ACADEMIC PRESS

Theoretical Population Biology 62 (2002) 339–348

Theoretical Population Biology

www.academicpress.com

A vectorized method of importance sampling with applications to models of mutation and migration

Montgomery Slatkin

Department of Integrative Biology, University of California, Berkeley, California 94720-3140, USA

Received 4 October 2001

Abstract

An importance-sampling method is presented for computing the likelihood of the configuration of population genetic data under general assumptions about population history and transitions among states. The configuration of the data is the number of chromosomes sampled that are in each of a finite set of states. Transitions among states are governed by a Markov chain with transition probabilities dependent on one or more parameters. The method assumes that the joint distribution of coalescence times of the underlying gene genealogy is independent of the genetic state of each lineage. Given a set of coalescence times, the probability that a pair of lineages is chosen to coalesce in each replicate is proportional to the contribution that the coalescence event makes to the probability of the data. This method can be applied to gene genealogies generated by the neutral coalescent process and to genealogies generated by other processes, such as a linear birth–death process which provides a good approximation to the dynamics of low-frequency alleles. Two applications are described. In the first, the fit of allele frequencies at two microsatellite loci sampled in a Sardinian population to the one-step mutation model is tested. The one-step model is rejected for one locus but not for the other. The second application is to low-frequency alleles in a geographically subdivided population. The geographic location is the allelic state, and the alleles are assumed to be sufficiently rare that their dynamics can be approximated by a linear birth–death process in which the birth and death rates are independent of geographic location. The analysis of eight low-frequency allozyme alleles found in the glaucous-winged gull, *Larus glaucescens*, illustrates how geographically restricted dispersal can be detected.

Keywords: Coalescent theory; Likelihood; Microsatellites; Gene flow

1. Introduction

The explosive growth of molecular analysis of genetic variation in human and other populations has led to the development of new statistical methods of data analysis. One class of methods calculates the likelihood of one or more population genetic parameters as a function of the observed configuration of data. Such methods make full use of data rather than relying on summary statistics and provide a statistical framework within which to test hypotheses and estimate parameters. Except in a few special cases, likelihoods cannot be computed analytically, so the focus of recent theoretical efforts has been on the development of efficient computer-intensive methods that rely on randomly generated replicates of population genetic models. At present, these methods fall into two categories: methods based on the Metropolis–Hastings algorithm (MH) and methods based on importance sampling (IS). Both classes of methods rely on coalescent theory, and both have been called Markov chain Monte Carlo (MCMC) methods, although some reserve MCMC for MH methods only. The important distinction between MH and IS methods is that different replicates are independent in IS methods and are correlated in MH methods. Felsenstein and his collaborators (Kuhner et al., 1995, 2000; Beerli and Felsenstein, 1999) initiated the use of MH methods in population genetics and have developed several programs that analyze a variety of processes, including population growth, migration and recombination. Recently, several other papers employing MH methods have appeared (e.g., Nielsen, 2000; Pritchard et al., 2000; Rannala and Reeve, 2001).

Griffiths and Tavaré (1994a,b) introduced an IS method that has been widely used. Their method, which they called an MCMC method, chooses among

E-mail address: slatkin@socrates.berkeley.edu (M. Slatkin).

0040-5809/02/$ - see front matter
PII: S0040-5809(02)00007-2

coalescence and mutation events based on prior probabilities of occurrence. Stephens and Donnelly (2000) examined the general theory of IS as applied to the neutral coalescent and introduced another IS method that is more efficient than the one proposed by Griffiths and Tavaré. The Stephens–Donnelly method uses an approximation to the posterior probabilities of coalescence and mutation events as a guide to sampling gene genealogies. In this paper, I will introduce still another IS method which differs from those of Griffiths and Tavaré and of Stephens and Donnelly. In this method, for each replicate a set of coalescence times is randomly generated and then, given the coalescence times, a gene genealogy is generated by non-randomly choosing among coalescence events based on the contribution each event makes to the overall likelihood. Associated with each branch of the gene genealogy is a vector whose elements are the probabilities of being each of the genetic states. Transitions among states on each branch are modeled by taking the appropriate power of the Markovian transition matrix. Generating coalescence times separately allows this method to be easily applied to models other than the neutral coalescent. Any process, such as a linear birth–death process, for generating a random set of coalescence times can be used. Representing the state of each lineage by a vector and employing efficient methods of matrix multiplication make it possible to allow for an arbitrarily large number of transitions on each branch with no increase in running time.

In this paper, I will introduce the general method and then apply it to two data sets. The first application is to two samples of microsatellite alleles in a human population. Microsatellite alleles are assumed to be neutral and the coalescence times are generated by a neutral coalescent model in an exponentially growing population. The goal is to determine whether the data can be accounted for by a mutation model that assumes a change in only one repeat unit each generation (the one-step model) or whether multiple steps must be allowed for. In this case, the genetic state is the number of repeat units of the microsatellite motif and transitions among states are governed by a mutation matrix that allows for changes in allele size by one or more repeat units.

The second application is to the numbers of allozyme alleles in a geographically subdivided population. In this case, a birth–death model is used to approximate the dynamics of a rare allele. Because the birth–death model assumes that each copy of the allele reproduces independently of all others, allelic reproduction is independent of geographic location. The genetic state is the geographic location of each copy and transitions among states are modeled by a migration matrix. In this example, the goal is to determine whether the data show evidence of geographically restricted dispersal, i.e., whether the migration pattern differs from an island model of migration.

2. Theory

2.1. The model

The data consist of the genetic states of n chromosomes in a sample. Each chromosome can be in one of d states, so the data set is a set of numbers $D = \{i_1, \ldots, i_n\}$, where $1 \leqslant i_j \leqslant d$. For example, for a microsatellite locus at which allele sizes differ by the number of repeats units, i_j is the number of repeats of the allele on chromosome j. For microsatellite loci, the total number of states may not be known but d can be chosen to be sufficiently large that its value does not affect the results. The model assumes a single genetic locus at which transitions occur independently on each chromosome.

The transition from one state to another on each lineage is described by a Markov chain with transition matrix $\mathbf{F}$ which has elements F_{ij}, the probabilities that a chromosome will be in state j in generation $t+1$ given that it is in state i in generation t. Time dependence of $\mathbf{F}$ can be incorporated if necessary but that possibility will not be considered here. The stationary distribution of $\mathbf{F}$, if it exists, will be denoted by the row vector π: i.e., $\pi\mathbf{F} = \pi$.

The mathematical problem is to find the probability of the data given $\mathbf{F}$ and assumptions about the population from which the sample is drawn. As Felsenstein (1988) has pointed out, this problem can be expressed as a summation over all gene genealogies with n tips:

$$\Pr(D) = \sum_G \Pr(D \mid \mathbf{F}, G) \Pr(G), \tag{1}$$

where G is a gene genealogy. $\Pr(G)$ is the probability of G for the population from which the sample is drawn. For neutral alleles, $\Pr(G)$ depends only on the demographic history of the population, while $\Pr(D \mid \mathbf{F}, G)$ depends on mutation and other processes that create diversity among lineages.

For small n and simple assumptions about a population, Eq. (1) can be evaluated by direct summation and integration, but for larger n that becomes impossible because of the rapidly increasing number of gene genealogies. The method introduced in this paper relies on separating the summation over genealogies into two summations, one over the set of coalescence times and the other over the set of topologies:

$$\Pr(D) = \sum_{\mathbf{t}} \sum_B \Pr(D \mid \mathbf{F}, \mathbf{t}, B) \Pr(\mathbf{t}) \Pr(B), \tag{2}$$

where $\mathbf{t} = \{t_n, t_{n-1}, \ldots, t_2\}$ is the set of coalescence times of a gene genealogy and B is the topology (or branching

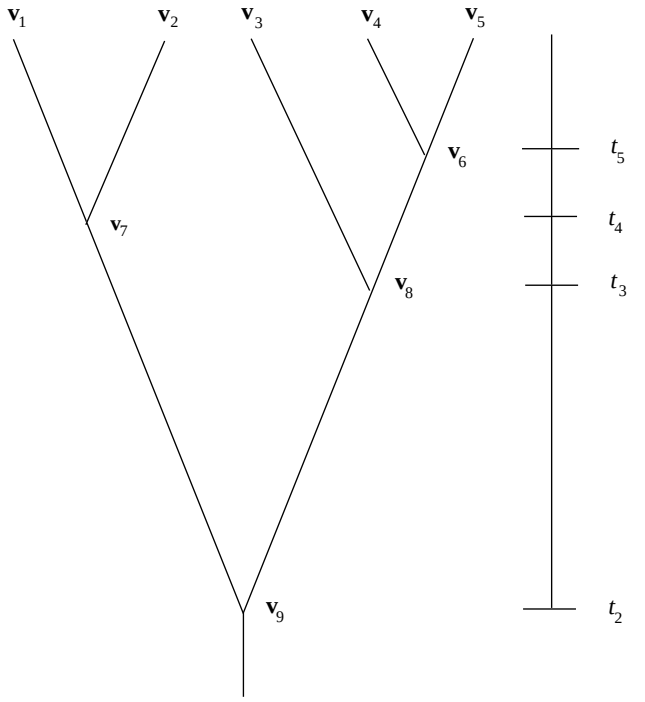

Fig. 1. Genealogy for a sample size of 5 illustrating the notation used in the text. The t_i are the coalescence times at which the number of lineages decreases from i to $i-1$. The $\mathbf{v}_i$ are the vectors indicating the probabilities that each lineage starts in one of the d allelic states.

pattern) of the genealogy. The time t_k is the time at which the number of descendent lineages increases from $k-1$ to k, as illustrated in Fig. 1.

2.2. Probability of the data

The first step is to apply the standard method of Felsenstein (1973) to compute the probability of the data, given the transition matrix, the coalescence times, and the topology $\Pr(D \mid \mathbf{F}, \mathbf{t}, B)$. I will describe this method in terms of vectors and matrices because that is the basis for the method of IS described later. A rooted genealogy with n tips has $2n-2$ branches, as illustrated in Fig. 1 for a tree with five tips. In this paper, the first n branches are the terminal branches. The internal branches will be numbered in increasing order as they are formed by coalescence events going backwards in time. With that convention, branch $n+1$ is created at t_n, branch $n+2$ is created at t_{n-1}, and so on. Branch $2n-1$ is created by the last coalescent event (at t_2) and represents the most recent common ancestor of the gene genealogy.

The branch lengths are also needed. Let u_i, be the length of branch i measured in generations. At the time branch i is formed, a vector $\mathbf{v}_i$ is associated with it. In what follows, the elements of $\mathbf{v}_i$ sum to 1, and they can be interpreted as being a vector of probabilities of being in each of the d states at the time the branch is formed. A vector $\mathbf{v}_{2n-1}$ is associated with the most recent common ancestor of the genealogy and represents the probabilities of the ancestral state. At the tips, $\mathbf{v}_1, \ldots, \mathbf{v}_n$ represent the data. In what follows the data will be assumed to be known perfectly, so the jth element of $\mathbf{v}_i$ will be 1 if that tip is in state j and 0 otherwise. The same formal theory can allow for the possibility that there is some uncertainty in the data, which might result from errors or intrinsic ambiguity in genotyping. In that case, $\mathbf{v}_1, \ldots, \mathbf{v}_n$ would indicate the probabilities of each tip being in each state.

At the other end of each branch, a vector $\mathbf{v}'_i = (\mathbf{F}^{\mathrm{T}})^{u_i}\mathbf{v}_i$ is associated, where $\mathbf{F}^{\mathrm{T}}$ is the transpose of $\mathbf{F}$. Note that $\mathbf{v}'_i$ is not necessarily normalized unless $\mathbf{F}$ is symmetric (i.e., if $\mathbf{F}^{\mathrm{T}} = \mathbf{F}$). When $\mathbf{F}$ is symmetric, $\mathbf{v}'_i$ can be interpreted as the vector of state probabilities immediately before that lineage joins another at a node. When $\mathbf{F}$ is not symmetric the $\mathbf{v}'_i$ do not have that meaning, but they are still the vectors that arise naturally in computing of the likelihood.

At each node, the vector $\mathbf{v}$ associated with the ancestral lineage is obtained by taking the Schur product and renormalizing so that the elements sum to 1. The jth element of the Schur product of vectors $\mathbf{x}$ and $\mathbf{y}$ is $x_j y_j$, where x_j and y_j are the jth elements of $\mathbf{x}$ and $\mathbf{y}$, and the normalization constant is the dot product $\mathbf{x} \cdot \mathbf{y} = \sum_{j=1}^{d} x_j y_j$. I will denote the normalized Schur product by $*$, so the jth element of $\mathbf{x} * \mathbf{y} = x_j y_j / (\mathbf{x} \cdot \mathbf{y})$. Although this notation is non-standard, it will simplify the description of the IS method presented later.

The probability of the data, i.e., the likelihood, is obtained by taking the dot product of the final Schur product with the assumed distribution of ancestral states. Often the ancestral distribution is assumed to be the stationary distribution, π, but in other cases it may be better to assume that the ancestral state is known.

To illustrate this method, assume that the data are represented by vectors $\mathbf{v}_1, \ldots, \mathbf{v}_5$ and the genealogy is as shown in Fig. 1. The probability of these data is obtained by working down the tree to obtain

$$\begin{aligned} \mathbf{v}_6 &= \mathbf{v}'_4 * \mathbf{v}'_5, \\ \mathbf{v}_7 &= \mathbf{v}'_1 * \mathbf{v}'_2, \\ \mathbf{v}_8 &= \mathbf{v}'_3 * \mathbf{v}'_6, \\ \mathbf{v}_9 &= \mathbf{v}'_7 * \mathbf{v}'_8, \end{aligned} \tag{3a}$$

and then multiplying the accumulated dot products,

$$\Pr(D \mid \mathbf{F}, \mathbf{t}, B) = (\mathbf{v}'_4 \cdot \mathbf{v}'_5)(\mathbf{v}'_1 \cdot \mathbf{v}'_2)(\mathbf{v}'_3 \cdot \mathbf{v}'_6)(\mathbf{v}'_7 \cdot \mathbf{v}'_8)(\mathbf{v}_9 \cdot \pi), \tag{3b}$$

if it is assumed that the ancestral state is randomly drawn from the stationary distribution.

In a phylogenetic context, each tip of a genealogy represents a different taxon, and the data vector associated with a tip represents the observed state of that taxon. The population genetic problem is different

because the assignment of the data vectors to tips is arbitrary. The algorithm described above gives the probability for a particular assignment of data to tips. To compute the overall probability of the data, the result for a single assignment must be multiplied by the number of distinguishable rearrangements of the data on the tips. The problem is the same as the one that arises in the derivation of the Ewens sampling formula (Ewens, 1972). For a given data set, α_k is the multiplicity of state k, and the number of distinguishable configurations of the data is

$$C_D = \frac{n!}{\alpha_1!\alpha_2!\cdots\alpha_n!}.$$

For example, one of the data sets analyzed later is of the number of repeat units of a microsatellite locus in a sample of size 10. The data vector is $\{8, 11, 11, 11, 11, 12, 12, 12, 12, 13\}$, where the numbers are the numbers of repeat units of each allele in the sample. In this case, $C_D = 6300$. To obtain the probability of the data, the probability obtained from a single assignment of the data to the tips, as calculated by the algorithm described here, has to be multiplied by C_D. When the probability of the data is used for the estimation of parameters by maximum likelihood, the value of C_D does not matter, but it is needed when results are compared to those obtained from other methods (cf. Fig. 2).

2.3. Generation of coalescence times

The joint distribution of coalescence times depends on what process is assumed to have generated the data. In this paper, I will consider two possibilities: the neutral coalescent and a linear birth–death process. For the neutral coalescent, a scaled time τ is defined by

$$\tau(t) = \int_0^t \frac{dt'}{2N(t')}. \tag{4}$$

In terms of τ, the distribution of times during which k ancestral lineages are present is exponential with mean $2/(k(k-1))$ and the distributions for different values of k are independent (Griffiths and Tavaré, 1994a). A random set of coalescence times is generated by drawing $(\tau_{k-1} - \tau_k)$ from an exponential distribution with the appropriate mean and then transforming the resulting τ_k back to the natural time scale using the inverse of Eq. (4).

A linear birth–death process provides an accurate approximation to the dynamics of a low-frequency allele, even if there is selection in favor of or against heterozygous carriers. Wiuf (2001) shows that if heterozygous carriers of an allele have a relative fitness $1+s$ compared to individuals lacking that allele, then a birth–death process with birth rate $\lambda = s + r + \frac{1}{2}$ and death rate $\mu = \frac{1}{2}$ approximates the dynamics of that allele in a

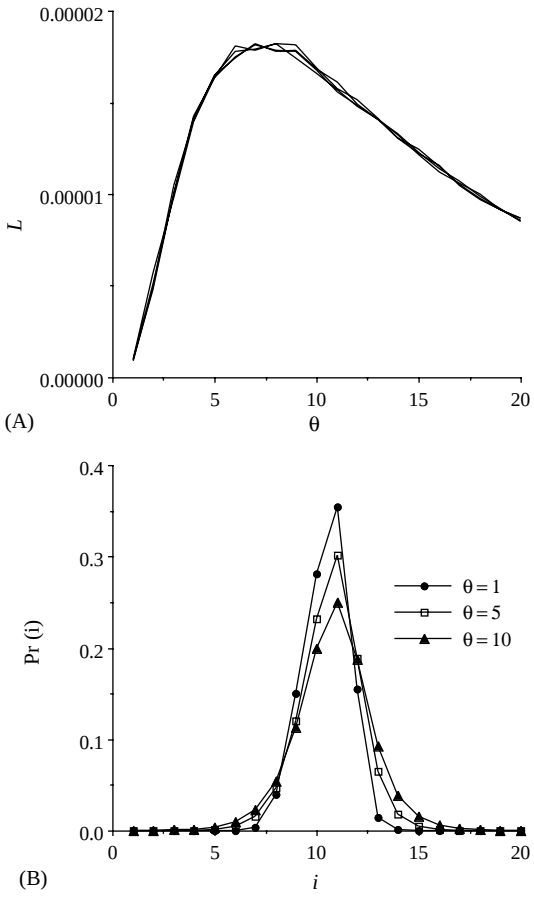

Fig. 2. (A) Likelihood curves for the sample microsatellite data set $\{8, 11, 11, 11, 11, 12, 12, 12, 12, 13\}$, assuming a one-step mutation model. Each curve is based on 10,000 replicates of the IS method described in the text. The results are not smoothed. These results are comparable to those of Stephens and Donnelly (2000, Fig. 3) if their results are multiplied by $\frac{1}{20}$, a factor that accounts for the assumed uniform initial probability. (B) Posterior distribution of i, the allelic state at t_2, under the assumption that the prior distribution is uniform on the integers between 1 and 20 (which was assumed in (A)).

population that has undergone exponential growth at rate r. If t_1 is the time at which the allele arose by mutation (i.e., the allele age), then the coalescence times are generated by drawing $n-1$ random variables independently from the kernel distribution,

$$b(t) = \frac{[P(0,t)]^2 e^{-\xi t_1}}{2[f - P(0,t_1)e^{-\xi t_1}]}, \tag{5}$$

on the interval $(0, t_1)$ where $\xi = r + s$ and $P(0,t) = 2f\xi/[f - (f - 2\xi)e^{-\xi t_1}]$, and arranging them in decreasing order (Slatkin, 2002).

There are two possibilities for choosing t_1. It may be given a specific value as it is, for example, in the case of linkage disequilibrium mapping of an allele found in an isolated population founded at time t_1 (Rannala and Slatkin, 1998). In that case, one copy of the causative allele is assumed to have been present in the population at the time of founding t_1 generations in the past. The

alternative is to assume that t_1 is randomly drawn from a prior distribution that depends on r, s, and the allele frequency. Wiuf (2001) and I (Slatkin, 2002) provide derivations of the prior distribution for a linear birth–death process. In the general case,

$$\Pr(t_1) = C\frac{e^{-rt_1}E(1-E)^{n-1}}{[f-(f-2\xi)E]^{n+1}} \tag{6a}$$

if $\xi = r + s > 0$, and

$$\Pr(t_1) = C\frac{e^{-rt_1}E(E-1)^{n-1}}{[E(f-2\xi)-f]^{n+1}} \tag{6b}$$

if $\xi < 0$, where n is the number of copies of the allele found in a sample, $E = e^{-\xi t_1}$, f is the fraction of the population sampled, and C is a normalization constant that can be expressed in terms of hypergeometric functions. The distributions for special cases ($r = 0$, $r + s = 0$, and $r = s = 0$) are obtained by taking the appropriate limits. There is an efficient rejection method to generate a random t_1 (Slatkin, 2002).

When applying this theory to a sample of N_{sam} individuals ($2N_{sam}$ chromosomes) from a population of N individuals, $f = N_{sam}/N$ and the allele frequency is $n/(2N_{sam})$. In most applications, N and N_{sam} are not known precisely and the analysis has to be done for a range of possible values of f. It is not necessary to assume that f is small although it usually is. If f is small, the $\Pr(t_1)$ and the joint distribution of coalescence times depend only weakly on f so any uncertainty in its values will not usually affect the results, although that has to be checked in each case.

The linear birth–death process also describes the numbers of copies of a low-frequency allele in a geographically subdivided population, provided that the same values of f, s, and r are appropriate for each subpopulation. Population subdivision can be ignored in this case because, when an allele is in low frequency, each copy replicates and survives independently of each other copy and follows the same rules in each subpopulation. Therefore, the geographic location of each copy does not affect the number of descendent copies. That restriction does not require that the subpopulations be of equal size or that migration among them is symmetric or conservative. One of the applications of the methods developed later is to low-frequency alleles found in geographically subdivided populations. In that case, the state of each copy is its geographic location and the transition matrix $\mathbf{F}$ is the migration matrix.

2.4. Sampling of topologies

One way of sampling topologies is random sampling (RS) in which topologies are generated by assuming that, when a coalescence occurs, each pair of lineages is equally likely to coalesce. With RS, the sum in Eq. (2) is computed approximately by generating for replicate h of H replicates a random set of coalescence times $\mathbf{t}_h$ and a randomly generated topology B_h, and then averaging over replicates:

$$\Pr(D) \approx \frac{1}{H}\sum_{h=1}^{H}\Pr(D \mid \mathbf{F}, \mathbf{t}_h, B_h). \tag{7}$$

RS performs well for small values of n because all topologies will be generated sufficiently often that all terms in the summand will be adequately represented. With larger values of n, however, RS does not perform well because relatively few topologies contribute significantly to the sum in Eq. (2), and RS does not find those topologies often enough to provide a good approximation.

An alternative to RS is IS. The idea is to sample topologies in such a way that those for which $\Pr(D \mid \mathbf{F}, \mathbf{t}, B)$ is largest and hence contribute most to the sum in Eq. (2) are sampled most often. When IS is used, the results for each replicate must be weighted by a factor that accounts for non-random sampling of topologies:

$$\Pr(D) \approx \frac{1}{H}\sum_{h=1}^{H} w_h \Pr(D \mid \mathbf{F}, \mathbf{t}_h, B_h). \tag{8}$$

As required by the general theory of IS (Stephens and Donnelly, 2000), the weighting factor has to be

$$w_h = \frac{\Pr_{RS}(B_h)}{\Pr_{IS}(B_h)}, \tag{9}$$

where the numerator is the probability of B_h under RS and the denominator is the probability under whatever method of importance sample is used.

To choose a way of non-randomly sampling the coalescence events, I use as a guide the algorithm for calculating the probability of the data. First, a set of coalescence times, $\mathbf{t}$, is generated. At t_k, each of the $k(k-1)/2$ pairs of lineages is considered. For lineage j ($1 \leqslant j \leqslant k$), the vector $\mathbf{v}'_j = (\mathbf{F}^{\mathrm{T}})^{u_j}\mathbf{v}_j$ is computed, where u_j is the length of branch j. In this discussion, j numbers the lineages present between t_{k+1} and t_k. The probability that lineages j and j' are chosen to coalesce is

$$\Pr(j, j') = \frac{\mathbf{v}'_j \cdot \mathbf{v}'_{j'}}{\sum_{j=1}^{k}\sum_{j'=1}^{j-1}(\mathbf{v}'_j \cdot \mathbf{v}'_{j'})}. \tag{10}$$

That is, the probability that a pair of lineages is chosen to coalesce is proportional to the contribution that the coalescence of those two lineages would make to the probability of the data (cf. Eq. (3)). When a coalescence occurs, the normalized Schur product $\mathbf{v}'_j * \mathbf{v}'_{j'}$ defined above is assigned to the ancestral lineage.

This algorithm provides a way to non-randomly sample the space of topologies. The weighting factor, w_r, in Eq. (9) is the product of weights from each of the

$n-1$ coalescent events:

$$w_r = \prod_{k=2}^{n} w_{r,k}, \tag{11}$$

where

$$w_{r,k} = \frac{2/k(k-1)}{(\mathbf{v}'_l \cdot \mathbf{v}'_l)/\sum_{l=1}^{k}\sum_{l'=1}^{l-1}(\mathbf{v}'_l \cdot \mathbf{v}'_{l'})} \tag{12}$$

and l and l' are the lineages that actually coalesce.

This method of IS is computationally convenient because the dot product in the denominator of Eq. (12), $\mathbf{v}_j \cdot \mathbf{v}_{j'}$, cancels the same dot product that appears in the calculation of $\Pr(D \mid \mathbf{F}, \mathbf{t}, B)$ (Eq. (3)). Furthermore, the product of $2/(k(k-1))$ is the same for every replicate, so for each replicate it is necessary only to multiply the terms corresponding to the double summation that appears in Eq. (12) for each node in the genealogy. Consequently, the calculations proceed relatively quickly and with little risk of overflow or rounding error.

The running time of this method increases rapidly with n, the sample size. For large n, it can become prohibitively slow. Running time can be decreased substantially by placing an upper limit on the number of lineages tested at each coalescent event, which I call the *span* S of the simulation. At each coalescent event, S lineages are chosen at random, all pairs of them are tested, and one pair is chosen to coalesce. If the number of lineages remaining is equal to or less than S, then all pairs are tested. The weight in Eq. (12) needs to be adjusted to account for this change in the algorithm.

3. Applications

The method described in the previous section provides a way to approximate the probability of the data as a function of parameters, i.e., the likelihood, from which a maximum likelihood estimate and support interval can be obtained. If a prior distribution of the parameter or parameters is assumed, then the likelihood provides the basis for computing the posterior distribution and carrying out a Bayesian analysis.

3.1. Microsatellite loci

At a microsatellite locus, alleles are distinguished by the number of tandemly repeated copies of a 2–6 base pair motif. The number of repeats is the allelic state. In principle, the number of states is infinite but, in practice, relatively few states are found and the total number of states, d, can be chosen to be large enough that its value does not affect the results. A simple model of mutation at a microsatellite locus and the one that is usually assumed is the symmetric one-step model: in each generation, an allele has a probability $\mu/2$ of increasing or decreasing in size by one repeat unit and a probability $1-\mu$ of remaining the same size. Other mutation models that relax the assumption of symmetry or the assumption of changes by only one repeat unit have been proposed.

To illustrate this method developed in this paper and to test its performance, I will reanalyze a small data set analyzed by Stephens and Donnelly (2000), a simulated data set of size 10, $\{8, 11, 11, 11, 11, 12, 12, 12, 12, 13\}$. They assumed a population of constant size, so the only free parameter is the product of the mutation rate and the population size, $\theta = 4N\mu$. Stephens and Donnelly (2000) in their Fig. 3 showed that their method produced likelihood curves that differed only slightly from one set of 10,000 replicates to another and that were nearly the same as a likelihood curve based on 10^7 replicates. In contrast, the IS method of Griffiths and Tavaré (1994b), which was first implemented by Nielsen (1997) for the analysis of microsatellite loci, yielded likelihood curves that differed substantially from one set of 10,000 replicates to another. Fig. 2A shows that the IS method described in this paper results in likelihood curves that are comparable to those obtained by Stephens and Donnelly. The weighted average of the ancestral vector provides the likelihood of the ancestral type for each θ, as shown in Fig. 2B. The simulations for each set of 10,000 replicates took roughly $1\frac{1}{2}$ min on a 550 MHz PC running Linux.

One question that can be asked of a microsatellite locus is whether the one-step mutation model is valid. As an alternative to the one-step model, I analyzed a symmetric geometric model in which the probability of an increase or decrease by i repeat units is $\mu(1-\alpha)\alpha^{i-1}/2$ if $0<\alpha<1$, which is interpreted as the one-step model when $\alpha = 0$.

To illustrate the use of the geometric mutation model, I reanalyzed data from two loci of the 10 examined by Di Rienzo et al. (1994) in a sample of 50 individuals from Sardinia. One locus, STS 287, was typed on 88 chromosomes. The data can be represented by the vector $\{41, 14, 27, 6\}$, meaning that 41 copies had the minimum number of repeats observed, 14 had one repeat more than the minimum, 27 had two repeats more, and six had three repeats more. There were no gaps in the distribution of allele sizes. The absolute number of repeats does not matter in this case because the geometric mutation model does not depend on the absolute repeat number. The other locus was AFM 158 which was typed on 96 chromosomes. The configuration was $\{36, 0, 42, 0, 8, 0, 0, 0, 10\}$. In analyzing both data sets, I added 5 states lower than the smallest repeat number and 5 larger than the largest repeat number in order to minimize the effect of the limits allele size, so $d = 14$ for STS 287 and $d = 19$ for AFM 158. The population of Sardinia is typical of European populations in that it shows evidence of past population

growth. In this analysis, I assumed a current population size of $N_0 = 10^7$ and a past rate of exponential growth of $r = 0.01$. The value of N_0 does not matter because, in the analysis, only the product $N_0\mu$ affects the results.

Fig. 3 shows the likelihood as a function of $\theta = 4N_0\mu$ for $\alpha = 0$, 0.1 and 0.2 for STS 287. There is little difference between the likelihood curves for $\alpha = 0$ and 0.1 but the curve for $\alpha = 0.2$ is substantially lower. These results suggest that the one-step mutation model is sufficient to account for these data and that it is possible to reject a model that assumes frequent multi-step mutations. Fig. 4, on the other hand, leads to a different conclusion for AFM 158. Smaller values of α (0, 0.2 and 0.4) result in much smaller values of the likelihood. Although there is no power to distinguish between $\alpha = 0.6$ and 0.8, this analysis strongly suggests that when a mutation occurs, changes in allele size by more than one step are relatively common. These results also suggest that θ is smaller for AFM 158 than for STS 287.

The likelihood curves shown in Figs. 3 and 4 are not as pleasingly smooth as those in Fig. 2A because of the larger sample sizes analyzed. Fig. 5 shows the variability among sets of replicates for $\alpha = 0$ for STS 287 (part A) and $\alpha = 0.6$ for AFM 158 (part B). Although there is considerable variation among sets of replicates, the underlying similarity provides confidence in the conclusions. These runs are relatively slow, roughly 42 h for 200,000 replicates for STS 287 to draw each curve and 48 h for 100,000 replicates for each curve for AFM 158, which is somewhat slower because d was larger. Using

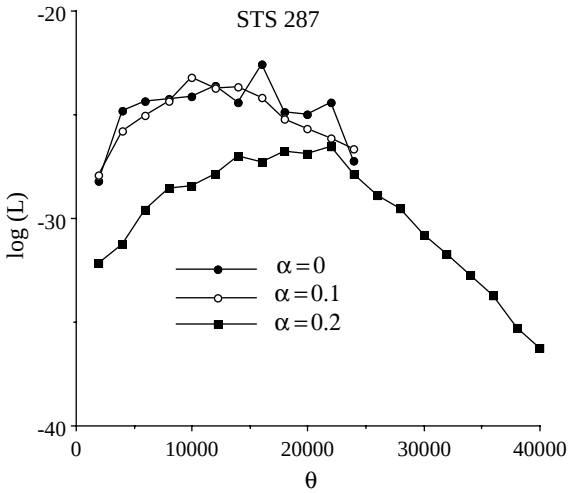

Fig. 3. Log-likelihood of the microsatellite data set $\{41, 14, 27, 6\}$ for the locus STS 287 studied by Di Rienzo et al. (1994) in a sample of Sardinians. The geometric mutation model with parameters $\theta = 4N_0\mu$ and α was assumed. The values for each point shown were obtained by averaging results from 200,000 replicates of the IS method (with a span of 30) described in the text. The demographic model assumed a current effective population size of $N_0 = 10^7$ and a past rate of exponential growth of $r = 0.01$.

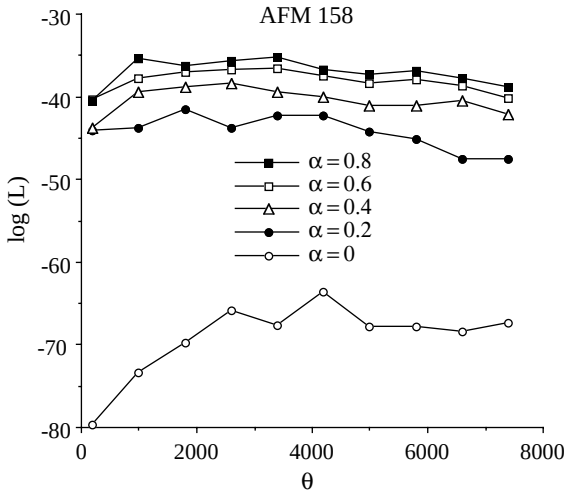

Fig. 4. Log-likelihood of the microsatellite data set $\{36, 0, 42, 0, 8, 0, 0, 0, 10\}$ for the locus AFM 158 studied by Di Rienzo et al. (1994) in a sample of Sardinians. As in Fig. 3, the geometric mutation model with parameters $\theta = 4N_0\mu$ and α and an exponential model of population growth with $N_0 = 10^7$ and $r = 0.01$ were assumed. The values for each point shown were obtained by averaging results from 100,000 replicates of the IS method (with a span of 30) described in the text.

computers and compilers optimized for floating point vector calculations could decrease the running time substantially.

I also examined how varying the span of the method affected the results. The span, as defined above, is the maximum number of lineages tested for possible coalescence. Fig. 6 shows the results for 100,000 replicates on AFM 158 for the case with $\alpha = 0.6$. These runs were made on a slightly faster computer (731 MHz). They took 13.46 h for a span of 10, 26.42 h for a span of 20, 38.74 h for a span of 30, and 50.04 h for a span of 40. Thus, the span has a major effect on the running time for this sample size ($n = 96$). Although there is considerable variability among the results, the variability among these four sets of 100,000 replicates is not obviously greater than among the four sets shown in Fig. 5B, which all used a span of 30. Therefore, with larger sample sizes, it seems best to use a small span for most cases. Conclusions based on those simulations can be verified by running a few cases with larger spans.

3.2. Migration rates and pattern in a subdivided population

If a population is divided into several discrete subpopulations, the geographic location of an allele can be regarded as its state and the migration matrix can be regarded as the transition matrix. The dynamics of an allele in low frequency can be approximated by a linear birth–death process even if the allele is not selectively

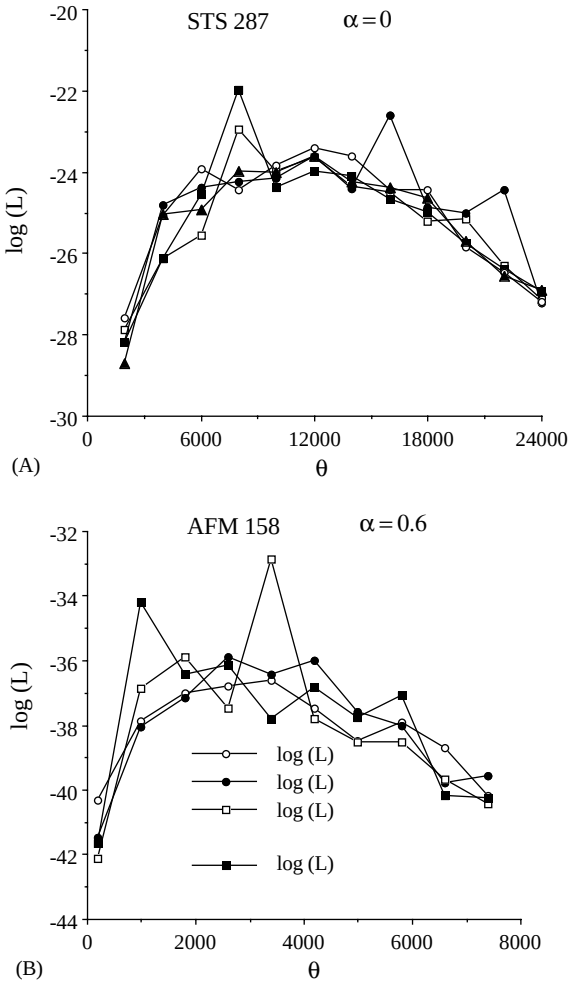

Fig. 5. (A) Four replicate sets of results for one of the cases shown in Fig. 3 for STS 287. Each point is based on the average of 200,000 replicates. (B) Four replicate sets of results for one of the cases shown in Fig. 4 for AFM 158. Each point is based on the average of 100,000 replicates with a span of 30.

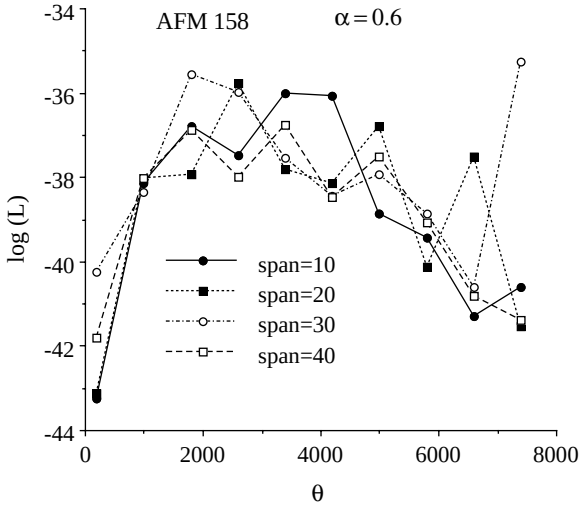

Fig. 6. Log-likelihood curves for AFM 158 with $\alpha = 0.6$ for different values of the span, which is the maximum number of lineages tested at each coalescence event. Each point is based on the average of 100,000 replicates.

neutral (Wiuf, 2001). The underlying assumption of a linear birth–death process is that each copy of the allele reproduces independently of all other copies. Therefore, each allele reproduces independently of the subpopulation in which it is present, provided that the growth rate and selection intensity are the same in each subpopulation and that the same fraction of each subpopulation has been sampled. Consequently, the distribution of intra-allelic coalescence times is independent of the migration matrix and the IS method described above can be used.

I will illustrate the use of this method by reanalyzing allozyme data from the study by Bell (1992, 1996) of the glaucous-winged gull, *Larus glaucescens*. Bell sampled 33 populations on the western coast of North America. My previous analysis of these data showed that pairwise estimates of F_{ST} exhibited a pattern of isolation by distance, suggesting that dispersal is geographically restricted (Slatkin, 1993). For application of the IS method in this paper, I chose the eight alleles that were found in 10 to 105 copies overall. These limits were chosen to ensure that alleles were common enough to be somewhat informative but rare enough that the birth–death approximation is valid. Table 1 shows the data for these eight alleles. All alleles except one at *CK*2 had frequencies between 0.167 and 0.333 in at least one subpopulation. The original data set is presented by Bell (1992) and a file containing the data was kindly provided by D.A. Bell.

In this example, subpopulations were treated as if they were evenly spaced and of equal size, something that is not true of the actual sampling locations in Bell's study. The intention is to illustrate the method in a simple context rather than to draw strong conclusions about dispersal tendencies of the glaucous-winged gull. Each allele was analyzed separately, using a geometric model of dispersal that is algebraically the same as the geometric model of mutation used for microsatellite loci. In this case, the number of states is the number of subpopulations sampled, $d = 33$. The model with $\alpha = 0$ represents the one-dimensional stepping-stone model with dispersal at rate $m/2$ between adjacent subpopulations. The model with $\alpha = 1$ corresponds to an island model of dispersal with a probability $1 - m$ of not dispersing and a probability $m/(d - 1)$ of dispersing to each other subpopulation. In this case, the question is whether these data allow rejection of the hypothesis that $\alpha = 1$, meaning that dispersal is geographically restricted.

Fig. 7 shows the results for each allele separately for $\alpha = 0.2$. To generate the intra-allelic coalescence times, it was necessary to assume a value of f, the fraction of the

Table 1
List of allele counts for eight alleles found in Bell's (1992) allozyme study of the Glaucous-Winged Gull, *Larus glaucescens*

ada	3	3	6	0	2	2	1	0	2	1	4	0	1	1	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ck2	0	0	0	0	0	0	0	0	2	2	3	0	0	2	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
eap1	0	0	0	0	0	0	0	2	3	5	6	1	1	8	2	5	6	5	7	2	6	2	8	6	1	3	3	1	5	1	13	3	0
est1	4	0	0	1	1	0	0	0	2	1	0	0	1	0	1	0	1	3	1	9	0	2	2	1	2	1	5	2	3	6	21	5	5
gpd	0	2	6	2	4	5	2	5	2	0	5	2	5	5	1	3	1	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0
icd1	1	0	3	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	10	1	2	0	0	0	5	0
pgm	5	0	1	2	1	0	1	1	2	4	4	0	1	0	0	1	0	0	1	1	0	3	4	1	0	1	1	1	0	0	3	1	0
pgm	2	3	3	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Note. These eight alleles were the only ones that had counts between 10 and 105. The numbers indicate the numbers of copies found at each of the 33 sampling locations, listed in the same order as in Bell's (1992) thesis. These data were extracted from the full data set that was kindly provided in electronic form by Bell. The locus names indicate the loci at which each allele was found. These data were analyzed using the importance sampling method described in the text to produce the likelihood curves shown in Figs. 7 and 8.

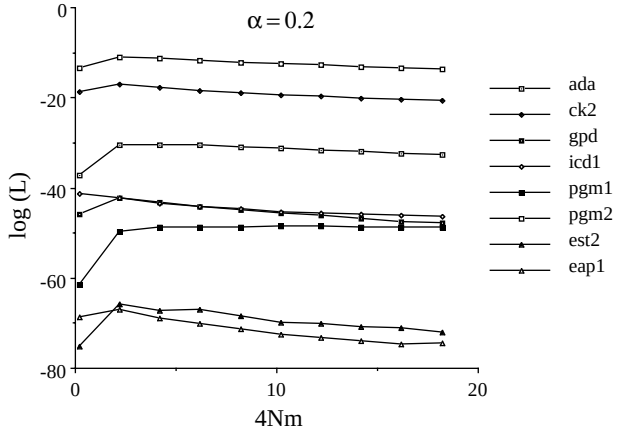

Fig. 7. Log-likelihood curves for each of the eight low-frequency alleles listed in Table 1, taken from the allozyme survey of the glaucous-winged gull, *Larus glaucescens*, of Bell (1992). The population was assumed to be of constant size and the alleles were assumed to be neutral and a fraction $f = 0.001$ of each subpopulation was assumed sampled. Results are plotted as a function of the equivalent value of $4Nm$, as explained in the text.

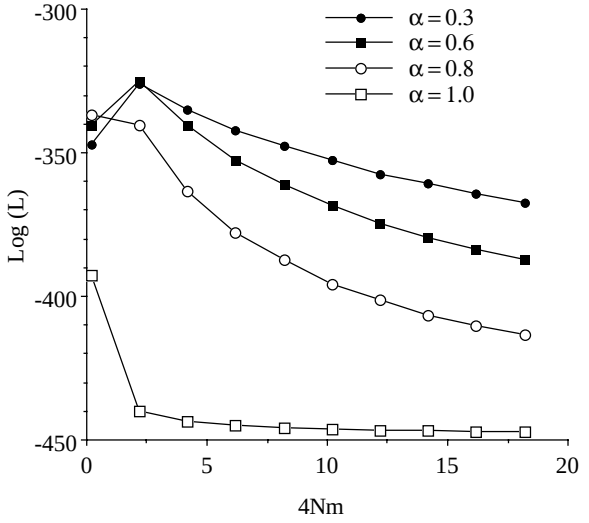

Fig. 8. Log-likelihood curves obtained for each value of α by adding log-likelihood values for each of the eight alleles listed in Table 1.

population sampled. The value of f is $n/(2N)$, where n is the number of copies of each allele and N is the number of individuals in each subpopulation (assumed to be the same for the analysis here). The value of N is unknown, but I have shown elsewhere (Slatkin, 2002) that the joint distribution of intra-allelic coalescence times depends only on the ratio f/n, which is $1/(2N)$ in this case. Hence, the results depend only on the product Nm, as is in other models of gene flow and genetic drift. The results in Figs. 7 and 8 are expressed as functions of Nm without having to assume a particular value of either f or N.

As Fig. 7 shows, there is little information in each allele, but because rare alleles are approximately independent of one another, the likelihoods can be multiplied to provide an overall likelihood, as shown in Fig. 8. Based on these data, there is support for a model of restricted gene flow because we can reject $\alpha = 1$ and 0.8. The rough estimate of the migration rates ($4Nm$) is consistent with estimates based on F_{ST} (Slatkin, 1993).

4. Discussion and conclusion

In many applications of population genetics theory to genetic data, the goal is to compute the likelihood of one or more parameters as a function of the data. The method presented in this paper employs importance

sampling (IS) to allow efficient approximation of the likelihood in cases in which numerous transitions among genetic states can occur. In choosing a computer-intensive method for approximating the likelihood, there are three considerations: confidence in the results, running time, and ease of implementation. At present, neither IS nor MH methods can ensure complete confidence. Convergence with increasing numbers of replicates and low variation among sets of replicates for the same parameter values usually indicate that accurate results have been obtained, but it is still possible that rare but important sample paths have been consistently missed. Conclusions obtained from application of any computer-intensive method, including the one described here, must be regarded as tentative.

The IS method presented here has the advantage of ease of implementation, but at the expense of longer running times than methods tailored to particular problems. A transition matrix for any type of Markov chain can be substituted for the transition matrix based on the geometric model used in the two examples. The averaging over coalescence times and the IS of genealogies proceeds in the same way in all cases. In practical terms, other models are analyzed by modifying the procedure in the C program that I freely distribute that calculates the elements of the Markovian transition matrix, **F**. The running speed does not depend on the complexity of the transition matrix because the matrix multiplication step relies on the diagonalization of the transition matrix, which is done only once for each set of parameter values.

As I have implemented this method, its application is limited in practice to sample sizes of 100 or smaller when the number of genetic states is on the order of 20. It is possible that a more efficient implementation could be made by using the same set of simulated topologies for more than one set of parameter values and that faster running times could be achieved by using a programming environment optimized for vector and matrix calculations. The goal of this paper has been to introduce a new method of IS and illustrate its application to two problems of biological interest.

Acknowledgments

This research was supported in part by NIH Grant R01-GM40282. I thank J. Felsenstein and M. Stephens for helpful discussions of this topic, and A. Di Rienzo and D.A. Bell for permission to use their data. D.A. Bell and two referees made helpful comments on an earlier version of this paper.

References

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

Bell, D.A., 1992. Hybridization and sympatry in the Western Gull/Glaucous-winged Gull Complex. Ph.D. Thesis, University of California, Berkeley.

Bell, D.A., 1996. Genetic differentiation, geographic variation and hybridization in gulls of the Larus glaucescens-occidentalis complex. Condor 98, 527–546.

Di Rienzo, A.D., Peterson, A.C., Garza, J.C., Valdes, A.M., Slatkin, M., et al., 1994. Mutational presses of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA 91, 3166–3170.

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

Felsenstein, J., 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22, 240–249.

Felsenstein, J., 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22, 521–565.

Griffiths, R.C., Tavaré, S., 1994a. Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. London B 344, 403–410.

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

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

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

Nielsen, R., 1997. A likelihood approach to populations samples of microsatellite alleles. Genetics 146, 711–716.

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

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Rannala, B., Reeve, J.P., 2001. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. Am. J. Hum. Genet. 69, 159–178.

Rannala, B., Slatkin, M., 1998. Likelihood analysis of disequilibrium mapping, and related problems. Am. J. Hum. Genet. 62, 459–473.

Slatkin, M., 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution 47, 264–279.

Slatkin, M., 2002. The age of alleles. In: Slatkin, M., Veville, M. (Eds.), Modern Developments in Theoretical Population Genetics. Oxford University Press, Oxford, pp. 233–259.

Stephens, M., Donnelly, P., 2000. Inference in molecular population genetics. J. R. Stat. Soc. B 62, 605–635.

Wiuf, C., 2001. Rare alleles with selection. Theor. Popul. Biol. 59, 287–296.