Theoretical Population Biology **62**, 215–229 (2002)
doi:10.1006/tpbi.2002.1601

TPB

Gene Mapping via the Ancestral Recombination Graph

Fabrice Larribe and Sabin Lessard

Département de Mathématiques et de Statistique, Université de Montréal, C.P. 6128, Succursale Centre-Ville, Montréal, Qué bec, Canada, H3C 3J7
E-mail: larribe@dms.umontreal.ca, lessards@dms.umontreal.ca

and

Nicholas J. Schork

Department of Epidemiology and Biostatistics, Case Western Reserve University, 2500 Metro-Health Drive, Cleveland, Ohio 44109-1998
E-mail: njs2@po.cwru.edu

Received December 26, 2001

We present a multilocus gene mapping method based on linkage disequilibrium, which uses the ancestral recombination graph to model the history of sequences that may harbor an influential variant. We describe the construction of a recurrence equation used to make inferences about the location of a trait-influencing mutation. We demonstrate how a Monte Carlo algorithm combined with a local importance sampling scheme can be used for mapping. We explain how to simulate the timing of events in the coalescent in the presence of recombination and mutation, which accomodates variable population size. We provide an example to illustrate the use of the method, which can be easily extended to more general situations. Although the method is computationally intensive and variation in the likelihood profiles can occur, the method offers a great deal of promise. © 2002 Elsevier Science (USA)

Key Words: **ancestral recombination graph (ARG); linkage disequilibrium; gene mapping; variable population size.**

1. INTRODUCTION

Some methods using linkage disequilibrium for fine mapping purposes have recently been proposed (Xiong and Guo, 1997; Rannala and Slatkin, 1998; Graham and Thompson, 1998; McPeek and Strahs, 1999; Service *et al*., 1999; Morris *et al*., 2000; Lam *et al*., 2000). Some of these methods, like those proposed by Xiong and Guo (1997), use pairwise statistics to form a likelihood and assume independence between marker loci. Others methods, such as those developed by Graham and Thompson (1998), Morris *et al*. (2000) and Rannala and Slatkin (1998), model the ancestry of sequences in a way that allows dependence between loci. The method of Rannala and Slatkin (1998) is of particular interest because of its direct use of a coalescent process to model the evolution of sequences. Although this method takes into account recombination in modeling the history of the sequences, it does assume elements of a standard coalescent process, i.e., without recombination. It is therefore of interest to design a gene mapping method that considers both coalescence events and recombination events, in order to derive a more powerful method which can work on larger chromosomal regions. In fact, this idea has begun to take shape recently in the literature (see, e.g., Marjoram *et al*., 2000).

0040-5809/02 $35.00

In this paper, we propose a method for mapping a gene influencing a certain trait by modeling the history of a sample of sequences with an ancestral recombination graph (Griffiths and Marjoram, 1996a, b). This is a true multilocus mapping method. We review the basics of the coalescent process and its extension to the ancestral recombination graph, and describe the construction of a recurrence equation used to make inferences about the location of the putative disease gene. More details about the coalescent and extensions taking into account recombination can be found in works of Griffiths (1981, 1991), and also Griffiths and Marjoram (1996b) who consider the problem from a more theoretical point of view.

2. THE ANCESTRAL RECOMBINATION GRAPH

The coalescent process (Kingman, 1982) provides a way to model the ancestry of DNA sequences without recombination. It should be understand that by "sequence" we mean an actual ordered set of DNA fragments or markers, some of which may be variable in the population at large. Both mutation and coalescence events can be considered and the outcome be represented by a tree. Excellent reviews of the subject are given by Hudson (1990), Nordborg and Tavaré (2002) and Nordborg (2001). Inference using the coalescent has been discussed by Griffiths (1989) and Griffiths and Tavaré (1994a–c, 1996, 1997). More recently, Stephens and Donnelly (2000) have worked to develop efficient importance sampling algorithms in the standard coalescent, followed by Fearnhead and Donnelly (2001) in the case of the coalescent with recombination.

An extension of the traditional coalescent, termed the ancestral recombination graph (ARG), accounts for recombination (Re), and it has been described by Griffiths and Marjoram (1996a, b). When recombination is considered in a coalescent model, a sequence is modeled as having one or two parental sequences in the previous generation (actually, two if Re occurred and one otherwise), and thus we have to consider a graph instead of a tree. Notice that considering recombination in modeling the evolution of a set of sequences is particularly important if one wants to develop a tool for gene mapping because recombination is fundamental in shuffling DNA from sequence to sequence as chromosomes are transmitted from generation to generation.

Assume a population of $2N$ sequences that are evolving in accordance with the Wright–Fisher model; in particular, the population size is assumed constant, generations are discrete, non-overlapping, and mating is at random. The ancestry of a sample of sequences is modeled back in time, starting from the current sample and until the most recent common ancestor (MRCA) of the sample is found. At each step in the graph, one of the following events can occur: (1) two sequences coalesce if they share a common ancestor; (2) one sequence mutates and then the genetic material at a single marker locus in one sequence is changed; or (3) one sequence recombines. When a recombination event occurs, a point of recombination is chosen randomly from a given distribution, and the sequence is separated into two parts coming from two parental sequences called the "left" and "right" parental sequences: the left parental sequence has the genetic material of the "child" from the beginning of the sequence to the point of recombination, and the right parental sequence has the same genetic material as the child from the point of recombination to the end of the sequence.

Time is measured in units of $2N$ generations, and N is assumed large. The mutation rate u per sequence per generation is scaled so that $\theta = 4Nu$. In the same way, the recombination rate r per sequence per generation is scaled so that $\rho = 4Nr$. Then, the number of ancestral sequences in the graph is a birth and death process with birth and death rates $k\rho/2$ and $k(k-1)/2$, respectively, when the current number of ancestral sequences is k. A coalescence event decreases the number of ancestors by one, while a mutation event does not change the number of ancestors, and a recombination event increases the number of ancestors by one. Since coalescence events occur at a quadratic rate, and recombination events occur at a linear rate, the number of ancestors remains finite, and the graph leads eventually to one ancestor, the MRCA of the sample.

A sample ancestral recombination graph, taken from Griffiths and Marjoram's (1996a) paper, is presented in Fig. 1. This graph shows explicitly the ancestral and non-ancestral material in a set of sequences. Going backward in time, a coalescence event occurs when two sequences join together to form a single one, a mutation event occurs when one marker allele in one sequence is changed, and a recombination event occurs when one sequence is cut into two (note that the number of the interval that contains the point of recombination is indicated in a small box in Fig. 1).

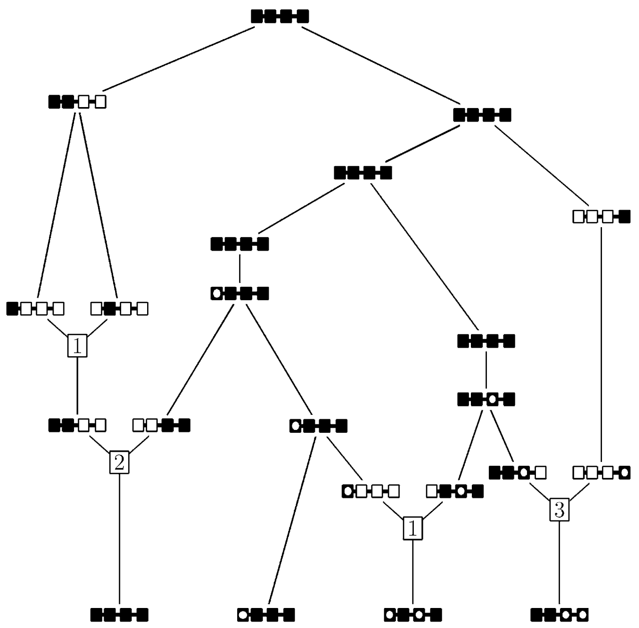


FIG. 1. An example of ARG. Symbols: ■, primitive ancestral marker; ▣, mutant ancestral marker; □, non-ancestral marker.

3. THE MODEL

We consider sequences of small physical length, so that we may assume no interference, and that the distances measured in Morgans are additive. Each generation, every sequence comes from two parental sequences with probability r and one parental sequence with probability $1 - r$. Here, since we assume no interference, r is the distance between the first marker and the last.

We consider a sample of cases and controls, who, respectively, exhibit (cases) and do not exhibit (controls) a given phenotype. The trait is assumed to be caused by a single mutation, and this mutation happened only once in the history of the sample (this is known as the infinitely many site model). Moreover, the sample is assumed to come from a young isolated population, and the mutation to have high penetrance. Our models and methods are meant to identify the position of that mutation.

We suppose n sequences of d possible different types, with n_i sequences of type i $(i = 1, \ldots, d)$. A sequence is made of L marker alleles. These markers are ordered and the exact locations of $L - 1$ of them are known. We denote by r_m the distance in Morgans between markers m and $m + 1$. The trait-influencing mutation (TIM) is itself a marker allele whose location is unknown, and the state of this allele can be inferred from the phenotype. This marker will be treated as the others in the construction of the sample genealogy. The TIM is supposed to be between the first marker and the last one, and as such, cannot be outside the observed sequence of markers. Moreover, the TIM is supposed to be at a distance r_T from the first marker. We derive a maximum likelihood estimate of r_T, and we find this estimate by calculating the maximum likelihood of the ancestral recombination graph conditional on the value of r_T. Let r be the length of the sequences, that is, $r = \sum_{p=1}^{L-1} r_p$. Let x_m be the location of the marker locus m, with the convention that the marker locus 1 is at the origin, that is,

$$\begin{cases} x_1 = 0, \\ x_m = \sum_{p=1}^{m-1} r_p, \quad 2 \leqslant m \leqslant L. \end{cases}$$

The sequences can be partitioned into intervals, where interval p is the segment between markers loci p and $p + 1$. A sequence is illustrated in Fig. 2. It is understood that a "sequence" is simply an ordered set of markers at known loci except the one associated to the TIM.

Each marker locus m has a coalescent tree $T(m)$ describing the history of the sample for this marker. To obtain $T(m)$, start from the initial sample, and for each sequence follow the edges of the ancestral recombination graph for marker m; when a recombination event occurs, take the left path if the recombination happened after m, otherwise take the right path. The set of all these edges defines $T(m)$. Figure 3 illustrates this concept, showing the partial trees for the ARG of Fig. 1. A marker m in a given sequence is ancestral if this marker is included in $T(m)$, otherwise, it is non-ancestral. Moreover, an ancestral marker can be of two different types: a primitive type (■), if it is a copy of an allele from the MRCA without mutation, or a mutant type (▣), if it is a mutant copy of an allele from the MRCA. A non-ancestral marker will be represented by □.

The τth historical event backward in time ($\tau = 0, \ldots, \tau^*$, where 0 corresponds to the initial sample and τ^* to the last coalescence to the MRCA) is assumed to occur at time t_τ. Let us denote by $\mathbf{H}_\tau$ the set of ancestral sequences at time t_τ just after τ events occurred. Then, $\mathbf{H}_\tau$ is a set of sets: a set of ancestral

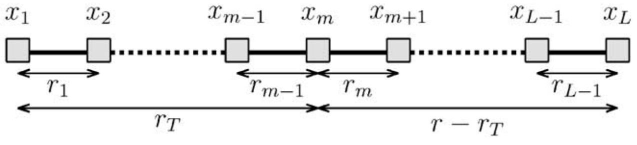
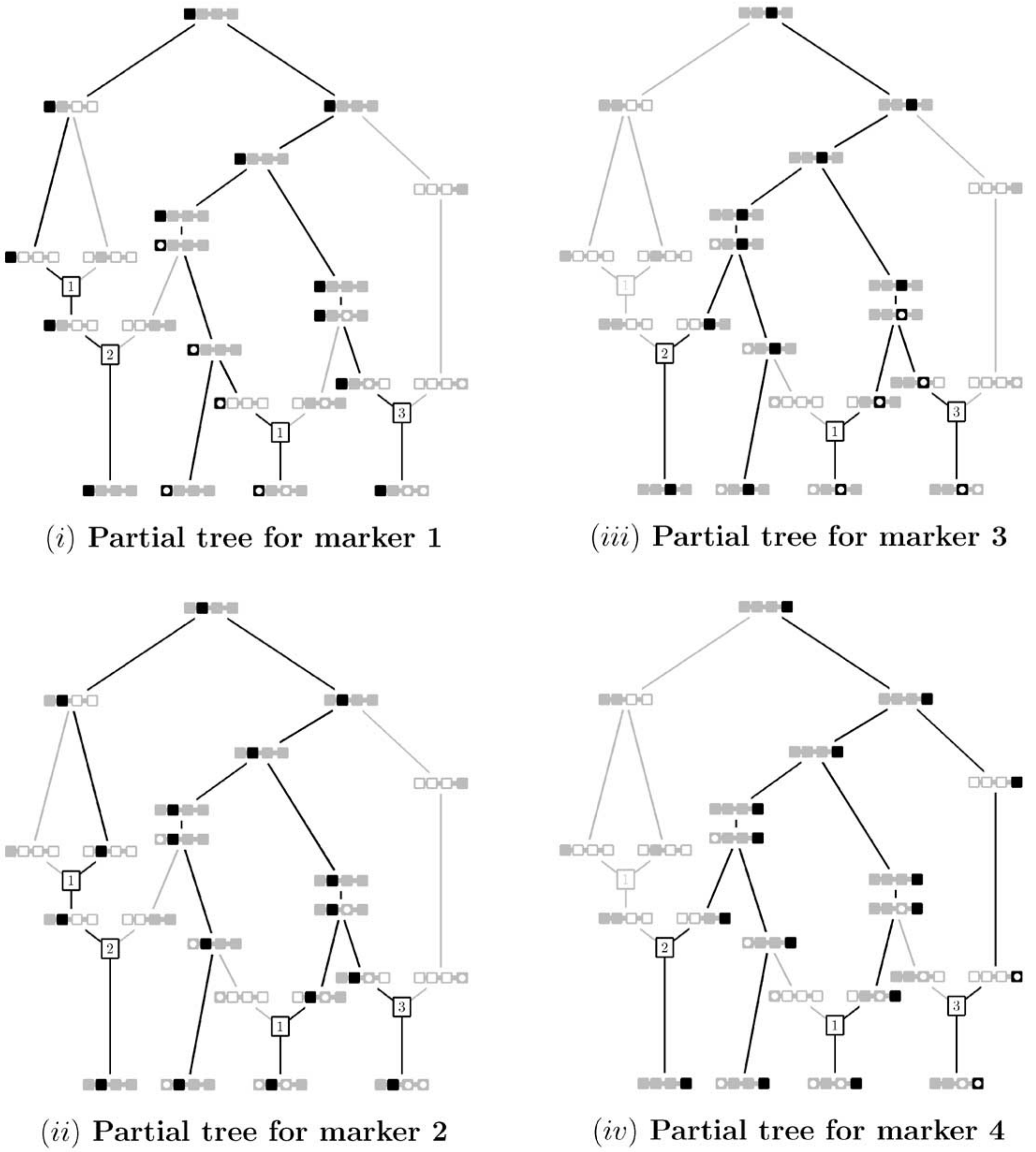


FIG. 2. Sequence configuration if the TIM is at position m.



(i) **Partial tree for marker 1**

(iii) **Partial tree for marker 3**

(ii) **Partial tree for marker 2**

(iv) **Partial tree for marker 4**

FIG. 3. Partial tree of ARG of Fig. 1 for each marker: partial tree for marker 1(i), 2(ii), 3(iii), and 4(iv).

markers, a set of mutations, a set of multiplicities, and a set of information about the trait, so that $\mathbf{H}_\tau = (A, M, n, T)$, following Griffiths and Marjoram's (1996a) notation. The following notation is used for the possible events as they occur backward in time:

$$\begin{cases} C_i & \text{Co of two identical sequences } i, \\ C_{ij}^k & \text{Co of sequences } i \text{ and } j \text{ into sequence } k, \\ M_i^j & \text{Mu from sequence } i \text{ to sequence } j, \\ R_i^{jk} & \text{Re of sequence } i \text{ into sequences } j \text{ and } k, \end{cases}$$

where Co means coalescence, Mu mutation and Re recombination.

Let $Q(\mathbf{H}_\tau)$ be the probability distribution of state $\mathbf{H}_\tau$. Then, $Q(\mathbf{H}_\tau)$ is a function of $Q(\mathbf{H}_{\tau+1})$. We will write, by convention, $\mathbf{H}_{\tau+1} = (\mathbf{H}_\tau + R_i^{jk})$, if $\mathbf{H}_\tau$ is modified by an event of recombination R_i^{jk} at time $t_{\tau+1}$, and so on.

If a recombination event occurs, the density $f_Z(z)$ of the point of recombination is assumed to be

$$f_Z(z) = \frac{1}{r} \quad \text{if } 0 < z < r.$$

Although the point of recombination can be anywhere in the sequence, only a recombination event between two ancestral markers will affect the original sample. Such a recombination event will be called an "ancestral recombination", as in Fearnhead and Donnelly (2001). Consider, for example, the two sequences of five markers in Fig. 4.

Following Fig. 4, if a recombination event occurs involving sequence 1 between markers 1 and 2, the two resulting sequences will be on one hand a sequence of five non-ancestral markers and on the other hand the same sequence as sequence 1. This illustrates the fact that this event does not modify the history of the sample nor does it bring any new information. Let γ_i (κ_i) be the number of the first (last) intervals of a sequence of type i where a recombination event can affect the ancestral material. For example, in Fig. 4, $\gamma_1 = 2$, $\kappa_1 = 3$ for sequence 1, and $\gamma_2 = 1$, $\kappa_2 = 2$ for sequence 2. Moreover, let

$$c_i = \frac{1}{r}[\max\{x_m;\ \text{marker } m \in A_i\} - \min\{x_m;\ \text{marker } m \in A_i\}],$$

where A_i represents the set of ancestral markers on sequence i. Then, c_i represents the proportion of sequence i for which a recombination event could affect its ancestral material. Moreover, let

$$b = \sum_{i=1}^{d} n_i(\max\{x_m;\ \text{marker } m \in A_i\} - \min\{x_m;\ \text{marker } m \in A_i\}).$$

Then, $0 \leqslant b \leqslant nr$ and b is the total sequence length over all sequences where a recombination event could affect the ancestral material, taking into account the multiplicities of the sequences.

Similarly, a mutation event does not necessarily affect the history of the sample, and we distinguish between ancestral and non-ancestral mutation events. Let us denote by $|A_i|$ the number of markers loci on sequence i where a mutation event is ancestral when it occurs, and by $a = \sum_{i=1}^{d} n_i|A_i|$, the corresponding total number of markers loci over all sequences ($n \leqslant a \leqslant nL$). In Fig. 4, for example, if the multiplicity of sequences 1 and 2 is one, then $a = 3 + 2$.

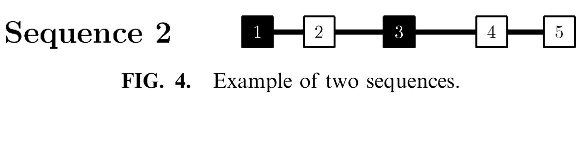


FIG. 4. Example of two sequences.

4. RECURSION PROBABILITIES

Consider the history of a sequence from state (A, M, n, T), and assume that events of coalescence, mutation, and recombination occur with the corresponding probabilities:

$$\begin{cases} n(n-1)/[n(n-1)+n\theta+n\rho] & \text{for a coalescence,} \\ n\theta/[n(n-1)+n\theta+n\rho] & \text{for a mutation,} \\ n\rho/[n(n-1)+n\theta+n\rho] & \text{for a recombination.} \end{cases}$$

Now, if a coalescence event occurs, there are $n-1$ sequences one step back in time. As each sequence has the same probability to coalesce, sequences of type i coalesce with probability $(n_i-1)/(n-1)$, since one step back in time, there are $n_i - 1$ sequences of type i. If two different types of sequence, i and j, coalesce into a new sequence k, we must take into account the possibility that types i and k, or j and k, are identical. Thus, the probability that i and j coalesce to k is $(n_k + 1 - \delta_{ik} - \delta_{jk})/(n-1)$, where $\delta_{ik} = 1$ if $i = k$, and 0 if $i \neq k$.

If the first event back in time is a mutation event, then there is a sequence i which comes from a sequence k with probability $(n_k+1)/n$, since one step back in time the number of sequences k is $n_k + 1$ (the actual number of sequences, k, plus the new one produced), and the total number of sequences is unchanged, and so this total remains equal to n. Moreover, a mutation event occurs in non-ancestral material with probability $(nL-a)/nL$.

A recombination event can occur in any of the $L-1$ intervals of any sequence. When this happens in a given interval of some sequence i, then the sequence i has the same genetic material as some sequence j on the left of the recombination interval, and the same genetic material as some sequence k on the right of the recombination interval. In this context, we thus consider ordered pairs of sequences j and k. One step back in time, we have $n_j + 1$ sequences of type j and $n_k + 1$ sequences of type k, since the total number of sequences is then $n+1$, and the total number of possible ordered pairs of sequences is $n(n+1)$. This event has probability $[(n_i+1)(n_j+1)]/[n(n+1)]$. Moreover, since $0 \leqslant b \leqslant nr$, a recombination event could happen without affecting the ancestral material with probability $(nr-b)/nr$.

These facts lead to a recursion, when time is measured in units of $2N$ generations, analogous to that discussed by Griffiths and Marjoram (1996a) in the derivation of their formula for a continuous model:

$$\begin{aligned}
Q(\mathbf{H}_\tau) = {} & \frac{n(n-1)}{D}\left[\sum_1 \frac{n_i-1}{n-1} Q(\mathbf{H}_\tau + C_i)\right. && \textcircled{1}\\
& \left. + 2\sum_2 \frac{n_k+1-\delta_{ik}-\delta_{jk}}{n-1} Q(\mathbf{H}_\tau + C_{ij}^k)\right] && \textcircled{2}\\
& + \frac{n\theta}{D}\left[\sum_3 \frac{n_j+1}{n} Q(\mathbf{H}_\tau + M_i^j)\right. && \textcircled{3}\\
& \left. + \frac{(nL-a)}{nL} Q(\mathbf{H}_\tau)\right] && \textcircled{4}\\
& + \frac{n\rho}{D}\left[\sum_{i=1}^{d}\left\{\sum_{p=\gamma_i}^{\kappa_i} \frac{r_p}{r}\frac{(n_j+1)(n_k+1)}{n(n+1)}\right.\right.\\
& \left. Q(\mathbf{H}_\tau + R_i^{jk}(p))\right\} && \textcircled{5}\\
& \left. + \frac{(nr-b)}{nr} Q(\mathbf{H}_\tau)\right], && \textcircled{6}\ (1)
\end{aligned}$$

where $D = [n(n-1) + n\theta + n\rho]$, and the numbers on the right refer to the following events:

① Coalescence of two sequences of the same type i,
② Coalescence of two sequences of different types, i and j, to one sequence of type k,
③ Mutation of sequence i to sequence j, where sequence j may already exist,
④ Mutation in non-ancestral material,
⑤ Recombination of sequence i in interval p that produces sequences j and k, where sequences j and k may already exist,
⑥ Recombination in non-ancestral material,
and where the numbers under the summation signs mean:

(1) Summation over types $i = 1, \ldots, d$; $\{i; n_i > 1\}$,
(2) Summation over unordered pairs i, j that possess the same set of mutations in the ancestral material,
(3) Summation over all singleton mutations.

Note that the parameter r_T is hidden in the preceding formula: for two values of p, corresponding to the intervals on both sides of the TIM, the values of r_p depend on the value of r_T. Let r_l and r_r be the lengths of the intervals on the left and the right of the TIM that depend only on r_T. Define:

$$\delta_p^l = \begin{cases} 1 & \text{if } \max\{x \in [x_p; x_{p+1}]\} = r_T, \\ 0 & \text{otherwise,} \end{cases}$$

$$\delta_p^r = \begin{cases} 1 & \text{if } \min\{x \in [x_p; x_{p+1}]\} = r_T, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, δ_p^l (δ_p^r) is equal to 1 if the TIM is the marker on the right (on the left) of interval p, and equal to 0 otherwise. With this in mind, the recombination term (line ⑤) in the preceding equation can be written in the form:

$$\sum_{i=1}^{d}\sum_{p=\gamma_i}^{\kappa_i}\left[\frac{r_p}{r}(1-\delta_p^l)(1-\delta_p^r) + \frac{r_l}{r}\delta_p^l + \frac{r_r}{r}\delta_p^r\right] \frac{(n_j+1)(n_k+1)}{n(n+1)} Q(\mathbf{H}_\tau + R_i^{jk}(p)). \tag{2}$$

Note that $n_j + 1$ and $n_k + 1$ are the numbers of sequences of types j and k existing one step back in time. In Griffiths and Marjoram's (1996a) model, assuming a continuous segment of DNA, they always have $n_j = n_k = 0$; but in the discrete implementation, n_j and n_k are not always zero.

5. ALGEBRAIC IDENTITIES AND IMPORTANCE SAMPLING

We now use a Monte Carlo strategy introduced for the first time by Griffiths and Tavaré (1994b). Using the notation $\alpha = a/(nL)$, $\beta = b/(nr)$, $D_{\mathbf{H}_\tau} = [n(n-1) + n\alpha\theta + n\beta\rho]$, and $S_{\mathbf{H}_\tau} = n\sum_1 (n_i - 1) + 2n\sum_2 (n_k + 1 - \delta_{ik} - \delta_{jk}) + \theta\sum_3 (n_j+1) + \sum_{i=1}^{d}\sum_{p=\gamma_i}^{\kappa_i} [\rho\frac{r_p}{r}]/[(n+1)]$, recursion (1) can be written in the form

$$\begin{aligned}
Q(\mathbf{H}_\tau) = {} & \sum_1 \frac{S_{\mathbf{H}_\tau}}{D_{\mathbf{H}_\tau}}\frac{n(n_i-1)}{S_{\mathbf{H}_\tau}} Q(\mathbf{H}_\tau + C_i)\\
& + \sum_2 \frac{S_{\mathbf{H}_\tau}}{D_{\mathbf{H}_\tau}}\frac{2n(n_k+1-\delta_{ik}-\delta_{jk})}{S_{\mathbf{H}_\tau}} Q(\mathbf{H}_\tau + C_{ij}^k)\\
& + \sum_3 \frac{S_{\mathbf{H}_\tau}}{D_{\mathbf{H}_\tau}}\frac{\theta(n_j+1)}{S_{\mathbf{H}_\tau}} Q(\mathbf{H}_\tau + M_i^j)\\
& + \sum_{i=1}^{d}\sum_{p=\gamma_i}^{\kappa_i} \frac{(n_j+1)(n_k+1)S_{\mathbf{H}_\tau}}{D_{\mathbf{H}_\tau}}\\
& \frac{\rho\frac{r_p}{r}}{S_{\mathbf{H}_\tau}(n+1)} Q(\mathbf{H}_\tau + R_i^{jk}(p)).
\end{aligned}$$

Let us define a Markov chain with transition probabilities from $\mathbf{H}_\tau$ to $\mathbf{H}_{\tau+1}$, denoted by $P(\mathbf{H}_{\tau+1}|\mathbf{H}_\tau)$, as follows. At time $\tau + 1$, a transition is made from $\mathbf{H}_\tau$ to

$$\begin{aligned}
& (\mathbf{H}_\tau + C_i) \quad \text{w.p. } n(n_i-1)/S_{\mathbf{H}_\tau},\\
& (\mathbf{H}_\tau + C_{ij}^k) \quad \text{w.p. } 2n(n_k+1-\delta_{ik}-\delta_{jk})/S_{\mathbf{H}_\tau},\\
& (\mathbf{H}_\tau + M_i^j) \quad \text{w.p. } \theta(n_j+1)/S_{\mathbf{H}_\tau},\\
& (\mathbf{H}_\tau + R_i^{jk}(p)) \quad \text{w.p. } \rho\tfrac{r_p}{r}/[(n+1)S_{\mathbf{H}_\tau}],
\end{aligned} \tag{3}$$

where w.p. means “with probability”. Let

$$f(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) = \begin{cases} S_{\mathbf{H}_\tau}/D_{\mathbf{H}_\tau} & \text{if Co or Mu,} \\ (n_j + 1)(n_k + 1)S_{\mathbf{H}_\tau}/D_{\mathbf{H}_\tau} & \text{if Re.} \end{cases} \qquad (4)$$

The state $\mathbf{H}_\tau$ is the state chain at step τ for $\tau = 0, \ldots, \tau^*$ (where τ^* is the absorption time). There is an absorbing state when a common ancestor is found for all sequences, and then the configuration is denoted by $\mathbf{H}_{\tau^*}$. Since the MRCA has primitive ancestral markers at all loci, $\mathbf{H}(\tau^*)$ is uniquely determined with probability 1. This means that $Q(\mathbf{H}_{\tau^*})$ is 1 for a single sequence and 0 for all others. Then, we have

$$Q(\mathbf{H}_0) = \sum_{\mathbf{H}_{\tau+1}} f(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) P(\mathbf{H}_{\tau+1}|\mathbf{H}_\tau) Q(\mathbf{H}_{\tau+1})$$

for $\tau = 0, \ldots, \tau^*$, from which

$$Q(\mathbf{H}_0) = \sum_{H_1} \sum_{H_2} \cdots \sum_{H_{\tau^*}} f(H_0) f(H_1) f(H_2) \cdots$$

$$f(H_{\tau^*-1}) P(H_1|H_0) P(H_2|H_1) \cdots$$

$$P(H_{\tau^*}|H_{\tau^*-1}) Q(H_{\tau^*})$$

and therefore

$$Q(\mathbf{H}_0) = E_P \left[\prod_{\tau=0}^{\tau^*-1} f(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) \right]. \qquad (5)$$

This is an importance sampling representation with proposal distribution P.

Let $\Theta = \{\theta, r_T\}$ be the set of the unknown parameters of the process. Then, given $\Theta_0 = (\theta_0, r_{T_0})$, an estimate of $Q_\Theta(\mathbf{H}_\tau)$ can be found for different values of Θ. We have

$$Q_\Theta(\mathbf{H}_\tau) = \sum_{\mathbf{H}_{\tau+1}} h_{\Theta\Theta_0}(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) P_{\Theta_0}(\mathbf{H}_{\tau+1}|\mathbf{H}_\tau) Q_\Theta(\mathbf{H}_{\tau+1}),$$

where

$$h_{\Theta\Theta_0}(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) = \frac{f_\Theta(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) P_\Theta(\mathbf{H}_{\tau+1}|\mathbf{H}_\tau)}{P_{\Theta_0}(\mathbf{H}_{\tau+1}|\mathbf{H}_\tau)},$$

which can be estimated by the expectation

$$E_{P_{\Theta_0}} \left[\prod_{\tau=0}^{\tau^*-1} h_{\Theta\Theta_0}(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) \right].$$

Denoting by r_{0l} and r_{0r} the parameters r_l and r_r under Θ_0, and with f_Θ as defined in (4) and P_Θ as in (3), the function $h_{\Theta\Theta_0}$ takes the form

$$h_{\Theta\Theta_0}(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) = \frac{S_{(\mathbf{H}_\tau, \Theta_0)}}{D_{(\mathbf{H}_\tau, \Theta)}} \phi(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}),$$

where $\phi(\mathbf{H}_\tau, \mathbf{H}_{\tau+1})$

$$= \begin{cases} \theta/\theta_0 & \text{if Mu,} \\ (n_j + 1)(n_k + 1) \dfrac{r_p(1-\delta^l_p)(1-\delta^r_p) + r_l\delta^l_p + r_r\delta^r_p}{r_p(1-\delta^l_p)(1-\delta^r_p) + r_{0l}\delta^l_p + r_{0r}\delta^r_p} & \text{if Re,} \\ 1 & \text{otherwise.} \end{cases}$$

It is important to note that if a recombination event occurs in an interval where the TIM does not exist, then $\phi(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) = (n_j + 1)(n_k + 1)$, and $f = h$. On the other hand, $\phi(\mathbf{H}_\tau, \mathbf{H}_{\tau+1})$ is different from 1 as soon as a recombination event occurs in the interval harboring the TIM, i.e., whenever δ^l_p or δ^r_p is different from 0, and then $f \neq h$.

The proposed method to evaluate the likelihood of r_T along a sequence involves evaluating Q_{Θ_0} for $\Theta_0 = \{\{\theta_0, r_{T_1}\}, \{\theta_0, r_{T_2}\}, \ldots, \{\theta_0, r_{T_{L-1}}\}\}$, i.e., the use of $L - 1$ driving sets. Note that only one driving value is used for the parameter θ, but several values could, in theory, be used in combination with different values of r_T. For the driving value r_{T_p}, we take the middle of interval p $(1 \leqslant p \leqslant L - 1)$. The likelihood in interval p for values other than r_{T_p} is evaluated by the importance sampling scheme described above: graphs constructed with a driving value r_{T_p} are used to evaluate the likelihood in the region (x_p, x_{p+1}) of the sequence.

6. SIMULATING RECOMBINATION

In the construction of the graph, we have to calculate at each step of the process the probabilities of all the events that can occur one step back in time. In order to simulate a recombination event, we can first choose a sequence at random, and then an interval of recombination at random on the chosen sequence, as in Griffiths and Marjoram (1996a), or equivalently, use the following method: first choose an interval of recombination at random and then choose a sequence at random that will recombine at a point in the chosen interval. Denote by n'_p the number of sequences for which interval p is ancestral for recombination. This is the case for type i if interval p is between the first marker (γ_i) and the last marker (κ_i) in the set of the ancestral markers. From this, a reformulation of the previous scheme can be derived:

$$
\begin{aligned}
&Q(\mathbf{H}_\tau)\\
&= \frac{1}{D_{\mathbf{H}_\tau}}\Bigg[n\sum_{1} (n_i-1)Q(\mathbf{H}_\tau+C_i)\\
&\quad + 2n\sum_{2}(n_k+1-\delta_{ik}-\delta_{jk})Q(\mathbf{H}_\tau+C_{ij}^k)\\
&\quad + \theta\sum_{3}(n_j+1)Q(\mathbf{H}_\tau+M_i^j)+\frac{\rho}{n+1}\\
&\quad \sum_{p=1}^{L-1} n'_p\frac{r_p}{r}(n_J+1)(n_K+1)Q(\mathbf{H}_\tau+R_I^{JK}(p))\Bigg],
\end{aligned}
$$

where I is a sequence chosen at random, and sequences J and K are determined by I and p. The last term of the previous equation can be written:

$$
\begin{aligned}
&\frac{\rho}{n+1}\sum_{p=1}^{L-1}\frac{n'_p}{d}\left[\frac{r_p}{r}(1-\delta^l_p)(1-\delta^r_p)+\frac{r_l}{r}\delta^l_p+\frac{r_r}{r}\delta^r_p\right]\\
&\quad\times(n_J+1)(n_K+1)Q(\mathbf{H}_\tau+R_I^{JK}(p)).
\end{aligned}
$$

Therefore, we will choose an interval of recombination at random, then a sequence I at random, and finally we will calculate the factor $(n_J+1)(n_K+1)$. With this scheme in mind, the recursion becomes

$$
\begin{aligned}
Q(\mathbf{H}_\tau) &= \sum_{1}\frac{S_{\mathbf{H}_\tau}}{D_{\mathbf{H}_\tau}}\frac{n(n_i-1)}{S_{\mathbf{H}_\tau}}Q(\mathbf{H}_\tau+C_i)\\
&\quad+\sum_{2}\frac{S_{\mathbf{H}_\tau}}{D_{\mathbf{H}_\tau}}\frac{2n(n_k+1-\delta_{ik}-\delta_{jk})}{S_{\mathbf{H}_\tau}}Q(\mathbf{H}_\tau+C_{ij}^k)\\
&\quad+\sum_{3}\frac{S_{\mathbf{H}_\tau}}{D_{\mathbf{H}_\tau}}\frac{\theta(n_j+1)}{S_{\mathbf{H}_\tau}}Q(\mathbf{H}_\tau+M_i^j)\\
&\quad+\sum_{p=1}^{L-1}\frac{(n_J+1)(n_K+1)S_{\mathbf{H}_\tau}}{D_{\mathbf{H}_\tau}}\\
&\quad\frac{\rho n'_p[r_p/r]}{S_{\mathbf{H}_\tau}(n+1)}Q(\mathbf{H}_\tau+R_I^{JK}(p)),
\end{aligned}
$$

which is of the form

$$Q(\mathbf{H}_\tau)=\sum_{\mathbf{H}_{\tau+1}} f(\mathbf{H}_\tau,\mathbf{H}_{\tau+1})P(\mathbf{H}_{\tau+1}\mid\mathbf{H}_\tau)Q(\mathbf{H}_{\tau+1}),$$

where $D_{\mathbf{H}_\tau}=[n(n-1)+n\alpha\theta+n\beta\rho]$ as previously defined and

$$
\begin{aligned}
S_{\mathbf{H}_\tau} &= n\sum_{1}(n_i-1)+2n\sum_{2}(n_k+1-\delta_{ik}-\delta_{jk})\\
&\quad+\theta\sum_{3}(n_j+1)+\sum_{p=1}^{L-1}\frac{\rho n'_p[r_p/r]}{(n+1)}.
\end{aligned}
$$

At time $\tau+1$, a transition is made from $\mathbf{H}_t$ to:

$$
\begin{aligned}
&(\mathbf{H}_\tau+C_i)\quad \text{w.p. } n(n_i-1)/S_{\mathbf{H}_\tau},\\
&(\mathbf{H}_\tau+C_{ij}^k)\quad \text{w.p. } 2n(n_k+1-\delta_{ik}-\delta_{jk})/S_{\mathbf{H}_\tau},\\
&(\mathbf{H}_\tau+M_i^j)\quad \text{w.p. } \theta(n_j+1)/S_{\mathbf{H}_\tau},\\
&(\mathbf{H}_\tau+R_I^{JK}(p))\quad \text{w.p. } \rho[n'_p[r_p/r]]/[(n+1)S_{\mathbf{H}_\tau}],
\end{aligned}
$$

where w.p. means “with probability”. The function f is of the same form:

$$
f(\mathbf{H}_\tau,\mathbf{H}_{\tau+1})=\begin{cases}
S_{\mathbf{H}_\tau}/D_{\mathbf{H}_\tau},\\
\quad\text{if Co or Mu},\\
(n_J+1)(n_K+1)S_{\mathbf{H}_\tau}/D_{\mathbf{H}_\tau},\\
\quad\text{if Re},
\end{cases}
$$

while the functions h and ϕ differ only by replacing the sequences i, j, k, assuming a recombination event, by I, J, K, that is,

$$h_{\Theta\Theta_0}(\mathbf{H}_\tau,\mathbf{H}_{\tau+1})=\frac{S_{(\mathbf{H}_\tau,\Theta_0)}}{D_{(\mathbf{H}_\tau,\Theta)}}\phi(\mathbf{H}_\tau,\mathbf{H}_{\tau+1}),$$

where $\phi(\mathbf{H}_\tau,\mathbf{H}_{\tau+1})$

$$
=\begin{cases}
\theta/\theta_0 & \text{if Mu},\\
(n_J+1)(n_K+1)\\
\dfrac{r_p(1-\delta^l_p)(1-\delta^r_p)+r_l\delta^l_p+r_r\delta^r_p}{r_p(1-\delta^l_p)(1-\delta^r_p)+r_{0l}\delta^l_p+r_{0r}\delta^r_p} & \text{if Re},\\
1 & \text{otherwise}.
\end{cases}
$$

Note again that if a recombination event happens in an interval where the TIM does not exist, then $\delta^l_p=\delta^r_p=0$, $\phi(\mathbf{H}_\tau,\mathbf{H}_{\tau+1})=(n_J+1)(n_K+1)$, and $f(\cdot\,,\cdot)=h(\cdot\,,\cdot)$. On the other hand, $\phi(\mathbf{H}_\tau,\mathbf{H}_{\tau+1})$ is different from 1, as soon as a recombination event happens in the interval harboring the TIM, i.e., whenever δ^l_p or δ^r_p is different from 0, and then $f(\cdot\,,\cdot)\neq h(\cdot\,,\cdot)$.

7. VARIABLE POPULATION SIZE

We have assumed until now that the population size is constant. For realistic applications however, it is useful to allow variable population size. In the coalescent without recombination (Donnelly and Tavaré, 1995; Griffiths and Tavaré, 1997; Norborg, 2001), the change in population size causes a change of scale in the coalescent tree. For example, if a population experiences an exponential expansion, the more one goes back into the past, the more a coalescence event is likely to occur, since the population gets smaller and it takes less time for a coalescence event to occur. With recombination, a variable population size does not only change the scale of the ancestral graph, but also its topology. In the case of an exponential expansion, for example, the more one goes back into the past, the more a coalescence event is likely to occur instead of a recombination event.

Let W_τ be the time of occurrence of the τth event ($\tau = 0, \ldots, \tau^*$), and $S_\tau = W_\tau - W_{\tau-1}$. If the population size is $2N$ and the number of sequences in the sample is n, then the probabilities of coalescence, recombination and mutation in one generation are, respectively, $\binom{n}{2}/2N$, βr and αu. Moreover, the time until an event occurs is exponentially distributed with parameter:

$$\begin{aligned}1 - \left(1 - \frac{n(n-1)}{4N}\right)&(1-\beta r)^n (1-\alpha u)^n \\ &\approx \frac{n}{4N}(n - 1 + \alpha\theta + \beta\rho).\end{aligned}$$

Therefore, the expected time to the next event in units of $2N$ generations (Griffiths and Marjoram, 1996a) is

$$E(S_\tau | W_\tau) = 2/[n(n - 1 + \alpha\theta + \beta\rho)]. \tag{6}$$

Let us now suppose that the population size varies with time in a deterministic fashion. We will suppose that we know $2N(t)$, the size of the population at time t. Let us denote by $v(t)$ the ratio of the population size at time t and 0, so that $v(t) = N(t)/N(0)$, where $N(0) = N$. Let $\lambda(t) = 1/v(t)$. Given a sample of n sequences at time t, we have $P(\text{Co}) = \binom{n}{2}/2N(t) = [n(n-1)]/[4v(t)N]$, $P(\text{Re}) = nr$ and $P(\text{Mu}) = nu$. Therefore, the probabilities of these events, given that at least one of them occurs, are:

$$\begin{aligned}P(\text{Co}|\cdot) &= \frac{[n(n-1)]/[4v(t)N]}{[n(n-1)]/[4v(t)N] + nr + nu} \\ &= \frac{n(n-1)\lambda(t)}{n(n-1)\lambda(t) + n\rho + n\theta},\end{aligned}$$

$$\begin{aligned}P(\text{Re}|\cdot) &= \frac{nr}{[n(n-1)]/[4v(t)N] + nr + nu} \\ &= \frac{n\rho}{n(n-1)\lambda(t) + n\rho + n\theta},\end{aligned}$$

$$\begin{aligned}P(\text{Mu}|\cdot) &= \frac{nu}{[n(n-1)]/[4v(t)N] + nr + nu} \\ &= \frac{n\theta}{n(n-1)\lambda(t) + n\rho + n\theta}.\end{aligned}$$

Then, the recurrence equation for $Q(\mathbf{H}_\tau)$ is

$$\begin{aligned}Q(\mathbf{H}_\tau) = \int_{W_\tau}^{\infty} \Bigg[&\frac{(nL-a)\theta/L + (nr-b)\rho/r}{[n(n-1)\lambda(W_{\tau+1}) + n\theta + n\rho]} Q(\mathbf{H}_\tau) \\ &+ \frac{n(n-1)\lambda(W_{\tau+1})}{D'} \sum_1 \frac{(n_i - 1)}{(n-1)} Q(\mathbf{H}_\tau + C_i) \\ &+ \frac{2n(n-1)\lambda(W_{\tau+1})}{D'} \sum_2 \frac{(n_k + 1 - \delta_{ik} - \delta_{jk})}{(n-1)} \\ &Q(\mathbf{H}_\tau + C_{ij}^k) + \frac{n\theta}{D'} \sum_3 \frac{(n_j+1)}{n} \\ &Q(\mathbf{H}_\tau + M_i^j) + \frac{n\rho}{D'} \sum_{i=1}^{d} \sum_{p=\gamma_i}^{\kappa_i} \frac{r_p}{r} \frac{(n_j+1)(n_k+1)}{n(n+1)} \\ &\times Q(\mathbf{H}_\tau + R_i^{jk}(p)) \Bigg] g(W_{\tau+1}|W_\tau)\, dW_{\tau+1},\end{aligned}$$

where $D' = [n(n-1)\lambda(W_{\tau+1}) + n\theta + n\rho]$, and $g(W_{\tau+1}|W_\tau)$ is the distribution of the time of occurrence of the $(\tau + 1)$th event, given the time of occurrence of the τth event. The probability that $W_{\tau+1}$ exceeds s given W_τ is the probability that no coalescence, recombination, or mutation event occurs in the time interval S_τ, and is given by

$$P(W_{\tau+1} > s | W_\tau) = \exp\left(-\int_{W_\tau}^{s} \Phi(u)\, du\right),$$

where

$$\Phi(u) = \frac{n(n-1)}{2}\lambda(u) + \frac{n\alpha\theta}{2} + \frac{n\beta\rho}{2}.$$

Now, in order to take into account time and variable population size in the Monte Carlo algorithm, we can simulate the time to the next event according to the right distribution (Donnelly and Tavaré, 1995; Griffiths and Tavaré, 1996). Suppose that $\lambda(u) = \exp(\kappa u)$, so that the population grows exponentially fast. Donnelly and Tavaré (1995) have suggested simulating time in the following way: let $\{U_0, U_1, \ldots\}$ be a sequence of mutually independent uniform random variables, and then solve the following equation:

$$\begin{aligned}1 - P(W_{\tau+1} \leqslant s | W_\tau) &= U_\tau, \\ \Leftrightarrow \exp\left(-\int_{W_\tau}^{s} \Phi(u)\, du\right) &= U_\tau,\end{aligned}$$

$$\Leftrightarrow -\int_{W_\tau}^{s}\left\{\frac{n(n-1)}{2}\exp(\kappa u)+\frac{n\alpha\theta}{2}+\frac{n\beta\rho}{2}\right\}du=\log(U_\tau)$$

$$\Leftrightarrow \frac{-n(n-1)}{2\kappa}\exp(\kappa s)-s\left(\frac{n\alpha\theta}{2}+\frac{n\beta\rho}{2}\right)$$

$$+\frac{n(n-1)}{2\kappa}\exp(\kappa W_\tau)+W_\tau\left(\frac{n\alpha\theta}{2}+\frac{n\beta\rho}{2}\right)$$

$$-\log(U_\tau)=0.$$

There is no direct solution to express the unknown variable s as a function of the other parameters. However, in the particular case where $\theta = \rho = 0$, the above equation becomes

$$s=\frac{1}{\kappa}\log\left[-\frac{2}{n(n-1)}\kappa\log(U_\tau)+\exp(\kappa W_\tau)\right]. \quad (7)$$

This result can be found in Griffiths and Tavaré (1998) and Griffiths (2001). Going back to the general case, we have to find the solution of

$$a\exp(bs)+sc+d=0,$$

which can be derived using the algebraic software package `Maple`:

$$s=-\frac{1}{bc}\left[W\left(\frac{1}{c}ab\exp\left(-\frac{db}{c}\right)\right)c+db\right], \quad (8)$$

where W is the *Lambert W* function defined by

$$W(x)=\sum_{n=1}^{\infty}\frac{(-1)^{n-1}n^{n-2}}{(n-1)!}x^n.$$

This function is the inverse of the function we^w (Corless *et al.*, 1996). Note that the above series oscillates between large negative and positive values for real $x \geqslant 4$ (Weissten, 2000). However, an algorithm based on Halley iteration converges rapidly for all valid x (Briggs, 1998). The algorithm produces the value $W_\tau = W(x)$. Using this strategy, we can now model variable population sizes of exponential growth.

We now show an example for the distribution of the time to the next event. Simulating 80,000 variates distributed uniformly on [0, 1], the observations were regrouped to approximate the densities of s_c (coalescence only) using (7) and s_{crm} (coalescence, recombination and mutation) using (8), and corresponding means μ_c and μ_{crm}. Three cases were considered with $n = 88, \kappa = 1700, t = 10^{-5}$ and the following values for the other parameters:

(a) $\theta = 10$, $\rho = 10$, in which case $\mu_c = 2.01 \times 10^{-4}$ and $\mu_{crm} = 1.78 \times 10^{-4}$;
(b) $\theta = 100$, $\rho = 100$, in which case $\mu_c = 2.01 \times 10^{-4}$ and $\mu_{crm} = 8.55 \times 10^{-5}$;
(c) $\theta = 358$, $\rho = 358$, in which case $\mu_c = 2.01 \times 10^{-4}$ and $\mu_{crm} = 3.81 \times 10^{-5}$.

The densities for the three cases are illustrated in Fig. 5.

In case (a), where θ and ρ are small, the two densities are almost identical. In case (b), we see a significant difference in the distributions: the "next event" will occur sooner. The third case, (c), illustrates a situation in which recombination and mutation rates are higher (corresponding to the case of EPM disease, see Virtaneva *et al.*, 1996). The larger θ and ρ are, the smaller the mean time to the next event, and it thus becomes more important to take into account variable population size. A larger value for the parameter of recombination ρ is especially important in view of studies of sequences of moderate length.

8. COMPUTER IMPLEMENTATION

A computer program, written in `C++`, implements the above procedure in the case of a constant population

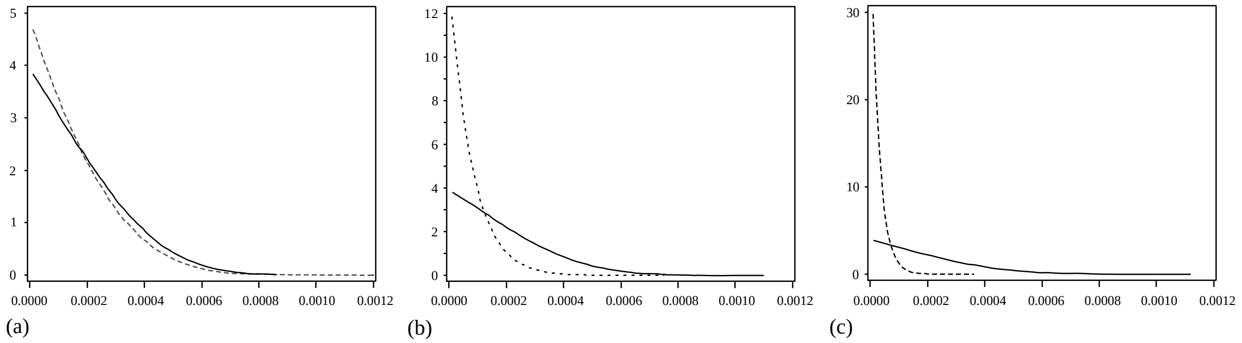


FIG. 5. Densities for the three cases (a), (b) and (c). Plain line is for s_c and dash line for s_{crm}.

size, and also for variable population size. The program compiles adequately with the standard GNU compiler on several operating systems (e.g., Windows, Linux, Solaris, Irix).

However, there are numerical problems related to the computation of the likelihood value. As we have seen (Eq. (5)), the likelihood value is estimated by

$$L(r_T) = E_P\left[\prod_{\tau=0}^{\tau^*-1} f(\mathbf{H}_\tau, \mathbf{H}_{\tau+1})\right]. \tag{9}$$

Usually the function f at each step of the graph is less than one, while the number of steps, τ^*, is large, particularly if the number and the length of the sequences are large. Therefore, the product in (9) can be very small, in fact smaller than the precision of the computer. One approach to solve this problem is to rescale the likelihood value such that λL, for some λ, is estimated for K simulations:

$$\lambda\frac{1}{K}\sum_{i=1}^{K}\left[\prod_{\tau=0}^{\tau^*-1} f(\mathbf{H}_\tau, \mathbf{H}_{\tau+1})\right] = \frac{1}{K}\sum_{i=1}^{K}\lambda\left[\prod_{\tau=0}^{\tau^*-1} f(\mathbf{H}_\tau, \mathbf{H}_{\tau+1})\right].$$

This approach can work, but it is still limited by the precision of the computer. Another approach is to use an extended numeric range floating point type. The `EXTNUM` package can be used (NBIC, 2001). An `EXTNUM` number can take any (positive or negative) value between $1 \times 10^{-646,456,993}$ and $2 \times 10^{646,456,992}$, and therefore solve this problem.

Another important point is to have a good random number generator, which should be fast, and have a long period. We have used a `C++` implementation (Wagner, 2001) of the Mersenne Twister (Matsumoto and Nishimura, 1998), which has a period of $2^{19937} - 1$.

The program takes ASCII files for input (one data file and one file with the parameter values), and generates ASCII files for output. Graphs can be made with this output directly by the `GNUPLOT` program, or, for example, the `SAS` program.

9. AN EXAMPLE

In order to verify the proposed theory and its implementation in a computer program, we consider an example. This example also illustrates the results produced by the program.

The data is generated according to a Wright–Fisher neutral model by the `ms` program from Hudson (2001), using a constant population size. A sample of 30 sequences are generated with $\rho = 60$, in order to have five segregating sites in the sequence (using an option in the program). To illustrate the method, we will suppose that the third site is unknown and in complete linkage with a gene causing a disease, and that we have full penetrance. In assessing the data under these assumptions, we have also removed information on the third site, but consider information on disease status. The position of the TIM is then in the second interval.

The data consists of seven different sequence types and four marker loci, and information on the trait. We have 20 control sequences and 10 case sequences. The data are shown in Fig. 6.

We set $\rho = 60$, and $\theta = 60$ for analyzing the data. Given an effective population size of 10,000, we have $r = 0.15$ cM. The distance between marker loci was assumed to be 0.05 cM for each of the three intervals. Results are shown in Fig. 7. The scale is the same on all the graphs to compare the heights of the curves: the y-axis is the logarithm of the likelihood, and the x-axis is the position of the TIM on the sequence (r_T), where 0 is the position of the first marker of the sequence. The first graph (Fig. 7a) shows likelihood profiles of two independent runs of 16 millions iterations (plain lines), and the combined likelihood on these 32 millions iterations (dash line).

The maximum likelihood estimate of r_T is $\hat{r}_T \approx$ 0.078 cM. The two profiles are very similar, even if a difference of height in the likelihood is observed. If fact, each of these two runs of 16 millions is constructed on three independent runs: one of 1 million, a second one of 5 millions and a third run of 10 millions. To see variability in the profiles, Fig. 7b–d shows likelihood profiles for the two runs of 1, 5 and 10 million iterations,

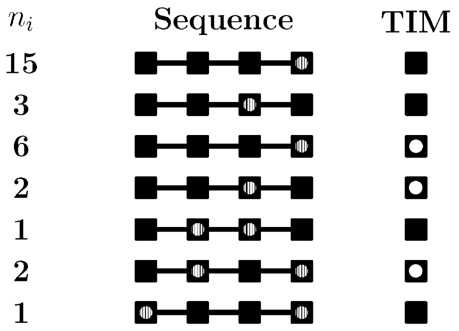


FIG. 6. Simulated data. Multiplicity of each sequence (n_i), marker loci and information on the trait (TIM).

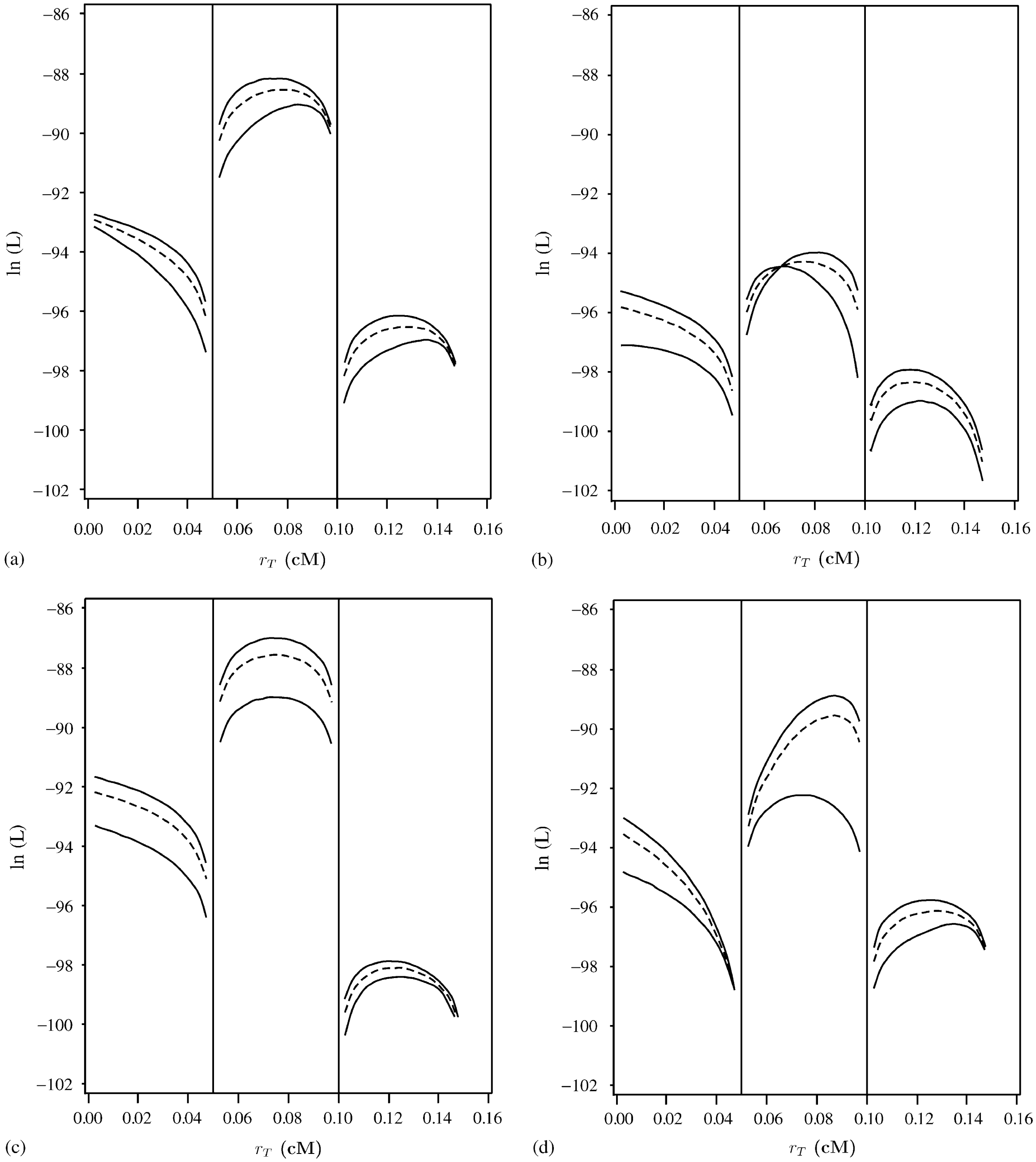


FIG. 7. Likelihood profile obtained with data from a simulated example of 30 sequences. Plain lines are independent results for a defined number of iterations of the process, and dashed lines are the combined likelihood of these two runs. (a) 16 × 10^6 iterations run twice (b) 1 × 10^6 iterations run twice (c) 5 × 10^6 iterations run twice (d) 10 × 10^6 iterations run twice.

respectively. We can see that the value of the maximum likelihood is higher when more iterations of the process are done. In the three cases, the maximum likelihood of r_T is in the second interval, as expected. Even though the likelihood profiles are similar, there are non-negligible variations in the likelihood curves. To observe two similar replications of the likelihood curves, more iterations are probably needed.

With a high-performance desktop computer (Pentium III 1 GHz), a mean time of approximately 7 h is needed to do 1 million iterations for this data. A total of 9 days of computing time have been necessary to do the 32 millions iterations. Of course, several machines were used and the time needed to analyze data is greatly reduced from a few days to a day as a result. But, it is clear that a high performance computer is a prerequisite if one is to use the method with real data.

In the implementation of the method, there is an option to abort construction of graphs that take too much time, in case the process is "lost". In such cases, the likelihood returned by the program for the graph is 0. Let v be the maximum number of events allowed in a construction of a graph. For the above example, $v = 1000$, which in reality is just a precaution, because none of the 32 millions graphs have been aborted. When making millions of graphs, there are large differences in likelihood heights between different graphs, and as such, only a few graphs contribute significantly to the likelihood. Therefore, even if there is no direct relationship between the number of events in a graph and the likelihood, we can improve the speed of the process by aborting early graphs that contain too many events. Note that a similar process is used by Griffiths and Marjoram (1996a), but the switch to abort the construction of a graph is based on a combinatorial constant.

We have arbitrarily set v in order to have between 5% and 10% of the graphs really constructed; this can easily be done after a few short trials. We chose $v = 75$ to match (approximately) our proposed requirements. We observed that the percentage of aborted graphs is 91%, 94% and 95% in the first, second and third interval, respectively. Results of a process of 20 millions iterations is shown in Fig. 8a. As expected, the likelihood profile is almost the same as the likelihood on the 32 millions iterations with $v = 1000$, and we obtain $\hat{r_T} = 0.083$, slightly higher than our previous estimation. The difference is in the computation time used to run the process: only 4.5 h are necessary to do 1 million iterations, instead of 7 h when v was 1000.

Another parameter we can use to improve the speed of the process is the probability of recombination. We know that the larger the sequence, the longer the time needed to relate the observed sequences to a common ancestor. We introduce now the factor ξ which is a recombination weight. First divide and multiply by ξ the probability of recombination in Eq. (1), such that the probability of the event $R_i^{jk}(p)$ is now

$$\frac{\xi \rho r_p / r}{S_{\mathbf{H}_\tau}(n+1)}$$

and the function $f(\mathbf{H}_\tau, \mathbf{H}_{\tau+1})$ associated with this event is now

$$\frac{(n_j+1)(n_k+1)S_{\mathbf{H}_\tau}}{\xi D_{\mathbf{H}_\tau}}.$$

We considered another estimation with $\xi = 0.33$ and $v = 60$, and the likelihood profile based on 20 millions iterations and the results are shown if Fig. 8b. The estimate of r_T is now 0.072 cM. We can see that the

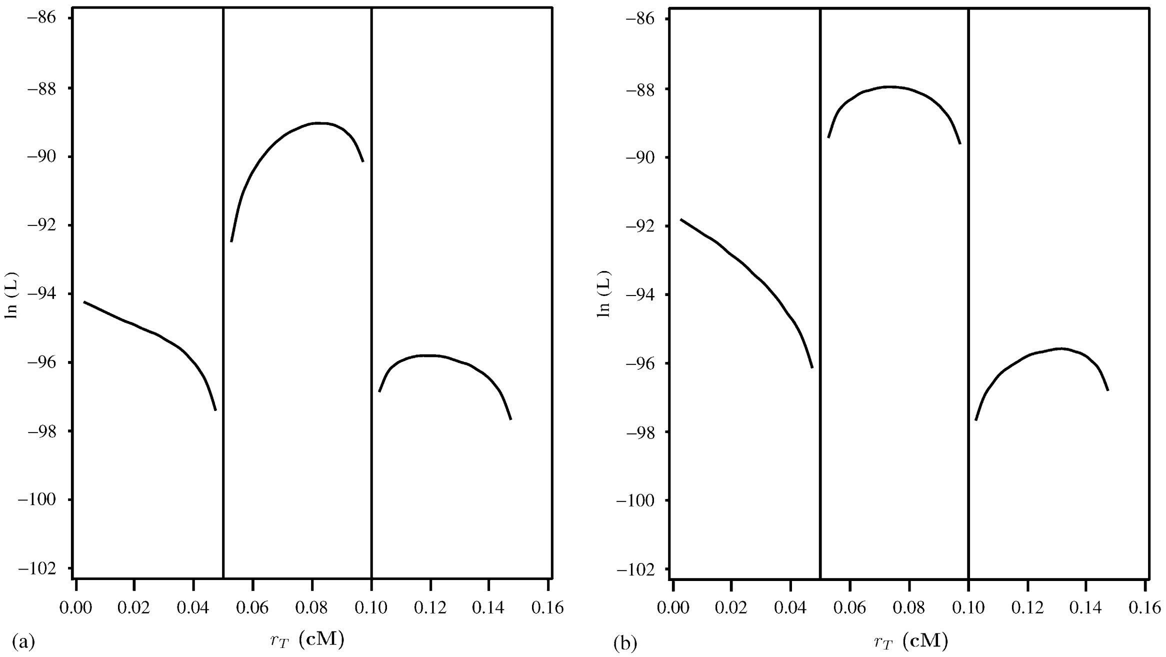


FIG. 8. Likelihood profile based on 20 millions iterations a) with $v = 75$, b) with $\xi = 0.33$ and $v = 60$.

likelihood is very similar to the other estimations, particularly the original estimation based on 32 millions iterations. There is a significant difference in computing time however: a mean time of 1 h 51 m was necessary for each of the million iterations, so the process is almost four times faster than the first set of estimations.

Of course, a value for ξ is difficult to choose. If ξ is too small, not enough recombination events will occur, and the estimation of r_T will become difficult. Also, if ξ is too small, the likelihood in a single interval will be flat. Therefore, if the ξ parameter is used (i.e., $\xi \neq 1$), great caution should be used to interpret the results. The use of the ξ parameter in the above example does not provide proof of its applicability, but is only shown to make clear that its use can greatly help to shorten the speed of the process. Also, it demonstrates the facility to work with the ancestral recombination graph and the Griffiths–Tavaré's (1994b) method of estimation.

10. DISCUSSION

We have developed a framework for gene mapping using linkage disequilibrium, by modeling the history of observed sequences collected on cases and controls via the ancestral recombination graph. The likelihood of the location of the disease gene is estimated by a Monte Carlo method based on a recurrence equation involving probability distributions. Several driving values for the Markov chain are used, and an importance sampling scheme is developed to obtain the likelihood for other values inside a single interval between two markers. One strength of this method is that it is a true multilocus method. Sequences of any length could be studied by this method, as long as interference can be ignored. We think that this method can be extended to more general situations and have a lot of potential. The method is designed to use markers with low mutation rates, but can be adapted easily to any type of marker, as well as any kind of mutation process. We have shown how to simulate the time to the next event in the presence of coalescence, mutation and recombination, and have also developed our method to accommodate variable population size. As pointed out by an anonymous referee and already mentioned by R.C. Griffiths (pers. comm.), an alternative way is to simulate two independent random variables X and Y such that

$$P(X > x) = \exp\left(- \int_t^x \binom{2}{n} \exp(\kappa u)\, du \right),$$

$$P(Y > y) = \exp\left(-\frac{n}{2}(y - t)(\alpha\theta + \beta\rho) \right),$$

and then take $W_{\tau+1} = \min(X, Y)$. We think that our proposed method to map a disease gene has potential, and it is our hope that its description will spark further interest in the subject.

ACKNOWLEDGMENTS

We thank Robert C. Griffiths and two anonymous referees for their comments on an earlier version of this paper. This work was supported by FCAR, CRSNG and SignalGene. N.J. Schork was supported in part by NIH Grants NHLB1 HL54998-01 and HL094-01. Fabrice Larribe was funded through SignalGene Inc, 8475, Avenue Christophe-Colomb, bureau 1000, Montréal, Qué, Canada H2M 2N9.

REFERENCES

Briggs, K. M. 1998. W-ology, or, some exactly solvable growth models, http://www.btexact.com/people/briggsk2/W-ology.html.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., and Knuth, D. E. 1996. On the Lambert W Function, *Adv. Comput. Math.* **5**, 329–359.

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

Fearnhead, P., and Donnelly, P. 2001. Estimating recombination rates from population genetic data, *Genetics* **159**, 1299–1318.

Graham, J., and Thompson, E. A. 1998. Disequilibrium likelihoods for fine-scale mapping of a rare allele, *Am. J. Hum. Gen.* **63**, 1517–1530.

Griffiths, R. C. 1981. Neutral two-locus multiple allele models with recombination, *Theor. Popul. Biol.* **19**, 169–186.

Griffiths, R. C. 1989. Genealogical-tree probabilities in the infinitely-many-site model, *J. Math. Biol.* **27**, 667–680.

Griffiths, R. C. 1991. The two-locus ancestral graph, *in* "Selected Proceedings of the Symposium on Applied Probability" (I. V. Basawa and R. L. Taylor, Eds.), Sheffield, 1989, Institute of Mathematical Statistics. IMS Lecture Notes-Monograph.

Griffiths, R. C. 2000. Ancestral inference from gene trees, *in* "Genes, Fossils, and Behaviour: An Integrated Approach to Human Evolution" (P. Donnelly and R. Foley, Eds.), IOS Press, to appear.

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

Griffiths, R. C., and Marjoram, P. 1996b. An ancestral recombination graph, *in* "IMA Volume on Mathematical Population Genetics" (P. Donnelly and S. Tavaré, Eds.), pp. 257–270, Springer-Verlag, New York.

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

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

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

Griffiths, R. C., and Tavaré, S. 1996. Markov chain inference methods in population genetics, *Math. Comput. Model.* **23**, 141–158.

Griffiths, R. C., and Tavaré, S. 1997. Computational methods for the coalescent, *in* "Progress in Population Genetics and Human Evolution" (P. Donnelly and S. Tavaré, Eds.), IMA Volumes in Mathematics and its Applications, Vol. 87, pp. 165–182, Springer-Verlag, Berlin.

Griffiths, R. C., and Tavaré, S. 1998. The age of mutation in a general coalescent tree, *Stochast. Models* **14**, 273–295.

Hudson, R. R. 1990. Gene genealogies and the coalescent process, *in* "Oxford Surveys in Evolutionary Biology" (D. Futuyma and J. Antonovics, Eds.), Vol. 7, pp. 1–44, Oxford Univ. Press, Oxford, UK.

Hudson, R. R. 2001. Generation of samples of gametes. Available at: http://home.uchicago.edu/ ~ rhudson1/source/mksamples.html.

Kingman, J. F. C. 1982. The coalescent, *Stochast. Process. Appl.* **13**, 235–248.

Lam, J. C., Roeder, K., and Devlin, B. 2000. Haplotype fine mapping by evolutionary trees, *Am. J. Hum. Gen.* **66**, 659–673.

Marjoram, P., Markovtsova, L., and Tavaré, S. 2000. I see dead people: Gene mapping via ancestral inference, *Genetic Analysis Workshop* 12. Available at: http://www-hto.usc.edu/papers/abstracts/GAW12.html.

Matsumoto, M., and Nishimura, T. 1998. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator, *ACM Trans. Modeling Comput. Simul.* **8**(1), 3–30.

McPeek, M. S., and Strahs, A. 1999. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping, *Am. J. Hum. Gen.* **65**, 858–875.

Morris, A. P., Whittaker, J. C., and Balding, D. J. 2000. Bayesian fine-scale mapping of disease loci, by hidden Markov models, *Am. J. Hum. Gen.* **67**, 155–169.

NBIC (National Biotechnology Information Facility)/Regents of the New Mexico State University, 2001. *EXTented NUMeric* (*EX-TNUM*) *range floating point type*. http://www.nbif.org/products/bioinfo/extnum/extnum.php.

Nordborg, M. 2001. Coalescent theory. Chapter 7, *in* "Handbook of Statistical Genetics" (D. Balding, M. Bishop, and C. Cannings, Eds.), pp. 179–212, Wiley, Chichester.

Nordborg, and M., Tavaré, S. 2002. Linkage disequilibrium: What history has to tell us, *Trends Genet.* **18**, 83–90.

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

Service, S. K., Temple Lang, D. W., Freimer, N. B., and Sandkuijl, L. A. 1999. Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder population, *Am. J. Hum. Gen.* **64**, 1728–1738.

Stephens, M., and Donnelly, P. 2000. Inference in molecular population genetics, *J. Roy. Statist. Soc. Ser. B* **62**, 605–655.

Stephens, M. 2001. Inference under the coalescent. Chapter 8, *in* "Handbook of Statistical Genetics" (D. Balding, M. Bishop, and C. Cannings Eds.), pp. 213–238, Wiley, Chichester.

Virtaneva, K., Miao, J., Träskelin, A.-L., Stone. N., Warrington, J. A., Weissenback, J., Myers, R. M., Cox, D. R., Sistonen, P., de la Chapelle, A., and Lehesjoki, A.-E. 1996. Progressive myoclonus epilepsy EMP1 locus to a 175kb interval in distal 21q, *Am. J. Hum. Gen.* **58**, 1247–1253.

Wagner, R. 2001. Mersenne twister random number generator, http://www-personal.engin.umich.edu/ wagnerr/MersenneTwister.html.

Weissten, E. 2000. Wolfram Research, http://math-world.wolfram.-com/LambertsW-Function.html.

Xiong, M., and Guo, S. W. 1997. Fine-scale genetic mapping based on linkage disequilibrium: Theory and applications, *Am. J. Hum. Gen.* **60**, 1513–1531.