{\beta\alpha}. \qquad (3)$$

This can be simplified because

$$\sum_{k=0}^{\infty} \frac{j}{j+\theta}\left(\frac{\theta}{j+\theta}\right)^k P^k_{\beta\alpha} = (1-\lambda_j)(1-\lambda_j P)^{-1}_{\beta\alpha},$$

where $\lambda_j = \theta/(j+\theta)$.

Thus, it is necessary to calculate the matrices $(1-\lambda_j P)^{-1}$, for $\lambda_1, \ldots, \lambda_n$, only once. Once these have been calculated, evaluating (3) for any α and H is computationally inexpensive. (Equation 3 can be applied directly for single-locus models; for multilocus models Stephens and Donnelly 2000 suggest a numerical integration technique.)

Approximating $\pi(\alpha|H)$ when $\rho \neq 0$: We now extend the approximation of Stephens and Donnelly (2000) to the case where there is recombination. This involves two extra complications. The first is how to deal with nonancestral loci, and the second is to allow for the effect of recombination.

Our approximation to $p(\alpha|H)$ is just the probability of a chromosome of type α under a stochastic process, which is simpler than the ARG but retains most of the important properties of the ARG. In the infinite-sites case, the simplified process is as follows.

Initialization: Assume there are j chromosomes in H. Let s be the number of segregating sites in the infinite-sites case. These occur at positions $x_1, \ldots, x_s$ in order. Denote the $s-1$ midpoints of consecutive positions by y_i, $i = 1, \ldots, s-1$. That is $y_i = (x_{i+1} + x_i)/2$. Further, let $z_i = x_{i+1} - x_i$, for $i = 1, \ldots, s-1$.

Recombinations: For $i = 1, \ldots, s-1$, generate a recombination at point y_i with probability $z_i\rho/(z_i\rho + j)$, independently of recombinations at y_l, $l \neq i$. Let k be the number of recombinations generated. These k recombinations split the chromosome into $k+1$ intervals. Denote these by $\mathbf{r} = \{r_1, \ldots, r_{k+1}\}$.

Imputation: Impute types at nonancestral segregating sites in H. If p_i is the proportion of a chromosome in H that is of type i at a specific site, then a type i is imputed at that site with probability p_i. All imputations are done independently of each other.

Mutations: Treat each r_i, $i = 1, \ldots, k+1$ independently; for each r_i simulate the type of the new chromosome, α, on that region according to the approximation in *Approximating $\pi(\alpha|H)$ when $\rho = 0$* (that is, simulate the type of a complete chromosome from Equation 3,

and then use only the value of this type on the interval r_i). The type of α is given by the union of its type on each r_i.

The process for the finite sites case is similar, with sites being replaced by loci.

The approximation is a sum over all possible imputations and recombinations. This summation can be evaluated efficiently, using dynamic programming. See APPENDIX A for details.

Proposal density: We have suggested an approximation to $p(\alpha|H)$, which then defines an approximation to the optimal proposal density for our problem. However, this approximation to the proposal density does not have attractive computational properties. In particular, to sample from this density, it is necessary to calculate the transition probabilities for every possible transition.

We can substantially reduce the computation involved in sampling from our proposal density by making a simple change. We choose a proposal density that can be sampled from by first choosing a chromosome, using some simple probability rule, and then choosing an event to occur to this chromosome. The probability that a chromosome is chosen will be proportional to the rate at which an event occurs to that chromosome, when the information of the types of the chromosomes is ignored. If a_i is the fraction of loci of chromosome i that are ancestral, b_i is the fraction of loci of chromosome i that are between ancestral loci, and there are j chromosomes in the current configuration of the ancestral history, then chromosome i is chosen with probability proportional to $(j - 1) + a_i\theta + b_i\rho$.

Once a chromosome has been chosen, we use our approximation to $p(\alpha|H)$ and the optimal proposal density (2) to calculate the probabilities of each possible event occurring to that chromosome. For a mathematical formulation of our proposal density see APPENDIX B.

A further improvement can be obtained by noting that if the most recent common ancestor at a locus has been found, then the probability of the type at that locus is independent of the types of the chromosomes at all other loci and is distributed as a draw from the mutation stationary distribution (FEARNHEAD 2001). Thus, the probability of our current state can be factorized into the probability of the most recent common ancestor of that locus and the probability of the types of the chromosomes at all other loci. As the former is known, we need only approximate the latter. Thus as we simulate the ancestral history back further in time, we need not simulate it at that locus.

Implementation

We have considered estimating the likelihood for a single value of (ρ, θ), using importance sampling, and have suggested a sensible proposal density for this by approximating the optimal proposal density. We now consider how we can extend this to estimating the likelihood surface over a grid of (ρ, θ) points.

Estimating the likelihood surface: While our proposal density was derived for estimating the likelihood at a specific (ρ, θ) value, it can be used to estimate the likelihood at other (ρ, θ) values, using an idea from GRIFFITHS and TAVARÉ (1994a). We can think of a proposal density being chosen for estimating the likelihood at a specific value (ρ_1, θ_1), called the driving value, but being used as a proposal density to estimate the likelihood at other values of (ρ, θ). Since the computational cost of sampling from the proposal density is significantly higher than the cost of calculating the prior probability of an ancestral history, using a sample from one proposal density to estimate the likelihood at a large number of points is computationally efficient.

However, the proposal density is chosen to approximate the posterior density of ancestral histories given the data and the driving value of the parameters. If the value of the parameters (ρ, θ) where the likelihood is being estimated is sufficiently different from the driving value, then this posterior distribution is likely to be quite different from the optimal proposal density for estimating the likelihood at (ρ, θ). As a result the estimate of the likelihood could be poor.

Bridge sampling: One way around this is to use a set of driving values $(\rho_1, \theta_1), \ldots, (\rho_k, \theta_k)$. A number of ancestral histories, say N, could be simulated from each of the proposal densities, $q_{(\rho_i,\theta_i)}(G)$. Each of these kN histories could be used to estimate the likelihood surface. If the set of driving values covers the area of the parameter space of interest, then we could combine the estimates of the likelihood surface to produce an estimate that is accurate over the whole area of interest and not just around one point. A suitable method for combining these estimates is bridge sampling (MENG and WONG 1996).

The basic idea of bridge sampling is to generate our sample from the mixture

$$\frac{1}{k}\sum_{j=1}^{k} p(G|\rho_j, \theta_j, A_n) = \begin{cases} \frac{1}{k}\sum_{j=1}^{k} c_j p(G|\rho_j, \theta_j), & \text{if } G \in \mathcal{G}, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $c_1, \ldots, c_k$, are (unknown) normalizing constants and $\mathcal{G}$ is the set of histories that are consistent with the data. Given a sample, $\{G_1, \ldots, G_{kN}\}$, from (4),

$$\frac{1}{N}\sum_{i=1}^{kN} \frac{p(G_i|\rho, \theta, A_n)}{\sum_{j=1}^{k} c_j p(G|\rho_j, \theta_j)} \tag{5}$$

is an estimate of $p(A_n|\rho, \theta)$.

Unfortunately we cannot sample directly from (4), so instead we generate a weighted sample from (4) using importance sampling. We use

$$\frac{1}{k}\sum_{j=1}^{k} q_{(\rho_j, \theta_j)}(G)$$

as our proposal density in the importance sampling.

A further complication is that we do not know the

normalizing constants $c_1, \ldots, c_k$. Due to using importance sampling to sample from (4), we need only know these normalizing constants up to a common factor. Also, $c_j = p(A_n|\rho_j, \theta_j)$, and so we can estimate these normalizing constants using (5). As c_j appears in (5), we end up with a set of $k - 1$ simultaneous equations for c_j/c_1, for $j = 2, \ldots, k$. These can be solved iteratively (see Meng and Wong 1996).

To use this iterative method for estimating the ratios c_j/c_1, we need to store the histories G_i, $i = 1, \ldots, kN$. For most problems, where kN is of the order of millions, this is impracticable. Instead, we store only the first M histories. These histories are used to estimate the ratios c_j/c_1. Conditional on these estimates of c_j/c_1, $j = 2, \ldots, k$, we can calculate our estimate of the likelihood surface sequentially (that is without needing to store the histories). Thus, no further histories are stored, and the estimates of the ratios c_j/c_1 obtained from the first M histories are used to calculate the estimate of the likelihood surface on the basis of the set of kN histories.

Kuhner *et al.* (1995) used a related idea (that of Geyer 1991) to calculate likelihood curves for θ (in the absence of recombination) from k independent MCMC samplers, each run with a different driving value. However, the importance sampling schemes of Griffiths and Tavaré (1994c), Griffiths and Marjoram (1996a), and Stephens and Donnelly (2000) and the MCMC schemes of Kuhner *et al.* (1998), Beerli and Felsenstein (1999), and Kuhner *et al.* (2000) attempt to estimate the likelihood curve (or surface) using a single driving value. These methods will all suffer from the problem that, regardless of how efficient their algorithm is, the estimate of the likelihood for values of the parameters away from the driving value may well be poor. It should be possible to apply the idea of bridge sampling, using multiple driving values, to each of these methods. This may allow each method to estimate the likelihood curve accurately over a larger region of parameter space.

Comparison with existing methods

We compare our new method with the existing importance sampling method of Griffiths and Marjoram (1996a) and the MCMC method of Kuhner *et al.* (2000). Each of these methods provides an approximation for the joint likelihood surface for ρ and θ and then estimates ρ and θ by calculating the value of ρ and θ for which this approximate likelihood is maximized [the maximum-likelihood estimate (MLE) for the approximated likelihood surface]. Hence there are two possible comparisons to make:

i. Compare each method on how accurately it approximates the likelihood surface.
ii. Compare each method on the properties of its estimates of ρ and θ.

Our perspective is that the first comparison is more fundamental (although Kuhner *et al.* 2000 test the performance of their method on the basis of the second criteria). The likelihood surface contains all the information about ρ and θ that is contained in the data, and thus being able to estimate this likelihood surface accurately should be the basis for good estimation of ρ and θ.

As we note later, it is not obvious whether the general statistical theory that usually makes maximum likelihood the preferred method of estimation applies in this context. Nonetheless, short of a fully Bayesian approach (see discussion), estimation via maximum likelihood deserves serious consideration. It may happen that a method that estimates the likelihood poorly actually results in an estimate that is closer to the truth for some simulated data sets. But it is hard to imagine this effect being systematic, so that among "maximum-likelihood" methods, choosing the method that most accurately approximates the likelihood would seem prudent.

Sensible interval estimation is a more important goal than providing only point estimates with no accompanying measure of their uncertainty. Although the usual theory linking the shape of the likelihood surface to confidence intervals is also not obviously applicable here, we provide encouraging empirical evidence in *Properties of the maximum-likelihood estimator*. Thus inaccurate estimates of the likelihood surface could also lead to inappropriate confidence intervals for parameters of interest.

As a result of these considerations, we compare the three methods solely on the basis of (i). Note also that designing a suitable simulation study to compare the methods on the basis of (ii) is difficult. Our perspective on implementing a computationally intensive method for approximating the likelihood surface is that you should not run the method for an arbitrary (large) number of iterations and then use the MLEs for ρ and θ. Instead, you should run the method for sufficiently many iterations until you have confidence in the final approximation of the likelihood surface; only then should this surface be used for inference (*e.g.*, by using it to calculate the MLEs for ρ and θ). For certain data sets, accurate estimation of the likelihood surface requires much more computation than for others. As we discuss in more detail below, for some methods, and in particular for ours, it is possible to get a reasonable indication from the output of the program as to whether or not the likelihood is being approximated well. If it is not, the method should be run longer. A simulation study that fixes in advance the number of iterations is thus not the most practically useful comparison of different methods.

Diagnostics: An important consideration for computationally intensive statistical methods is diagnosing, from the output of the program, how accurately the likelihood curve is being estimated and hence how long the program needs to be run to obtain accurate estimates.

For importance sampling, the estimate of the likeli-

hood for any (ρ, θ) value is just the average of the weights associated with each sampled history. The variance of the estimator based on N runs of the importance sampler is σ_W^2/N, where σ_W^2 is the variance of the importance weights. (Note that for any proposal density, the mean of the associated importance weights is the value of the likelihood we are trying to estimate. Thus the key to choosing a good proposal density is to choose one for which the variance of the importance weights is not too large.)

A related measure of the performance of an importance sampling method is the so-called effective sample size (ESS). It is defined as

$$\text{ESS} = N\,\mu_W^2/(\sigma_W^2 + \mu_W^2),$$

where μ_W is the mean of the associated importance weights (in our case, the likelihood that we are trying to estimate).

A helpful informal interpretation of the ESS is that if after N runs of the importance sampling method, the ESS equals M, then the accuracy of our estimate will be (approximately) the same as if we had been able to take M independent samples from the appropriate distribution. (Note that if we fix N, then the larger the variance of the importance weights, the smaller the ESS.)

In principle, one advantage of importance sampling over MCMC is that the independence of the samples makes it straightforward to assess the accuracy of the likelihood estimates. Thus, if we have N runs producing importance weights $w_1, \ldots, w_N$, we can estimate σ_W^2 by the sample variance of the importance sampling weights, s_W^2. We can then estimate the variance of our estimator of the likelihood by s_W^2/N and estimate the ESS as

$$\frac{(\sum_{i=1}^N w_i)^2}{\sum_{i=1}^N w_i^2}. \tag{6}$$

Care must be taken, as s_W^2 can substantially underestimate σ_W^2, and (6) overestimate the true ESS, if sufficient runs are not used. See the DISCUSSION for a more detailed review of this problem.

Diagnostics for an MCMC method are less straightforward. Because the recombinant genealogies sampled by the Markov chain are not independent, the sample variance of the likelihood across these genealogies is no longer a suitable measure of accuracy (as it ignores the positive correlation between sampled recombinant genealogies). There is a considerable literature on diagnostics for MCMC (for example, see BROOKS and ROBERTS 1998), but these diagnostics use the raw output of the Markov chain, which is not available to the user of the MCMC method of Kuhner and colleagues.

Implementation: We make comparisons for both sequence and microsatellite data. For sequence data we compared

i. our method (implemented by a program called Infs).
ii. the importance sampling method of GRIFFITHS and MARJORAM (1996a; implemented by a program called Recom58, which was kindly provided by the authors), and
iii. the MCMC method of KUHNER *et al.* (2000; implemented by their program Recombine, which is available from http://www.evolution.genetics.washington.edu/lamarc.html).

For the microsatellite data, we compared

i. our method (implemented by a program called Fins) and
ii. the importance sampling method of GRIFFITHS and MARJORAM (1996a; implemented by the program Twoloc, which was kindly provided by the authors).

Currently there is not a version of the MCMC method of KUHNER *et al.* (2000) available for microsatellite data.

For sequence data, we based our comparisons on the data sets simulated in WALL (2000; kindly provided by the author). These are samples of 50 chromosomes, simulated under the infinite-sites model with $\rho = 3.0$ and $\theta = 3.0$. For ease of comparison, we consider solely estimating the likelihood curve for ρ (and hence the MLE for ρ), conditional on the true value of θ. For approximating the likelihood curves, Infs used five driving values (with bridge sampling): $\theta = 3.0$, and $\rho = \{1.0, 2.0, 4.0, 6.0, 9.0\}$; while Recom58 (which allows only a single driving value to be specified) used the driving value of $\theta = 3.0$ and $\rho = 3.0$. Also, Recom58 requires knowledge of the type of the most recent common ancestor, and the true type was used. In practice this would either have to be inferred from an outgroup, guessed, or the program would need to be run for a number of different, plausible, values of the type of the common ancestor.

The program Recombine uses MCMC to simulate from the posterior distribution of recombinant genealogies (which are similar to our ancestral histories but do not include mutations), conditional on a (ρ, θ) value (the driving value). Likelihood surfaces are then calculated using importance sampling. As with Infs, the accuracy of the likelihood surface may be poor away from the driving value. The authors suggest (see KUHNER *et al.* 2000) that Recombine should be implemented with a number of short runs, followed by a long run of the MCMC sampler. After each short run, the driving value for the following run is set to be the MLE for the current run. The final likelihood surface is based solely on the sample of ancestral histories obtained from the final long run. The idea behind this is that the short runs should enable Recombine to find a good driving value, which is then used for the long run.

We initially implemented Recombine as suggested by the authors. However, we found that often Recombine

would estimate θ as <0.2, which is clearly inconsistent with the number of segregating sites in the data sets. The reason for this appeared to be that for successive short runs of the MCMC sampler the estimated MLE for θ would become smaller. Thus the final long run of the MCMC sampler would have a driving value close to 0, and the final MLE for θ is biased toward this driving value. To overcome this problem, we "cheated" and set Recombine to use θ = 3.0 as its driving value. Short runs were still used for Recombine to find a good driving value for ρ. While the true value of θ would not be known in practice, a sensible approach would be to fix the driving value of θ to be Waterson's estimate of θ (Waterson 1975), which is based on the number of segregating sites in the sample, and which in the authors' experience is usually close to the MLE for θ. This approach substantially reduced, but did not eradicate, the proportion of times that Recombine gave implausibly low estimates for the MLE for θ. (One possible reason for the inaccuracies of Recombine is that the single long run uses only one driving value; Stephens 1999 shows that for Coalesce, the analogous MCMC program for models with no recombination, the estimate of the likelihood will have infinite variance for values of θ more than twice the driving value.)

Initially we ran Infs, Recom58, and Recombine on the 20 data sets from Table 2 of Wall (2000), using 1 million iterations in each case. For a more detailed comparison, we then ran each method six independent times (five runs of 1 million and one run of 10 million) on a single data set (data set 19; the data set for which there was least agreement between the three methods, with estimates of ρ ranging from 0.6 to 4.8). The computing time for Infs and Recom58 varies considerably between data sets (it depends a lot on the number of segregating sites in the data set). For each data set, the computing time for a million iterations of each of the two methods was comparable and varied between 2 and 12 hr on a Pentium 400-MHz PC. The computing time for Recombine depended less on the data set but often varied considerably (by up to an order of magnitude) between independent runs on the same data set. In particular, the computing time depends a lot on the driving value for ρ that is used in the MCMC sampler (which can differ across independent runs on the same data set). The average computing time for Recombine was comparable to the computing time for runs of the same length using the other two methods.

For the microsatellite data set we simulated a sample of 50 chromosomes at two microsatellite loci; the data were simulated with ρ = 2.0, and for each locus we assumed a symmetric stepwise mutation model with θ = 2.0. Again we only considered inference for ρ conditional on the true value for θ.

Both importance sampling programs Fins and Twoloc were run six independent times on this data set: five short runs of length 1 million and one long run of length 10 million. For both programs, we used a different single driving value for each of the five short runs. All runs used a θ driving value of 2.0, but the driving values of ρ varied (ρ = 1.0, 2.0, 3.0, 4.0, and 5.0). The long runs were an attempt to obtain an accurate estimate of the true likelihood surface. For Fins, four driving values (θ = 2.0 and ρ = {1.0, 2.0, 4.0, 8.0}) and bridge sampling were used, while for Twoloc (which only allows a single driving value), the driving value θ = ρ = 2.0 was used. A run of length 1 million from Fins took 1.5 hr on a 400-MHz Pentium PC, while the same length run on Twoloc took twice as long.

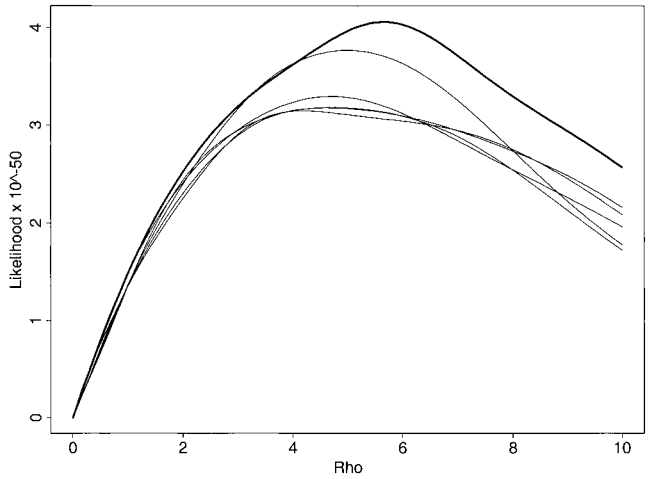

Figure 2.—Comparison of six estimates of the likelihood surface for data set 19 in Wall (2000). Each estimate is based on an independent run of our method, Infs. Five estimates are based on runs of length 1 million, and one (shown by the thick line) is based on a run of length 10 million.

Results of comparison: As there is no quantitative measure of accuracy that is obtainable from the output of Recombine, we can only compare the performance of this MCMC method with the importance sampling methods of Infs and Recom58 by looking at approximated likelihood curves for each method on the same data set. Figures 2–4 each show six approximations of the likelihood curve for data set 19; for each of the figures, the approximations are obtained from independent runs of one of the three methods. While the approximated likelihood curves that were obtained from the six runs of Infs are very similar (see Figure 2), those obtained by the MCMC method, Recombine, vary dramatically (see Figure 3). The value of ρ for which the maximum of the likelihood curve is attained varies from 4.2 to 5.7 for the curves obtained by Infs but from 0.0 to 4.8 for Recombine. (Note that Recombine allows for the possibility of repeat mutations, while Infs does not, so comparing the estimated likelihood curves obtained by the two programs for the same data set may be inappropriate. The variability of approximations of the likelihood across independent runs of a single program gives a measure of the computational efficiency of that

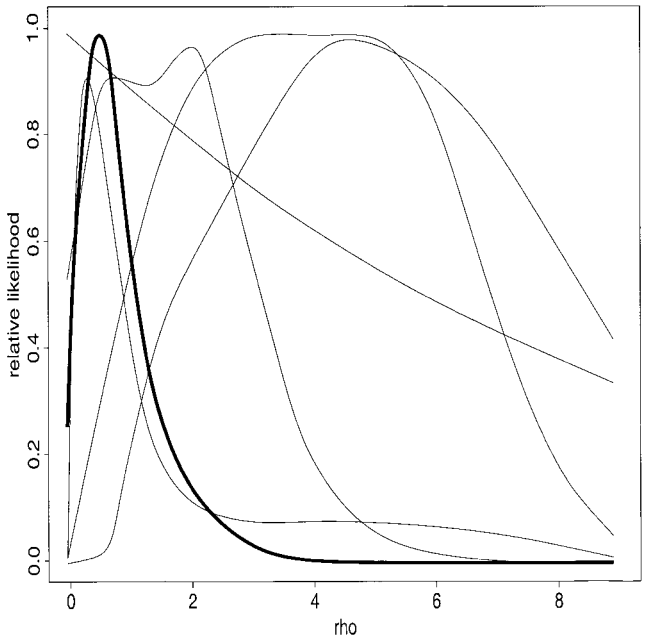

FIGURE 3.—Comparison of six estimates of the relative likelihood surface for data set 19 in WALL (2000). Each estimate is based on an independent run of the method of KUHNER *et al.* (2000), Recombine. Five estimates are based on runs of length 1 million, and one (shown by the thick line) is based on a run of length 10 million.

method, which can be meaningfully compared across different programs.)

In many of the examples we considered, the behavior of Recombine was poor: Similar results were obtained from Recombine on the other data sets where the MLE for ρ is nonzero (data not shown). For example, for data set 13, the likelihood curves from Recombine varied so much that the MLE for ρ varied from 0.0 to 4.8 across six independent runs (each of length 1 million). WALL (2000) also reported large variation in the value of the MLE estimated by Recombine from independent runs when analyzing data set 11. For data sets where the MLE for ρ is zero, Recombine did generally estimate the MLE as zero, but the slope of the estimated likelihood curve (which contains information about how accurate the MLE is) varied considerably across independent runs.

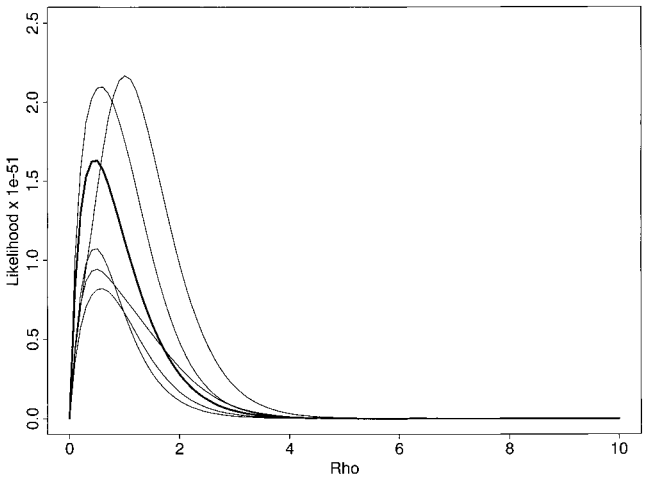

FIGURE 4.—Comparison of six estimates of the likelihood surface for data set 19 in WALL (2000). Each estimate is based on an independent run of the method of GRIFFITHS and MARJORAM (1996a), Recom58. Five estimates are based on runs of length 1 million, and one (shown by the thick line) is based on a run of length 10 million.

TABLE 1

Comparison of the accuracy of likelihood curves approximated by Infs and Recom58

	Method	
Data set	Infs (ESS)	Recom58 (ESS)
1	610	5
2	38	22
3	37	2
4	190	4
5	1	1
6	25	31
7	340	77
8	430	23
9	200	2
10	7000	2
11	470	6
12	5000	80
13	8	1
14	23	23
15	41	8
16	17	4
17	1500	2
18	21	9
19	1000	4
20	3500	51

Shown is a comparison of our method, Infs, and the method of GRIFFITHS and MARJORAM (1996a), Recom58, on 20 simulated data sets. For each data set the estimated ESS for the two importance sampling methods is given. The larger the true ESS the more accurate the estimate of the likelihood is (see text for more details).

In contrast we found that the estimates of the likelihood curves obtained by Infs were generally similar across multiple independent runs on the same data set, particularly for data sets where the estimated ESS was large. The exceptions were the two data sets with smallest ESS (see Table 1): data sets 5 and 13. For both these data sets, the magnitude of the likelihood surface varied by up to a factor of 10 across six independent runs of Infs. For data set 13, the maximum-likelihood estimates of ρ varied from 2.5 to 5.3, and 95% confidence intervals for ρ varied between [0.8, 8.8] and [0.9, 12]. The estimated relative-likelihood curves for data set 5 are shown in Figure 5. In this case, the maximum-likelihood estimate for ρ varied from 2.3 to 4.0, and 95% confidence intervals varied between [1.0, 4.6] and [1.0, 8.9]. For both these data sets the inaccuracies of the estimated

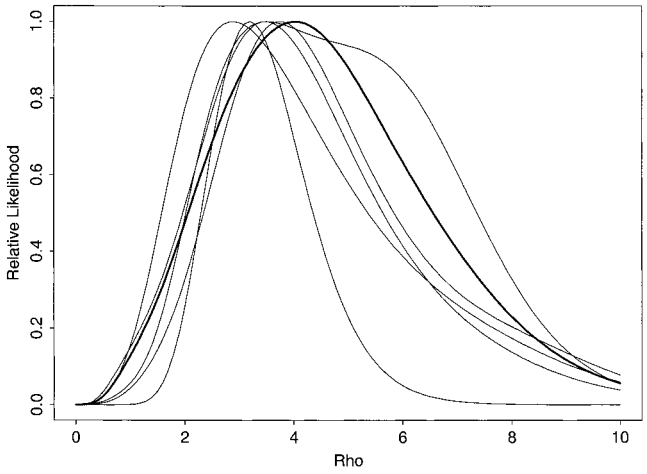

FIGURE 5.—Comparison of six estimates of the likelihood surface for data set 5 in WALL (2000), one of the two data sets for which Infs performed worst. Each estimate is based on an independent run of our method, Infs. Five estimates are based on runs of length 1 million, and one (shown by the thick line) is based on a run of length 10 million.

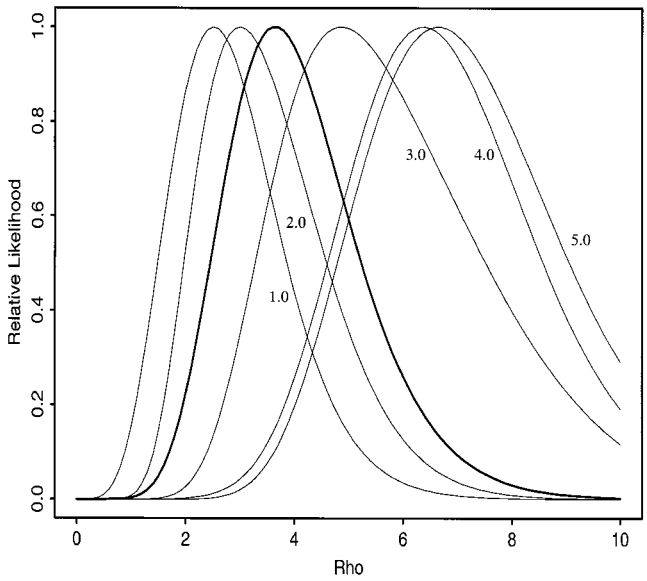

FIGURE 6.—Comparison of estimates of the relative likelihood curve (*i.e.*, each curve has been normalized so that its maximum value is 1.0; the unnormalized curves vary by more than a factor of 500) for ρ (conditional on θ = 2.0) for microsatellite data, obtained from the method of GRIFFITHS and MARJORAM (1996a), Twoloc. The thin lines are estimates obtained from runs of length 1 million, each with a different driving value for ρ (each curve is labeled with its driving value of ρ). The thick line is an estimate based on a single run of length 10 million.

likelihood surfaces could be diagnosed from the ESSs, which were small for all runs of Infs.

For data set 19, the general shapes of the likelihood curves approximated by Recom58 are similar (see Figure 4), but the curves themselves are of different magnitudes. In contrast, the approximations of the likelihoods obtained by Infs are similar both in shape and magnitude (see Figure 2). A comparison of Figures 2 and 4 shows that Infs and Recom58 are producing substantially different approximations of the shape of the likelihood curve. By comparing the ESSs for both programs (see Table 1) we have strong evidence that it is Infs that is producing the more accurate approximation. For the runs of length 10 million, the ESS of Infs was ~10,000, while that of Recom58 was still <10 (results not shown). (This misbehavior of Recom58 is analagous to that of the Griffith and Tavaré importance sampling method; see STEPHENS and DONNELLY 2000.) Further support is given by the results of Section 5.2 of STEPHENS and DONNELLY (2000): They show that poor estimation of the likelihood curve often takes the form of underestimation. For our example the estimate of the likelihood curve obtained by Recom58 is an order of magnitude smaller than that of Infs.

A more detailed comparison of the performance of the two importance sampling methods can be carried out because a natural comparison of importance sampling methods is via their ESSs. Table 1 gives the estimated ESS from Infs and Recom58 for each of the 20 data sets. These give a gauge as to how accurately the likelihood curve is estimated by the two importance sampling methods. As noted above, if too small a sample is drawn from the proposal density, then the estimate of the ESS can overestimate the true value (NEAL 1998). In particular this means that if the estimated ESS is small (which suggests that the sample from the proposal density is too small), then it is possible that it is an overestimate of the true ESS. Thus, sensible comparisons of the ESSs of the two importance sampling methods are possible only for the data sets where the estimated ESSs are large. For example, for data sets 10, 12, 19, and 20, the estimated ESSs for Infs are >1000 and are likely to be an accurate estimate of the true ESSs. For each of these data sets, the estimated ESS of Recom58 is between two and four orders of magnitude smaller (and is, if anything, likely to be an overestimate of the true ESS for Recom58). This increase in efficency by four orders of magnitude is typical in our experience.

Now consider the microsatellite data set. The estimates of the relative likelihood curve for ρ (conditional on θ = 2.0) obtained from the importance sampling method of GRIFFITHS and MARJORAM (1996a), Twoloc, are shown in Figure 6. In this figure each likelihood curve has been normalized so that it has a maximum relative likelihood of one. The unnormalized estimates of the likelihood from the six independent runs vary by more than a factor of 500. Figure 6 shows that there is no agreement between the estimates of the relative likelihood across the independent runs. In particular the estimated MLE of ρ appears to be affected by the driving value of ρ. For the five short runs, the estimate of the MLE of ρ increases as the value of the driving value for ρ increases.

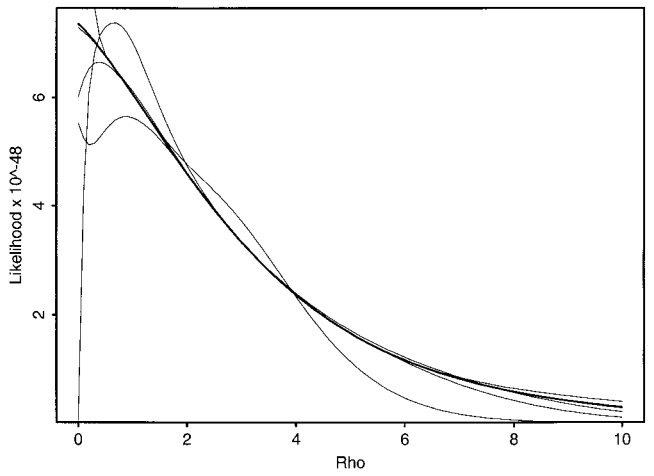

FIGURE 7.—Comparison of estimates of the likelihood curve for ρ (conditional on θ = 2.0) for microsatellite data, obtained from our method, Fins. The thin lines are estimates obtained from runs of length 1 million, each with a different driving value for ρ (ρ = 1.0, 2.0, 3.0, 4.0, and 5.0). The thick line is an estimate based on a single run of length 10 million, obtained using four driving values and bridge sampling.

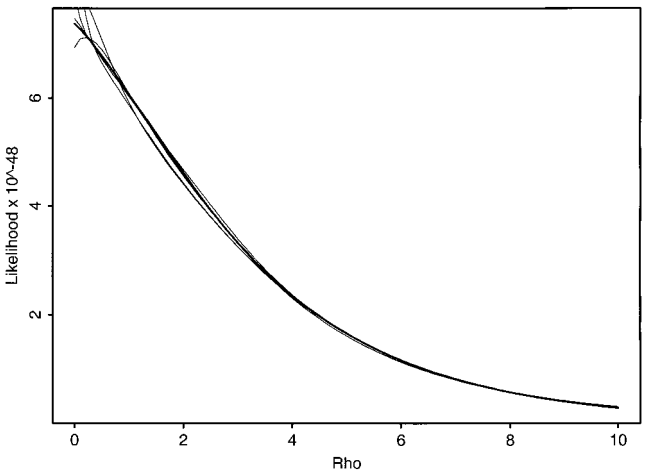

FIGURE 8.—Comparison of estimates of the likelihood curve for ρ (conditional on θ = 2.0) for microsatellite data, obtained from our method, Fins. All estimates are obtained using four driving values and bridge sampling. The thin lines are estimates based on runs of length 100,000, and the thick line is an estimate based on a run of length 10 million.

The estimates of the absolute likelihood curve for ρ (conditional on θ = 2.0) obtained from our importance sampling method, Fins, are shown in Figure 7. There is considerably more agreement between the estimates of the likelihood for ρ across the independent runs. This is not only agreement about the shape of the likelihood curve but also agreement on the magnitude of the likelihood.

The estimated ESS for the long run is large (>10,000) for almost all values of ρ, which suggests that this is an accurate estimate of the true likelihood curve. For each of the short runs the likelihood is accurately estimated close to the driving value. However, the likelihood tends to be poorly estimated away from the driving value. If bridge sampling is used with four driving values (θ = 2.0 and ρ = {1.0, 2.0, 4.0, 8.0}), accurate estimates of the likelihood curve for all values of ρ between 0.0 and 10.0 can be obtained using runs of length 100,000 (see Figure 8).

Properties of the maximum-likelihood estimator

We carried out a simulation study to analyze the performance of maximum-likelihood estimation of ρ and θ for sequence data. We consider properties of the MLEs for ρ, θ, and ρ/θ. The reason for estimating ρ/θ is that often, by using comparisons with other species, we can estimate μ , so (as $r = \mu\rho/\theta$) to estimate r we only need an estimate of ρ/θ.

We considered two cases, first looking at the sampling properties of the MLE and coverage properties of confidence intervals, when our modeling assumptions are correct. Second we considered the robustness of the MLE and associated confidence intervals under deviations from the assumptions of a constant population size and random mating.

The authors believe that certain sample configurations could result in MLEs for ρ that are infinite and hence that the mean and variance of the estimates of ρ and ρ/θ are infinite. Thus we summarize our results in terms of the median estimate and the summary statistic g of WALL (2000): the proportion of times the MLE is within a factor of 2 of the truth.

Confidence intervals were calculated using a chi-square approximation to the likelihood-ratio statistic (as used in KUHNER *et al.* 1998). This approximation is based on an asymptotic result for independent data. The result states that (under certain regularity conditions, which include the independence of each element of the sample) the distribution of the likelihood-ratio statistic tends to a chi-square distribution as the sample size tends to infinity. For population genetics data the chromosomes are not independent, and so this result need not apply. Nonetheless, one aim of our simulation was to assess the coverage properties of intervals constructed as if this theory did apply.

Our simulation study is based on generating 100 samples from a chosen model (the simulations were carried out using a program by R. Hudson, and because of the potential application to gene mapping, we chose models and parameter values that are plausible for humans) and then running the program Infs (using bridge sampling) on each of these samples to obtain estimates of the MLE of ρ and θ and approximate confidence intervals. The accuracy of each estimate of the likelihood was gauged using its ESS, and, where necessary, using multiple independent runs with different driving values. The number of runs required increases quickly with both the number of segregating sites and the amount

TABLE 2

Effect of sample size on sampling properties of MLEs

Sample size	Mean ($\hat{\theta}$)	SD ($\hat{\theta}$)	θ in C.I.	Med ($\hat{\rho}$)	$g(\hat{\rho})$	ρ in C.I.	Med ($\hat{\rho}/\hat{\theta}$)	$g(\hat{\rho}/\hat{\theta})$	ρ/θ in C.I.
50	0.99	0.54	96	0.55	0.23	100	0.67	0.24	99
100	0.95	0.41	96	0.65	0.20	99	0.48	0.13	99
200	1.00	0.44	97	0.2	0.22	100	0.25	0.25	99

Shown are estimated means and standard deviations of the maximum-likelihood estimates of θ; estimated median of maximum-likelihood estimates of ρ and ρ/θ; *g*, the proportion of times that the MLE is within a factor of 2 of the truth; and coverage properties of an ~95% confidence interval (calculated using a chi-square approximation to the likelihood ratio statistic). Results are based on 100 samples (simulated with $\rho = \theta = 1.0$) for each of three sample sizes.

of recombination. When analyzing a sample of size 50, simulated with $\rho = \theta = 1.0$, ~400,000 iterations were required. This took between ½ and 1 hr on a modern PC (with an Intel Pentium II 400-MHz CPU). By comparison, when analyzing a sample of size 50 generated with $\rho = \theta = 3.0$, on average 5 million iterations were used, which took 1 day's computing time using a single 400-MHz PC. (The computational burden of obtaining the MLE for a single data set means that a detailed simulation study of properties of the estimator is computationally extremely demanding.)

Sampling properties of the MLE: First we considered the effect of the sample size on the accuracy and coverage properties of the MLE. We simulated samples of sizes 50, 100, and 200 from an infinite-sites model with $\rho = \theta = 1.0$, assuming a constant population size and random mating. For example, for humans, assuming an effective population size of 10,000 diploid individuals, these values of ρ and θ correspond to a 2.5-kb sequence (assuming genome-wide average rates for mutation and recombination).

A summary of the results is given in Table 2, and histograms of the estimates of θ and ρ obtained from samples of size 50 are given in Figure 9 (the histograms for samples of size 100 and 200 are similar; also the histogram for estimates for ρ/θ is similar to that for ρ). The MLE performs well at estimating θ. The distribution of MLEs of ρ is highly skewed in all cases. The median MLE of ρ is significantly less than the truth, whereas the average value is close to, and slightly larger than, the truth for all three simulation studies (results not shown; remember the theoretical mean is likely to be infinite).

Increasing the sample size has no noticeable effect on the performance of the MLE. The randomness in the simulation study is much larger than any effect that increasing the sample size has. There is significant evidence that the putative 95% confidence intervals for ρ and ρ/θ are conservative for all three sample sizes (P values = 0.006 and 0.037 for the one-sided test).

Second, we considered the effect of sequence length on the MLE. We simulated samples of size 50 with $\rho = \theta = 1.0$, 2.0, and 3.0 (corresponding for humans, again assuming an effective population size of 10,000, and genome-wide average rates for mutation and recombination, to sequences 2.5, 5.0, and 7.5 kb long). A summary of the results is given in Table 3, and histograms of the estimates for θ and ρ are given in Figure 9 (again, the histograms for ρ/θ are similar to those for ρ). Increasing the size of the sequence does improve the accuracy of the MLE. In particular, it appears to increase the median of the MLE for ρ and ρ/θ and also increase the values of *g*. These are both consequences of the fact that the skewness in the distribution of MLEs for ρ reduces as the sequence length increases. For both $\rho = \theta = 2.0$ and $\rho = \theta = 3.0$, the simulation results are consistent with the approximate confidence intervals having the correct coverage probabilities.

Robustness properties of the MLE: To test the robustness properties of the MLE, we considered two deviations from our underlying model: nonconstant population sizes and nonrandom mating (the models we used, and their effect on the underlying genealogical tree, are discussed in Donnelly and Tavaré 1995). In both cases, ρ and θ are not natural or well-defined quantities (they depend on how the effective population size is defined). By comparison, ρ/θ is a natural parameter (it is equal to r/μ). Therefore, we focus solely on how well ρ/θ is estimated.

We considered three models with exponential population growth. The first model assumes that t generations in the past, the effective population size is $N_t = N_0\exp\{-(\beta t/4N_0)\}$, where N_0 is the current effective population size and β is a parameter that governs the speed of growth. We took $\beta = 0.7$ (the MLE from the β-globin data set of Harding *et al.* 1997). If $N_0 = 10{,}000$ diploid individuals, then this corresponds to the effective population size halving every 20,000 generations.

The second and third models assume a constant population size, followed by a recent sudden expansion. From mitochondrial DNA, it has been suggested that a recent sudden human population expansion occurred between 33,000 and 150,000 years ago (Sherry *et al.* 1994; Rogers and Jorde 1995). For both models we

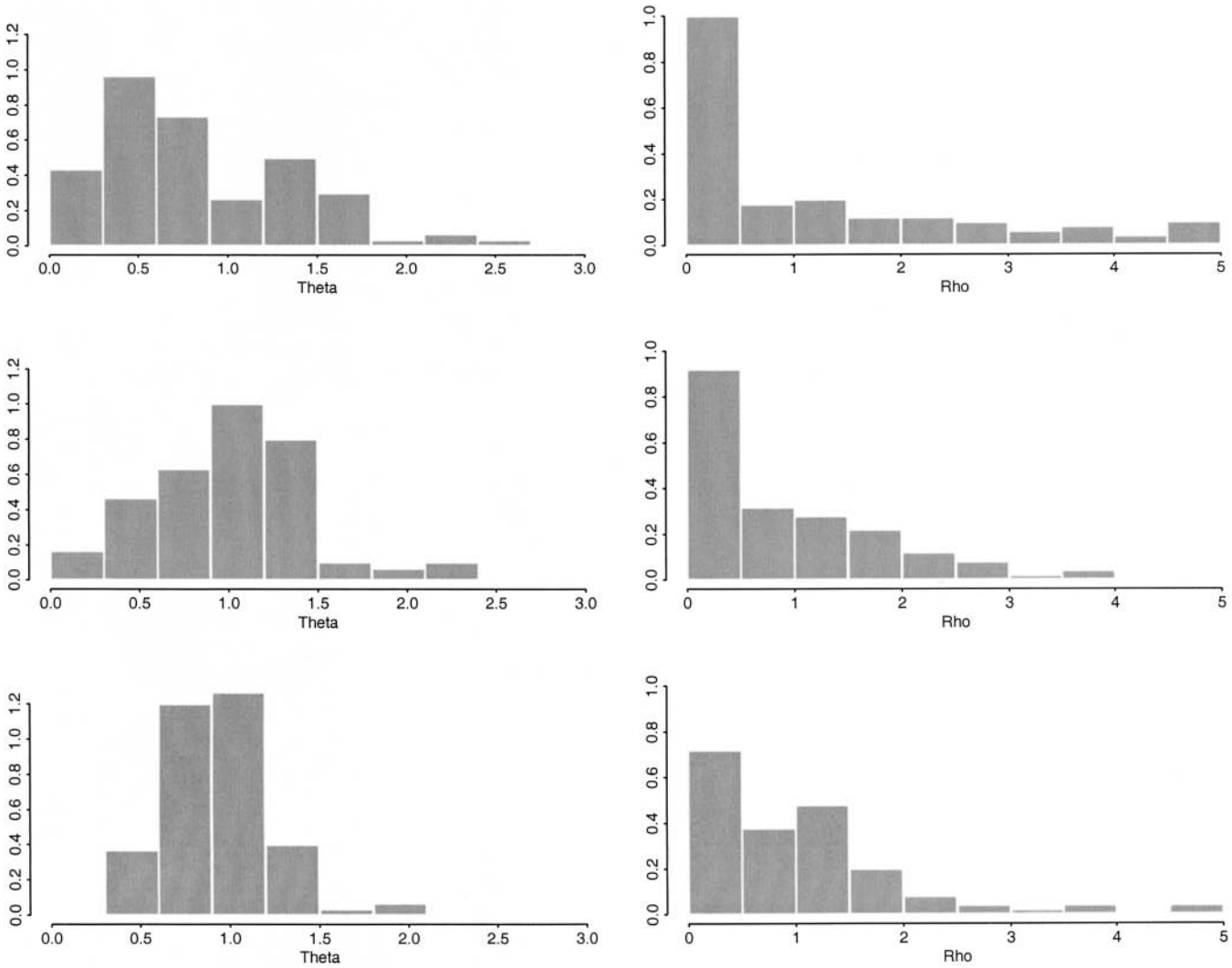

FIGURE 9.—Histograms of $\hat{\theta}/\theta$ (left column) and $\hat{\rho}/\rho$ (right column) for different sequence lengths. The top row is for $\rho = \theta = 1.0$, the middle row for $\rho = \theta = 2.0$, and the bottom row for $\rho = \theta = 3.0$. All histograms are based on estimates from 100 data sets of size 50.

assumed an effective population size of 10,000 diploid individuals, followed by an exponential expansion to a current effective population size of 5,000,000 diploid individuals. In the second model the expansion started 1600 generations ago, and in the third it started 4000 generations ago. (This model is related to the one used in KRUGLYAK 1999 to study the extent of linkage disequilibrium in the human genome).

TABLE 3

Effect of sequence length on sampling properties of MLEs

$\theta (= \rho)$	Mean $(\hat{\theta}/\theta)$	SD $(\hat{\theta}/\theta)$	θ in C.I.	Med $(\hat{\rho}/\rho)$	$g(\hat{\rho}/\rho)$	ρ in C.I.	Med $(\hat{\rho}/\hat{\theta})$	$g(\hat{\rho}/\hat{\theta})$	ρ/θ in C.I
1	0.99	0.54	96	0.55	0.23	100	0.67	0.24	99
2	1.00	0.43	95	0.65	0.41	98	0.71	0.35	97
3	1.00	0.31	100	0.8	0.50	94	0.8	0.50	92

Shown are estimated means and standard deviations of the maximum-likelihood estimates of θ; estimated median of maximum-likelihood estimates of ρ and ρ/θ; g, the proportion of times that the MLE is within a factor of 2 of the truth; and coverage properties of an $\sim$95% confidence interval (calculated using a chi-square approximation to the likelihood ratio statistic). Results are based on 100 samples (each of size 50) for each of the three sequence lengths.

TABLE 4

Robustness to population growth

Model	Med ($\hat{\rho}/\hat{\theta}$)	$g(\hat{\rho}/\hat{\theta})$	ρ/θ in C.I.
1	0.4	0.19	100
2	0.6	0.24	99
3	0.4	0.28	97

Shown are the median of the MLE of ρ/θ; *g*, the proportion of times the MLE is within a factor of 2 of the truth; and coverage properties for the ~95% confidence interval for ρ/θ under population growth. Model 1 has slow exponential growth (population doubling every 20,000 generations to a current size of 10,000). Models 2 and 3 have recent rapid exponential growth from an effective population of 10,000–5,000,000. In model 2 the growth started 1600 generations ago, and in model 3 it started 4000 generations ago. Results are based on 100 samples for each model, with $r = \mu = 2.5 \times 10^{-5}$, so the true value of ρ/θ is 1.

TABLE 5

Robustness to population structure

Migration rate	Med ($\hat{\rho}/\hat{\theta}$)	g ($\hat{\rho}/\hat{\theta}$)	ρ/θ in C.I.
6.25×10^{-6a}	0.27	0.24	73
6.25×10^{-6b}	0.40	0.25	86
2.5×10^{-5a}	0.63	0.46	96
2.5×10^{-5b}	0.40	0.29	91
2.5×10^{-4a}	0.48	0.33	96
2.5×10^{-4b}	0.43	0.37	97

Shown are the median of the MLE of ρ/θ; *g*, the proportion of times the MLE is within a factor of 2 of the truth; and coverage properties for the ~95% confidence interval for ρ/θ under population growth. Samples were simulated under a two-island model, with a diploid population size of 10,000 on each island, with migration rates of 6.25×10^{-6}, 2.5×10^{-5}, and 2.5×10^{-4} (these represent the fraction of each generation that are migrants).

[a] Twenty-five chromosomes sampled from each population.

[b] Fifty chromosomes sampled from a single population. Results are based on 100 simulated samples from each model, with $r = \mu = 2.5 \times 10^{-5}$, so the true value of $\rho/\theta = 1$.

For all three models, samples were generated with $r = \mu = 2.5 \times 10^{-5}$, which corresponds to a 2.5-kb region (assuming genome-wide average recombination and mutation rates). The results of the simulation studies are shown in Table 4.

For all three models, the performance of the MLE for ρ/θ is comparable to its performance in the constant population size case. Despite the model misspecification, there is no evidence to support the hypothesis that the putative 95% confidence intervals for ρ/θ are anticonservative. In fact, there is significant evidence that they are conservative for model 1 ($P = 0.006$) and model 2 ($P = 0.037$). In the former case, this is due to the lack of information in the data resulting in extremely large confidence intervals.

To test robustness under nonrandom mating, we used an island model, with random mating within the populations of each island and migration between the islands. For simplicity, we considered a model with just two islands and assumed that the population size of each island was identical (10,000 diploid individuals). We considered three migration rates (m): They corresponded to 6.25×10^{-6}, 2.5×10^{-5}, and 2.5×10^{-4} of each generation being migrants (the scaled migration rates, $4Nm$, are 0.25, 1, and 10, respectively). For all cases we simulated data with $r = \mu = 2.5 \times 10^{-5}$, which corresponds to a 2.5-kb region (assuming genome-wide average recombination and mutation rates). For each migration rate we simulated both a sample of size 50 from a single population and one consisting of samples of size 25 from each population.

The results are given in Table 5. Once again the median values of the MLE for ρ/θ underestimate the truth (in this case the mean values were also <1, perhaps indicative that population structure creates patterns consistent with no or little recombination). For the two larger migration rates, the performance of the MLE, as measured by *g*, is better than in the random-mating case. As might be expected, if the migration rate is small, then the performance of the MLE for ρ/θ from data where all chromosomes are sampled from a single population is better than the MLE from data with chromosomes sampled from both populations.

For the two models with the larger migration rates, the simulation results are consistent (at the 95% level) with the approximate confidence intervals having the correct coverage probabilities. However, when $4Nm = 0.25$ the confidence intervals are anticonservative. This is because patterns in the data simulated under this model are consistent with little recombination and the MLE often severely underestimates the true value. In fact, of the 41 data sets for which the confidence intervals did not contain the true value of ρ/θ, only once was the value of ρ/θ overestimated.

A common measure of population structure is F_{ST} (Wright 1951; Cavalli-Sforza *et al.* 1994, pp. 29–30, for a definition). F_{ST} can be related to the population-scaled migration rate (Hudson *et al.* 1992): For a two-island model, assuming a small mutation rate, $F_{ST} \approx 1/(1 + 16Nm)$. So our simulations refer to populations with F_{ST} of 0.5, 0.2, and 0.024. For humans, observed values of F_{ST} are on the order of 0.01 for closely related populations and 0.1–0.3 for distantly related populations (Cavalli-Sforza *et al.* 1994). Our results thus suggest that the MLE for ρ/θ performs adequately and that the approximate confidence interval is robust for levels of population structure consistent with most human populations.

DISCUSSION

Estimation of recombination rates from population data is an important and challenging problem. Methods

based on summary statistics do not use the full information contained in the data, while full-likelihood-based methods struggle due to the difficulty in accurately estimating the likelihood curve. We developed an importance sampling algorithm that can estimate the joint likelihood of ρ and θ. Existing methods have been developed by Griffiths and Marjoram (1996a) and Kuhner *et al.* (2000). Nielsen (2000) considers a related problem of estimating recombination rates from single nucleotide polymorphism data.

We noted that there are two different sorts of comparisons possible between these methods. The possibilities are to compare the methods on how well they approximate the likelihood surface and to compare the methods on the properties of the resulting "maximum-likelihood" estimates of the recombination and mutation rates. We have taken the view that, as the principal aim of each method is to approximate the likelihood surface, the most appropriate comparison is the first of these.

For the two importance sampling methods, a natural comparison is via the ESS, defined in *Diagnostics*. (The ESS is directly related to the variance of the estimate of the likelihood.) We calculated these for 20 different sets of sequence data (see Table 1; each data set was simulated with $\rho = \theta = 3.0$ under an infinite-sites model). The results suggest that the new importance sampling method is up to four orders of magnitude more efficient than the importance sampling method of Griffiths and Marjoram (1996a). Further results showed that our method is also substantially more efficient at analyzing microsatellite data.

To compare our method with the MCMC method of Kuhner *et al.* (2000), we concentrated on looking at multiple independent approximations of the likelihood curve for ρ for a given data set (no simple comparison, like that based on estimated ESS values, is possible in this case). The results for one data set are given in Figures 2 and 3. Figure 3 shows considerable variation across the independent approximations of the likelihood curve by the MCMC method. In contrast, the independent approximations obtained by our importance sampling method are very similar (see Figure 2).

Similar results were obtained for different data sets (see *Results of comparison* for more details). In fact, the performance of Recombine was poor on the majority of data sets to which we applied it. In Kuhner *et al.* (2000), the average performance of the MLE for ρ/θ was recorded for simulated data sets with various different parameter values. The results there were encouraging and at first seemed to contradict the results we obtained from using Recombine. One reason for this may be the substantially different parameter regimes considered by Kuhner *et al.* (2000) compared to those we considered. Kuhner *et al.* (2000) simulated data where the rate of recombination was small compared to the mutation rate (ρ/θ was between 0.00 and 0.08), and Recombine may perform better for these parameter values. However, for most organisms the recombination rate is the same order of magnitude as (or greater than) the mutation rate (*e.g.*, Drosphila and bacteria: see Feil *et al.* 1999; Andolfatto and Przeworski 2000).

An additional difference is that even for the parameter values they considered, Kuhner *et al.* (2000) did not attempt to assess how well their method approximated the likelihood surfaces. Instead, they examined the average behavior, across many simulated data sets, of their MLE. This average behavior was encouraging, as it was also in the study of Wall (2000) for $\rho = \theta = 3.0$. Nonetheless, it remains disconcerting that for Recombine, as seen in Figure 3, the likelihood surface, and hence the MLE, obtained for a particular data set depends crucially on the choice of random number seed.

All the methods for estimating recombination rates from population data are very computationally intensive. As such, gains in efficiency have real practical value. The computational time and cost of analyzing real data may still seem large, but in most cases it will be much smaller than the time and cost of collecting the data in the first place. One word of caution, though, is that the computational time increases rapidly with the size of data set (particularly as the length of sequence, as measured by ρ and θ, increases). Further research is still needed for implementing full-likelihood (or maybe approximate-likelihood) methods for large data sets. One example of current research in this area is the method proposed by Wall (2000).

One novelty of our method is the use of bridge sampling. This enables multiple driving values to be used, with the results for these driving values being combined in a sensible manner. Using only a single driving value can result in a poor estimate of the likelihood surface away from this driving value. This can result in the likelihood (away from the driving value) being underestimated with a high probability, which in turn results in a bias of the MLE toward the driving value. By using multiple driving values this problem can be substantially overcome. Bridge sampling could also be applied to both the importance sampling method of Griffiths and Marjoram (1996a) and the MCMC method of Kuhner *et al.* (2000) and may enable each method to approximate the likelihood surface well over a larger grid of (ρ, θ) values.

It is important, when using any computationally intensive statistical method, to check the accuracy of the estimate of the likelihood curve before using it for inference. For importance sampling methods, one measure of the accuracy for importance sampling schemes is the ESS (see *Diagnostics*). If the ESS could be calculated exactly, then it would give a direct measure of the accuracy of the approximation of the likelihood surface. A 100-fold increase in ESS equates to a 10-fold increase in accuracy: An ESS of 100 suggests the approximation is accurate to within 10%, and an ESS of 10,000 suggests the approximation is accurate to within 1%. (In practice

the positive correlation between estimates of the likelihood at similar parameter values will make the approximation of the relative likelihood much more accurate than is suggested by the ESSs.)

Unfortunately, the ESS cannot be calculated analytically and has to be estimated. Often the estimated ESS can be significantly larger than the true ESS. So, while a low estimated ESS is indicative of a poor estimate, a large estimated ESS does not guarantee an accurate one. An improvement on using the ESS from a single run is to track the value of the ESS as the number of iterations increases. If the estimate of the ESS is accurate, then this should increase linearly with the number of iterations. If this happens, (and particularly if the ESS is large) the estimated ESS should be close to the truth. A parallel approach is to use multiple runs with different driving values. If the results from these independent runs are consistent, then that is evidence that the results are accurate. (See Stephens and Donnelly 2000 for more discussion on convergence diagnostics.)

Diagnostics for MCMC methods are less straightforward, as the values simulated by the Markov chain (in the case of Recombine these "values" are recombinant genealogies) are not independent. However, there is considerable literature on MCMC convergence diagnostics, and many of these diagnostics could be applied to an MCMC method for the problem we consider here. These diagnostics require the raw output of the Markov chain, which is not available from Recombine. Thus currently the only method for assessing the accuracy of the approximated-likelihood surface obtained by Recombine is by running Recombine a number of independent times on the same data set and directly comparing the approximations of the likelihood surface that are produced.

Another potential problem with any computationally intensive method is the possibility of an undetected bug in the computer code. This is particularly important for a problem like estimating recombination rates, when little is known about the sampling properties of the MLE (and hence bugs cannot be detected by comparing the output of the program with theoretical expectations). One diagnostic check that the authors found useful (and that helped to find a bug in an early version of one of the programs) is to run the program on a large number of independent data sets, each data set simulated with the same parameter values. If the approximated likelihoods for each data set are multiplied together, we obtain a "composite" likelihood function. By construction, this composite likelihood function is the likelihood of independent identically distributed sets of data, and the usual asymptotic theory will apply. In particular, the likelihood-ratio statistic based on this composite likelihood should be approximately chi square (with the degrees of freedom equal to the number of parameters), so that checking its empirical distribution provides a useful diagnostic tool.

A consideration in using full-likelihood-based methods is that little is known theoretically about the behavior of the MLE. It is not known whether the MLE for ρ is consistent or whether, asymptotically, the likelihood-ratio statistic has a chi-square distribution. Even if these asymptotic results apply, it will still not be clear whether the resulting approximations are accurate for real data sets whose size is far from the asymptotic limit.

In *Properties of the maximum-likelihood estimator* we used a simulation study to analyze the sampling properties of the MLE for ρ and θ and to consider the robustness of the MLE to the demographic model. The amount of information in the data depends primarily on the length of the sequence being analyzed and only very slightly on the number of chromosomes in the sample. For small sequences ($\theta = 1.0$, which models an "average" 2.5-kb sequence of human DNA), the MLEs for ρ and ρ/θ performed poorly. For larger sequences ($\theta = 2.0$ and 3.0), their performance improved substantially. The MLE for θ performs better than the MLE for ρ (uniformly across all the parameter values we chose). The usual theory for obtaining confidence intervals for MLEs does not apply in this setting. Nonetheless, approximate 95% confidence intervals, suggested by that theory, for both ρ and θ showed no evidence for being anticonservative.

We analyzed the robustness properties of the MLE and associated confidence intervals under models of population growth and population substructure. We solely consider the MLE for ρ/θ, as neither ρ nor θ are well defined for these models (in contrast ρ/θ is just the ratio of the probability of recombination to the probability of mutation). The simulation results are broadly encouraging: They suggest that the MLE performs satisfactorily and that confidence intervals are robust to population growth and structure. With the exception of data simulated under a model with very strong population structure (more than is consistent with human population genetic data), the confidence intervals for ρ/θ showed no significant evidence for being anticonservative.

In considering parameter estimation, for ease of comparison with published work, we concentrated on maximum-likelihood estimation. However, having obtained the likelihood surface it would also be straightforward to adopt Bayesian approaches to estimation. These can also be undertaken directly in MCMC methods, *e.g.*, as in Wilson and Balding 1998, without explicitly generating a likelihood surface.) As noted elsewhere (Tavaré *et al.* 1997; Wilson and Balding 1998) there are natural advantages in Bayesian methods in this context. First, they allow other information (for example, existing estimates of the effective population size or calibration of mutation rates by comparisons with other species) to be incorporated. Perhaps more importantly, they allow a sensible incorporation of uncertainty about relevant

qualities and hence sensible measures of uncertainty in final estimates.

Throughout this article we focused on the problem of estimating likelihood surfaces for unknown parameters. However, the method that we present here can be used not only to estimate the recombination and mutation rates but also to perform "ancestral inference," by which we mean inference about events that have occurred in the history of the sample of chromosomes. This might include estimating the time of the most recent common ancestor at a specific site in the chromosome, the time since a specific mutation occurred, or the number of recombination events that have taken place in a given region of the chromosome. In each case one can ask about the conditional distribution of the unknown quantities in the light of the observed data. For any given (ρ, θ) value, our importance sampling method generates a sample of weighted ancestral histories, which can be viewed as an approximation to the conditional distribution of ancestral histories given the data. From this it is straightforward to obtain an approximation for the marginal conditional distribution of any quantity of interest. For example, consider the time of a specific mutation. If $w_1, \ldots, w_N$ are the weights assigned to N sampled ancestral histories, and $t_1, \ldots, t_N$ the times of the mutation in each of these sampled histories, then we can approximate the conditional (continuous) distribution of the time of the mutation by the discrete distribution that assigns probability $w_i/(\Sigma_{j=1}^{N} w_j)$ to time t_i. Alternatively, an approximation to the conditional distribution of the number of recombination events in a specific region is given by the discrete distribution, which assigns to the number i a probability mass that is proportional to the sum of the weights of sampled ancestral histories that have i recombination events in that region. A natural point estimate for these unknown quantities is the mean of the corresponding conditional distribution.

We express our gratitude to Matthew Stephens for useful discussions and comments. This work was supported by UK Engineering and Physical Sciences Research Council (EPSRC) grant GR/M14197 and UK Biotechnology and Biological Sciences Research Council (BBSRC) grant 43/MMI09788. The programs Infs and Fins are available from http://www.stats.ox.ac.uk/mathgen/software.html.

LITERATURE CITED

Andolfatto, P., and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156:** 257–268.

Beaumont, M., 1999 Detecting population expansion and decline using microsatellites. Genetics **153:** 2013–2029.

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.

Brooks, S. P., and G. O. Roberts, 1998 Assessing convergence of Markov chain Monte Carlo algorithms. Stat. Comput. **8:** 319–335.

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

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

Evans, G., 1993 *Practical Numerical Integration.* John Wiley & Sons, New York.

Fearnhead, P., 2001 Haplotypes: the joint distribution of alleles at linked loci. Technical report, Department of Statistics, University of Oxford (available from http://www.stats.ox.ac.uk/mathgen/publications.html).

Feil, E. J., M. C. J. Maiden, M. Achtman and B. G. Spratt, 1999 The relative contributions of recombination and mutation to the divergence of clones of Neisseria meningitidis. Mol. Biol. Evol. **16:** 1496–1502.

Geyer, C. J., 1991 *Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo.* Tech. Rep. 568, School of Statistics, University of Minnesota, Minneapolis/St. Paul.

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

Griffiths, R. C., and P. Marjoram, 1996b An ancestral recombination graph, pp. 257–270 in *IMA Volume on Mathematical Population Genetics,* edited by P. Donnelly and S. Tavaré. Springer-Verlag, Berlin/Heidelberg/New York.

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

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

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

Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J. Cox *et al.*, 1997 Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60:** 772–789.

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

Hudson, R. R., 1987 Estimating the recombination parameter of a finite population without selection. Genet. Res. **50:** 245–250.

Hudson, R. R., and N. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

Hudson, R. R., M. Slatkin and W. P. Maddison, 1992 Estimation of levels of gene flow from DNA-sequence data. Genetics **132:** 583–589.

Kaplan, N., and R. R. Hudson, 1985 The use of sample genealogies for studying a selectively neutral m-loci model with recombination. Theor. Popul. Biol. **28:** 382–396.

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

Kingman, J. F. C., 1982b Exchangeability and the evolution of large populations, pp. 97–112 in *Exchangeability in Probability and Statistics,* edited by G. Koch and F. Spizzichino. North Holland, Amsterdam.

Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common diseases. Nat. Genet. **22:** 139–144.

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

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

Kuhner, M. K., J. Yamato and J. Felsenstein, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393–1401.

Meng, X., and W. H. Wong, 1996 Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Stat. Sinica **6:** 831–860.

Neal, R. M., 1998 *Annealed Importance Sampling.* Tech. Rep. 9805, Department of Statistics and Department of Computing Science, University of Toronto, Toronto, Ontario, Canada.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.

Rogers, A. R., and L. B. Jorde, 1995 Genetic evidence on modern human origins. Hum. Biol. **67:** 1–36.

Sherry, S. T., A. R. Rogers, H. Harpending, H. Soodyall, T. Jenkins *et al.*, 1994 Mismatch distribution of mtDNA reveals recent human population expansions. Hum. Biol. **66:** 761–775.

Stephens, M., 1999 Problems with computational methods in population genetics, contribution to the 52nd session of the International Statistical Institute (available from http://www.stats.ox.ac.uk/~stephens/group/publications.html).
Stephens, M., and P. Donnelly, 2000 Inference in molecular population genetics (with discussion). J. R. Stat. Soc. Ser. B **62:** 605–655.
Tavaré, S., D. J. Balding, R. C. Griffiths and P. Donnelly, 1997 Inferring coalescent times from DNA sequence data. Genetics **145:** 505–518.
Wakeley, J., 1997 Using the variance of pairwise differences to estimate the recombination rate. Genet. Res. **69:** 45–48.
Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.
Waterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.
Wilson, I. J., and D. J. Balding, 1998 Genealogical inference from microsatellite data. Genetics **150:** 499–510.
Wright, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

Communicating editor: D. Charlesworth

APPENDIX A: DYNAMIC PROGRAMMING CALCULATION OF $p(\alpha|H)$

Here we describe the dynamic programming technique for the finite-sites case, with L loci. A K-allele model is assumed at each locus, with mutation transition matrix P and mutation rate θ/L. We use a numerical integration scheme similar to that of Stephens and Donnelly (2000). Assume suitable k, t_m, and w_m have been chosen (for example, using Gaussian quadrature; see Evans 1993), so that for a general function $f(\cdot)$,

$$\int_0^\infty \exp\{-t\} f(t)\,\mathrm{d}t \approx \sum_{m=1}^{k} w_m f(t_m).$$

Assume further that the current sample configuration, H, contains j chromosomes. Finally let π_i, $i = 1, \ldots, L$ be a K-vector of the proportions in H of each of the K types at locus i, and define $Q(t) = \exp\{\theta\, t(P - I)/L\}$.

Now $p(\alpha|H)$ is calculated recursively. Let $p_i(\alpha)$ denote the probability that the type of the new chromosome generated under our approximation will be the same as α at the first i loci. Further, let $p_i(\alpha|s, t)$ denote the same probability conditional on the ith locus being obtained by mutating the ith locus of chromosome s in H, with the number of mutations being Poisson with rate θt. Now, we use the approximation (obtained via numerical integration) that

$$p_i(\alpha) = \sum_{m=1}^{k} \sum_{s=1}^{j} w_m p_i(\alpha|s, t_m/j)/j. \tag{A1}$$

Let $q_i = z_i\rho/(j + z_i\rho)$, the probability of a recombination between the ith and $(i + 1)$th loci, and let α_{i+1} be the type of α at the $(i + 1)$th locus. If the sth chromosome in H is ancestral at the $(i + 1)$th locus, then denoting $H_{s(i+1)}$ to be its type,

$$p_{i+1}(\alpha|s, t) = \{(1 - q_i)p_i(\alpha|s, t) + q_i p_i(\alpha)\} Q_{H_{s(i+1)}\alpha_{i+1}}(t). \tag{A2}$$

If it is not ancestral then

$$p_{i+1}(\alpha|s, t) = \{(1 - q_i)p_i(\alpha|s, t) + q_i p_i(\alpha)\}(\pi_{i+1}Q(t))_{\alpha_{i+1}}. \tag{A3}$$

Initially, $p_0(\alpha|s, t) = 1$ for all s and t. Equations A1–A3 can be used to calculate $p_i(\alpha)$ recursively for $i = 1, \ldots, L$. However, $p_L(\alpha) = p(\alpha|H)$, the required quantity. This computation is linear in kL. An equivalent formulation is possible for the infinite-sites case.

APPENDIX B: MATHEMATICAL FORMULATION OF PROPOSAL DENSITY

Here we give a mathematical formulation of the proposal density we use for the finite-sites model [see also *Approximating* $\pi(\alpha|H)$ *when* $\rho \neq 0$ and *Proposal density*; the mathematical formulation of the proposal density for an infinites sites model follows analogously]. We describe the procedure for simulating a single event in the ARG.

Consider an L locus model, with K alleles at each locus. As before, let P be the $K \times K$ mutation matrix, and let θ and ρ be, respectively, the mutation and recombination rates across the region of interest.

Assume that there are currently j branches in the ARG. For $i = 1, \ldots, j$ let a_i be the proportion of loci at which branch i is ancestral and b_i be the distance between the extreme ancestral loci of branch i. (If branch i has l ancestral loci at positions $x_1 < x_2, \cdots < x_l$, then $a_i = l/L$ and $b_i = x_l - x_1$.) For a haplotype α let j_α be the number of the j branches in the ARG that are of type α. (To have the same type requires the branch to have the same ancestral loci as α and at each ancestral locus to have the same allele as α.) Finally let H be the set of haplotypes of the j branches, and let $H - \alpha$ denote the set of $j - 1$ haplotypes obtained by removing haplotype α from H.

We simulate the next event in the ARG via a two-stage process:

1. Choose branch i with probability proportional to $(j - 1 + a_i\theta + b_i\rho)$. Denote the type of the chosen branch to be α.
2. Choose an event to occur to the chosen branch with the following probabilities. [Throughout, C is a normalizing constant, chosen so that the probability of all possible events sums to one; the probabilities, $p(\cdot|\cdot)$, are defined in *Approximating* $\pi(\alpha|H)$ *when* $\rho \neq 0$ and calculated as described in appendix a.]
   i. A coalescence with another branch of type α. This occurs with probability $C(j_\alpha - 1)/p(\alpha|H - \alpha)$.
   ii. A coalescence with a branch of type β. If γ is the haplotype produced by coalescing α with β, then this event occurs with probability

$$C j_\beta \left(\frac{p(\gamma|H - \alpha - \beta)}{p(\alpha|H - \alpha)p(\beta|H - \alpha - \beta)} \right).$$

Such an event can occur for any haplotype β that

shares the same alleles as α at common ancestral loci.

iii. A mutation at locus l, mutating allele α_l to β_l. If β is the new haplotype produced by this mutation, then this event occurs with probability

$$CP_{\beta_l\alpha_l}\theta\left(\frac{p(\beta|H-\alpha)}{p(\alpha|H-\alpha)}\right).$$

Such an event can occur for all loci l that are ancestral in α and for $\beta_l = 1, \ldots, K$.

iv. A recombination between two neighboring loci. If the positions of the loci are x_l and x_{l+1}, and the recombination produces haplotypes β and γ, then this event occurs with probability

$$C(x_{l+1} - x_l)\rho\left(\frac{p(\beta|H-\alpha)p(\gamma|H-\alpha+\beta)}{p(\alpha|H-\alpha)}\right).$$

Such an event can occur for $l = 1, \ldots, L-1$, providing x_l and x_{l+1} lie between (or are) the extreme ancestral loci. Conditional on such an event, the position of the recombination breakpoint is generated uniformly on the interval $[x_l, x_{l+1}]$.