Statistical Applications in Genetics and Molecular Biology

Volume 7, *Issue* 1 2008 *Article* 27

A Composite-Conditional-Likelihood Approach for Gene Mapping Based on Linkage Disequilibrium in Windows of Marker Loci

Fabrice Larribe* Sabin Lessard†

*Université du Québec à Montréal, larribe.fabrice@uqam.ca

†Université de Montréal, lessards@dms.umontreal.ca

A Composite-Conditional-Likelihood Approach for Gene Mapping Based on Linkage Disequilibrium in Windows of Marker Loci*

Fabrice Larribe and Sabin Lessard

Abstract

A composite-conditional-likelihood (CCL) approach is proposed to map the position of a trait-influencing mutation (TIM) using the ancestral recombination graph (ARG) and importance sampling to reconstruct the genealogy of DNA sequences with respect to windows of marker loci and predict the linkage disequilibrium pattern observed in a sample of cases and controls. The method is designed to fine-map the location of a disease mutation, not as an association study. The CCL function proposed for the position of the TIM is a weighted product of conditional likelihood functions for windows of a given number of marker loci that encompass the TIM locus, given the sample configuration at the marker loci in those windows. A rare recessive allele is assumed for the TIM and single nucleotide polymorphisms (SNPs) are considered as markers. The method is applied to a range of simulated data sets. Not only do the CCL profiles converge more rapidly with smaller window sizes as the number of simulated histories of the sampled sequences increases, but the maximum-likelihood estimates for the position of the TIM remain as satisfactory, while requiring significantly less computing time. The simulations also suggest that non-random samples, more precisely, a non-proportional number of controls versus the number of cases, has little effect on the estimation procedure as well as sample size and marker density beyond some threshold values. Moreover, when compared with some other recent methods under the same assumptions, the CCL approach proves to be competitive.

KEYWORDS: gene mapping, linkage disequilibrium, composite likelihood, conditional likelihood, ancestral recombination graph, importance sampling

---

*This research was supported in part by the Natural Sciences and Engineering Research Council of Canada. We thank Gabrielle Boucher and Marie Forest for their contribution to simulations.

1 Introduction

Linkage disequilibrium (LD) refers to the non-random association of variants along a DNA sequence. Suppose that, some generations ago, a mutation creating or influencing a given trait and responsible for a disease occurred on a certain sequence. Such a mutation will be called a trait-influencing mutation (TIM). If this mutation is rare enough so that it occurred only once in the history of the current population, all the individuals showing the trait, the cases, not only bear the disease allele from the original bearer, but also share some genetic material identical by descent (IBD) in the vicinity of the TIM locus. The original sequence on which the mutation occurred will not descend to the current cases in one invariant block however; this sequence will be broken into pieces by recombination events and altered by mutation events elsewhere along the sequence. The linkage disequilibrium pattern around the TIM locus is the result of all such events and, conversely, this pattern gives information about the exact position of the TIM.

As noted by Nordborg and Tavaré (2002), the main difference between LD analysis and linkage analysis is found in the type of sample that is under study rather than in the methodology that is used. Whereas linkage analysis considers sequences taken in closely related individuals whose recent genealogies are known, at least in part, to evaluate the likelihood of the position of the TIM, LD analysis concerns sequences that are chosen at random but are nonetheless related by their unknown ancestry since they come from the same population. Due to the larger number of recombination events that can occur when one considers the whole ancestry of the sample, mapping methods based on linkage disequilibrium in random samples may achieve a higher resolution than methods based on linkage analysis on family data whose limits have been pointed out by Boehnke (1994). For this reason, one might prefer the former methods for gene mapping at a finer scale.

Several measures of pairwise linkage disequilibrium used for fine-scale mapping have been studied and compared (see, e.g., Devlin and Risch 1995, and references therein). Such measures provide a level of association between a marker locus and a disease mutation locus. Most methods for gene mapping which use statistics based on single markers one at a time, or combinations of single markers (Terwilliger 1995, Xiong and Guo 1997, Collins and Morton 1998), ignore or underestimate the dependence of linked markers. As an illustration of this point, consider Figure 1 which shows association between 30 marker loci and a disease mutation locus as measured by r^2, the square of the correlation coefficient between alleles at two loci, for three data sets identified by A, B and F (see Section 4 for details). A high level of association around

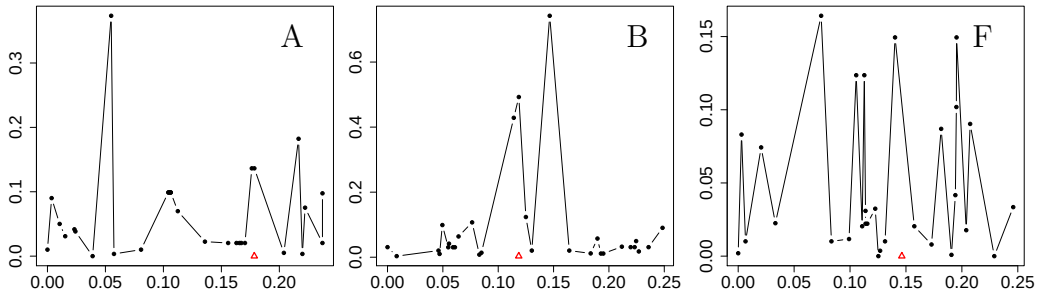
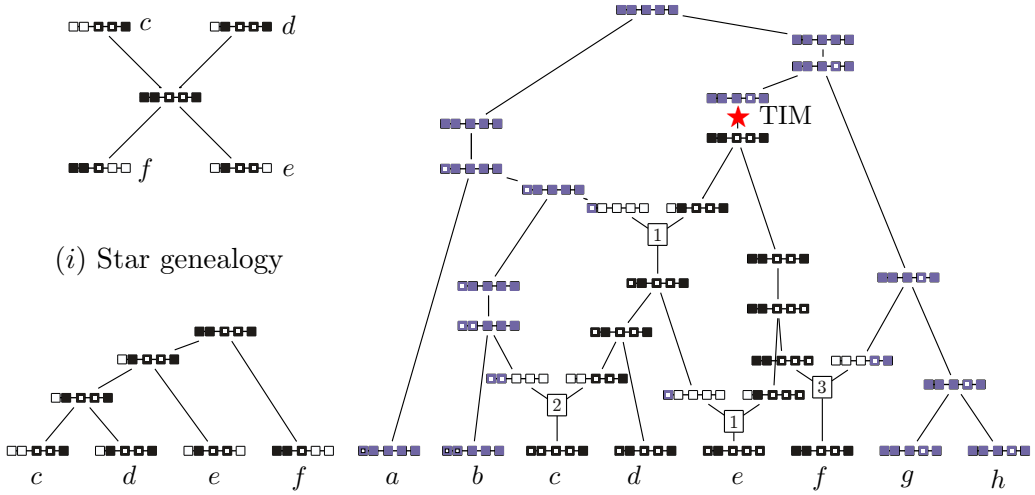


Figure 1. Association between 30 marker loci and a disease mutation locus for three data sets, A, B and F. The horizontal axis is for the position in the sequence (in cM) and the vertical axis for the value of r^2. A red triangle indicates the position of the putative mutation.

the disease mutation locus is observed for the data set B, while association for the data sets A and F is almost nonexistent so that it is useless to locate the disease mutation. The variability of such statistics has already been pointed out (see, e.g., Nordborg and Tavaré 2002). Likelihood methods based on multilocus linkage disequilibrium are generally more difficult to develop and implement, but they are also expected to give more precise estimates for the location of a TIM. Notice that other multilocus measures, e.g., the level of homozygosity (Sabatti and Risch 2002) or entropy (see, e.g., Nothnagel *et al.* 2002), have also been proposed.

One important aspect of any LD-based likelihood method is how the genealogy is modeled. The simplest assumption is to suppose a star genealogy for the cases (Figure 2, *i*). All lineages at the TIM locus directly coalesce backward in time to the most recent common ancestor (MRCA). This is the assumption made in McPeek and Strahs (1999), Morris *et al.* (2000) and Liu *et al.* (2001), for example, in which the portion of the original sequence around the TIM locus inherited by the current cases is inferred from mutation and sharing patterns at given marker loci. A star genealogy has the advantage to simplify the analysis since then, given the MRCA, all sampled sequences are independent and some correlation can be taken into account by introducing correcting terms. Such a genealogy appears to be appropriate for a very fast expanding population. Another assumption is to suppose a tree genealogy for the cases with a succession of coalescence events at the TIM locus from the sampled sequences to the MRCA as in Rannala and Slatkin (1998), Rannala and Reeve (2001), Garner and Slatkin (2002)(see Figure 2, *ii*). Morris *et al.* (2002) uses a tree genealogy for the cases and a star genealogy for the con-

trols. A tree genealogy does not assume independence between the sampled sequences given the MRCA, but it ignores recombination events between them and between any two ancestral sequences. If the sample size is very small and recombination events very rare, this assumption makes sense, but in general we have to consider the possibility that an ancestral sequence comes from two parental ancestral sequences that have recombined. This will be the case if the recombination event occurs within the ancestral material of the sequence. Going backward in time, such a sequence will have two parental sequences bearing some material ancestral to the sample, and not only one. Thus, this leads to a graph genealogy, called an ancestral recombination graph, ARG (Griffiths 1981, Hudson 1983, Griffiths and Marjoram 1996) (see Figure 2, *iii*). The incorporation of such a graph in gene mapping based on a finite set of markers observed in cases as well as in controls is the objective of the present work. Zollner and Pritchard (2005) considered a local approximation of the ARG at a focal point along the sequence. We will consider the ARG at multiple points at a time.



(*i*) Star genealogy

(*ii*) Tree genealogy

(*iii*) ARG genealogy

Figure 2. Schematic representation of three genealogies for a sample of 8 sequences at 5 loci: four cases in black that bear a disease allele at locus 3 (c, d, e and f) and four controls in blue (a, b, g and h); ■ for a primitive allele at an ancestral locus; ◘ for a derived allele at an ancestral locus; □ for a non-ancestral locus; ▣ for a recombination event between loci m and $m+1$. *(i)* Star genealogy for cases; *(ii)* Tree genealogy for cases; *(iii)* Ancestral recombination graph for cases and controls.

The ARG is a birth and death process that generalizes the coalescent (Kingman 1982) by allowing for recombination events when one follows the ancestry of sampled sequences backward in time. Excellent reviews and perspectives on the subject can be found in the literature (Hudson 1990, Nordborg 2001, Tavaré and Zeitouni 2004, Nordborg and Tavaré 2002, Hein *et al.* 2005, Wakeley 2008). The approach can be used to deduce a recurrence equation backward in time for the ancestral material of sampled sequences under recombination and mutation according, e.g., to an infinitely-many-sites mutation scheme. Then, the likelihood of the sampled sequences can be estimated using importance sampling to reconstruct the unknown genealogy. This has been done, e.g., to evaluate the recombination rate (Griffiths and Marjoram 1996, Fearnhead and Donnelly 2001) and the mutation rate (Stephens and Donnelly 2000), or to map a TIM locus based on single nucleotide polymorphisms (SNPs) as markers with a primitive type and a derived type at each locus (Larribe *et al.* 2002). The primitive type refers to the allele of the most recent common ancestor, and has not the same meaning as an ancestral locus, as we shall see in section 2. The derived type corresponds to the mutant allele. Other approaches to reconstruct unknown genealogies and estimate unknown parameters include Markov chains Monte Carlo (MCMC) methods and Bayesian inference (Kuhner *et al.* 2000, Nielsen 2000). Unfortunately, the statistical procedures to deal with missing data, that is, importance sampling and MCMC, require so intensive computing that they are untractable in most practical cases in estimating a full-likelihood function based on all the information available in sampled DNA sequences. Recently, Zollner and Pritchard (2005) proposed a local approximation of the ARG in the vicinity of a putative disease mutation locus to simplify gene mapping in a case-control context. On the other hand, Minichiello and Durbin (2006) developed a heuristic algorithm to infer plausible ancestral recombination graphs. More recently, Wu (2007) developed an association method using ancestral recombination graphs that minimize the number of recombination events.

A composite-likelihood approach (see, e.g., Lindsay 1988, Varin and Vidoni 2005, Varin 2008) has already been proposed to tackle the problem of excessive computing time and it has been applied to the estimation of the recombination rate (Hudson 2001, Fearnhead and Donnelly 2002, McVean *et al.* 2002, Wall 2004). Basically, a composite likelihood function is some average of marginal likelihood functions, each one using only part of all the information available and requiring less computing time to estimate than the full-likelihood function. In the case of the recombination rate, for instance, it may be a product of likelihood functions for two or three sites at a time. Li and Stephens (2003) proposed a product of approximate conditional (PAC) likelihoods for

haplotypes sampled sequentially one at a time.

Other tractable methods to estimate the recombination rate include summary statistics including bounds for the minimum number of recombination events required to explain the history of the sample (see, e.g., Hudson and Kaplan 1985, Hudson 1987, Wall 2000, Myers and Griffiths 2003, and references therein).

In this paper, we present and extend a maximum-likelihood approach to estimate the position of a TIM from observed markers in cases and controls that is based on the ARG coupled with importance sampling originally developed by Griffiths and Marjoram (1996) to estimate the recombination rate. In section 2, we describe in detail the full-likelihood procedure for a rare recessive disease allele and a given set of SNPs as markers. The recurrence equation backward in time for the ancestral material of sampled sequences and the likelihood function with a given proposal distribution for their history are deduced. The mutation term in the recurrence equation and some misprints in the likelihood function given in Larribe *et al.* (2002) are corrected. In section 3, we propose a composite-conditional-likelihood (CCL) function which is a weighted product of conditional likelihood functions associated to windows of k contiguous marker loci, the condition being the observed sample configuration at those marker loci and the weight being inversely proportional to the number of windows of that size encompassing the TIM locus. Then, the method is applied to a range of data sets in section 4. Numerical comparisons with other recent methods are presented in section 5. The results and limitations of the method are discussed in section 6.

2 Method for single nucleotide polymorphisms

We assume that the disease allele under study is rare and recessive: a case individual is supposed to carry the disease allele twice (the risk allele, also called derived allele) at the TIM locus, whereas a control individual is supposed to carry the non-disease allele twice (also called primitive) at the TIM locus. There is a one-to-one correspondence between the individual phenotype and the set of alleles at the TIM locus, so that the allele at this locus carried by sequences sampled in cases and controls is known with certainty. This is a basic genetic model to test the method proposed, but it can be extended to deal with more general situations allowing for partial penetrance. In such a case, the alleles at the TIM locus have to be inferred using, e.g., the frequency of the disease in the whole population and a Bayesian approach. Essentially,

the method proposed is based on reconstructing the genealogy of sampled sequences knowing the alleles at given marker loci and a TIM locus. This is precisely what characterizes the present method.

A sequence of type i with respect to L loci is described by an ordered L-tuple $\mathbf{s}^{(i)} = (s_1^{(i)}, ..., s_L^{(i)})$ where $s_m^{(i)}$ represents the genetic material at the m-th locus from the beginning (left end by convention) to the end (right end by convention) of a DNA string. One of these loci is the unknown TIM locus, whose exact position has to be estimated, while the others are known marker loci. The distance between loci m and $m+1$, which corresponds to the length of segment m, is represented by the recombination rate between these two loci and is denoted by r_m. The corresponding distance from the beginning of the sequence to the TIM locus is denoted by r_T. If the TIM locus corresponds to locus m, then the distances between the TIM locus and the first loci to its left and right, r_{m-1} and r_m, are represented by r_l and r_r, respectively. Interference is ignored, which means that the distances r_m for $m = 1, ..., L$ are supposed to be additive. This is a reasonable assumption if the distances are small enough. In particular, the total length of the sequence, from the first locus to the last one, is $r = \sum_m r_m$. It is assumed throughout that the TIM locus does not correspond to the first or last locus so that $0 < r_T < r$. Figure 3 illustrates this situation, where $x_1, \ldots, x_L$ refer to the positions of the L loci with the convention $x_1 = 0$.

The mutation rate per sequence per generation at locus m is denoted by u_m. It is assumed throughout this section that mutation is so rare that it occurred only once in the coalescent tree of sampled sequences at every polymorphic locus. This is a reasonable assumption for SNPs. Moreover, the older and the younger types, called the primitive and derived types and represented by 0 and 1, respectively, are present in the sample and supposed to be known (these types can be inferred from others species, for instance). Therefore, we have $s_m^{(i)} = 0$ or 1 for every i and m considered.

We will trace back the history of a sample of sequences from the time of the sampling to the time of the most recent common ancestor (MRCA).

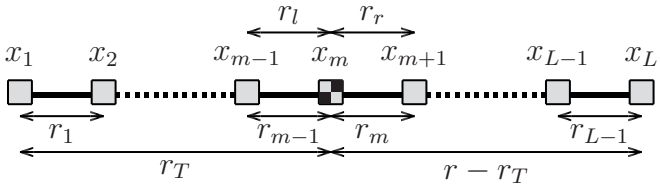


Figure 3. Illustration of a sequence of L loci with the TIM locus corresponding to locus m.

10.2202/1544-6115.1298

A sequence $\mathbf{s}^{(i)} = (s_1^{(i)}, ..., s_L^{(i)})$ ancestral to the sample is characterized by a subset of loci at which it is ancestral, the subset of loci that are copied in the sample, denoted by $A^{(i)}$, and the genetic material it carries at these loci, that is, $s_m^{(i)}$ for all m in $A^{(i)}$. These define the type of an ancestral sequence i. Moreover, let $B^{(i)}$ be the set of all loci m such that $m_1 \leq m < m_2$ for some m_1 and m_2 in $A^{(i)}$. Notice that a mutation event at locus m of a given sequence of type i affects the ancestral material if m belongs to $A^{(i)}$, while a recombination event in segment m between loci m and $m+1$ does it if m belongs to $B^{(i)}$. The sets $A^{(i)}$ and $B^{(i)}$ represent the ancestral markers and segments, respectively, of a sequence of type i.

A mutation event can occur backward in time at locus m in $A^{(i)}$ of an ancestral sequence of type i only if there is a single ancestral sequence of type i and a single type i in the ancestral material with $s_m^{(i)} = 1$, which means $s_m^{(\cdot)} = \sum_i n_m^{(i)} s_m^{(i)} = 1$, where $n_m^{(i)}$ is the multiplicity of the sequences of type i ancestral at locus m. Then, the locus m will be said unique and the parental sequence of the sequence of type i will be of type j satisfying $s_m^{(j)} = 0$ and $s_l^{(j)} = s_l^{(i)}$ for all $l \neq m$ in $A^{(j)} = A^{(i)}$.

On the other hand, a recombination event in segment m of an ancestral sequence i with $B^{(i)}$ containing m will produce backward in time two parental sequences of types j and k satisfying $s_l^{(j)} = s_l^{(i)}$ for $l \leq m$ in $A^{(i)}$, which defines $A^{(j)}$, and $s_l^{(k)} = s_l^{(i)}$ for $l \geq m+1$ in $A^{(i)}$, which defines $A^{(k)}$.

Finally, a coalescence event between two ancestral sequences of types i and j can occur backward in time only if these are compatible in the sense that they are of the same type, that is, $i = j$, or they are of different types, that is, $i \neq j$, but then $s_m^{(i)} = s_m^{(j)}$ for all m in both $A^{(i)}$ and $A^{(j)}$. Then, the result of the coalescence event will be a parental sequence of type k satisfying $s_m^{(k)} = s_m^{(i)}$ if m belongs to $A^{(i)}$ and $s_m^{(k)} = s_m^{(j)}$ if m belongs to $A^{(j)}$, with the parental sequence being ancestral exactly at all these markers.

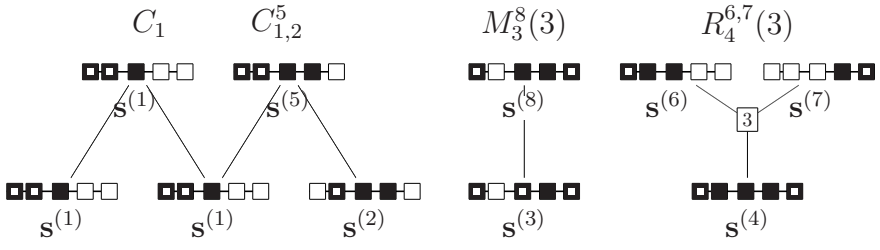


Figure 4. Examples of events backward in time affecting the ancestral material of sequences.

We will use the following notation for the different events of coalescence (C), mutation (M) and recombination (R) backward in time that can affect the ancestral material of the initial sample:

- C_i for C of two sequences of type i,
- C_{ij}^k for C of two sequences of types i and j, $i \neq j$, to a parental sequence of type k,
- $M_i^j(m)$ for M at locus m of a sequence of type i to a parental sequence of type j,
- $R_i^{jk}(m)$ for R anywhere in segment m of a sequence of type i to two parental sequences of types j and k.

We refer to Figure 4 for examples of events backward in time.

Time is measured in units of $2N(0)$ generations, where $N(0)$ is the population size at the time of sampling. The τ-th event backward in time affecting the ancestral material occurs at time back t_τ, with $\tau = 0$ corresponding to the initial sampling and $\tau = \tau^*$ corresponding to the coalescence event to the MRCA.

The population size at time back t is assumed to be given by the equation $N(t) = N(0)\exp(-\kappa\, t)$, where κ is positive if the population is growing exponentially fast forward in time or negative if it is declining. If $\kappa = 0$, then the population size is constant and denoted by N. The ratio $\lambda(t) = N(0)/N(t)$ is used throughout.

Let $\rho_m = 4N(0)r_m$ be the scaled recombination rate per generation between loci m and $m+1$. For the whole sequence, the scaled recombination rate per generation is $\rho = \sum_m \rho_m$.

Similarly, let $\theta_m = 4N(0)u_m$ be the scaled mutation rate at locus m. For the L loci, the scaled mutation rate is $\theta = \sum_m \theta_m$. If $u_m = u$ for $m = 1, ..., L$, then $\theta_m = \theta/L$ for $m = 1, ..., L$ where $\theta = 4N(0)Lu$ is the scaled mutation rate per generation for the whole sequence.

The configuration of the ancestral material at time back t_τ is given by the types of the ancestral sequences with their multiplicities following the τ-th change in the ancestral material and is denoted by $\mathbf{H}_\tau$. In particular, $\mathbf{H}_0$ at time $t_0 = 0$ is the set of the sampled sequences, and $\mathbf{H}_{\tau^*}$ at time t_{τ^*} contains only one sequence, the sequence carried by the MRCA. Owing to the assumptions made, this sequence has 0 at every locus.

If there are n_i ancestral sequences of type i at time back t_τ for a total number of ancestral sequences $n = \sum_i n_i$, then the rates of change in the ancestral material by coalescence, mutation and recombination, respectively, at a subsequent time back t, given no change from t_τ to t, are $n(n-1)\lambda(t)/2$, $n\alpha\theta/2$ and $n\beta\rho/2$, respectively, where α and β are the probabilities that a mutation event and a recombination event, respectively, affect the ancestral material, that is,

$$\alpha = \sum_i \frac{n_i}{n} \sum_{m \in A^{(i)}} \frac{\theta_m}{\theta}$$

and

$$\beta = \sum_i \frac{n_i}{n} \sum_{m \in B^{(i)}} \frac{\rho_m}{\rho}.$$

The total rate of change in the ancestral material is then $nD(t)/2$, where

$$D(t) = (n-1)\lambda(t) + \alpha\theta + \beta\rho.$$

The time of occurrence of the $(\tau+1)$-th change in the ancestral material given the time of occurrence of the τ-th change has some density function determined by the total rate of change, and represented by $g(t_{\tau+1}|t_\tau)$, that can be simulated using mutually independent uniform random variables (Donnelly and Tavaré 1995, Griffiths and Tavaré 1996, Larribe *et al.* 2002). The probability that the $(\tau+1)$-th change in the ancestral material is caused by a coalescence event, a mutation event and a recombination event, respectively, given the time of occurrence of the τ-th change and $\mathbf{H}_\tau$, is then

$$P_\tau(C) = \int_{t_\tau}^{\infty} \left[\frac{(n-1)\lambda(t_{\tau+1})}{D(t_{\tau+1})}\right] g(t_{\tau+1}|t_\tau)\ dt_{\tau+1},$$

$$P_\tau(M) = \int_{t_\tau}^{\infty} \left[\frac{\alpha\theta}{D(t_{\tau+1})}\right] g(t_{\tau+1}|t_\tau)\ dt_{\tau+1},$$

$$P_\tau(R) = \int_{t_\tau}^{\infty} \left[\frac{\beta\rho}{D(t_{\tau+1})}\right] g(t_{\tau+1}|t_\tau)\ dt_{\tau+1},$$

respectively, by conditioning on the value taken by $t_{\tau+1}$ given the value of t_τ. In the case of a constant population size, $\lambda(t) = 1$ and the above probabilities reduce to

$$P_\tau(C) = \frac{n-1}{n-1+\alpha\theta+\beta\rho},$$

$$P_\tau(M) = \frac{\alpha\theta}{n-1+\alpha\theta+\beta\rho},$$

$$P_\tau(R) = \frac{\beta\rho}{n-1+\alpha\theta+\beta\rho}.$$

Considering only one mutation event at every polymorphic locus with 0 and 1 being the primitive and derived types, respectively, and assuming a random sample of ancestral sequences at every step of change backward in time by coalescence, mutation or recombination, the likelihood of the configuration $\mathbf{H}_\tau$ is related to the likelihood of all previous possible configurations $\mathbf{H}_{\tau+1}$ by the recurrence system of equations

$$\begin{aligned} Q(\mathbf{H}_\tau) &= P_\tau(C)\sum_{n_i>1}\frac{(n_i-1)}{(n-1)}Q(\mathbf{H}_\tau+C_i) \\ &+ P_\tau(C)\sum_{\substack{i\neq j\\ \text{compatible}}}\frac{2(n_k+1-\delta_{ik}-\delta_{jk})}{(n-1)}Q(\mathbf{H}_\tau+C_{ij}^k) \\ &+ P_\tau(M)\sum_i\sum_{\substack{m\in A^{(i)}\\ \text{unique}}}\frac{\theta_m}{\alpha\theta}\frac{(n_j+1)}{n}Q(\mathbf{H}_\tau+M_i^j(m)) \qquad (1) \\ &+ P_\tau(R)\sum_i\sum_{m\in B^{(i)}}\frac{\rho_m}{\beta\rho}\frac{(n_j+1)(n_k+1)}{n(n+1)}Q(\mathbf{H}_\tau+R_i^{jk}(m)). \end{aligned}$$

This recurrence system of equations simplifies and corrects an expression given in Larribe *et al.* (2002) in the case of a finite set of L loci with a uniform mutation rate, that is, $\theta_m = \theta/L$. We refer to Griffiths and Marjoram (1996) for the case of a continuous set of loci.

In the above system of equations, the notation $\mathbf{H}_\tau + C_i$ is used for the configuration obtained from the configuration $\mathbf{H}_\tau$ and the coalescence of two ancestral sequences of type i, and similarly for the other possibilities. The equations are obtained by conditioning on the kind of event changing the ancestral material at time back $t_{\tau+1}$. Given a mutation event in the ancestral material, for instance, whose probability is $P_\tau(M)$, the configuration $\mathbf{H}_\tau$ is compatible only with a configuration $\mathbf{H}_{\tau+1} = \mathbf{H}_\tau + M_i^j(m)$ with one less sequence of type i but one more of type j for some sequence of type i that is ancestral at locus m, and the only one to carry 1 at this locus in $\mathbf{H}_\tau$, and then the mutation event must occur at locus m of a parental sequence of type j, whose probability is $(n_j+1)/n$ times θ_m/θ divided by α, the probability of the condition that the mutation event occurs in the ancestral material. Similarly, given a coalescence event, whose probability is $P_\tau(C)$, the configuration $\mathbf{H}_\tau$ is compatible with either (a) $\mathbf{H}_{\tau+1} = \mathbf{H}_\tau + C_i$ containing one less sequence whose

type is i, and then one sequence of this type must be the parental sequence of the two that coalesce, whose probability is $(n_i - 1)/(n-1)$, or (b) $\mathbf{H}_{\tau+1} = \mathbf{H}_\tau + C_{ij}^k$ containing one less sequence of type i and one less of type $j \neq i$ but one more of type k, and then one sequence of type k must be the parental sequence of two sequences of types i and j, whose probability is $(n_k + 1 - \delta_{ik} - \delta_{jk})/(n-1)$ times 2, since the sequences are not ordered. Finally, given a recombination event, whose probability is $P_\tau(R)$, the configuration $\mathbf{H}_\tau$ is compatible only with a configuration $\mathbf{H}_{\tau+1} = \mathbf{H}_\tau + R_i^{jk}(m)$ with one less sequence of type i but one more of type j and one more of type k for some sequence of type i with m in $B^{(i)}$, and then the recombination event must occur in segment m of two parental sequences of types j and k to produce a sequence of type i, whose probability is $[(n_j+1)(n_k+1)]/[n(n+1)]$ times ρ_m/ρ divided by β, the probability of the condition that the recombination event affects the ancestral material.

The above system of equations is in the form

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

where

$$a(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) = \begin{cases} \frac{P_\tau(C)(n_i-1)}{(n-1)} & \text{if } \mathbf{H}_{\tau+1} = \mathbf{H}_\tau + C_i, \\ \frac{2P_\tau(C)(n_k+1-\delta_{ik}-\delta_{jk})}{(n-1)} & \text{if } \mathbf{H}_{\tau+1} = \mathbf{H}_\tau + C_{ij}^k, \\ \frac{P_\tau(M)\theta_m(n_j+1)}{\alpha\theta n} & \text{if } \mathbf{H}_{\tau+1} = \mathbf{H}_\tau + M_i^j(m), \\ \frac{P_\tau(R)\rho_m(n_j+1)(n_k+1)}{\beta\rho n(n+1)} & \text{if } \mathbf{H}_{\tau+1} = \mathbf{H}_\tau + R_i^{jk}(m). \end{cases}$$

This system cannot be used directly to reconstruct the history of the sample backward in time. A proposal distribution for the conditional probabilities backward in time $P(\mathbf{H}_{\tau+1} \mid \mathbf{H}_\tau)$ is needed. We choose

$$P(\mathbf{H}_{\tau+1} \mid \mathbf{H}_\tau) = \frac{b(\mathbf{H}_\tau, \mathbf{H}_{\tau+1})}{f(\mathbf{H}_\tau)}, \tag{2}$$

where

$$b(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) = \begin{cases} \frac{a(\mathbf{H}_\tau, \mathbf{H}_{\tau+1})}{(n_j+1)(n_k+1)} & \text{if } \mathbf{H}_{\tau+1} = \mathbf{H}_\tau + R_i^{jk}(m), \\ a(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) & \text{otherwise,} \end{cases}$$

while $f(\mathbf{H}_\tau)$ is a normalizing factor, that is,

$$f(\mathbf{H}_\tau) = \sum_{\mathbf{H}_{\tau+1}} b(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}).$$

Therefore, we get

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

where

$$f(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) = \begin{cases} (n_j+1)(n_k+1) f(\mathbf{H}_\tau) & \text{if } \mathbf{H}_{\tau+1} = \mathbf{H}_\tau + R_i^{jk}(m), \\ f(\mathbf{H}_\tau) & \text{otherwise.} \end{cases} \quad (3)$$

Notice that the value of $b(\mathbf{H}_\tau, \mathbf{H}_{\tau+1})$ when $\mathbf{H}_{\tau+1} = \mathbf{H}_\tau + R_i^{jk}(m)$ reduces to $[P_\tau(R)\rho_m]/[\beta\rho n(n+1)]$, which does not depends on i, j, k. This simplifies the procedure since we just have to take into account the possibility of a recombination event between loci m and $m+1$ first, and then choose at random the triplet i, j, k if such an event occurs.

Iterating from $\tau = 0$ to $\tau = \tau^* - 1$, we have

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

where the expectation is on P, the proposal distribution defined previously, given $\mathbf{H}_0$. This can be seen as an importance sampling procedure (see, e.g., Stephens 2001, and references therein). The proposal distribution depends on unknown parameters, including the parameter to estimate, r_T. Using a proposal distribution P_0 that uses a trial value r_0 for r_T and known estimates for the other parameters (mutation rates, initial population size and population growth rate), the likelihood of the initial sample takes the form

$$Q(\mathbf{H}_0) \;=\; E_{P_0} \left[Q(\mathbf{H}_{\tau^*}) \prod_{\tau=0}^{\tau^*-1} \frac{f(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) P(\mathbf{H}_{\tau+1}|\mathbf{H}_\tau)}{P_0(\mathbf{H}_{\tau+1}|\mathbf{H}_\tau)} \right].$$

The assumptions are such that the configuration $\mathbf{H}_{\tau^*}$ must contain only the sequence with 0 at every locus whose likelihood is 1, that is, $Q(\mathbf{H}_{\tau^*}) = 1$.

The trial value r_0 for the position of the TIM is usually chosen in the middle of two adjacent marker loci, and the $L-2$ pairs of such marker loci

are considered in turn to cover the whole sequence. Assuming that the TIM locus corresponds to locus m, then r_0 is chosen in the middle of the interval $[x_{m-1}, x_{m+1}]$, which will be referred to as interval m. Then, an estimate of the likelihood function for r_T in interval m, represented by $L_m(r_T)$ and given by $Q(\mathbf{H}_0)$ for r_T in interval m and 1 elsewhere, is

$$\hat{L}_m(r_T) = \frac{1}{K}\sum_{k=1}^{K}\left[\prod_{\tau=0}^{\tau^*-1}\frac{f(\mathbf{H}_\tau^{(k)}, \mathbf{H}_{\tau+1}^{(k)})P(\mathbf{H}_{\tau+1}^{(k)}|\mathbf{H}_\tau^{(k)})}{P_0(\mathbf{H}_{\tau+1}^{(k)}|\mathbf{H}_\tau^{(k)})}\right],$$

where $\mathbf{H}_\tau^{(k)}$ for $\tau = 0, ..., \tau^*$ and $k = 1, ..., K$ are K histories of the sample independently simulated backward in time using the proposal distribution P_0 for a TIM at locus m. By convention, $\hat{L}_m(r_T) = 1$ for r_T outside interval m. The estimation procedure can be repeated with the maximum point of $\hat{L}_m(r_T)$ in interval m as trial value for r_T. Finally, after repeating the same procedure for $m = 2, ..., L-1$, the parameter r_T is estimated by the maximum point of

$$\hat{L}(r_T) = \prod_{m=2}^{L-1}\hat{L}_m(r_T) \tag{4}$$

over the whole sequence, which is an estimate of the likelihood function $L(r_T) = \prod_{m=2}^{L-1} L_m(r_T)$. The maximum point is represented by $\hat{r_T}$.

Notice that we could have chosen $b(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) = a(\mathbf{H}_\tau, \mathbf{H}_{\tau+1})$ in the proposal distribution for reconstructing the history of a sample backward in time, in which case the expression for the likelihood of the sample configuration would have taken a simpler form with $f(\mathbf{H}_\tau, \mathbf{H}_{\tau+1}) = f(\mathbf{H}_\tau)$. However, this choice would require to compute all the products $(n_j+1)(n_k+1)$ at every step of the reconstructing procedure even if no recombination event occurs, which would significantly increase the simulation time.

Here is a sketch of the algorithm, called MapARG, to obtain an estimate of the likelihood $L(r_T)$.

Algorithm A_1:
for $m = 2, \ldots, L-1$ **do** ⌊ *for each interval m* ⌉
  for $k = 1, \ldots, K$ **do** ⌊ *K ancestral recombination graphs* ⌉
    insert TIM in interval m.
    for $\tau = 0, \ldots, \tau^\star - 1$ **do** ⌊ *simulation of each step of the history* ⌉
      1. evaluate all possible states $\mathbf{H}_{\tau+1}$;
      2. evaluate $P_0(\mathbf{H}_{\tau+1} \mid \mathbf{H}_\tau)$ for these states with eq. (2);
      3. choose randomly an event among all possible events;
      4. evaluate $f(\mathbf{H}_\tau, \mathbf{H}_{\tau+1})$ using eq. (3);

10: 5. evaluate $w_\tau(r_T) = f(\mathbf{H}_\tau, \mathbf{H}_{\tau+1})P(\mathbf{H}_{\tau+1} \mid \mathbf{H}_\tau)/P_0(\mathbf{H}_{\tau+1} \mid \mathbf{H}_\tau)$
11: **end for**
12: evaluate $y_k(r_T) = \prod_{\tau=0}^{\tau^\star-1} w_\tau(r_T)$.
13: **end for**
14: $\hat{L}_m(r_T) = \frac{1}{K}\sum_{k=1}^{K} y_k(r_T)$. ⌊ *likelihood in interval* m ⌉
15: **end for**
16: 1. evaluate $\hat{L}(r_T) = \prod_{m=2}^{L-1} \hat{L}_m(r_T)$. ⌊ *global likelihood* ⌉
17: 2. $\hat{r}_T$ corresponds to the maximum of $\hat{L}(r_T)$. ⌊ *MLE* ⌉

3 Composite conditional likelihood

We propose to consider marginal likelihood functions based on the genetic material carried by sampled sequences in cases and controls at a fixed number d of contiguous marker loci, called windows of size d, as illustrated in Figure 5. The total number of such windows is $G = L - d$, where L is the total number of marker loci, including the TIM locus, and therefore $L - 1$ is the number of marker loci. Going from the left to the right of the sequence, these windows are numbered from 1 to G. Notice that interval m between the $(m-1)$-th marker locus and the next is included in window g if and only if g is comprised between

$$\underline{g}(m) = \max(1, m + 1 - d)$$

and

$$\overline{g}(m) = \min(m - 1, L - d).$$

In such a case, let $L_{m,g}(r_T)$ represent the marginal likelihood function for the position of the TIM in interval m, as defined above, which uses only the information in window g, that is, which ignores all marker loci outside window g. Moreover, $L_{m,g}(r_T) = 1$ by convention if r_T falls outside interval m.

Then, a composite-likelihood (CL) function for the position of the TIM along the whole sequence which uses windows of size d and gives the same weight to every interval can be defined as

$$CL_d(r_T) = \prod_{m=2}^{L-1} \left(\prod_{g=\underline{g}(m)}^{\overline{g}(m)} L_{m,g}(r_T) \right)^{w_m},$$

where

$$w_m = \frac{1}{\overline{g}(m) - \underline{g}(m) + 1}.$$

10.2202/1544-6115.1298

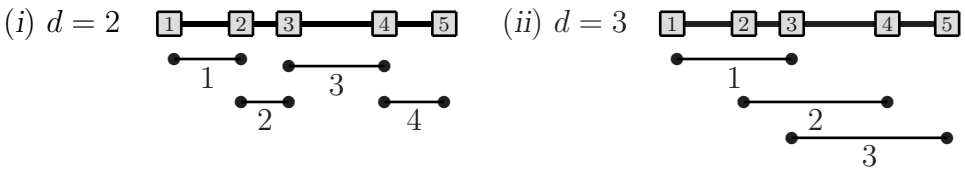


Figure 5. Examples of windows with 5 marker loci: (i) four windows of size $d = 2$; (ii) three windows of size $d = 3$.

Such a function does not correspond to an exact likelihood, since the events occurring in different windows are not independent, but it makes sense as an average.

Actually, the above composite-likelihood function is not well defined. The problem is that $L_{m,g}(r_T)$ does not use the same information from one window g to another. As a matter of fact, this function is the likelihood of the sample configuration at the TIM locus including its position r_T in interval m, represented by $\mathbf{H}_0^{r_T}$, jointly with the sample configuration at the d marker loci belonging to window g, represented by $\mathbf{H}_0^g$. This can be written as

$$L_{m,g}(r_T) = Q(\mathbf{H}_0^{r_T}, \mathbf{H}_0^g).$$

But the likelihood of $\mathbf{H}_0^g$ may differ from one g to another. Therefore, we propose to use the conditional marginal likelihood defined as

$$L_{m,g}(r_T|\mathbf{H}_0^g) = \frac{Q(\mathbf{H}_0^{r_T}, \mathbf{H}_0^g)}{Q(\mathbf{H}_0^g)} \tag{5}$$

and the corresponding composite-conditional-likelihood (CCL) function

$$CCL_d(r_T) = \prod_{m=2}^{L-1} \left(\prod_{g=\underline{g}(m)}^{\overline{g}(m)} L_{m,g}(r_T|\mathbf{H}_0^g) \right)^{w_m}. \tag{6}$$

If $d = L - 1$, that is, when one considers a single window for the whole set of markers, then the function $CCL_{L-1}(r_T)$ is proportional to $CL_{L-1}(r_T)$, which is equal to $L(r_T)$. Therefore, this case corresponds to the full-likelihood (FL) procedure.

When windows of size d are considered, the likelihood function is estimated $d-1$ times in all intervals but the first $d-2$ and the last $d-2$, which is almost all intervals if the number of markers is large compared to the window size. This procedure can be compared to the full-likelihood procedure repeated the

same number of times, that is, $d-1$ times, with the composite-full-likelihood (CFL) function

$$CFL_{d-1}(r_T) = \prod_{g=1}^{d-1} \left(L^{(g)}(r_T)\right)^{1/(d-1)} = \prod_{m=2}^{L-1} \left(\prod_{g=1}^{d-1} L_m^{(g)}(r_T) \right)^{1/(d-1)},$$

where $L^{(g)}(r_T)$ for $g = 1, ..., d-1$ are independent copies of the likelihood function $L(r_T)$ and similarly $L_m^{(g)}(r_T)$ for the likelihood function $L_m(r_T)$ in interval m.

Since the proposal distribution used is far from being the ideal distribution (see, e.g., Stephens and Donnelly 2001), it is challenging to find histories of the subsample associated with high contributions to its likelihood. Recall that this likelihood is an expectation with respect to the proposal distribution; the likelihood increases significantly when a history associated with a high value is found, and then decreases slowly as more histories associated with lower values are simulated. The overall result is that the height of the likelihood function fluctuates as the number of simulations increases. The CFL function, which corresponds to a geometric average over several runs of simulations, should help resolve this issue.

Here is the sketch of the algorithm used to obtain the composite conditional likelihood. The expressions for $Q(\mathbf{H}_0^g)$ and $Q(\mathbf{H}_0^{r_T}, \mathbf{H}_0^g)$ are evaluated independently; the evaluation of $Q(\mathbf{H}_0^{r_T}, \mathbf{H}_0^g)$ proceeds as in the previous section (Algorithm A_1), with the difference that it is done with d marker loci at a time; the evaluation of $Q(\mathbf{H}_0^g)$ is also similar, but with two differences: (1) it is done with d marker loci at a time, and (2) the TIM locus is not inserted into the sequence.

1: Algorithm A_2:
2: **for** $g = 1, \ldots, G$ **do** ⌊ *for each window g* ⌉
3: evaluate $Q(\mathbf{H}_0^g)$ ⌊ *denominator of eq. (5)* ⌉
4: **for** each interval in window g **do**
5: evaluate $Q(\mathbf{H}_0^{r_T}, \mathbf{H}_0^g)$. ⌊ *numerator of eq. (5)* ⌉
6: **end for**
7: evaluate $L_{m,g}(r_T \mid \mathbf{H}_0^g)$. ⌊ *eq. (5)* ⌉
8: **end for**
9: 1. evaluate $C\hat{C}L_d(r_T)$. ⌊ *eq. (6)* ⌉
10: 2. $\hat{r}_T$ corresponds to the maximum of $C\hat{C}L_d(r_T)$. ⌊ *MLE* ⌉

10.2202/1544-6115.1298

4 Results on simulated data sets

A C++ program, named MapARG, implements the method described in the previous section. Our objective here is to ascertain its validity and its accuracy on a limited number of simulated data sets but under a variety of conditions. The method is still computationally intensive, and it would take a long time to get results for a large number of samples as in Morris *et al.* (2002), for one thousand data sets.

Using the `ms` program of Hudson (2002) for generating sequences under recombination in a neutral population of constant size, we have generated 16 samples of 10 000 sequences for a fixed value of the scaled recombination rate corresponding to a fixed number of sites along the whole sequence (chosen to be $\rho = 100$ for 250Kb and $\rho = 500$ for 1.25Mb), and a given number of segregating loci (chosen to be 80). Despite the finite number of loci, the mutation events are assumed to occur according to the infinitely-many-sites model, which implies no recurrent mutation, at positions from the beginning to the end of the sequence comprised between 0 to ρ in units of scaled recombination rate. Therefore, the segregating loci are different from one sample to another. Moreover, the scaled recombination rate ρ_m between loci m and $m+1$ is transformed into a genetic scale, using $r_m = \rho_m/(4N)$ and a constant population size $N = 10000$, which corresponds to the effective population size for humans (see, e.g., Wall 2003, and references therein); this gives an approximation for the genetic distance in cM between adjacent segregating loci.

Notice that the sample size corresponds to half the population size, since $2N = 20000$, while the program in Hudson (2002) uses approximations based on the assumption of a small sample in a large population inherent to the coalescent.

In each sample, we have chosen the most polymorphic loci (namely, those showing the highest heterozygosity) for the markers and the TIM. Then, the TIM locus has been chosen between the first and third quartiles of the sequence (in order to have marker loci on both sides of the TIM locus), such that the frequency of the derived type at the TIM locus is as close as possible to a given value ξ. This value has been chosen to be 0.1 in half of the samples with sequences of total length 0.25cM (samples A, B, C, D, E, F, G, H), which corresponds to a common disease, and 0.01 in the other half in the case of sequences of total length 1.25cM (samples S, T, U, V, W, X, Y, Z), which models a rare disease. The sequences that carry the derived type at the TIM locus are considered to be cases, and the others that carry the primitive type to be controls.

For comparison, Morris *et al.* (2002) considered frequencies for the disease

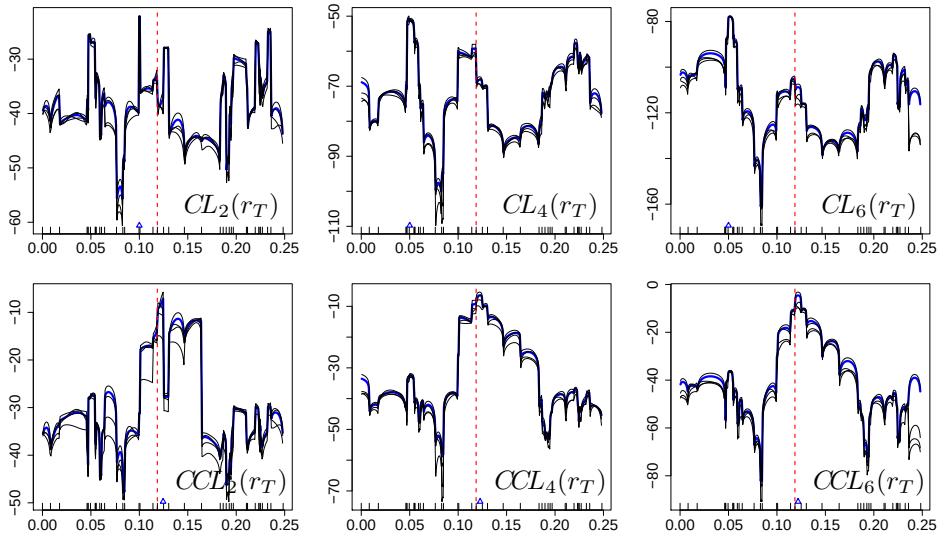


Figure 6. Comparison of estimates based on $CL_d(r_T)$ and $CCL_d(r_T)$ for sample B. The horizontal axis is for r_T (in cM) and the vertical axis for the log-likelihood. A thin line is for one run of simulations ($K = 1t$ for each interval) and a thick blue line for four runs combined ($K = 4t$ for each interval). A dashed vertical line indicates the real position of the TIM locus and a triangle the value of the estimate $\hat{r_T}$. Small vertical lines on the horizontal axis indicate marker loci. Results obtained with 40 loci.

allele comprised between 0.01 and 0.025 and sequences of length 2.25Mb. On the other hand, Zollner and Pritchard (2005) worked with sequences of 1Mb. The ancestral recombination graph becomes very complex, and the computation time necessary for simulations very long, when dealing with genetic lengths that large. Then, there is almost independence between the first segregating locus and the last one, while the adjacent segregating loci are closely linked if the segregating loci are dense enough.

The analysis for the position of the TIM is done using a subsample taken in the larger sample which is assumed to be representative of the whole population. Formally, the method requires a random subsample. In such a case however, the subsample size has to be large to contain a significant number of cases when the disease allele frequency is low. Fixing the numbers of cases and controls in the subsample introduces a selection bias but ensures a minimum of information on both groups. To do our first analysis, a balanced subsample with 50 cases and 50 controls drawn at random without replacement in each sample has been used.

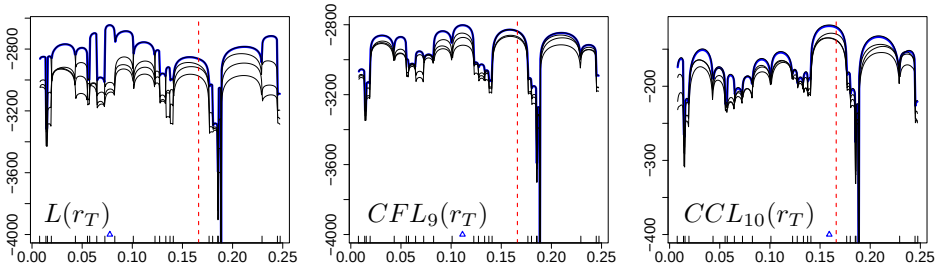


Figure 7. Comparison of estimates based on $L(r_T)$, $CFL_9(r_T)$ and $CCL_{10}(r_T)$ for sample D. Results obtained with 25 loci.

In order to reconstruct the history of the subsample, we have arbitrarily chosen a mutation rate at every polymorphic locus equal to $u_m = 5 \times 10^{-5}$. This assumption does not affect the estimation procedure as long as the rate chosen is low enough not to allow for recurrent mutations (see the paragraph about the effect of the mutation rate in the Discussion).

Figure 6 compares estimated profiles of the composite-conditional-likelihood function, $CCL_d(r_T)$, and the composite-likelihood function, $CL_d(r_T)$, for $d = 2, 4$ and 6, obtained from simulations using the same data set (sample B). The results show that the former keep the same general shape for the different values of d, while this is not the case for the latter. The maximum likelihood estimate of r_T based on $CCL_d(r_T)$ proves to be robust, differing little with different values of d. One might still expect a more precise estimate when a larger value for d is used, since then less information is lost, but the value of d may not have to be so large. Of course, the estimation procedure is faster when d is smaller, since then less events have to be considered to reconstruct the history of the subsample backward in time. To get the estimated profile of $CCL_d(r_T)$ for $d = 2$ in Figure 6, the computation time was 3m 3s, but 17m 58s for $d = 4$ and 58m 5s for $d = 6$. Note however that the number of times each interval between marker loci is considered, and therefore the number of ancestral recombination graphs simulated, is not the same with different window sizes, being approximately three times larger with $d = 4$, and five times larger with $d = 6$, than with $d = 2$.

Figure 7 compares the composite-conditional-likelihood procedure with windows of size 10 to the full-likelihood procedure with 9 repetitions and the composite-full-likelihood procedure with 10 repetitions. The CFL method appears to be less variable than the simple FL method, the estimates obtained being less dependent on the number of simulations. This is also the case with the CCL approach, which proves to be as efficient as the CFL method in

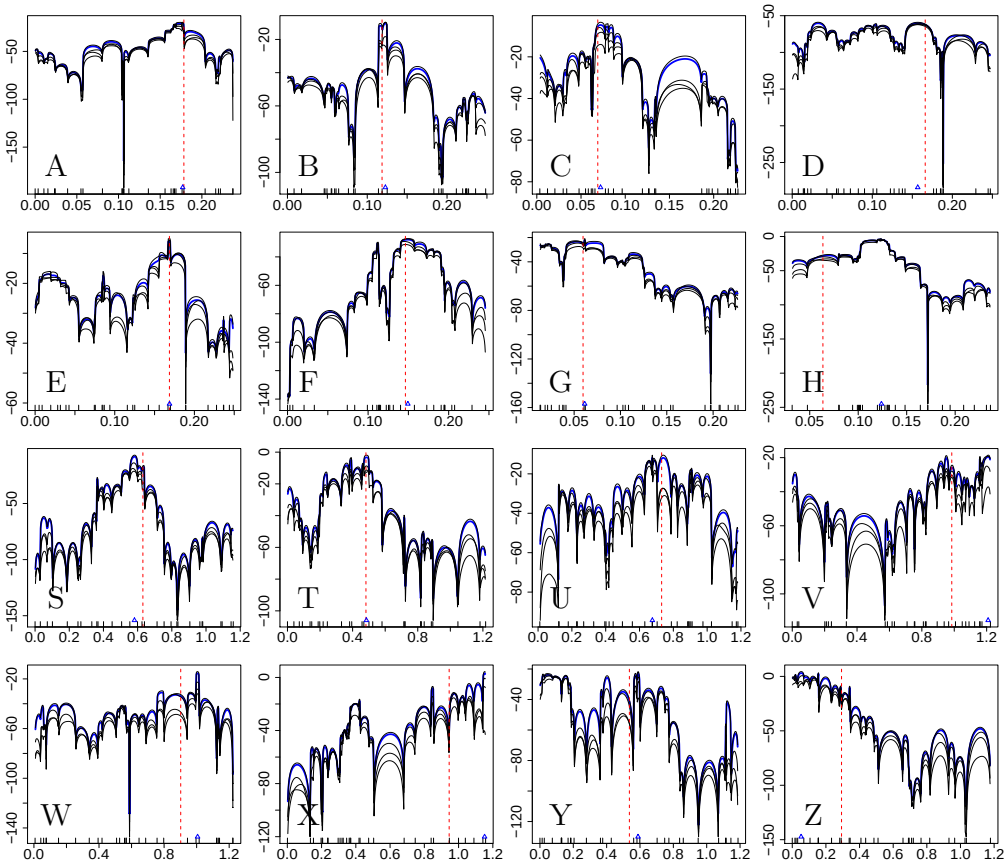


Figure 8. Estimates based on $CCL_6(r_T)$ with 30 loci, for samples A to H, and S to Z.

locating the TIM. A comparison with the results obtained in Larribe *et al.* (2002) using the FL function shows also the superiority of the CCL approach. From now on, we will report only results based on the composite-conditional-likelihood function $CCL_d(r_T)$.

The composite-conditional-likelihood function, $CCL_d(r_T)$, has been estimated four times for each of the simulated data sets using windows of size 6 ($d = 6$) and one thousand ($K = 1000$ or $1t$) simulated histories of the subsample for each of the 29 intervals each time. The results are presented in Figure 8. The maximum likelihood estimate $\hat{r_T}$ is most often in the right interval and very close to the real position of the TIM locus. We can see some variability between the four estimates obtained with different runs of simulations, but the results are usually consistent from one run to another.

Different subsamples have been drawn from one of the sample to study the

10.2202/1544-6115.1298

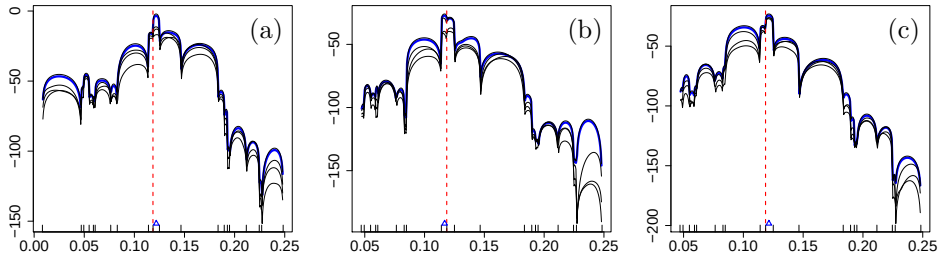


Figure 9. Effect of non-random selection of cases and controls on estimates based on $CCL_5(r_T)$ with 20 loci for sample B: (a) 50 cases and 50 controls (about 25 minutes in computation time); (b) and (c) 48 cases and 451 controls, for a frequency of the disease allele around 0.1 in the population (more than 4 hours in computation time).

effect of subsample size and non-random selection of cases and controls. A balanced subsample (50 random cases, 50 random controls) and two proportional subsamples of size 500 (48 random cases and 451 random controls, giving a frequency of cases in the subsample around the frequency of the disease allele in the whole sample, which is equal to 0.1) have been considered. The values of all other parameters are as previously. Figure 9 shows the simulation results. We can see some differences in the likelihood profiles, especially at the beginning of the sequence, but the conclusions remain essentially similar. We have also analyzed the same data using balanced subsamples of size 20, 50, 200 and 500 (this is the total number of sequences, so in each case half are cases and the other half are controls). The simulation results (not shown) show very little difference in the likelihood profiles. The main difference lies in the computation time required to do the analysis: it is (roughly) proportional to the subsample size.

Notice that only the most polymorphic loci have been used for the analysis. This does not guarantee that the whole sequence is covered equally with markers. The number of markers has to be large enough in order to avoid big gasps between adjacent marker loci. Figure 10 shows simulation results using the 20 (a), 30 (b) and 35 (c) most polymorphic loci in sample H and windows of size $d = 15$. The position of the TIM in this example is in the first quarter of the sequence, and even with the 30 most polymorphic loci, we have little information on the beginning of the sequence and the TIM locus cannot be found. With 35 loci however, this region is covered and the position of the TIM is estimated accurately. Considering an increasing number of marker loci,

from 2 to 70, in sample B (results not shown), we have obtained likelihood profiles converging rather rapidly and showing little difference beyond 16 loci, and almost none beyond 30. These numbers may of course vary from one example to another.

It is interesting to compare the likelihood profiles in Figure 10 to the one presented in Figure 8 for the same sample H, but with a smaller window size of 6. We can see that increasing the window size improves the estimation procedure. This has been verified also using sample A (Figure 11). Note however that a window size $d = 4$ gives already a precise estimate in this case, showing a substantial improvement from results obtained with $d = 2$ and almost nothing to gain in increasing the window size to $d = 14$.

5 Comparisons

An interesting aspect of the composite-conditional-likelihood approach is that it allows us to analyze long sequences. Longer is the sequence, more intricate is the history of the sample. The analysis becomes rapidly untractable, in which case the use of windows of marker loci is appealing. Consider for instance simulated sequences of length 33.27Mb, which corresponds roughly to the length of the shortest human chromosome, with a disease allele of frequency equal to 2.5%. Figure 12 shows the likelihood profile obtained with 200 loci and windows of length five. It is maximum at a point very close to the position of the disease mutation.

It is of interest to compare the proposed method with other recent methods. We have to emphasize that this is a difficult task, since each method has its strengths and weaknesses. Computational time involved in getting results also imposes an important limitation. Zollner and Pritchard (2005) have compared

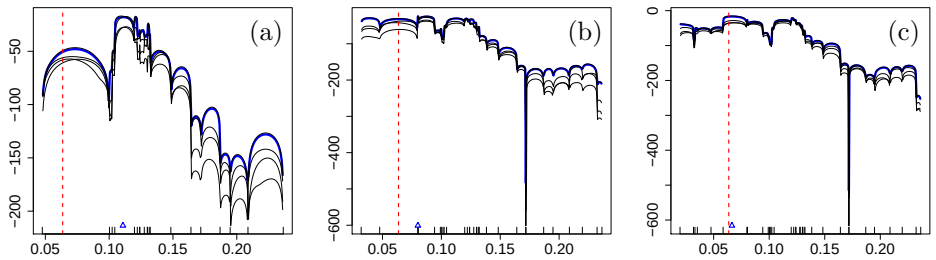


Figure 10. Effect of increasing the number of loci on estimates based on $CCL_{15}(r_T)$ for sample H: (a) 20 loci; (b) 30 loci; (c) 35 loci.

10.2202/1544-6115.1298

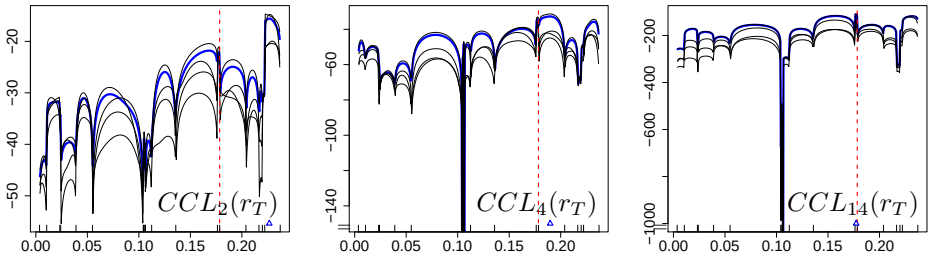


Figure 11. Effect of the window size on estimates based on $CCL_d(r_T)$ for sample A (20 loci used).

their method based on a local approximation of the ARG near the putative location of the disease mutation, called LATAG, with the method of McPeek *et al.* (1999) based on the decay of haplotype sharing, called DHSMAP, using 50 data sets. In order to do something similar, we have applied DHSMAP to the data sets used in the previous section. Results are shown in Figure 13. Comparison with Figure 8 reveals that the CCL approach gives point estimates for the TIM locus closer to the real value for most of the data sets (A, C, D, F, S, T, U, V, Y) and is outperformed by the DHSMAP method for only one data set (H). Notice that both methods underperform for the same data sets (W, X, Z).

We have also made a comparison with a more recent method proposed by Minichiello and Durbin (2006), called Margarita and based on a heuristic inference of plausible ARGs. Results are shown in Figure 14. This time, the shapes of the curves and maximum points obtained are closer to those in Figure 8, but the CCL method is clearly not outperformed and does surprisingly well

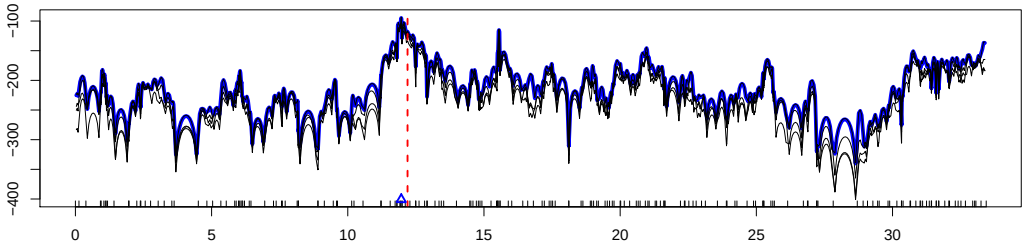


Figure 12. $CCL_5(r_T)$ for a sequence of length 33.37Mb, with 200 markers.

in comparison with the other method in the case of two data sets (A, B).

Finally, we have compared the three methods with some simple statistics, e.g., r^2 (results not shown). Generally speaking, if the disease mutation locus can be found by simple statistics, both DHSMAP and CCL with small windows give an accurate estimate for the TIM locus. When the simple statistics do not work, the CCL method gives generally an estimate closer to the real value.

Notice however that DHSMAP involves less calculations than the CCL approach and both DHSMAP and Margarita might prove to be more robust.

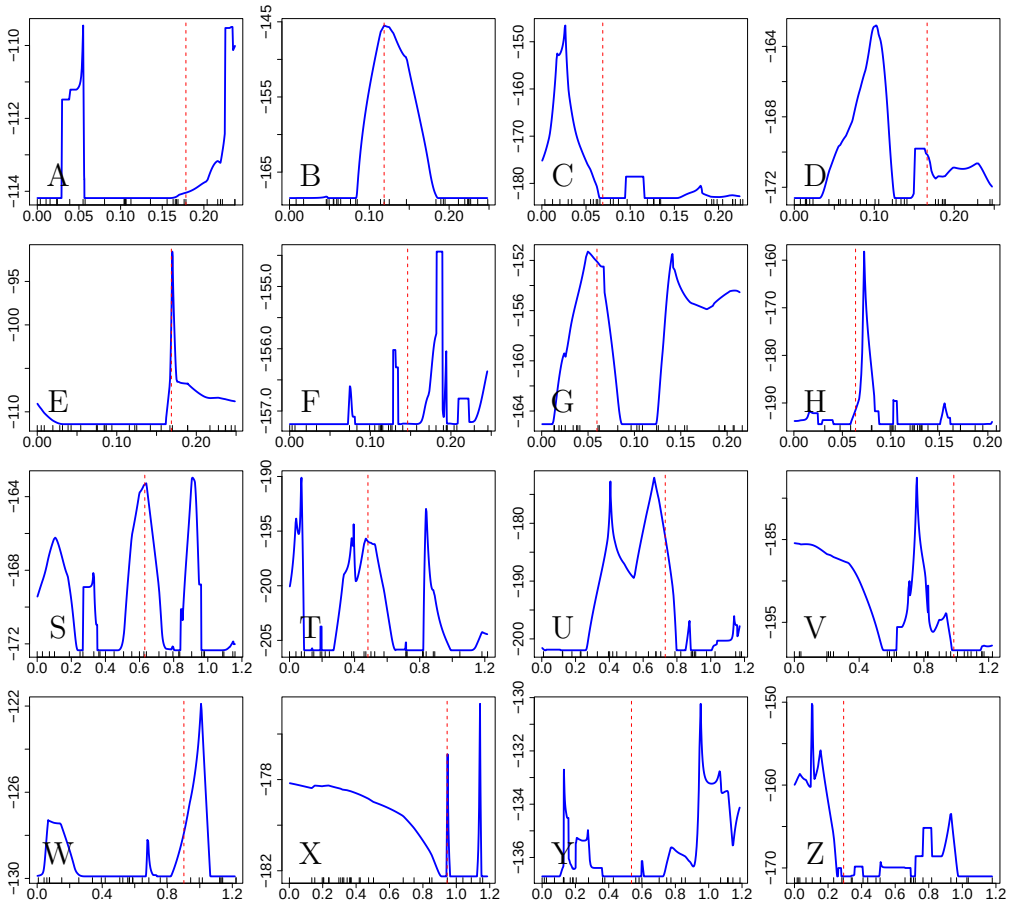


Figure 13. Likelihood profiles by the DHSMAP method of McPeek *et al.* (1998) with 30 loci, for samples A to H, and S to Z.

6 Discussion

We have extended and refined a full-likelihood method introduced in Larribe *et al.* (2002) for mapping a TIM locus based on observed markers in cases and controls. The mutation term in the recurrence equation (1) for the likelihood of sample configurations corrects the expression given in Larribe *et al.* (2002) in the special case of a uniform mutation rate, that is, $\theta_m = \theta/L$. The factor $1/L$ was absent in this paper, but the results presented there remain fully valid since the system of equations for the likelihood function

$$Q^*(\mathbf{H}_\tau) = L^{s_\tau} Q(\mathbf{H}_\tau),$$

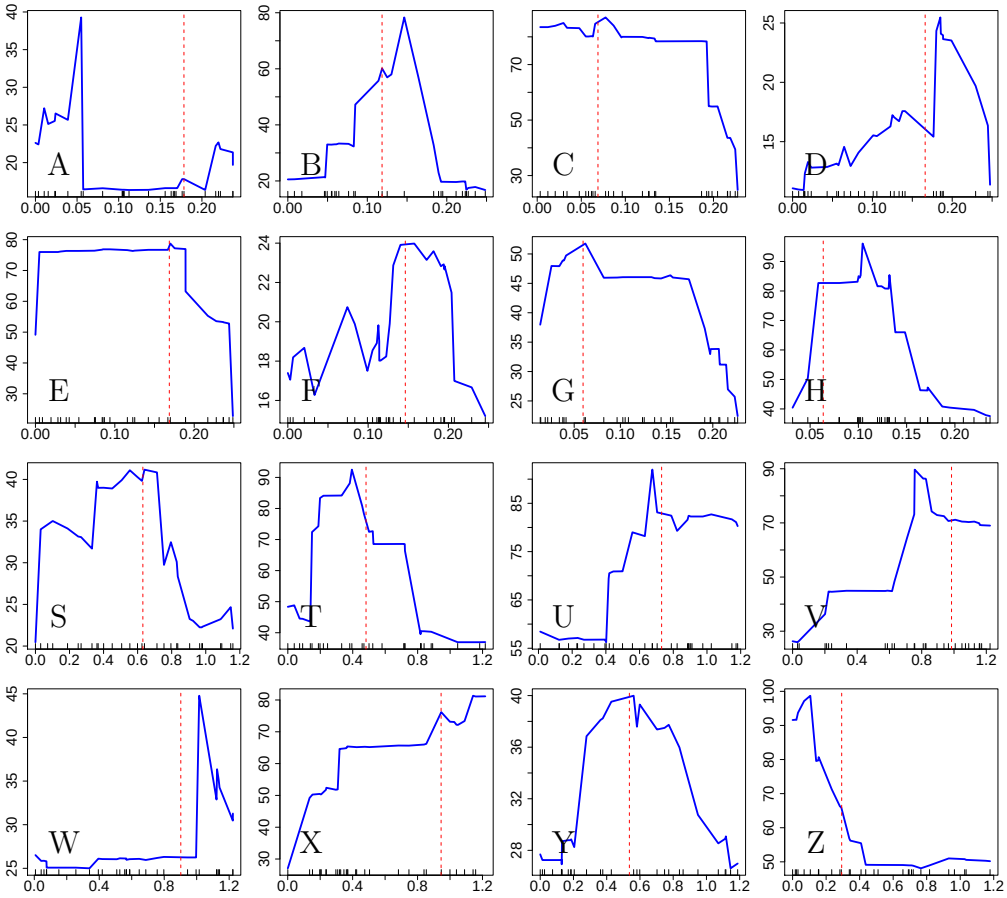


Figure 14. "ARG MAP SCORE" by the Margarita method of Minichiello and Durbin (2006) with 30 loci, for samples A to H, and S to Z.

where s_τ designates the number of polymorphic loci in $\mathbf{H}_\tau$, is the same with θ instead of θ/L. This is the case because this number is reduced in $\mathbf{H}_\tau$ only if the $t_{\tau+1}$-th event of change in the ancestral material is an event of mutation. Since $Q^*(\mathbf{H}_0)$ differs from $Q(\mathbf{H}_0)$ only by a multiplicative factor L^n, then the maximum likelihood procedure leads to the same estimation for the location of the TIM.

The main difficulties in the full-likelihood method proposed in Larribe *et al.* (2002) were the computing time and the variability of the likelihood profiles. These issues have been addressed by considering a composite-likelihood approach based on windows of a given number of contiguous marker loci. The likelihood function for the position of the TIM is simpler to estimate if one considers a smaller number of markers at a time and a weighted geometric average over subsets of marker loci is easy to compute. Some care must be taken, however, since the likelihood of the sample configuration in a window of marker loci may greatly differs from one window to another. One way to eliminate this bias effect is to consider the conditional likelihood function for the position of the TIM, given the sample configuration at the marker loci in the window. This leads to the proposed composite-conditional-likelihood (CCL) function.

We have shown that the method can be applied to SNPs as markers, with known primitive and derived types, but it can be extended to microsatellites which are highly polymorphic, showing multiple types corresponding to numbers of repetitions of DNA patterns with no type in particular being identified as primitive to the others. Of course, the mutation term in the recurrence equation has to be changed accordingly.

The method is based on samples of haplotypes. If enough genotyping has been done on family members, then haplotypes can be obtained. Otherwise, they have to be inferred from genotypes by statistical methods (see, e.g., Fallin and Schork 2000, or Stephens *et al.* 2001). Such an inference procedure has not been incorporated into the method, but this could be done. Morris *et al.* (2004), for instance, did it to improve their original method (Morris *et al.* 2002); Rannala and Reeve (2001) considered that haplotypes were known, while Zollner and Pritchard (2005) used haplotypes estimated from the method of Stephens *et al.* (2001). Another related sampling issue that has not been taken into account is incomplete penetrance and phenocopy. All these aspects could be incorporated using, e.g., the current frequency of the disease mutation and a Bayesian approach.

More importantly, the method assumes a random sample in the whole population. In most genetic studies, samples are ascertained by the disease status: cases and controls are sampled in similar numbers, particularly when

the disease is rare since then a random sample has to be very large to include a significant number of cases. Fortunately, our simulations have shown that the bias caused by non-random selection of cases and controls has little effect on the estimation results. This is also the case for the sample size and the density of markers once some threshold value is reached.

As noted in Nordborg (2001), a sample of sequences taken in a population can only give information about the genealogy of that population. This genealogy may be informative, and then an inference method can work, or it may be non-informative, and then there is little we can do to change that situation. This may explain the general pattern of our simulation results which are very clear for some data sets, and more or less conclusive for others. Similar results are found in Zollner and Pritchard (2005). Notice also that different independent subsamples taken from the same sample (see, e.g, Figure 9 (b) and (c)) lead to similar conclusions.

The method assumes that the mutation rate is known. In the case of SNPs with the same mutation rate at every locus, however, the likelihood function for the position of the TIM differs only by a multiplicative factor from one mutation rate to another. Therefore the maximun-likelihood estimate does not depend on the mutation rate and it does not have to be known.

The proposal distribution to reconstruct the history of sampled sequences is arbitrary but it is advantageous to choose one as close as possible to the real distribution (Stephens and Donnelly 2000, Fearnhead and Donnelly 2001), which is unknown, or as fast as possible to simulate. We have chosen a proposal distribution which is uniform over all possible recombination events between two adjacent marker loci for this reason. For the same reason, in the case of microsatellites with a large number of mutation events to consider, the proposal distribution would be chosen uniform over all mutation events at a polymorphic locus. Of course, such choices affect the value of the variable we have to simulate, but they proved to reduce significantly the simulation time without reducing the accuracy of the whole estimation procedure.

References

Boehnke, M. (1994). Limits of resolution of linkage studies: Implications for the positional cloning of disease genes. *Am. J. Hum. Genet.* 55, 379–390.

Collins, A. and Morton, N.E. (1998). Mapping a disease locus by allelic association. *Proc. Natl. Acad. Sci.* 95, 1741–1745.

Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29, 311–322.

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

Fallin, D. and Schork, N. J. (2000). Accuracy of haplotype frequency estimation for biallelic loci, via the EM algorithm for unphased diploid genotype data. *Am. J. Hum. Genet.* 67, 947–959.

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

Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *J. R. Stat. Soc. Ser. B* 64, 657–680.

Garner, C. and Slatkin, M. (2002). Likelihood-based disequilibrium mapping for two-marker haplotype data. *Theor. Popul. Biol.* 61, 153–161.

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

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

Griffiths, R. C. and Tavaré, S. (1996). Monte Carlo inference methods in population genetics. *Math. Comput. Model.* 23, 141–158.

Hein, J., Schierup, M. and Wiuf, C. (2005). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford University Press, Oxford.

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

Hudson, R. R. (1987). Estimating the recombination parameter of a finite population model without selection. *Genet. Research* 50, 245–250.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In: *Oxford Surveys in Evolutionary Biology*, Vol. 7. D. Futuyma and J. Antonovics (Eds). Oxford University Press, New York, pp. 1–44.

Hudson, R. R. (2001). Two-locus sampling distribution and their application. *Genetics* 159, 1805–1817.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18, 337–338.

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

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

Kuhner, M. K., Yamamoto, J. and Felsenstein, J. (2000). Maximum likelihood estimation of recombination rates from population data. *Genetics* 156, 1393-1401.

Larribe, F., Lessard, S. and Schork, N. J., (2002). Gene mapping via the ancestral recombination graph. *Theor. Popul. Biol.* 62, 215–229.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165, 2213–2233.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* 80, 221–239.

Liu, J. S., Sabatti, C., Teng, J., Keats, B. J. and Risch, N. (2001). Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.* 11, 1716-1724.

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

McVean, G., Awadalla, P. and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160, 1231-1241.

Minichiello, M. J. and Durbin, R. (2006). Mapping trait loci using inferred ancestral recombination graphs. *Am. J. Hum. Genet.* 79, 910–922.

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. Genet.* 67, 155–169.

Morris, A. P., Whittaker, J. C. and Balding, D. J. (2002). Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am. J. Hum. Genet.* 70, 686–707.

Morris, A. P., Whittaker, J. C. and Balding, D. J. (2004). Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *Am. J. Hum. Genet.* 74, 945–953.

Myers, S. R. and Griffiths, R. C. (2003). Bounds on the minimum number of recombination events in a sample history. *Genetics* 163, 375–394.

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

Nordborg, M. (2001). Coalescent theory. In: *Handbook of Statistical Genetics*. D. Balding, M. Bishop and C. Cannings (Eds.). Wiley, Chichester, U.K., pp. 179–212.

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

Nothnagel, M., Fürst, R. and Rohde, K. (2002). Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Human Heredity* 54, 186–198.

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

Sabattti, C. and Risch, N. (2002). Homozygosity and linkage disequilibrium. *Genetics* 160, 1707–1719.

Stephens, M. (2001). Inference under the coalescent. In: *Handbook of Statistical Genetics*. D. Balding, M. Bishop and C. Cannings (Eds.). Wiley, Chichester, U.K., pp. 213–238.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *J. R. Stat. Soc. Ser. B* 62, 605–655.

Stephens, M., Smith, N. J. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68, 978–989.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *J. R. Stat. Soc. Ser. B* 62, 605–655.

Tavaré, S. and Zeitoumi, O. (2004). *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXXI – 2001*. J. Picard (Ed.). Lecture Notes in Mathematics, Vol. 1837. Springer-Verlag, Berlin and Heidelberg.

Terwilliger, J. D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci, *Am. J. Hum. Genet.* 56, 777–787.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis* 92, 1–28.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* 92, 519–528.

Wakeley, J. (2008). *Coalescent Theory: An Introduction.* Roberts & Company Publishers, Greenwood Village, Colorado.

Wall, J. D. (2000). A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.* 17, 156–163.

Wall, J. D. (2003). Estimating ancestral population sizes and divergence times. *Genetics* 163, 395–404.

Wall, J. D. (2004). Estimating recombination rates using three-site likelihoods. *Genetics* 167, 1461–1473.

Wu, Y. D. (2007). Association mapping of complex diseases with ancestral recombination graphs: Models and efficient algorithms. In: *Research in Computational Molecular Biology.* T. Speed and H. Huang (Eds.). Proceedings Series: Lecture Notes in Computer Science , Vol. 4453. Springer, Berlin and Heidelberg, pp. 488–502.

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

Zollner, S. and Pritchard, J. K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* 169, 1071–1092.