Multipoint Fine-scale Linkage Disequilibrium Mapping: Importance of Modeling Background LD

Andrew L. Strahs and Mary Sara McPeek

Abstract

In linkage disequilibrium (LD) mapping, use of information on multiple markers simultaneously is expected to lead to greater power to detect association and smaller confidence intervals (CIs) for the location of the variant of interest than would be obtained from single-point analysis. Among the important challenges facing case-control LD mapping methods are (i) even when an appropriate control sample is available, there may be background LD in the control sample which must be taken into account in the analysis, especially when fine-scale data are collected, and (ii) in practice, genotype rather than haplotype data are often available, limiting the applicability of some methods. Furthermore, in cases when genotype data can, in principle, be incorporated, it can be computationally challenging. We focus on simultaneous solution of these problems in the context of the Decay of Haplotype Sharing (DHS) method. We develop a computationally efficient method that allows for genotype or haplotype data on many loci and incorporates background LD based on a Markov model of order η. The case of a Markov model of order 2 is implemented in free software. In addition, we demonstrate that failure to adequately model background LD can potentially have a major effect on the analysis, and we develop and apply methods for assessing the adequacy of the model for background LD.

Keywords: Decay of Haplotype Sharing; linkage disequilibrium; fine-scale mapping; background linkage disequilibrium; cystic fibrosis; hidden Markov model

1 Introduction

Linkage disequilibrium (LD) has been shown to be useful for fine-mapping of trait-associated variants [6, 10, 11, 15]. While early approaches generally treated each marker separately, haplotype-based LD mapping methods have the potential to provide considerable additional information when dense marker data are available in a region. There are several approaches that combine results across loci in various ways without explicitly modeling dependence among loci [4, 7, 17, 23, 31, 32]. Among approaches that explicitly model dependence across loci, Service *et al.* [29] and MacLean *et al.*

[20] perform haplotype-based tests for association, in which they use multilocus models for haplotypes descended from an ancestor, taking into account recombination and mutation. They require haplotype data, assume background linkage equilibrium, and do not consider effects of population structure. There are several methods that perform haplotype-based tests of association in trios consisting of parents and an affected offspring, conditional on the transmitted and non-transmitted haplotypes in the parents (*e.g.* [2, 3, 36]). Lam *et al.* [16] use a parsimony method to build an evolutionary tree of disease haplotypes assuming a disease mutation occurs in an intermarker interval. They then compute the likelihood of the tree using a model for recombination and mutation. They obtain a posterior distribution for the location of the variant. Their method assumes haplotype data. Background linkage disequilibrium is taken into account by a Markov-type method in which the lag at any stage is chosen to coincide with the longest match in the control database.

McPeek and Strahs [22] form a confidence interval (CI) for the location of the variant, in which they make use of a multilocus decay-of-haplotype-sharing (DHS) model for haplotypes descended from an ancestor, taking into account recombination and mutation. They propose a quasi-likelihood approach to take into account population structure in the affecteds. Assuming a conditional coalescent model for the population structure, McPeek and Strahs [22] derive an approximate correction factor for the likelihood, and they model background LD by a Markov chain with lag 1. Morris *et al.* [25] concentrate on biallelic markers in a Bayesian framework and obtain the posterior distribution of the location of the trait-associated variant using Markov chain Monte Carlo (MCMC). They use a similar approach to that of McPeek and Strahs [22] to correct for population structure. Rannala and Reeve [28] also employ a Bayesian framework and obtain the posterior density of the position of the trait-associated variant by employing MCMC to integrate over coalescent genealogy trees, using biallelic marker data and information about candidate genes from an annotated human genome sequence. Neither Morris *et al.* [25] nor Rannala and Reeve [28] consider background LD.

Liu *et al.* [18] also obtain the posterior distribution of the location of the trait-associated variant using MCMC. Their model for population structure groups the disease haplotypes into clusters corresponding to different ancestral haplotypes and assumes a star-shaped genealogy for the haplotypes within each cluster given the ancestral haplotype. They model background LD by a Markov chain with lag 1. Zhang and Zhao [34] extend McPeek and Strahs [22] by implementing a stepwise mutation model for mutation in microsatellite markers. They extend the conditional coalescent model of McPeek and Strahs [22] to allow variable population size. Morris *et al.* [24] obtain the posterior distribution of location of the trait-associated variant and incorporate a shattered coalescent model for genealogies of the disease haplotypes using MCMC. They also model background LD using a Markov chain with lag 1.

In this study, we simultaneously tackle two of the major difficulties that arise in multipoint LD mapping with data on random samples of cases and controls: (1) LD is generally present in the controls as well as in the affecteds, and this background LD

can have a major impact on the analysis; (2) data are typically in the form of genotypes with unknown phase, rather than haplotypes. Dealing with both of these issues simultaneously presents particular computational challenges, and a focus of our work has been development of an efficient algorithm to handle them.

To deal with the second problem, that data are typically in the form of (unphased) genotypes, one possible solution is to try to reconstruct haplotypes based on population information, using one of the available methods [9, 12, 18, 19, 30]. We prefer instead to incorporate uncertainty about the haplotypes into the analysis. The likelihood framework of DHS makes an extension from haplotype to genotype data straightforward in principle: one need only sum the likelihoods of all possible sets of haplotypes compatible with the observed genotype data. This approach is implemented by Zhang and Zhao [35]. However, with more than a small number of loci, this straightforward approach quickly becomes computationally infeasible. We introduce a more computationally efficient approach using a hidden Markov model (HMM), in which we incorporate a Markov model, with lag η, for LD in the controls.

Methods for LD mapping generally consider two statistical problems, detection of association (*i.e.* hypothesis testing) and localization (*i.e.* construction of a CI for the variant of interest). If background LD, *i.e.* LD present in the controls as well as in the affecteds, is not adequately captured by the model, it may be falsely attributed to the presence of a variant associated with the trait. For the detection problem, unmodeled background LD could result in excess false positive detections of association. For the localization problem, unmodeled background LD could result in CIs that fail to have the appropriate probability of covering the true location. For the detection problem, a number of approaches have been developed that aim to produce valid hypothesis tests in the presence of background LD [2, 8, 27, 36]. Here we instead focus on the localization problem, which is not treated by these papers. In this context, McPeek and Strahs [22] model background LD in control haplotypes by use of a Markov chain model of lag $\eta = 1$. In analyzing the data set of Kerem *et al.* [15], we find that a Markov chain of lag $\eta = 2$ is preferable, as shown in the subsection "Importance of modeling background LD" of the Results section. Incorporation of Markov models for background LD is more challenging when genotype data are used instead of haplotype data, because implementation of a Markov model requires one to keep track of phase information. In this study, we devise a HMM to simultaneously incorporate genotype data and a Markov model with lag $\eta = 2$ for background LD.

These new methods make it feasible to perform multipoint LD mapping on data sets consisting of (unphased) genotypes for a large number of markers. We use simulated examples to compare fine-mapping based on genotype and haplotype data, and we use the CF data set [15] to demonstrate the importance of the improved modeling of background LD.

2 Methods

DHS method for haplotype data

As developed in McPeek and Strahs [22], the DHS likelihood under the model for a single observed haplotype $\mathbf{h}_{\text{obs}}$ drawn from the population of affecteds, when there is one ancestral haplotype $\mathbf{h}_{\text{anc}}$, is

$$L(x, \mathbf{h}_{\text{anc}}, \tau^{-1}, p; \mathbf{h}_{\text{obs}}) = (1-p)\tilde{L}(x, \mathbf{h}_{\text{anc}}, \tau^{-1}; \mathbf{h}_{\text{obs}}) + pP_{\text{null}}(\mathbf{h}_{\text{obs}}), \tag{1}$$

where x is the location of the variant; τ is number of generations to the ancestor, or, equivalently, τ^{-1} is a measure of the amount of linkage disequilibrium and is equal to the expected genetic distance from the variant to either edge of the ancestral segment in an observed haplotype; p is a heterogeneity parameter representing the probability that the haplotype $\mathbf{h}_{\text{obs}}$ is not descended from the ancestral haplotype $\mathbf{h}_{\text{anc}}$; and $P_{\text{null}}(\mathbf{h})$ is the frequency of haplotype $\mathbf{h}$ in the control population. Furthermore,

$$\begin{aligned}\tilde{L}(x, \mathbf{h}_{\text{anc}}, \tau^{-1}; \mathbf{h}_{\text{obs}}) \;=\; & \sum_{i=0}^{l_{re}} \sum_{j=0}^{l_{le}} \Bigg[g(\tau^{-1}, -j, i) \times \prod_{k=-j}^{i} m(k, \tau, \mathbf{h}_{\text{anc}}(k), \mathbf{h}_{\text{obs}}(k)) \times \\ & P_{null}(\mathbf{h}_{\text{obs}}(i+1), \mathbf{h}_{\text{obs}}(i+2), \ldots, \mathbf{h}_{\text{obs}}(l_{re})) \times \\ & P_{null}(\mathbf{h}_{\text{obs}}(-l_{le}), \mathbf{h}_{\text{obs}}(-l_{le}+1), \ldots, \mathbf{h}_{\text{obs}}(-j-1)) \Bigg]\end{aligned}$$

is the likelihood assuming that observed haplotype $\mathbf{h}_{\text{obs}}$ is a τth-generation descendent of ancestral haplotype $\mathbf{h}_{\text{anc}}$. In the above expression, $m(k, \tau, \alpha, \beta)$ models the mutation process; it is the probability that allele β is observed at locus k, given that the haplotype's τth generation ancestor at locus k had allele α. i and j index markers, with marker 0 corresponding to the putative location x of the variant, and with markers on, say, the distal side of x numbered with consecutive negative integers decreasing in the direction away from the centromere and with markers on the proximal side of x numbered with consecutive positive integers increasing in the direction of the centromere. (Note that the integer labels for the markers are defined relative to the putative position x of the variant, which varies across the region during the analysis.) Here, $-l_{le}$ is the index corresponding to the "left edge" of the data set (*i.e.* the marker farthest from the centromere), and l_{re} is the index corresponding to the "right edge" of the data set (*i.e.* the marker closest to the centromere). In the above expression, we sum over all possible choices of the marker intervals containing the two (unobserved) breakpoints of the ancestral segment. Moreover,

$$g(\tau^{-1}, -j, i) = e^{-\tau d_{-j,i}}(1 - e^{-\tau d_{-j-1,-j}})(1 - e^{-\tau d_{i,i+1}})$$

is the probability that $\mathbf{h}_{\text{obs}}$ inherits the variant and the ancestral segment, intact, between loci $-j$ and i inclusive but that it is no longer intact at locus $-j-1$ nor at locus $i+1$.

Here, $d_{i,j}$ is the genetic distance between loci i and j. McPeek and Strahs [22] discuss how to incorporate multiple ancestral haplotypes into this likelihood expression.

To combine likelihoods across observed haplotypes, one must make some assumptions about how the haplotypes are related. Under the assumption of independent recombinational histories (*i.e.* a star-shaped phylogeny), one can multiply the likelihoods across haplotypes. This approach is generally anti-conservative when this assumption does not hold. McPeek and Strahs [22] propose a quasi-likelihood approach to take into account population structure, which in principle could be applied to any chosen population model. For the case in which the affecteds are presumed to be only very distantly related with little else known about the population structure, McPeek and Strahs [22] propose a conditional coalescent model for the phylogeny relating the affected individuals, conditional on the time to the common ancestor. With complete data, they calculate and maximize a quasi-likelihood with respect to this model, and with incomplete data, they calculate and maximize a similar expression with the complete data likelihoods replaced by incomplete data likelihoods. We employ the same approach here. In the case of the conditional coalescent model or any other exchangeable population model, the parameter estimates obtained in this way are the same as under the assumption of independence, but with the standard errors for the parameters inflated and the log-likelihood deflated. When DHS is used to fine-map, this widens the CI for the location of the trait-associated variant. In practice, then, to implement the conditional coalescent model, we proceed as if the observations were independent, and then implement the appropriate correction to the log-likelihood and standard errors at the end. The approximate correction factor in the conditional coalescent case is

$$\sum_{k=1}^{n-2}\left(2(n-2)!(n+1)/[(n-1)(n-k+1)(n-k+2)(n-k)(k-1)!(n-k-2)!]\times \sum_{i=1}^{\infty}(-1)^{i+1}\binom{n+i-1}{n-k}^{-1}\right),$$

which corrects a typo in McPeek and Strahs [22] (factor of $(-1)^i$ vs. $(-1)^{i+1}$). The DHS model can also be extended to the case when population structure is known [33]. In that case, the shape of the likelihood curve and, in particular, the maximum, will generally not be the same when population structure is taken into account as when independence is assumed.

The formulae of this section give a mathematical representation of the likelihood. However, for computational efficiency in calculating and maximizing the likelihood, we reformulate the probability model as a hidden Markov model (HMM) in the subsection "HMM for haplotype data, with Markov(η) model for background LD" below.

Uncertainty in ancestral haplotype is incorporated in CI construction

To construct a CI for the location of the variant, McPeek and Strahs [22] invert an

(approximate) likelihood-ratio test. At each putative location x, their approximate log-likelihood is maximized over $\mathbf{h}_{\text{anc}}$, $1/\tau$ and p, assuming that the variant is located at x. (The properties of the profile likelihood are discussed in McCullagh and Nelder [21].) The CI is then based on a comparison of the highest maximized log-likelihood to the maximized log-likelihoods at other locations. We emphasize that inference about variant location is **not** performed conditional on the maximizing value of ancestral haplotype. The mapping approach of McPeek and Strahs [22] does, in fact, take into account the uncertainty in ancestral haplotype.

Mode of inheritance, mutation, and background LD

Implicit in the method given in the previous sub-section is the assumption of a multiplicative model for the mode of inheritance, similar to that described by Morris *et al.* [25]. Where β_k corresponds to allele k,

$$P\{\text{affected} \mid (G_1, G_2) = (i, j)\} = \beta_i \beta_j$$

for an individual with genotype (i, j) at the variant. The multiplicative model has the recessive model as a special case, but also allows heterogeneity. This model is convenient when one does not have the information of how the haplotypes are paired. When that information is available, one could easily implement some other mode of inheritance in the analysis.

The mutation model we use is the same as that given by McPeek and Strahs [22]. For biallelic loci, this amounts to assuming the same rate of mutation between the two alleles. We assume the same mutation rate at all markers. These assumptions can be easily modified [34].

When choosing mutation rates to use in the DHS analysis, it may not be appropriate to use a rate as low as the value of $\approx 10^{-9} - 10^{-8}$ given for SNP loci by Nielsen [26]. The reason is that only SNPs that are polymorphic across the individuals in the data set are chosen for analysis. Therefore, the ascertainment process for the data set insures the existence of at least 1 mutation at the SNP within the time-frame of the coalescence of the study sample at that SNP. Thus, conditional on a SNP being in the data set, its mutation rate over the time since the most recent common ancestor of the variant is substantially increased over the unconditional mutation rate given by Nielsen [26]. The extent of the increase depends on assumptions about the population history, but the conditional mutation rate could be several orders of magnitude larger than the unconditional mutation rate. Specification of a larger mutation rate would be expected to lead to a more conservative analysis. In our analysis of the CF data set, we use a mutation rate of 1×10^{-4} mutations per meiosis per marker. Note that in contrast to SNPs, microsatellites would be less affected by this selection effect. The mutation rate of a microsatellite is typically sufficiently high that its conditional mutation rate, over the time period since the most recent common ancestor of the variant, conditional on it being polymorphic in the study sample is very close to its unconditional mutation rate.

In expression (1), the model for background LD enters the likelihood through $P_{null}(\mathbf{h})$, which gives the frequency of haplotype $\mathbf{h}$ in the control population. In principle, one could think of leaving the control haplotype frequencies unconstrained (beyond the requirement that they sum to 1) when estimating them from data. However, with m loci, the number of parameters is $2^m - 1$ for SNP data (with many more for microsatellite data), and the size of the control sample available to estimate these parameters is typically small. Our approach is to constrain the control haplotype frequency distribution to be Markov (η), *i.e.* for a given lag η, we require $P_{null}\{\mathbf{h}(t) = i_t | \mathbf{h}(t-s) = i_{t-s}, \ldots, \mathbf{h}(t-1) = i_{t-1}\} = P_{null}\{\mathbf{h}(t) = i_t | \mathbf{h}(t-\eta) = i_{t-\eta}, \ldots, \mathbf{h}(t-1) = i_{t-1}\}$ for all s such that $\eta \leqslant s \leqslant t + l_e$, and all choices $i_{t-s}, \ldots, i_t$ for alleles, where $\mathbf{h}(t)$ is the allele at locus t in the haplotype $\mathbf{h}$. Such a Markov model can be useful as a simple tool for capturing the local dependence structure among loci on haplotypes randomly selected from a control population. McPeek and Strahs [22] implement the case $\eta = 1$ when complete haplotype data are available. Here, we implement the cases $\eta = 1$ and $\eta = 2$, when either haplotype or genotype data are available, by means of a HMM.

HMM for haplotype data, with Markov(η) model for background LD

For computational efficiency in calculating and maximizing the likelihood, we reformulate, as a HMM [1], the probability model of sub-section "DHS method for haplotype data". The HMM we give here differs from the one given in McPeek and Strahs [22]. Although the underlying likelihood is the same, the HMM given here has computational advantages over the previous version. These advantages are related to the extension of the HMM to allow background LD to be modeled by a Markov model with lag $\eta > 1$, which is given at the end of this sub-section. In section "Extension of HMM to genotype data", we extend this HMM to the case when only genotype data are available.

Suppose that, as in the previous subsection, the putative location of the variant of interest is labeled marker 0, and the markers are numbered with consecutive positive integers increasing in the direction of the centromere and consecutive negative integers decreasing in the direction away from the centromere. We define a discrete-time Markov chain $\{Q_l, 0 \leqslant l \leqslant l_{re}\}$, where l indexes loci on the centromeric side of the variant. The state space of $\{Q_l\}$ is $\{A, N\}$, where A stands for "ancestral" and N stands for "non-ancestral." We define the event $\{Q_l = A\}$ to occur when the entire segment between locus l and the variant of interest (locus 0) is inherited, unbroken by crossovers, from the ancestral haplotype. Note that $\{Q_l = A\}$ holds in this case even if one or more mutations have occurred at locus l (or elsewhere in the segment) in the time since the ancestor. We define $\{Q_l = N\} = \{Q_l = A\}^c$. The initial distribution of $\{Q_l\}$ is $P\{Q_0 = N\} = p = 1 - P\{Q_0 = A\}$. The transition probability matrix for $\{Q_l\}$ is given in Table 1a. In fact, we find it convenient to reverse the conditioning of Q. That is, we define the initial distribution to be $P\{Q_{l_{re}} = A\} = (1-p)e^{-\tau d_{0,l_{re}}} = 1 - P\{Q_{l_{re}} = N\}$, and we use the one-step transition probability matrix $P\{Q_l | Q_{l+1}\}$ given in Table 1b.

The resulting process $\{Q_l\}$ has the same distribution as before. The computational convenience of the reverse conditioning is related to the modeling of background LD and is explained below after the observation distribution is introduced. We can define a mirror-image Markov chain for the loci on the distal side of the variant, with the two chains conditionally independent given Q_0.

Table 1a: Transition probability matrix $P(Q_l|Q_{l-1})$

Current State	Probability that Next State Entered Is	
	A	N
A	$e^{-\tau d_{l,l+1}}$	$1-e^{-\tau d_{l,l+1}}$
N	0	1

Table 1b: Transition probability matrix $P(Q_l|Q_{l+1})$

Current State	Probability that Next State Entered Is	
	A	N
A	1	0
N	$\frac{(1-p)(e^{-\tau d_{l+1,0}}-e^{-\tau d_{l,0}})}{1-(1-p)e^{-\tau d_{l,0}}}$	$\frac{1-(1-p)e^{-\tau d_{l+1,0}}}{1-(1-p)e^{-\tau d_{l,0}}}$

Consider the observation sequence $\{O_l, 0 \leqslant l \leqslant l_{re}\}$ associated with the Markov chain $\{Q_l, 0 \leqslant l \leqslant l_{re}\}$, where O_l is the observed allele at locus l. Our formulation of the distribution of O_l conditional on Q_l depends on our model for background LD as well as on our model for mutations. For simplicity of exposition, we first assume background linkage equilibrium. In that case,

$$P\{O_l \mid Q_l\} = \begin{cases} m(l, \tau, \mathbf{h}_{anc}(l), O_l) & \text{if } Q_l = A \\ f_l(O_l) & \text{if } Q_l = N, \end{cases} \tag{2}$$

where $f_l(\alpha)$ is the frequency of allele α at locus l in the controls. We can allow the observed allele at locus l to be missing by setting $P\{O_l \mid Q_l\} = 1$ when O_l is missing. This will yield the appropriate likelihood calculation for the case when the event that O_l is missing is independent of Q_l.

We now relax the assumption of background linkage equilibrium. Assuming a Markov(1) model for background LD, the observation distribution for $0 \leqslant l < l_{re}$ is given by

$$P\{O_l \mid Q_l, O_{l+1}\} = \begin{cases} m(l, \tau, \mathbf{h}_{anc}(l), O_l) & \text{if } Q_l = A \\ f_{l,l+1}(O_l \mid O_{l+1}) & \text{if } Q_l = N, \end{cases} \tag{3}$$

where $f_{l,l+1}(\alpha|\beta)$ is the conditional frequency, in the controls, of allele α at locus l given allele β at locus $l+1$. If O_l is missing, we set $P\{O_l \mid Q_l, O_{l+1}\} = 1$, and if O_{l+1}

is missing, we set $P\{O_l \mid Q_l, O_{l+1}\}$ equal to expression (2). $P\{O_{l_{re}} | Q_{l_{re}}\}$ remains the same as in expression (2). Under our model, when $0 \leqslant l < l_{re}$, $P\{O_l \mid Q_l, Q_{l+1}, O_{l+1}\} = P\{O_l \mid Q_l, O_{l+1}\}$, which does not depend on Q_{l+1}. Note that if we had conditioned in the other direction, $P\{O_l \mid Q_l, Q_{l-1}, O_{l-1}\}$ would depend on Q_{l-1}, and this is the reason for our choice of the direction of conditioning. This is particularly useful for implementing a Markov model of lag $\eta > 1$ for the background LD. In that case, the observation distribution for $0 \leqslant l \leqslant l_{re} - \eta$ becomes

$$P\{O_l \mid Q_l, O_{l+1}, \ldots, O_{l+\eta}\} = \begin{cases} m(l, \tau, \mathbf{h}_{anc}(l), O_l) & \text{if } Q_l = A \\ f_{l,\ldots,l+\eta}(O_l \mid O_{l+1}, \ldots, O_{l+\eta}) & \text{if } Q_l = N, \end{cases} \quad (4)$$

and $P\{O_l \mid Q_l, Q_{l+1}, \ldots, Q_{l+\eta}, O_{l+1}, \ldots, O_{l+\eta}\} = P\{O_l \mid Q_l, O_{l+1}, \ldots, O_{l+\eta}\}$ does not depend on $Q_{l+1}, \ldots, Q_{l+\eta}$. This allows us to extend the model for background LD to Markov of lag $\eta > 1$ without increasing the size of the state space of the hidden Markov chain.

The joint process $\{Q_l, O_l\}$ was so far defined for $0 \leqslant l \leqslant l_{re}$. There is a corresponding mirror-image process defined on $-l_{le} \leqslant l \leqslant 0$. When background linkage equilibrium is assumed, these two processes are conditionally independent given $\{Q_0, O_0\}$. When background LD is modeled by a Markov model of lag $\eta > 0$, the two processes are conditionally independent given $\{Q_0, O_k, \ldots, O_{k+\eta-1}\}$, for any choice of k with $-\eta + 1 \leqslant k \leqslant \eta - 1$. In practice, however, we generally take the position of the variant ($l = 0$) to be in between markers, rather than at a marker, so that O_0 is always missing (this is discussed further in Appendix A). We have developed extensions to the Baum algorithms for likelihood calculation and maximization that are applicable to our model, as outlined in Appendix A.

Extension of HMM to genotype data

In practice, unambiguously-determined haplotype data are often unavailable. Instead, genotype data, in which phase is unknown, are commonly available. We describe an extension of the HMM of the previous sub-section to this case, which allows computationally efficient analysis of data sets involving genotype data on many loci.

Consider the model for multilocus genotype data from a single individual. To simplify the exposition, we first assume background linkage equilibrium. In that case, we consider the Markov chain $\{R_l^G, 0 \leqslant l \leqslant l_{re}\} = \{(Q_l^M, Q_l^P), 0 \leqslant l \leqslant l_{re}\}$, where $\{Q_l^M, 0 \leqslant l \leqslant l_{re}\}$ is the Markov chain of the previous subsection defined for the individual's maternally-inherited haplotype, while $\{Q_l^P, 0 \leqslant l \leqslant l_{re}\}$ is the Markov chain of the previous subsection defined for the individual's paternally-inherited haplotype, with $\{Q_l^M\}$ independent of $\{Q_l^P\}$. The state space of $\{R_l\}$ is $\{A, N\}^2$, the transition probabilities of $\{R_l\}$ are the products of the transition probabilities for the independent chains $\{Q_l^M\}$ and $\{Q_l^P\}$, and the initial distribution of $\{R_l\}$ is similarly obtained from the initial distributions of $\{Q_l^M\}$ and $\{Q_l^P\}$. The observation O_l^G is the genotype data

for the individual at locus l. The two possible phases for the genotype are *a priori* equally likely, so the observation distribution takes a simple form, given in Appendix B. As before, there is a corresponding mirror image process to $\{Q_l, O_l\}$ that is defined on $-l_{le} \leqslant l \leqslant 0$, with the two processes conditionally independent given $\{Q_0, O_0\}$.

We now extend the method to allow the background LD to be modeled by a Markov chain with lag $\eta \geq 1$. In order to implement a Markov model of lag η, we need to retain, at each locus l, the information of the phase of the genotype at l with respect to the genotypes at $l+1, \ldots, l+\eta$. When the genotype at locus l is recorded in a computer file, an arbitrary order of the two alleles at locus l is chosen, and we introduce the random variable Φ_l which represents the arbitrary order of the alleles in the recorded genotype. We assume that $\Phi_l = (M,P)$ or (P,M) with chance 1/2 each, where $\{\Phi_l = (M,P)\}$ denotes the event that the first allele listed in the file is the maternal allele and the second allele listed is the paternal allele, and vice versa for $\{\Phi_l = (P,M)\}$. For each l we define $(Q_l^1, Q_l^2) = (Q_l^M, Q_l^P)$ if $\Phi_l = (M,P)$ and $(Q_l^1, Q_l^2) = (Q_l^P, Q_l^M)$ if $\Phi_l = (P,M)$. That is, Q_l^i is the ancestral state corresponding to the ith recorded allele in the genotype, $i = 1,2$. Furthermore, we define $I_{l,l+1}$ to be the indicator of the event $\{\Phi_l = \Phi_{l+1}\}$, that is, the indicator of the event that the genotypes at loci l and $l+1$ happen to be recorded so that their phase with respect to one another is correct. We define the hidden Markov chain $\{R_l^G, 0 \leqslant l \leqslant l_{re}\}$ by $R_{l_{le}}^G = (Q_{l_{re}}^1, Q_{l_{re}}^2) \in \{A,N\}^2$ and $R_l^G = (Q_l^1, Q_l^2, I_{l,l+1})$ for $0 \leqslant l < l_{re}$, except that when $Q_l^1 = Q_l^2 = A$, the information of $I_{l,l+1}$ is not needed, so we collapse the two states $(A,A,1)$ and $(A,A,0)$ into a single state. The state space for $\{R_l^G, 0 \leqslant l < l_{re}\}$ is thus $\{(A,A)\} \cup [\{(A,N),(N,A),(N,N)\} \times \{0,1\}]$. The initial distribution and transition probabilities for this chain are easily determined, assuming that $\{Q_l^M\}$ and $\{Q_l^P\}$ are independent copies of the Markov chain in the previous sub-section, with the Φ_l i.i.d as given above. The observation distribution is given in Appendix B.

Note that in order to accommodate a Markov model of lag 1 for the background LD, we have increased the size of the state space from 4 to 7. Interestingly, we are able to accommodate a Markov model of lag 2 without any change in the state space. This is because, in the calculation of the observation distribution by the Baum algorithm, both R_l^G and R_{l+1}^G are available to condition on, as well as O_{l+1}^G and O_{l+2}^G, so there is no need to store extra information. As in sub-section "HMM for haplotype data, with Markov(η) model for background LD," we extend to our model the Baum algorithms for likelihood calculation and maximization (see Appendix A).

Choice of η for modeling background LD

For modeling background LD, we have defined a nested sequence of Markov models indexed by the lag η. The question arises as to how η should be chosen in practice. The usual trade-offs apply. In our case, if η is too small, background LD may be erroneously identified as LD with a trait-associated variant, while if η is too large, overfitting will quickly become a problem, as the number of parameters increases exponentially with η.

We view the choice of η as a problem of model selection. We consider two criteria, the Akaike information criterion (AIC) (Akaike 1972) and the Bayesian information criterion (BIC) (Schwartz 1978). Comparable formulations of these criteria are $AIC = -2L + 2k$ and $BIC = -2L + k\log n$, where L is the log-likelihood and k is the number of parameters. For AIC and BIC, the model that minimizes the criterion would be selected. The likelihood component L will always increase with η; both procedures include a penalty for the number of parameters, which offers some protection against overfitting.

In addition to these generic model selection techniques, we also perform an informal diagnostic, suggested by Paul Van Eerdewegh (personal communication), which is more specific to our method. To perform this diagnostic, which we call "mapping in controls", we plug the control haplotypes/genotypes into the mapping program in place of the affecteds' haplotypes/genotypes. We use the same control haplotypes to fit the Markov(η) model for background LD. To assess the results of mapping in controls, we generate the resulting log profile likelihood plot for the location of the variant. Because the same data are used both to estimate the parameters of the model for background LD and for mapping, if the model for background LD is adequate, the procedure should "recognize" these data as fitting the model. The existence of a pronounced peak in the resulting plot would suggest the presence of LD in the controls that is not adequately modeled by the Markov (η) model. If this peak coincides with the peak in the affecteds, then this suggests that the peak in the affecteds may be spurious or at least higher than is warranted. To remedy this, we would try increasing the lag η to capture more of the background LD.

3 Results

Importance of modeling background LD

We demonstrate the importance of modeling background LD in the CF data set of Kerem *et al.* [15]. The data set includes 94 haplotypes from affected individuals and 92 haplotypes from normal individuals. Pairs of haplotypes in individuals are not identified. Each haplotype consists of 23 biallelic markers within a 2-Mb region covering the gene. All physical distances are converted to genetic distances by use of the equivalence 1Mb $\approx$ 1 cM. Note that if the mutation rate were assumed to be 0, then the DHS results would be invariant under a rescaling of distance, *i.e.* 1Mb $\approx k$ cM for any k. When the mutation rate is low, the DHS results would be expected to be robust to deviation of k from 1. We further note that experiments (results not shown) have demonstrated that the method is relatively robust to misspecifications of genetic distances while errors in map order of the marker loci may have a more serious effect.

To demonstrate the importance of modeling background LD in the CF data set of Kerem *et al.* [15], we first perform mapping in the controls. Figure 1 gives the profile log-likelihood curves for the control haplotypes for LE ($\eta = 0$), $\eta = 1$ and $\eta = 2$. For

both $\eta = 0$ and $\eta = 1$, the plot for mapping in the controls is sharply peaked, suggesting that background LD that is present in the controls could be driving some of the mapping results in the affecteds. In contrast, the plot for $\eta = 2$ is completely flat, suggesting that there is little unmodeled background LD detected in the controls, but leaving open the possibility that the Markov(2) model could be overfitting the data. Table 2 gives the results of the AIC and BIC model selection procedures, where we have added constants to AIC and BIC so that each has value 0 for $\eta = 0$. The results indicate that the model with $\eta = 2$ is strongly preferred over $\eta = 0$ and $\eta = 1$ by both the AIC and BIC criteria. This leaves open the possibility that a value of $\eta > 2$ may be optimal according to the AIC and/or BIC criteria. However, the results of our diagnostic in Figure 1 suggest that additional unmodeled background LD, if present, is having little effect on the procedure.

Table 2: Model Selection for background LD in CF data set of Kerem *et al.* [15]

η	log-lik.	# param.	AIC	BIC
0	-1208.325	23	0	0
1	-739.951	45	-892.748	-837.269
2	-628.621	87	-1031.408	-870.013

Figure 2 shows the results of LD mapping when each of the three models for background LD is used. In the case of $\eta = 0$ (linkage equilibrium), the resulting profile log-likelihood curve for the affecteds resembles the profile log-likelihood for the controls given in Figure 1, suggesting that the mapping results in this case may be misleading due to unmodeled background LD. In fact, neither the 95% CI assuming independence of recombinational histories nor that assuming the conditional-coalescent model contain the true location of the variant. The resemblance between the curves for cases and controls is less strong for the case of $\eta = 1$ and non-existent for the case of $\eta = 2$. In both of these cases, the CIs cover the true location of the variant. In this data set, LD around the Δ508 mutation in the affecteds is very strong relative to the background LD, but if the LD signal were weaker, as might be expected in many data sets, the model for background LD would presumably become even more critical.

Mapping results for genotype vs. haplotype data

The cystic fibrosis (CF) data set of Kerem *et al.* [15] has almost complete haplotype information for affecteds and controls. By randomly combining haplotypes into genotypes and then throwing away the phase information, we can compare the results of LD mapping based on haplotype data with the results when only genotype data are available. Figure 3 gives the profile likelihood curve for one random pairing of affecteds' haplotypes to form genotypes for 47 individuals, where we treat phase as unknown. Here, a Markov model with lag $\eta = 2$ is used to capture background LD in the analysis. For comparison, figure 2 gives the same plot assuming haplotype data are available.

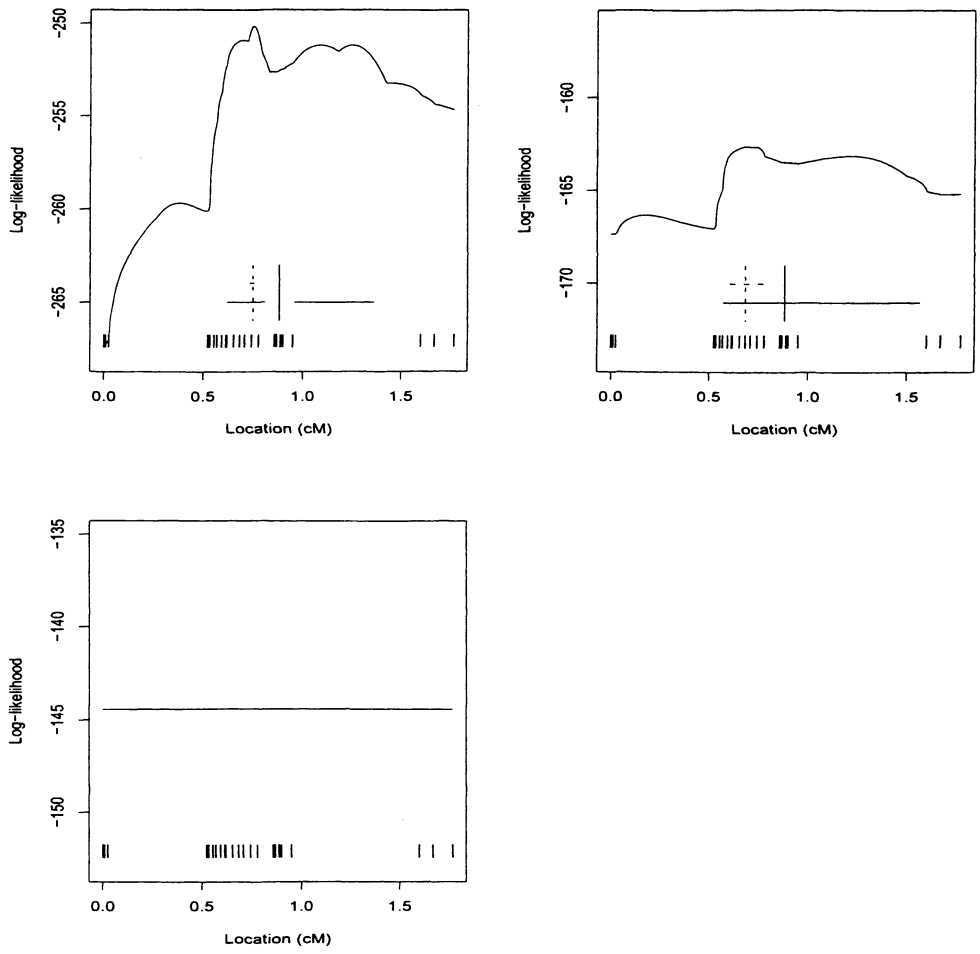

Figure 1: **Results of mapping-in-controls diagnostic** for DHS analysis of CF data set of Kerem *et al.* [15], where background LD is modeled as Markov(η) with $\eta = 0$ (upper left), $\eta = 1$ (upper right), $\eta = 2$ (lower left). Curve gives log profile likelihood vs. (putative) location x of variant, where x is expressed as distance from D21S1885. The dotted vertical line is the estimated variant location in controls, the unbroken vertical line is the true variant location, the dotted horizontal line is the 95% CI when independence of recombinational histories is assumed, and the unbroken horizontal line is the 95% CI when a conditional-coalescent model is assumed. The assumed mutation rate is 10^{-4} mutations per marker per meiosis. The hash marks give the locations of the biallelic markers. (Because the curve is flat for $\eta = 2$, we omit the CIs, both of which cover the entire region.)

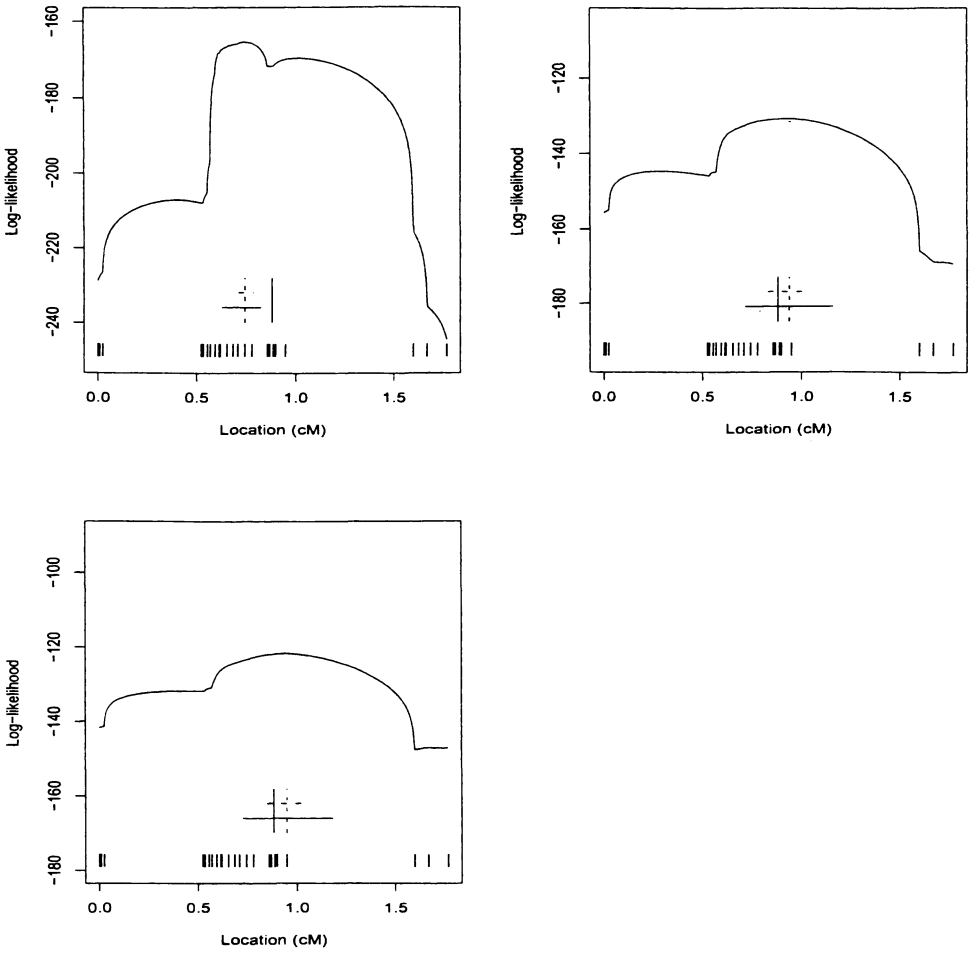

Figure 2: **Results of LD mapping for CF haplotype data** by the DHS method, where background LD is modeled as Markov(η) with $\eta = 0$ (upper left), $\eta = 1$ (upper right), $\eta = 2$ (lower left). Curve gives log profile likelihood vs. (putative) location x of variant, where x is expressed as distance from D21S1885. The dotted vertical line is the estimated variant location based on affecteds' haplotypes. The unbroken vertical line, unbroken and dotted horizontal lines, and the hash marks have the same meaning as in Figure 1.

The curves are nearly identical (after rescaling of the vertical axis), and the CIs for genotype data are only slightly wider, reflecting a slight decrease in information about the location of the variant due to the missing phase. This is expected because CF has a recessive mode of inheritance and low heterogeneity. This lack of heterogeneity means that many loci are homozygous and little information about phase is lost.

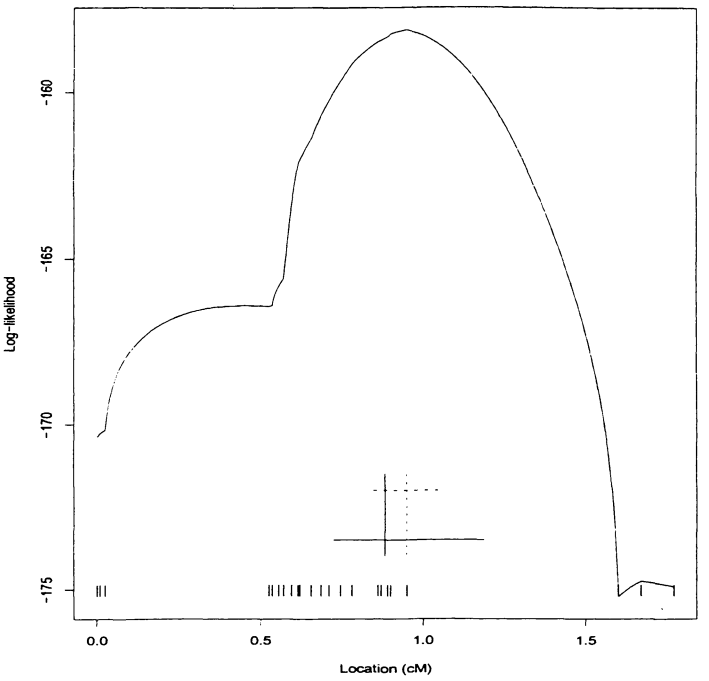

Figure 3: **Results of LD mapping for CF genotype data based on a random pairing of affecteds' haplotypes** Curve gives log profile likelihood vs. (putative) location x of variant, where x is expressed as distance from D21S1885, and where background LD is modeled as $\eta = 2$. The dotted vertical line is estimated location based on affecteds' unphased genotype data. The unbroken vertical line, unbroken and dotted horizontal lines, and the hash marks have the same meaning as in Figure 1.

To add more heterogeneity, we sample 47 affecteds' haplotypes without replacement and pair each with a randomly chosen control haplotype. This mimics the case of a rare dominant trait. We then assume we have only (unphased) genotype information for these 47 pseudo-individuals. The resulting log-likelihood curve, and the log-likelihood curve for the same data assuming haplotypes are available, are given in figure 4. In this case, the difference in the CIs between the cases when only genotype data are available and when haplotype data are available is somewhat more noticeable.

For this simulated example, the assumption of a multiplicative model for the mode of inheritance does not hold, but, at least in this case, the DHS method appears to be relatively robust to the deviation from that assumption.

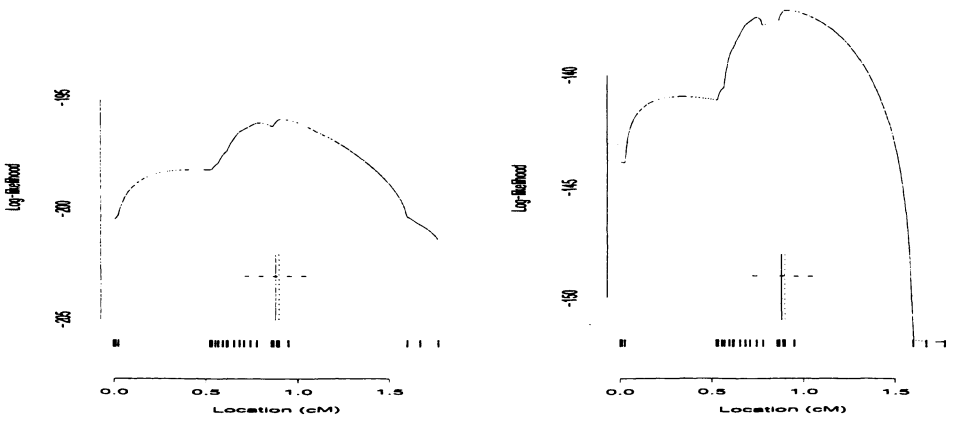

Figure 4: **Results of LD mapping for CF genotype data with added heterogeneity** (left) with results assuming haplotype data with added heterogeneity shown (right) for comparison. In each case, the curve gives log profile likelihood vs. (putative) location x of variant, where x is expressed as distance from D21S1885, and where background LD is modeled as $\eta = 2$. In the top plot, the dotted vertical line is estimated location based on unphased genotype data, while in the bottom plot, the dotted vertical line is estimated location based on haplotype data. The unbroken vertical line, unbroken and dotted horizontal lines, and the hash marks have the same meaning as in Figure 1. The details of the added heterogeneity are given in the text.

4 Discussion

In LD mapping, the presence of unmodeled background LD can have potentially serious consequences. LD that is common to both cases and controls may be mistaken for LD in cases that is due to the presence of a trait-associated variant. This is particularly likely to happen when markers are densely-spaced. For instance, in the CF data set of Kerem *et al.* [15], when background LD is inadequately modeled ($\eta = 0$ or 1) and the mapping-in-controls diagnostic is performed, the highest peak based on the DHS analysis corresponds to the most densely-genotyped region. That background LD would tend to be stronger among closely-spaced markers is to be expected under a model in which LD is broken up by recombination, leading to a sharp drop-off in LD with distance. As a result, if background LD is not appropriately modeled when performing localization, then the tendency will be for the estimated location of the

variant to fall in the most densely-genotyped region or in a region of very low marker information adjacent to the most densely-genotyped region. In addition to the effects of marker spacing, the degree of polymorphism of the markers is also a factor. The lag η of the Markov model may need to be greater with less polymorphic loci. Thus, for example, larger η may be required to adequately model background LD in SNP data than in microsatellite data of the same marker density.

The conclusions regarding importance of adequately modeling background LD would be expected to hold not only for the DHS method, but for other methods with a similar case-control approach, including those of Service *et al.* [29], Morris *et al.* [25], Liu *et al.* [18], and Morris *et al.* [24]. Our results address the effects of background LD on the localization problem. For the problem of detecting association with a variant, when such methods are applied, the presence of unmodeled background LD could be expected to increase the chance of false positive detection of association.

We extend the DHS methodology of McPeek and Strahs [22], who allow for haplotype data and model background LD using a Markov model with lag 1. For haplotype data, we develop a computationally efficient method that allows a Markov model for background LD with lag η. We also extend the method to allow use of (unphased) genotype data, which are much more commonly available than haplotype data. In practice, the order of the Markov model for background LD is limited by the size of the sample of controls needed to estimate the frequency parameters and by the increasing computational demands of higher order models, especially for genotype data. We have implemented, in free software, the DHS method for both haplotype and genotype data, with background LD modeled as Markov(η) where $\eta \leq 2$. These methods are implemented, for both haplotype and genotype data, using an efficient HMM. The genotype-data DHS HMM incorporates uncertainty about phase into the likelihood and allows the method to operate on data sets with a large numbers of marker loci. In addition to the CF data set which includes 23 markers, we have applied the methodology to data sets with (unphased) genotype data on 80+ markers (data not shown).

We demonstrate the importance of modeling background LD using the CF data set. For that data set, our results indicate that when background LD is assumed absent ($\eta = 0$) or is modeled by a Markov ($\eta = 1$) model, additional unmodeled background LD present in the controls could be driving some of the mapping results in the affecteds. The model selection criteria AIC and BIC both prefer the Markov model with $\eta = 2$ to those with $\eta = 1$ or $\eta = 0$. Based on the mapping-in-controls diagnostic, there is no detectable unmodeled background LD in the controls when the model with $\eta = 2$ is used. When mapping (with affecteds) is performed, the 95% CI for location does not cover the true variant when $\eta = 0$ is used, while the CIs do cover the true location in the cases $\eta = 1$ and $\eta = 2$. In the CF example, there is little heterogeneity among the affecteds, so the model for background LD plays less of a role in the analysis than it would in a situation of greater heterogeneity. Thus, we might reasonably expect the effects of background LD to be more important in other data sets. In practical situations, one could apply the AIC and BIC model selection criteria to compare models

of background LD, and one could apply the mapping-in-controls diagnostic to assess the adequacy of the chosen model.

We have extended our methods to allow for unphased genotype data as well as for haplotype data. While in some cases, genotype data on close family members may provide considerable haplotype information, our extension to unphased genotype data may be particularly useful when it is difficult to obtain genetic data from close relatives, as may happen when studying diseases with a late age of onset, such as Type 2 diabetes. In addition to allowing for haplotype or unphased genotype data, the DHS method can also be extended to allow for trio data [33].

There have been some interesting recent results regarding the possible nature of background linkage disequilibrium [5, 14]. These results suggest that high-resolution haplotype structure, at least in certain regions of the human genome, takes a relative simple form. This consists of disjoint haplotype blocks (of tens to hundreds of kb), where within each block there is very strong LD with only a few (*e.g.* ~2-7) commonly-occurring haplotypes. Between the blocks are regions over which there is lower LD (possibly representing recombination hotspots, at least in some cases [13]). Many questions remain about the extrapolation of these observations to the human genome as a whole and to various human populations, from the select regions, populations, and data sets that have so far been studied. There is currently some interest in a large-scale effort to explore this hypothesis and to take advantage of it for LD mapping (*e.g.* see `http://www.genome.gov/page.cfm?pageID=10001676`). The Markov models considered by [5] are extensions of the models we consider here. In order to characterize this block structure, if it exists, a tremendous amount of data would need to be collected and an enormous number of parameters estimated (including start and end points of blocks, common haplotypes in blocks and their frequencies, associations between common haplotypes in different blocks, and also characteristics of the regions of low LD between blocks). Were such information available, these more detailed models for background LD could be incorporated in a natural way in the DHS model. Furthermore, the DHS likelihood for a single haplotype could itself be modified to incorporate a model of block structure for fine-resolution haplotypes.

Another interesting extension would be to combine the DHS method with a method such as the structured association method of Pritchard *et al.* [27], which uses genotypes at unlinked markers to infer population substructure which is then used to test association at the locus of interest. The information of population substructure could presumably also be used for the localization problem with multilocus data. Alternatively, an idea similar to that of genomic control [8] might be adaptable to the localization problem.

Electronic-Database Information

Software for mapping with haplotype or genotype data and $\eta = 0, 1, 2$ is freely available at `http://galton.uchicago.edu/~mcpeek/software/dhsmap`.

Acknowledgments

This work is supported by National Institutes of Health grants DK55889 and HG01645. We are grateful to Nancy Cox and Jian Zhang for helpful discussions, Ken Wilder for his assistance with the software, and Paul Van Eerdewegh for his suggestion of the mapping-in-controls diagnostic.

Dedication

We dedicate this paper to Terry Speed on the occasion of his 60th birthday, with gratitude for his help, support, and encouragement.

Appendix A: Likelihood calculation and maximization

We have developed extensions to the Baum algorithms for likelihood calculation and maximization that are applicable to our model. We first define $\gamma_l(i)$, for a given sampled haplotype, as the probability that $Q_l = i$ at locus l, conditional on the observed haplotype and the parameter values (all of the following probabilities are conditional on the parameter values), *i.e.*,

$$\gamma_l(i) = P\{Q_l = i \mid \mathbf{O}\},$$

which by the definition of conditional probability is $P\{Q_l = i, \mathbf{O}\}/P\{\mathbf{O}\}$. The numerator is computed as the product of two complementary recursively generated variables, a "forward variable" α and a "backward variable" β. For l, $-l_{le} \leqslant l < 0$,

$$\begin{aligned} P\{Q_l = i, \mathbf{O}\} &= P\{O_{-l_{le}}, O_{-l_{le}+1}, \ldots, O_l, Q_l = i\} \times \\ &\quad P\{O_{l+1}, \ldots, O_{l_{re}-1}, O_{l_{re}} \mid O_{-l_{le}}, \ldots, O_l, Q_l = i\} \\ &= P\{O_{-l_{le}}, O_{-l_{le}+1}, \ldots, O_l, Q_l = i\} \times \\ &\quad P\{O_{l+1}, \ldots, O_{l_{re}-1}, O_{l_{re}} \mid O_l, \ldots, O_{l-\eta+1}, Q_l = i\} \\ &= \alpha_l(i)\beta_l(i), \end{aligned}$$

where

$$\alpha_l(i) = P\{O_{-l_{le}}, O_{-l_{le}+1}, \ldots, O_l, Q_l = i\}$$

and

$$\beta_l(i) = P\{O_{l+1}, \ldots, O_{l_{re}} \mid O_l, O_{l-\eta+1}, Q_l = i\}.$$

The definitions for the forward and backward variables on the centromeric side of locus 0 are mirror images of the previous case, *i.e.*, for l, $0 < l \leqslant l_{re}$,

$$\alpha_l(i) = P\{O_{-l_{le}}, O_{-l_{le}+1}, \ldots, O_{l-1} \mid O_l, \ldots, O_{l+\eta-1}, Q_l = i\}$$

and

$$\beta_l(i) = P\{O_l, \ldots, O_{l_{re}-1}, O_{l_{re}}, Q_l = i\}.$$

We note that the left and right sides of the chain are dependent, conditional on Q_0, only for $Q_0 = N$. In this case, the likelihood of the haplotype is just $P_{null}(\mathbf{h}_{obs})$. For $\eta = 2$,

$$\begin{aligned} P_{null}(\mathbf{h}_{obs}) \;=\; & f_{-l_{le},-l_{le}+1}(\mathbf{h}(-l_{le}), \mathbf{h}(-l_{le}+1)) \times \\ & \left(\prod_{l=-l_{le}, l\neq -2,-1,0}^{l_{re}-2} f_{l,l+1,l+2}(\mathbf{h}(l+2) \mid \mathbf{h}(l), \mathbf{h}(l+1)) \right) \times \\ & f_{-2,-1,1}(\mathbf{h}(1) \mid \mathbf{h}(-2), \mathbf{h}(-1)) \times \\ & f_{-1,1,2}(\mathbf{h}(2) \mid \mathbf{h}(-1), \mathbf{h}(1)) \end{aligned}$$

We note that we "skip over" locus 0 in this product, *e.g.*, we include (recall that O_0 is missing) $P\{O_1 \mid O_0, O_{-1}, \ldots, O_{-\eta}\}$ rather than $P\{O_1 \mid O_0, O_{-1}, \ldots, O_{-\eta-1}\}$, as a Markov model of order η generally implies. Were we to include the latter, the likelihood given $Q_0 = N$ would depend on the marker interval within which the variant is assumed to lie, although, according to our model, the haplotype is drawn from the normal population. Thus, we use

$$\begin{aligned} & P\{Q_l = i, \mathbf{O}\} = \\ & \quad P\{O_{-l_{le}}, O_{-l_{le}+1}, \ldots, O_{-1} \mid O_0, O_1, \ldots, O_\eta, Q_0 = i\} P\{O_0, \ldots, O_{l_{re}-1}, O_{l_{re}}, Q_0 = i\} \end{aligned}$$

to compute $\gamma_0(i)$.

Where $\mathbf{h}$ indexes the sampled haplotypes, $c_l^* = \sum_{\mathbf{h}} \gamma_{l,\mathbf{h}}(A)$ and $b_l^* = \sum_{\mathbf{h}} \gamma_{l,\mathbf{h}}(A) \times 1_{\mathbf{h}(l)\neq \mathbf{h}_{anc}(l)}$, then $(c_{-l_{le}}^*, b_{-l_{le}}^*, \ldots, c_{-1}^*, b_{-1}^*, c_0^*, c_1^*, b_1^*, \ldots, c_{l_{re}}^*, b_{l_{re}}^*)$ is the conditional expectation of the complete data sufficient statistic for $(1/\tau, p)$ given the data and current model. The model parameters are then re-estimated by maximizing the complete data log-likelihood, substituting this statistic for the complete data sufficient statistic.

The extension to genotype data is straightforward. The primary differences are (1) the state space of the Markov chain is larger and (2) c_l^* is formed by summing over the genotypes, rather than the haplotypes.

Appendix B: Genotype HMM Observation Distributions

Assuming linkage equilibrium,

$$\begin{aligned} P\{O_l^G = (\alpha, \beta) \mid R_l^G\} = 1/2 P\{O_l^M = \alpha \mid Q_l^M\} P\{O_l^P = \beta \mid Q_l^P\} \\ + 1/2 P\{O_l^M = \beta \mid Q_l^M\} P\{O_l^P = \alpha \mid Q_l^P\} \end{aligned}$$

for the case when both $\alpha \neq \beta$ and $Q_l^M \neq Q_l^P$, and

$$P\{O_l^G = (\alpha,\beta) \mid R_l^G\} = P\{O_l^M = \alpha \mid Q_l^M\} P\{O_l^P = \beta \mid Q_l^P\}$$

if either $\alpha = \beta$ or $Q_l^M = Q_l^P$, where O_l^M is defined to be the allele at locus l on the maternally-inherited haplotype, O_l^P is defined to be the allele at locus l on the paternally-inherited haplotype, and $P\{O_l^M \mid Q_l^M\}$ and $P\{O_l^P \mid Q_l^P\}$ are given by expression (2).

Let $O_l^G = (O_l^1, O_l^2)$ be the alleles of the genotype at locus l, given in order Φ_l. For a Markov model of lag 1, the observation distribution is given by

$$\begin{aligned} P\{O_l^G \mid R_l^G, O_{l+1}^G\} = {} & P(O_l^1 \mid Q_l^1, O_{l+1}^1) P(O_l^2 \mid Q_l^2, O_{l+1}^2) I_{l,l+1} \\ & + P(O_l^1 \mid Q_l^1, O_{l+1}^2) P(O_l^2 \mid Q_l^2, O_{l+1}^1)(1 - I_{l,l+1}) \end{aligned}$$

for $0 \leqslant l < l_{re}$, where $P(O_l \mid Q_l, O_{l+1})$ is given by expression (3). When $Q_l^1 = Q_l^2 = A$, this equation reduces to $P(O_l^G \mid Q_l^1 = Q_l^2 = A, O_{l+1}^G) = P(O_l^1 \mid Q_l^1 = A) P(O_l^2 \mid Q_l^2 = A)$. For a Markov model of lag 2, the observation distribution is given by

$$\begin{aligned} P\{O_l^G \mid R_l^G, R_{l+1}^G, O_{l+1}^G, O_{l+2}^G\} = {} & P(O_l^1 \mid Q_l^1, O_{l+1}^1, O_{l+2}^1) P(O_l^2 \mid Q_l^2, O_{l+1}^2, O_{l+2}^2) I_{l,l+1} I_{l+1,l+2} \\ & + P(O_l^1 \mid Q_l^1, O_{l+1}^1, O_{l+2}^2) P(O_l^2 \mid Q_l^2, O_{l+1}^2, O_{l+2}^1) I_{l,l+1} (1 - I_{l+1,l+2}) \\ & + P(O_l^1 \mid Q_l^1, O_{l+1}^2, O_{l+2}^2) P(O_l^2 \mid Q_l^2, O_{l+1}^1, O_{l+2}^1)(1 - I_{l,l+1}) I_{l+1,l+2} \\ & + P(O_l^1 \mid Q_l^1, O_{l+1}^2, O_{l+2}^1) P(O_l^2 \mid Q_l^2, O_{l+1}^1, O_{l+2}^2)(1 - I_{l,l+1})(1 - I_{l+1,l+2}) \end{aligned}$$

for $0 \leqslant l < l_{re} - 1$, where $P(O_l \mid Q_l, O_{l+1}, O_{l+2})$ is as given by expression (4). When $Q_l^1 = Q_l^2 = A$, this equation reduces to

$$P(O_l^G \mid Q_l^1 = Q_l^2 = A, R_{l+1}^G, O_{l+1}^G, O_{l+2}^G) = P(O_l^1 \mid Q_l^1 = A) P(O_l^2 \mid Q_l^2 = A).$$

Andrew L. Strahs, Department of Biostatistics, Harvard School of Public Health, Boston,
astrahs@hsph.harvard.edu

Mary Sara McPeek, Department of Statistics, University of Chicago, Chicago,
mcpeek@galton.uchicago.edu

References

[1] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[2] C. Bourgain, E. Genin, H. Quesneville, and F. Clerget-Darpoux. Search for multifactorial disease susceptibility genes in founder populations. *Annals of Human Genetics*, 64:255–265, 2000.

[3] D. Clayton and H. Jones. Transmission/disequilibrium tests for extended marker haplotypes. *American Journal of Human Genetics*, 65:1161–1169, 1999.

[4] A. Collins and N. E. Morton. Mapping a disease locus by allelic association. *American Journal of Human Genetics*, 95:1741–1745, 1998.

[5] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29:229–232, 2001.

[6] A. de la Chappelle and F. A. Wright. Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. *Proceedings of the National Academy of Sciences, USA*, 95:12416–12423, 1998.

[7] B. Devlin, N. Risch, and K. Roeder. Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics*, 36:1–16, 1996.

[8] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55:997–1004, 1999.

[9] L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12:921–927, 1995.

[10] J. Hästbacka, A. de la Chapelle, I. Kaitila, P. Sistonen, A. Weaver, and E. Lander. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genetics*, 2:204–211, 1992.

[11] J. Hästbacka, A. de la Chapelle, M. M. Mahanti, G. Clines, M. P. Reeve-Daly, M. Daly, B. A. Hamilton, K. Kusumi, B. Trivedi, and A. Weaver. The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell*, 78:1073–1087, 1994.

[12] M. Hawley and K. Kidd. Haplo: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *Journal of Heredity*, 86:409–411, 1995.

[13] A. J. Jefffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, 29:217–222, 2001.

[14] G. C. L. Johnson, L. Esposito, B. J. Barratt, A. N. Smeith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. C. J. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto,

S. C. L. Gough, D. G. Clayton, and J. A. Todd. Haplotype tagging for the identification of common disease genes. *Nature Genetics*, 29:233–237, 2001.

[15] B. Kerem, J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald, and L.-C. Tsui. Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245:1073–1080, 1989.

[16] J. C. Lam, K. Roeder, and B. Devlin. Haplotype fine mapping by evolutionary trees. *American Journal of Human Genetics*, 66:659–673, 2000.

[17] L. Lazzeroni. Linkage disequilibrium and gene mapping: an empirical least-squares approach. *American Journal of Human Genetics*, 62:159–170, 1998.

[18] J. S. Liu, C. Sabatti, J. Teng, J. B. Keats, and N. Risch. Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Research*, 11:1716–1724, 2001.

[19] J. C. Long, R. C. Williams, and M. Urbanek. An E-M algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics*, 56:799–810, 1995.

[20] C. J. MacLean, R. B. Martin, P. C. Sham, H. Wang, R. E. Straub, and J. S. Kendler. The trimmed-haplotype test for linkage disequilibrium. *American Journal of Human Genetics*, 66:1062–1075, 2000.

[21] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.

[22] M. S. McPeek and A. L. Strahs. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *American Journal of Human Genetics*, 65:858–875, 1999.

[23] A. P. Morris and J. C. Whittaker. Fine scale association mapping of disease loci using simplex families. *Annals of Human Genetics*, 64:223–237, 2000.

[24] A. P. Morris, J. C. Whittaker, and D. J. Balding. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *American Journal of Human Genetics*, 70:686–707, 2000.

[25] A. P. Morris, J. C. Whittaker, and D. J. Balding. Bayesian fine-scale mapping of disease loci, by Hidden Markov Models. *American Journal of Human Genetics*, 67:155–169, 2002.

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

[27] J. K. Pritchard, M Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *American Journal of Human Genetics*, 67:170–181, 2000.

[28] B. Rannala and J. Reeve. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *American Journal of Human Genetics*, 69:159–178, 2001.

[29] S. Service, D. Temple Lang, N. Freimer, and L. Sandkuijl. Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *American Journal of Human Genetics*, 64:1728–1738, 1999.

[30] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction. *American Journal of Human Genetics*, 68:978–989, 2001.

[31] J. D. Terwilliger. Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. *American Journal of Human Genetics*, 56:777–787, 1995.

[32] M. Xiong and S.-W. Guo. Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. *American Journal of Human Genetics*, 60:1513–1531, 1997.

[33] J. Zhang. *Linkage disequilibrium mapping by the decay of haplotype sharing in a founder population*. PhD thesis, University of Chicago, 2001.

[34] S. Zhang and H. Zhao. Linkage disequilibrium mapping in populations of variable size using the decay of haplotype sharing and a stepwise-mutation model. *Genetic Epidemiology*, 19(Suppl 1):S99–S105, 2000.

[35] S. Zhang and H. Zhao. Linkage disequilibrium mapping with genotype data. *Genetic Epidemiology*, 22:66–77, 2002.

[36] H. Zhao, S. Zhang, K. R. Merikangas, M. Trixler, D. B. Weldenauer, F. Sun, and K. K. Kidd. Transmission/disequilibrium tests using multiple tightly linked markers. *American Journal of Human Genetics*, 67:936–946, 2000.