Assessment of Linkage Disequilibrium by the Decay of Haplotype Sharing, with Application to Fine-Scale Genetic Mapping

Mary Sara McPeek and Andrew Strahs

Department of Statistics, University of Chicago, Chicago

Summary

Linkage disequilibrium (LD) is of great interest for gene mapping and the study of population history. We propose a multilocus model for LD, based on the decay of haplotype sharing (DHS). The DHS model is most appropriate when the LD in which one is interested is due to the introduction of a variant on an ancestral haplotype, with recombinations in succeeding generations resulting in preservation of only a small region of the ancestral haplotype around the variant. This is generally the scenario of interest for gene mapping by LD. The DHS parameter is a measure of LD that can be interpreted as the expected genetic distance to which the ancestral haplotype is preserved, or, equivalently, 1/(time in generations to the ancestral haplotype). The method allows for multiple origins of alleles and for mutations, and it takes into account missing observations and ambiguities in haplotype determination, via a hidden Markov model. Whereas most commonly used measures of LD apply to pairs of loci, the DHS measure is designed for application to the densely mapped haplotype data that are increasingly available. The DHS method explicitly models the dependence among multiple tightly linked loci on a chromosome. When the assumptions about population structure are sufficiently tractable, the estimate of LD is obtained by maximum likelihood. For more-complicated models of population history, we find means and covariances based on the model and solve a quasi-score estimating equation. Simulations show that this approach works extremely well both for estimation of LD and for fine mapping. We apply the DHS method to published data sets for cystic fibrosis and progressive myoclonus epilepsy.

Received December 21, 1998; accepted for publication June 23, 1999; electronically published August 11, 1999.

Address for correspondence and reprints: Dr. Mary Sara McPeek, Department of Statistics, University of Chicago, 5734 South University Avenue, Chicago, IL 60637. E-mail: mcpeek@galton.uchicago.edu

0002-9297/1999/6503-0031$02.00

Introduction

Gametic association or linkage disequilibrium (LD) is potentially useful for fine-scale genetic mapping as well as for the study of population history and dynamics. Rather than consider all types of LD, we focus attention on LD that is due primarily to the introduction of a variant on an ancestral haplotype that is then partially preserved in descendants. (We do allow for some proportion of the variant-containing haplotypes to be unrelated to the others.) We consider two distinct problems.

Problem 1: Assessment of the Magnitude of LD around a Variant

Suppose that a sample of haplotypes sharing a specific genetic variant are collected from a given population. Assuming that the haplotypes provide considerable marker information quite close to the variant, how can we quantify the degree of LD around the variant in the population? In principle, we view this problem as equivalent to estimation of the age of the variant, although in practice it should usually be regarded as equivalent to estimation of the time to the most recent common ancestor of the variant in the sample. (This distinction is discussed in detail by Rannala and Slatkin [1998].) In this problem, we assume that the genetic locations of the variant and of the markers are known. We also require additional information such as either a sample of control haplotypes from the same population or marker-allele frequencies for the population.

Problem 2: Fine-Scale Genetic Mapping by LD

Now suppose that a sample of haplotypes likely to have a specific genetic variant are collected from a given population but that this time the location of the variant is unknown and the goal is to estimate the location. We presume that strong linkage between the variant and a chromosomal region has been established and that the haplotype data consist of a considerable amount of marker information in a small region tightly linked to the variant. As above, we assume that the locations of the markers are known and that information on frequencies of haplotypes or alleles in an appropriate control population are available.

In both problems, we seek to use information on LD from multiple markers simultaneously, but most of the commonly used measures of LD are applicable only to pairs of loci, many only to biallelic loci. We note that to consider only pairwise information on LD among loci when extensive multilocus haplotypes are available is a tremendous waste of valuable information. Suppose that allele A at locus 1 and allele B at locus 2 seem to appear together on haplotypes disproportionately often. Certainly the presence or absence of shared alleles at several polymorphic loci between A and B on the A-B haplotypes would provide much additional evidence about the strength and nature of the association. Bennett (1954) and Slatkin (1972) have each given formulations of LD that are based on multiple biallelic markers, essentially looking at interactions of every order among the loci. For more than a handful of loci, the higher-order interactions become numerous and exceedingly complicated. These approaches are not tailored to the particular situation of inheritance by descent of ancestral haplotypes and do not provide a summary of LD in a form that would be useful for solving either of the problems mentioned above.

Likelihood approaches to problems 1 and 2, using pairs of loci and taking into account population structure, have been described by Kaplan et al. (1995), Graham and Thompson (1998), and Rannala and Slatkin (1998). Graham and Thompson (1998) also extend their implementation to interval mapping. Several methods for multipoint LD mapping have been proposed. The multipoint method of Terwilliger (1995) and the multipoint method of Xiong and Guo (1997), using their first-order approximation, are based on combination of single-point likelihoods. They take into account neither the dependence across loci within a haplotype nor the dependence across haplotypes that is due to population structure. Devlin et al. (1996) model dependence due to population structure, using a gamma model validated by simulations, although they do not model the dependence across markers within a haplotype. Lazzeroni (1998) develops a method for biallelic markers, in which transformed pairwise disequilibrium measures are combined across loci, by application of nonlinear regression. Rather than model dependence across markers within a haplotype and dependence due to population structure, Lazzeroni (1998) uses a bootstrap estimate of the covariance matrix, which can take into account covariance conditional on the realized population, although not unconditional covariance across possible realizations of the population.

For fine-scale mapping, the dependence across loci within an individual haplotype is expected to be extremely high. Therefore, we take the approach of modeling it explicitly. To take into account population structure, we use covariances calculated under a population model and apply a quasi-likelihood approach. The population model that we use is a variation on the coalescent model of Kingman (1982), in which we condition on the time to the most recent common ancestor.

Methods

We argue that in many contexts it is useful and natural to think of LD as occurring around a particular variant in a population, rather than as a function of a pair of loci. Thus, we propose to assess the distribution of the extent of a region of shared haplotype around a variant. We model the rate at which the number of individuals still sharing a haplotype decreases with increasing distance.

We begin by describing in detail the likelihood for a single observed haplotype under our model. Then we show how to form the joint likelihood or quasi-score function for all of the observations, depending on the assumed model for the descent relationships among the individual haplotypes. Initially we assume that the location of the variant is known. When the location of the variant is unknown, we consider location as a parameter and maximize the likelihood (or quasi-likelihood) over location, simultaneously with the other parameters. This is done by maximization of the fixed-location likelihood for each possible location of the variant within a fine grid, with the maximum-likelihood estimate taken to be the location, along with the corresponding maximizing parameters, for which the maximized likelihood is largest. Thus, all development of the likelihood for the location-known problem is immediately applicable to the location-unknown problem.

First consider the dense-marker case, for simplicity. Suppose that we observe a single haplotype that is a τth-generation descendant of a given ancestor and that inherits from this ancestor a particular variant at locus 0. If we assume that there is no selection at any loci other than locus 0 and that there is no interference, then the distances (in Morgans) from locus 0 to the right and left breakpoints of the ancestral segment in the descendant are distributed as independent exponential random variables with rate τ (see Appendix A). With interference, this result is approximate for small genetic distance and large τ. (In Appendix A, we show how interference can be incorporated into the model.) Then, the length (in Morgans) of the inherited ancestral segment has the distribution of a gamma$(2,\tau)$ random variable. The expected length of the ancestral segment is $2/\tau$ Morgans, and the expected genetic distance from the variant to either edge of the ancestral segment is τ^{-1} Morgans. A larger value of τ^{-1} corresponds to a larger inherited ancestral segment (i.e., higher LD), whereas $\tau^{-1} = 0$ corresponds to no LD. To solve problem 1 (i.e., to assess the magnitude of LD around a variant), we take τ^{-1} as a measure of the LD around the variant and proceed to estimate it from multilocus haplotype data.

Note that when the ancestral haplotype is known, τ is the age of the variant. Aside from the fact that τ^{-1} is a more natural measure of LD than is τ, we prefer to estimate τ^{-1} rather than τ because, in small samples, the sampling distribution is closer to normality, so the confidence intervals (CIs) based on the Fisher information and the normal approximation have closer to the true coverage for τ^{-1} than for τ (results not shown). To obtain an estimate and CI for τ, we simply invert the estimate and confidence bounds for τ^{-1}.

The idea behind our approach to problem 2, fine-scale genetic mapping by LD, is the same principle that is used in multipoint linkage mapping; that is, a likelihood is specified that describes a particular pattern of localized excess sharing that would be observed among those sharing a variant by descent. The likelihood typically has one or more parameters that specify the degree of excess sharing at the given location. The likelihood is maximized over these parameters at each location, and the location with the largest maximized likelihood is selected. (In multipoint linkage analysis and in the LD mapping method that we describe, the null likelihood is the same at every location.) For instance, in parametric linkage analysis, parameters of the genetic model for the trait may be maximized over at each point. In the semiparametric linkage analysis described by Whittemore (1996) and Kong and Cox (1997), the parameter δ specifying the degree of excess sharing is maximized over at each point. The analogous parameter in our formulation of the LD mapping problem is τ^{-1}. Thus, we maximize the likelihood over τ^{-1} at each location in the fine-mapping region, to determine the location whose maximized likelihood is largest. Intuitively, a large maximized likelihood at a particular location indicates a deviation, from null sharing, that is well explained by exponential decay of a shared haplotype centered around the given location.

Suppose that we can make observations only at marker loci, which are assumed to be very closely spaced. At this point, we still assume that we have perfect information about identity by descent from the ancestral haplotype at each marker; this assumption will be discarded later. As before, suppose that the inherited variant is at locus 0, with loci 1,2,3,... at increasing distance to one side and with loci −1,−2,−3,... at increasing distance to the other side. Let x_i be the signed genetic distance (in Morgans) of locus i from locus 0, and let $d_{j,i} = x_i - x_j$ for $j < i$. Suppose that the τth-generation descendant haplotype, call it "$\boldsymbol{h}_{\mathrm{obs}}$," inherits the variant and the ancestral block, intact, between loci $-k$ and j inclusive but that it is no longer intact at locus $-k-1$ nor at locus $j+1$. Then, letting $L(\tau^{-1};\boldsymbol{h}_{\mathrm{obs}})$ be the likelihood of the observed haplotype and letting the function g represent the portion of the likelihood involving τ^{-1}, we have

$$L(\tau^{-1};\boldsymbol{h}_{\mathrm{obs}}) \propto g(\tau^{-1},-k,j) = e^{-\tau d_{-k,j}}(1-e^{-\tau d_{-k-1,-k}})(1-e^{-\tau d_{j,j+1}}) \ . \qquad (1)$$

Here, the first factor represents the probability that there are no crossovers between loci $-k$ and j during τ generations, the second factor represents the probability that there is at least one crossover between loci $-k$ and $-k-1$ during τ generations, and the last factor represents the probability that there is at least one crossover between loci j and $j+1$ during τ generations. If locus j were to represent the edge of the observed haplotype, then the term $(1-e^{-\tau d_{j,j+1}})$ would not appear, and a similar situation would obtain for locus $-k$. We assume that adjacent markers are close enough that unobserved double recombinants would be too rare to be of any consequence in the analysis.

For a single τth-generation descendant haplotype, consider the function R of chromosomal location x (expressed as signed distance from the variant), which assigns the value $R(x) = A$ ("ancestral") to all locations x in the largest intact ancestral block surrounding the variant and which assigns value $R(x) = N$ ("nonancestral") to all other locations. Then, on the basis of formula (1), it can be seen that $R(x)$ is equivalent to a pair of continuous-time Markov chains indexed by position, both starting at the variant but going in opposite directions, with $P\{R(x+t) = A|R(x) = A\} = \exp(-\tau|t|)$, $P\{R(x+t) = N|R(x) = A\} = 1 - \exp(-\tau|t|)$, and $P\{R(x+t) = N|R(x) = N\} = 1$, for x and t either both >0 or both <0. With incomplete data—that is, when (*a*) the markers provide information only on identity by state, rather than identity by descent, with the ancestral haplotype, (*b*) some marker information is missing, (*c*) mutations are allowed, and (*d*) genotype data provide only partial information on haplotypes—we will make use of this Markov-chain formulation to create a hidden Markov model, which allows for maximization of the likelihood by the Baum/expectation-maximization (EM) algorithm (Baum 1972; Dempster et al. 1977).

Estimation of the Ancestral Haplotype

In practice, the ancestral haplotype is not known, and we estimate it from the data by maximum likelihood. To do this, we must include in the likelihood a component giving the probability that particular alleles will be observed, given that they are nonancestral; that is, in place of formula (1), the likelihood of a single observed haplotype can be written

$$L(\boldsymbol{h}_{\mathrm{anc}},\tau^{-1};\boldsymbol{h}_{\mathrm{obs}}) = g(\tau^{-1},-k,j) \times P_{\mathrm{null}}[h_{\mathrm{obs}}(j+1),h_{\mathrm{obs}}(j+2),\ldots] \times P_{\mathrm{null}}[h_{\mathrm{obs}}(-k-1),h_{\mathrm{obs}}(-k-2),\ldots]\ , \qquad (2)$$

where $\boldsymbol{h}_{\mathrm{anc}}$, the ancestral haplotype, is now a parameter in the likelihood and $\boldsymbol{h}_{\mathrm{obs}}$, the haplotype of the observed τth-generation descendant, is a data point. $g(\tau^{-1}, -k,j)$ is given in formula (1), $h_{\mathrm{obs}}(i)$ represents the allele at locus i in the haplotype $\boldsymbol{h}_{\mathrm{obs}}$, $P_{\mathrm{null}}[h_{\mathrm{obs}}(j+1),h_{\mathrm{obs}}(j+2),\ldots]$ is the joint probability that the alleles $h_{\mathrm{obs}}(j+1),h_{\mathrm{obs}}(j+2),\ldots$ occur in a nonancestral haplotype, and $P_{\mathrm{null}}[h_{\mathrm{obs}}(-k-1),h_{\mathrm{obs}}(-k-2),\ldots]$ is defined similarly. The form of this probability depends on the information that is available on control haplotypes. If only the allele frequencies from control haplotypes are available, as in the published EPM1 data set of Virtaneva et al. (1996), then $P_{\mathrm{null}}[h_{\mathrm{obs}}(j+1),h_{\mathrm{obs}}(j+2),\ldots]$ could be taken to be the product of the allele frequencies for the observed alleles—that is, $\Pi_{l\geqslant j+1}f[h_{\mathrm{obs}}(l)]$, where $f[h_{\mathrm{obs}}(l)]$ is the frequency of allele $h_{\mathrm{obs}}(l)$ in an appropriate control population. If control haplotypes are available, as in the published cystic fibrosis (CF) data set of Kerem et al. (1989), then a model such as a kth-order Markov-chain model could be used to obtain $P_{\mathrm{null}}[h_{\mathrm{obs}}(j+1),h_{\mathrm{obs}}(j+2),\ldots]$. In practice, there will typically not be enough information to warrant the use of a Markov chain of order >1. In this case, instead of multiplying the frequencies of the marker alleles, we multiply their one-step conditional frequencies, $f[h_{\mathrm{obs}}(l+1)|h_{\mathrm{obs}}(l)]$ = [no. of occurrences of the pair $(h_{\mathrm{obs}}(l+1),h_{\mathrm{obs}}(l))$ in the control population]/[no. of occurrences of $h_{\mathrm{obs}}(l)$ in the control population]. In practice, if any frequency estimate in the control population is zero, we adjust the distribution so that the frequency is given a low but nonzero value.

We maximize likelihood (2) over the ancestral haplotype and τ^{-1} simultaneously. In principle, this could be done by maximizing likelihood (2) over τ^{-1} for each possible ancestral haplotype, then choosing, as our maximum-likelihood estimate, the ancestral haplotype and corresponding τ^{-1} for which the maximized likelihood is highest. In practice, the number of possible ancestral haplotypes makes this approach infeasible. Instead, we implement a branch-and-bound procedure to maximize the likelihood without needing to maximize likelihood (2) for every possible ancestral haplotype. The procedure is described in more detail in Appendix B.

As one moves farther away from the variant, the estimation of the ancestral haplotype becomes progressively less certain, because the number of individuals still sharing the ancestral haplotype decreases. Thus, it is sensible to estimate the ancestral haplotype only out to a distance from the variant at which there is still a reasonable amount of sharing in the data set. We measure the amount of sharing of the ancestral haplotype out to a certain marker in the data set by the expected number of individuals still sharing the ancestral haplotype out to that point, conditional on the data and the parameter estimates (for details, see Appendix B). Through simulation, we find that stopping the ancestral-haplotype estimation at a distance at which the conditional expected number of individuals sharing the haplotype drops to <5 or <25% of the sample size, whichever is greater, provides an estimator of τ^{-1} that has little or no bias, even for small sample sizes (see "Simulation Results for Assessment of Magnitude of LD around a Variant" in Results section, below). On the other hand, for the multilocus mapping problem, it is necessary to include in the likelihood the same data for each possible location, and we are generally not concerned about bias in the estimation of τ^{-1} in that case. Thus, in the mapping problem, we do not impose any threshold for stopping the reconstruction of the ancestral haplotype.

Allowing for Chance Sharing of Alleles

In any assessment of LD, it is important to take into account chance sharing at the loci surrounding the variant. If we were to neglect the effects of chance sharing in our model, our estimate for the LD parameter τ^{-1} would tend to be biased upward, and results would tend to depend on the allele distributions at the markers that we happened to study. (Note, however, that, because we consider sharing at locus i, conditional on sharing at *all* loci between i and 0, chance sharing would have much less effect on the results than it does in assessments of LD using only pairs of loci.) To take into account chance sharing of alleles other than the variant, we simply sum the likelihood for the sampled haplotype over all possible breakpoints of the ancestral segment consistent with the data. Recall that the breakpoints of the ancestral segment are the nearest crossovers flanking the variant over the τ meioses occurring between the ancestral haplotype and the sampled haplotype. For instance, if the sampled haplotype matches the ancestral haplotype at loci $-k$ through j but not at locus $-k-1$ and locus $j+1$, then the left breakpoint of the ancestral segment could be anywhere between the variant and locus $-k-1$, and the right breakpoint of the ancestral segment could be anywhere between the variant and locus $j+1$. Thus, we replace the likelihood in equation (2) by

$$L(\boldsymbol{h}_{\mathrm{anc}},\tau^{-1};\boldsymbol{h}_{\mathrm{obs}}) = \sum_{i=0}^{j}\sum_{l=0}^{k} g(\tau^{-1},-l,i) \times P_{\mathrm{null}}[h_{\mathrm{obs}}(i+1),h_{\mathrm{obs}}(i+2),\ldots] \times P_{\mathrm{null}}[h_{\mathrm{obs}}(-l-1),h_{\mathrm{obs}}(-l-2),\ldots]\ . \qquad (3)$$

Allowing for Mutation

A descendant haplotype may vary from its ancestor not only because of the effects of recombination but also because of mutation. We modify likelihood (3) to take into account mutation as well. A rather general model for point mutation could include parameters m_l, giving the mutation rate per meiosis per generation at locus l, and a transition matrix P_l, with ijth entry giving the probability that the mutation is to allele j at locus l, conditional on the mutating allele being allele i at locus l. Allowing a nonzero diagonal for P_l allows different alleles at a locus to have different mutation rates. Under this model, the probability that a τth-generation descendant has allele j at locus l, given that the ancestral haplotype has allele i at locus l and conditional on no crossover events between the variant and locus l during the intervening τ meioses, is

$$m(l,\tau,i,j) = \sum_{k=0}^{\tau} \binom{\tau}{k} m_l^k (1-m_l)^{\tau-k} \left(P_l^k\right)_{(ij)} ,$$

where $(P_l^k)_{(ij)}$ is the ijth element of the kth power of the matrix P_l. For simplicity, in our present implementation we focus on the case when $m_l = m$ for all l and the ijth element of P_l is 0 if $i = j$ and is $1/(n_l - 1)$ otherwise, where n_l is the number of alleles at locus l; that is, the mutation rate is the same across all loci, and, when a mutation occurs, it is equally likely to change to any of the other alleles at that locus. Then

$$m(l,\tau,i,i) = (1 - m\frac{n_l}{n_l - 1})^{\tau} + \left[1 - (1 - m\frac{n_l}{n_l - 1})^{\tau}\right]\frac{1}{n_l} ,$$

and $m(l,\tau,i,j) = [1 - m(l,\tau,i,i)]/(n_l - 1)$, for $i \neq j$. Extension to more-elaborate models is feasible.

Allowing for mutation increases the number of possible intervals in which the breakpoints of the ancestral haplotype (i.e., crossover points flanking the variant after τ meioses) could occur in an observed descendant and still be consistent with the observed data. For instance, in likelihood (3), the right breakpoint of the ancestral haplotype can be assumed to occur somewhere between the variant and locus $j + 1$, because loci $1,\ldots,j$ match the ancestral haplotype by state whereas locus $j + 1$ does not match. However, when mutation is taken into account, it is possible that the right breakpoint of the ancestral haplotype is beyond locus $j + 1$ and that mutation rather than recombination causes locus $j + 1$ not to match the ancestral haplotype. Allowing for mutation, we replace likelihood (3) by

$$\begin{aligned} &L(\boldsymbol{h}_{\text{anc}},\tau^{-1};\boldsymbol{h}_{\text{obs}}) \\ &\quad = \sum_{i=0}^{l_{re}} \sum_{j=0}^{l_{le}} \Big\{ g(\tau^{-1},-j,i) \prod_{k=-j}^{i} m(k,\tau,h_{\text{anc},}(k)h_{\text{obs}}(k)) \\ &\quad\quad \times P_{\text{null}}[h_{\text{obs}}(i+1),h_{\text{obs}}(i+2),\ldots] \\ &\quad\quad \times P_{\text{null}}[h_{\text{obs}}(-j-1),h_{\text{obs}}(-j-2),\ldots]\Big\} , \qquad (4) \end{aligned}$$

where l_{re} and $-l_{\text{le}}$ are the indices of the markers at the right and left edges of the data set, respectively, and where $h_{\text{anc}}(k)$ and $h_{\text{obs}}(k)$ are the alleles at locus k in the ancestral and sampled haplotypes, respectively. Here we assume that the parameter m is known. Alternatively, it could be estimated by maximum likelihood, along with τ^{-1} and the ancestral haplotype.

Allowing for Multiple Origin of the Variant

It is possible that the variant under consideration may have arisen more than once and, thus, may lie on two or more ancestral haplotypes that may be partially preserved in the present sample. Unless the sample is extremely large, it is unlikely that many distinct ancestral haplotypes could be identified and estimated. First, suppose that there is only one ancestral haplotype that can be well estimated but that we wish to allow for some proportion of observed haplotypes that are not descended from this ancestor. We introduce a parameter p to represent the proportion of the variant haplotypes in the population that are not descended from this ancestral haplotype; and $1 - p$ is the proportion of variant haplotypes that are descended from the given ancestor. Then, conditional on a sample haplotype containing the variant, the likelihood of this observation is calculated as

$$(1-p)L(\boldsymbol{h}_{\text{anc}},\tau^{-1};\boldsymbol{h}_{\text{obs}}) + pP_{\text{null}}(\boldsymbol{h}_{\text{obs}}) ,$$

where $L(\boldsymbol{h}_{\text{anc}},\tau^{-1};\boldsymbol{h}_{\text{obs}})$ is given in equation (4) and $P_{\text{null}}(\boldsymbol{h}_{\text{obs}})$ is the probability of the haplotype $\boldsymbol{h}_{\text{obs}}$ estimated from the control population. As above, $P_{\text{null}}(\boldsymbol{h}_{\text{obs}})$ is estimated under a Markov model if control haplotypes are available, or it may be estimated by the product of the allele frequencies in the control population if control haplotypes are not available. When $p = 0$, we obtain equation (4). This model assumes that the proportion p of variants that are nonancestral are on haplotypes randomly chosen from the population. In "Simulation Results for Fine-Scale Genetic Mapping by LD" (see Results, below), we examine the effect on mapping when this model is used to analyze data in which there are actually two distinct ancestral haplotypes. For the cases that we have simulated, the model still seems to perform well.

If, in maximizing the likelihood over the ancestral haplotype, we identify two distinct ancestral haplotypes that both give high likelihoods, we could consider maximizing a likelihood of the form

$$p_1 L(\boldsymbol{h}_{\text{anc}1},\tau_1^{-1};\boldsymbol{h}_{\text{obs}}) + p_2 L(\boldsymbol{h}_{\text{anc}2},\tau_2^{-1};\boldsymbol{h}_{\text{obs}}) \\ +(1-p_1-p_2)P_{\text{null}}(\boldsymbol{h}_{\text{obs}}) \ ,$$

where $L(\boldsymbol{h}_{\text{anc}i},\tau_i^{-1};\boldsymbol{h}_{\text{obs}})$ is the likelihood calculated under ancestral haplotype i and where τ_i^{-1} is the parameter τ^{-1} corresponding to ancestral haplotype i, $i = 1,2$.

Independence of Recombinational Histories: Baum/EM Algorithm

So far, we have discussed only the likelihood for an individual observation. To combine the likelihood across individuals, we need a model for dependence among the individuals. First, consider the case in which variant-containing descendants of the ancestral haplotype are assumed to have independent recombinational histories—that is, the breakpoints of their ancestral blocks are independent. This is equivalent to assuming a star-shaped phylogeny, which corresponds to a limiting case of rapid growth of the variant population, either due to selection or by chance (although it does not necessarily imply rapid growth of the overall population). This could also be seen as a first-order approximation to the likelihood for more-general models involving dependence of haplotypes, which are discussed further in the next subsection, "Dependent Recombinational Histories."

In the case of independence, the likelihoods for the individual observations are simply multiplied to obtain the likelihood for the sample. Rather than summing each individual's likelihood over all possible breakpoints, as in equation (4), and over the possibilities of whether the variant is ancestral or nonancestral, and then multiplying these probabilities, we use the hidden Markov structure of the likelihood and employ the Baum/EM algorithm (Baum 1972) to maximize the likelihood. Below we present the hidden Markov model used to maximize the likelihood, taking into account chance sharing, mutation, and multiple origin of the variant. Modifications of this method allow incorporation of missing data and ambiguously determined haplotypes. The hidden Markov method that we describe assumes that both the location of the variant and the ancestral haplotype are known. To maximize the likelihood over ancestral haplotype and/or location of variant, in addition to the other parameters, we would, in principle, perform the hidden Markov method for every possible choice of ancestral haplotype and location and take the location, ancestral haplotype, τ^{-1}, and p for which the maximized likelihood is highest. In practice, we reduce the number of ancestral haplotypes considered, by using a branch-and-bound algorithm (Appendix B), and we consider a fine grid of possible variant locations.

Consider a discrete-time Markov chain $\{R_l\}$ with state space $\{A,M,N\}$, where A denotes "ancestral," M denotes "mutated," and N denotes "nonancestral." An allele in a haplotype is defined as being in the ancestral state if the entire segment between the site of that allele and the center locus is inherited unbroken by crossovers from the ancestral haplotype and if that site has not mutated to an allele different from that in the ancestral haplotype. Note that sites between that allele and the center locus may have mutated. An allele is defined as being in the mutated state if the entire segment between the site of that allele and the center locus is inherited unbroken by crossovers from the ancestral haplotype and if that site has mutated to an allele that does not match the ancestral haplotype. An allele is defined as being in the nonancestral state if, during the time since the ancestor, at least one crossover has occurred between the site of the allele and the center locus. Let l index the markers. For the moment, we can consider that we have two chains, one indexed by $l = 0,1,2,\ldots,l_{re}$ and one indexed by $l = 0,-1,-2,\ldots,-l_{le}$. As long as the value of the chain at $l = 0$ (variant) is not known, the two chains are dependent. The transition matrix for the right-hand chain (and for the left-hand chain but with $l+1$ replaced by $l-1$ in the latter chain) is

$$a_{(l,l+1)} = \\ \begin{array}{cc} & \begin{array}{ccc} A & M & N \end{array} \\ \begin{array}{c} A \\ M \\ N \end{array} & \left[\begin{array}{ccc} e^{-\tau d_{l,l+1}} m'(l,\tau) & e^{-\tau d_{l,l+1}}[1-m'(l,\tau)] & 1-e^{-\tau d_{l,l+1}} \\ e^{-\tau d_{l,l+1}} m'(l,\tau) & e^{-\tau d_{l,l+1}}[1-m'(l,\tau)] & 1-e^{-\tau d_{l,l+1}} \\ 0 & 0 & 1 \end{array}\right] \end{array} , \tag{5}$$

where $m'(l,\tau) = m(l,\tau,h_{\text{anc}}(l)h_{\text{anc}}(l))$, the chance that an allele has not mutated to a type that does not match the ancestor, conditional on no recombination between it and the center locus. In fact, we can reverse the direction of the left-hand chain and create a single nonhomogeneous Markov chain, $\{R'_l, l = -l_{le},\ldots,-1,0,1,\ldots,l_{re}\}$, with transition matrix (5) holding for $l \geqslant 0$, and with the following transition matrix holding for $l < 0$:

$$a_{(l,l+1)} = \\ \begin{array}{cc} & \begin{array}{ccc} A & M & N \end{array} \\ \begin{array}{c} A \\ M \\ N \end{array} & \left[\begin{array}{ccc} m'(l,\tau) & 1-m'(l,\tau) & 0 \\ m'(l,\tau) & 1-m'(l,\tau) & 0 \\ \beta(l,\tau,p)m'(l,\tau) & \beta(l,\tau,p)[1-m'(l,\tau)] & 1-\beta(l,\tau,p) \end{array}\right] \end{array} ,$$

where $\beta(l,\tau,p) = [(1-p)(e^{\tau x_{l+1}} - e^{\tau x_l})]/[1-(1-p)e^{\tau x_l}]$.

Consider the associated observation sequence $\{O_l, -l_{le} \leqslant l \leqslant l_{re}\}$, where O_l is the observed allele at locus l. If we use a first-order Markov chain to model nonancestral haplotypes, then we have $P\{O_l = j|R_{l-1}, R_l = A, O_{l-1}\} = 1$ if j is ancestral and $= 0$ otherwise; $P\{O_l = j|R_{l-1}, R_l = M, O_{l-1}\} = 1/(n_l - 1)$ if j is nonancestral and we use the mutation model described earlier, $P\{O_l = j|R_{l-1} = A, R_l = N, O_{l-1}\} = f(j)$, and $P\{O_l = j|R_{l-1} = N, R_l = N, O_{l-1} = i\} = f(i|j)$, where $f(j)$ is the control-population frequency of allele j at locus l, and $f(i|j) = f(i,j)/f(j)$, where $f(i,j)$ is the joint frequency of i at locus $l-1$ and j at locus l in the control population. Then $\{R_l, O_l\}$ is Markov. We apply the algorithm of Baum (1972) (an introduction is provided by Rabiner [1989]), which is both a precursor and a special case of the EM algorithm of Dempster et al. (1977). The incorporation of missing haplotype information and ambiguous haplotype determination is a more complicated variant of the above.

The Baum algorithm is easily adapted to provide derivatives of the log-likelihood, from which the observed Fisher information can be calculated to obtain standard errors for τ^{-1} and p. For the mapping problem, recall that, for each possible location in a fine grid, we use the Baum algorithm combined with a branch-and-bound algorithm (Appendix B), to maximize the likelihood over ancestral haplotype, τ^{-1}, and p. Then, to obtain a CI for the location of the variant, we invert the likelihood-ratio test, the same procedure that was followed by Devlin et al. (1996) and Lazzeroni (1998).

Dependent Recombinational Histories

As before, suppose that we have a sample of τth-generation descendant haplotypes. First, assume that only two of the haplotypes, H_1 and H_2, are related more recently than τ. Let t_a ("time apart") denote the time to the most recent common ancestor of the pair. Let $t_t = \tau - t_a$ ("time together"). If time is considered to run backward from the present, then from time 0 to time t_a the two haplotypes in the pair had independent recombinational histories, whereas from time t_a to time τ they had identical recombinational histories.

For $i = 1,2$, let R_i and L_i denote the right and left breakpoints, respectively, of the ancestral block within H_i. Then, under the given model, the joint distribution of (R_1, L_1, R_2, L_2) can be summarized into two cases: (i) with probability $t_t/(t_t + 2t_a)$, $R_1 = R_2$, which is exponentially distributed with rate $t_t + 2t_a$, and (ii) with probability $2t_a/(t_t + 2t_a)$, $\min(R_1,R_2)$ is exponentially distributed with rate $t_t + 2t_a$ and $\max(R_1,R_2) = \min(R_1,R_2) + E$, where E is exponentially distributed with rate τ, independent of $\min(R_1,R_2)$ (see Appendix C). (L_1,L_2) is independent of (R_1,R_2) and has the same distribution. This result can be directly used to compute the likelihood contribution of the dependent pair of haplotypes. It also follows that the correlation between R_1 and R_2, which is the same as both the correlation between L_1 and L_2 and that between $R_1 + L_1$ and $R_2 + L_2$, is

$$\frac{t_t}{t_t + 2t_a} = \frac{\tau - t_a}{\tau + t_a} \ . \qquad (6)$$

In principle, such calculations can be extended to larger numbers of haplotypes of known relationship and can be used to compute the likelihood of the data. However, these calculations very quickly become quite complicated as the number of dependent haplotypes increases. Furthermore, in many cases of interest, the exact relationship between individuals in the sample either is not known—for example, if "unrelated" individuals are sampled—or may be exceedingly complex—for example, in an inbred isolate. Then, the approach that we take is to obtain covariances between observations, on the basis of the available model for the ancestry of the sample, and to maximize the quasi likelihood (Wedderburn 1974) in the complete-data case, which we generalize to a quasi-score estimating equation in the incomplete-data case. We focus on the case when "unrelated" individuals are sampled and an exchangeable population model is assumed. Then, as shown below, the maximum quasi-likelihood estimator is the same as the maximum-likelihood estimator for the case of independence, but with the standard error inflated by a factor depending on the correlation.

Quasi-Score Estimating Equation

When observations are dependent, we use a quasi-score estimating equation as a way to obtain some of the desirable properties of maximum-likelihood estimation but use only the marginal likelihoods and the covariances between observations. We first introduce the quasi-score function for the complete-data case, then show the extension to the estimating equation that we use in the incomplete-data case. Finally, we focus on the case of "unrelated" individuals and assume an exchangeable population model for them.

In the simplest case of complete data with dense markers, the vector of observed lengths of ancestral segments $s = (s_1,\ldots,s_n)^T$ is a sufficient statistic for τ^{-1}, and the marginal distribution of each length is gamma$(2,\tau)$. As a consequence, if we let $\mu = 2/\tau$, the mean length of the ancestral segment, and $v(\mu) = \mu^2/2$, the variance of the length of the ancestral segment, then the derivative of the log-likelihood l with respect to μ, called the "score function," for a single observation is given by $\frac{\partial l}{\partial \mu}(\mu; s) = (s - \mu)/v(\mu)$. In the case of independent recombinational histories, the score function for a sample size of n haplotypes would be $\frac{\partial l}{\partial \mu}(\mu; \mathbf{s}) = \mathbf{1}^T(\mathbf{s} - \mu\mathbf{1})/v(\mu)$, where $\mathbf{1}$ is a column vector of 1's of length n. When the ancestral-segment lengths $\mathbf{s}$ are dependent with covariance matrix C_μ having (i,j)th entry equal to $\text{cov}_\mu(s_i,s_j)$, where the subscript μ indicates possible dependence on μ, then the quasi-score function is given by $Q(\mu; \mathbf{s}) = \mathbf{1}^T C_\mu^{-1}(\mathbf{s} - \mu\mathbf{1})$. This is the quasi-likelihood score function defined by Wedderburn (1974). When the observations are uncorrelated, the likelihood score function is obtained as before. The quasi-score function may be set equal to 0 and solved to obtain an estimator of μ (and, hence, of τ^{-1}). This estimator has many of the same desirable properties as a maximum-likelihood estimator. Its properties generally include approximate unbiasedness and asymptotic normality, with asymptotic standard error $(\mathbf{1}^T C_\mu^{-1}\mathbf{1})^{-\frac{1}{2}}$ (McCullagh and Nelder [1989] provide a detailed discussion).

With real data—that is, when observations are made on nondense markers with chance sharing possible—we extend the quasi-score estimating equation to the more general formulation $Q'(\mu; \text{data}) = \mathbf{1}^T K_\mu^{-1} \frac{\partial l}{\partial \mu}(\mu; \text{data})$, where K is the correlation matrix of $\frac{\partial l}{\partial \mu}(\mu; \text{ data})$; that is, Q' is a weighted sum of score functions across the individual haplotypes, where the weights are given by the inverse correlation matrix of the score functions. This is a special case of an estimating equation discussed by Lindsay (1988). In the complete-data case, this reduces to the previous formulation. The arguments for approximate unbiasedness and asymptotic normality, with asymptotic standard error $\{\mathbf{1}^T[K_\mu \text{var}(\dot{l})]^{-1}\mathbf{1}\}^{-\frac{1}{2}}$, where $\text{var}(\dot{l})$ is the variance of the score function for a single haplotype, follow the same lines as those for the quasi likelihood, but they depend crucially on the fact that the variances of the score functions for individual haplotypes, $\text{var}(\dot{l})$, are equal across observations.

The framework that has been discussed above is very general, and we now focus more specifically on its application in case the matrix C_μ (complete data) or K_μ (incomplete data) has all nondiagonal entries equal. In the complete-data case, this assumption is that the covariance between any pair of observations is assumed to be the same, say $(2/\mu^2)c_\mu > 0$, where c_μ is the correlation. This would occur when there is no information on the relatedness of the sampled haplotypes, beyond an assumed population model such as the coalescent model. Then $Q(\mu; \mathbf{s}) = [1 + (n - 1)c_\mu]^{-1}(\mu^2/2)\mathbf{1}^T(\mathbf{s} - \mu\mathbf{1})$. Solving $Q(\mu; \mathbf{s}) = 0$ is equivalent to solving $\mathbf{1}^T(\mathbf{s} - \mu\mathbf{1}) = 0$ (i.e., equivalent to maximizing the likelihood under independence). Thus, in the complete-data case with all covariances equal, the quasi-likelihood estimator is equal to the maximum-likelihood estimator in the case of independence, but the standard error of the estimator is inflated by the factor $\sqrt{1 + (n - 1)c_\mu}$. With incomplete data, assume that K_μ has all nondiagonal entries equal to k_μ. Then, $Q'(\mu; \text{data}) = [1 + (n - 1)k_\mu]^{-1}\mathbf{1}^T \frac{\partial l}{\partial \mu}(\mu; \text{data})$. Thus, solving $Q'(\mu; \text{data}) = 0$ is equivalent to solving $\mathbf{1}^T \frac{\partial l}{\partial \mu}(\mu; \text{data}) = 0$ (i.e., equivalent to maximizing the likelihood under independence). Thus, with incomplete data also, the quasi-score estimator equals the maximum-likelihood estimator in the case of independence, with the standard error of the estimator inflated by the factor $\sqrt{1 + (n - 1)k_\mu}$.

Example of the Conditional-Coalescent Model

We consider the case in which the ancestral process of the observed haplotypes is modeled by a coalescent conditional on time τ to the most recent common ancestor. To calculate the correlation between individual observations under this model, we use the result, given in equation (6), that the correlation is equal to $(\tau - t_a)/(\tau + t_a)$. With probability $2(n + 1)/[(n - 1)(n - j + 1)(n - j + 2)]$, a random pair will coalesce at the jth coalescent time, where time proceeds backwards. In that case, t_a would be equal to the jth coalescent time. The final ingredient required for the calculation is the distribution of the jth coalescent time, conditional on time τ to the most recent common ancestor; this is given in Appendix D, along with an expression for the correlation calculated under this model. In practice, we use an approximation to the correlation that is valid when τ is small relative to twice the effective population size, which is often taken to be 2×10^4. (In the EPM1 data set of Virtaneva et al. [1996], we estimate τ to be ~34; in the CF data set of Kerem et al. [1989], we estimate τ to be ~105.) This approximate correlation c_n, which is given in Appendix D, does not depend on the value of τ, although it does depend on the sample size n. With real data sets—that is, with incomplete data—the value required is k, the correlation between the score functions for two individuals, rather than c, the correlation between the segment lengths for the two individuals. Although, in principle, k could be obtained by simulation, we instead use the complete-data approximation $k \approx c$. Thus, using the quasi-likelihood estimating equation to estimate τ^{-1} under the coalescent model, we obtain the same estimate as was obtained in the independence case, but with the standard error inflated by the factor $\sqrt{1 + (n - 1)c_n}$. Similarly, the log-likelihood that we use is the same as that in the independence case but is multiplied by the factor $[1 + (n - 1)c_n]^{-1}$. Using these adjusted standard errors and log-likelihoods, we obtain CIs for τ^{-1} and location as described above for the independence case. Simulation results indicate that, in practice, this procedure gives CIs with essentially correct coverage for both τ^{-1} and location, under the conditional-coalescent model (see “Simulation Results for Assessment of Magnitude of LD around a Variant” and “Simulation Results for Fine-Scale Genetic Mapping by LD,” below).

Results

Simulation Studies

To evaluate the application of the decay of haplotype sharing (DHS) method to solution of both problem 1, assessment of the magnitude of LD around a variant, and problem 2, fine-scale genetic mapping by LD, we perform simulation studies. Several scenarios are simulated, including those of (*a*) a single ancestral haplotype with all individuals descended from it, (*b*) a single ancestral haplotype with some individuals not descended from it and instead having randomly generated haplotypes, and (*c*) two distinct ancestral haplotypes each having descendants in the sample. Each ancestral haplotype is generated randomly, and the time τ to the ancestor and the number of descendants of the ancestor in the sample are fixed. Simulations are performed under the independence assumption and under the conditional-coalescent assumption for the relationship among the descendants of the ancestor. When independence is assumed, mutations and recombinations are simulated for each descendant independently. When the conditional coalescent is assumed, a tree is simulated according to that distribution (see Appendix D), with mutations and recombinations generated along each branch.

We perform the simulations under somewhat unfavorable conditions—that is, low sample size and low marker density—compared with what is available in the two data sets that we analyze below. Most of our simulations involve sample sizes of 50 or 63 haplotypes, whereas the EPM1 data set of Virtaneva et al. (1996) has a sample size of 88 and the CF data set of Kerem et al. (1989) has a sample size of 94. In our simulations, the true τ is 100 generations (i.e., $\tau^{-1} = 1$ cM), and we consider the cases of a 0.5-cM map of microsatellites with heterozygosity .85 and a 0.2-cM map of biallelic markers with allele frequencies .7 and .3. One relevant quantity for comparison of marker resolution across data sets is $1 - \exp[-\tau(\text{intermarker distance, in Morgans})]$. This quantity is the expected proportion of haplotypes that experience a crossover between a given pair of adjacent loci. If this number is too high, then sharing by descent will drop off so rapidly between markers that there will not be sufficient resolution to estimate τ^{-1}. For our simulated microsatellite map this quantity is .39, whereas for our biallelic map it is .18. For comparison, the data set of Kerem et al. (1989) has biallelic markers with average distance between them of only 80 kb, which we convert to 0.0008 Morgans, and an estimated τ of 105, so $1 - \exp[-\tau(\text{intermarker distance})] = .08$. In the data set of Virtaneva et al. (1996), microsatellite markers are used with average intermarker distance of 224 kb and an estimated τ of 34, giving $1 - \exp[-\tau(\text{intermarker distance})] = .07$. Thus, the marker resolution in our simulations is inferior to that in the data sets.

Simulation Results for Assessment of Magnitude of LD around a Variant

One purpose of the first set of simulations (table 1) is to evaluate the effect that ancestral-haplotype reconstruction has on the estimate of τ^{-1}. Recall that we maximize the likelihood over the ancestral haplotype but that, for estimation of τ^{-1}, we estimate an ancestral haplotype only out to a distance from the variant at which the expected number of individuals still sharing the ancestral haplotype, conditional on the data, drops to <5 or <25% of the total sample size, whichever is greater, as described in “Estimation of the Ancestral Haplotype” (see Methods section above) and Appendix B. We wish to examine whether the estimation of an ancestral haplotype causes any serious bias in the estimation of τ^{-1}, or whether the uncertainty in that estimation can lead to undercoverage of the 95% CIs. For this reason, in table 1 we simulate haplotypes spanning a greater genetic distance than is spanned in either of the data sets, to ensure that a drop to only 25% of individuals sharing is achieved. This should accentuate any bias over the case in which the drop in sharing to the end of the haplotype is less severe. Data are analyzed under the same model (independence or conditional coalescent) that is used to generate them. The mutation rate and heterogeneity parameter are assumed to be 0, and the location of the variant is assumed to be known. For each

Table 1

Performance of DHS for Estimation of LD, When True Value of τ^{-1} = 1, p = 0, m = 0, Location of Variant Is Known, and Ancestral Haplotype Is Unknown

	Mean Estimate of τ^{-1} (Coverage of 95% CI)	
Sample Size and Map[a]	Independence	Coalescent
25:		
Biallelic	1.0 (.95)	1.1 (.96)
Microsatellite	1.0 (.95)	1.1 (.98)
50:		
Biallelic	1.0 (.95)	1.0 (.95)
Microsatellite	1.0 (.95)	1.0 (.96)
75:		
Biallelic	1.0 (.95)	1.0 (.95)
Microsatellite	1.0 (.95)	1.0 (.95)

[a] Biallelic markers have allele frequencies of 70% and 30% and 0.2-cM spacing; microsatellite markers have heterozygosity of 85% and 0.5-cM spacing.

case, 3,000 realizations are generated. For the case of independence, there is no detectable bias in the simulations, and the 95% CI covers the true value 95% of the time, even for a sample of size 25 (table 1). Under the conditional-coalescent model, for a sample size of 25, there is a slight bias and slight conservativeness of the CI, but they disappear by sample size 50 (table 1). Furthermore, we find that, even for small sample sizes, the sampling distribution of the estimated LD value $\hat{\tau}^{-1}$ is well approximated by the normal distribution (results not shown), for both the independence and conditional-coalescent cases.

In the second set of simulations (table 2), we consider performance of the DHS method for assessment of LD when there are mutations and when there are some haplotypes in the sample that are not descended from the ancestor. For the biallelic loci we assume a mutation rate of 5×10^{-7}, whereas for the microsatellites we assume a much higher mutation rate of 2×10^{-3}, where these rates are per locus per meiosis. For most of the simulations, we use a sample of 50 haplotypes descended from a common ancestor, with either 0 or 13 additional nonancestral haplotypes. In each case, the 50 haplotypes are simulated under either the independence or the conditional-coalescent model, with recombinations and mutations superimposed, whereas the nonancestral haplotypes are generated randomly and independently, with their distribution determined by the marker-allele frequencies. In each case, τ^{-1} and the ancestral haplotype are estimated. In the cases when nonancestral haplotypes are present, p is also estimated. The location of the variant and the mutation rate are assumed to be known. Data are analyzed under the same model (independence or conditional coalescent) that is used to generate them. For each case, 2,000 realizations are generated.

In our simulations, we find that estimation assuming no mutation and estimation assuming a mutation rate of 5×10^{-7} produce virtually identical results and that introduction of mutations into the data at such a low rate has almost no effect; however, this is no longer true at the higher rate of 2×10^{-3}. Table 2 shows that coverage of the 95% CI is quite good in all cases, with only slight overcoverage in the conditional-coalescent case. Inclusion of 20% nonancestral haplotypes does not appear to cause bias or to have much effect on the coverage. With the mutation rate of 2×10^{-3} and the conditional-coalescent model, the DHS method gives a slightly biased estimate of τ^{-1} with only 50 descendant haplotypes, but this slight bias disappears with 75 descendant haplotypes (see table 2).

Simulation Results for Fine-Scale Genetic Mapping by LD

We assess the performance of the DHS method for fine-scale genetic mapping in cases in which the ancestral haplotype is known and in cases in which it is unknown. When both the ancestral haplotype and the location of the variant are unknown, the DHS method becomes somewhat computationally intensive, especially with very polymorphic markers. Thus, we have performed only a small set of simulations of that case, limited to biallelic loci. Mapping simulations for the case when the ancestral haplotype is known can be performed much more readily. We have used such simulations to investigate a wider range of questions, including how mutations affect mapping using microsatellite markers, and also to identify, for biallelic loci, the scenarios that are the most challenging for the DHS method; these latter—namely, the conditional-coalescent case with either nonancestral haplotypes or multiple ancestors—are the scenarios that we have chosen to investigate in the more realistic simulations with ancestral haplotype unknown.

Table 3 shows the mapping results when independence holds and the ancestral haplotype is known; table 4 shows the same set of simulations for the conditional-coalescent model. In both tables, 1,000 realizations are generated for each case. Coverage of the 95% CI for location is very close to 95% in all cases. The median lengths of CIs for location in the independence case are

Table 2

Performance of DHS for Estimation of LD with Mutations and Nonancestral Haplotypes, When True Value of τ^{-1} = 1, Location of Variant Is Known, and Ancestral Haplotype Is Unknown

	No. (%) of Haplotypes				Mean Estimate of τ^{-1} (Coverage of 95% CI)	
m	Ancestral	Nonancestral	p Estimated?	Map[a]	Independence	Coalescent
5×10^{-7}	50	0	No	Biallelic	1.0 (.95)	1.0 (.95)
5×10^{-7}	50	13 (20.6)	Yes	Biallelic	1.0 (.95)	1.0 (.93)
2×10^{-3}	50	0	No	Microsatellite	1.0 (.95)	1.1 (.96)
2×10^{-3}	50	13 (20.6)	Yes	Microsatellite	1.0 (.94)	1.1 (.98)
2×10^{-3}	75	0	No	Microsatellite	1.0 (.95)	1.0 (.95)
2×10^{-3}	75	19 (20.2)	Yes	Microsatellite	1.0 (.95)	1.0 (.97)

[a] See footnote to table 1.

Table 3

Performance of DHS for Fine Mapping with Mutations and Nonancestral Haplotypes, Where True Value of τ^{-1} = 1, Independence Holds, Ancestral Haplotype Is Known, and Location and τ^{-1} Are Unknown, 50 Ancestral Haplotypes

Mutation Rate					95% CI for Location		Proportion of Location Estimates		
True	Assumed	No. (%) of Nonancestral Haplotypes	p Estimated?	Map[a]	Coverage	Median Length	In Correct Interval	≤1 Interval off Correct Interval	Mean Estimate of τ^{-1}
0	0	0	No	Biallelic	.94	.16 cM	.92	.997	1.0
0	0	0	No	Microsatellite	.96	.20 cM	1.00	1.00	1.0
0	0	13 (20.6%)	Yes	Biallelic	.94	.23 cM	.84	.996	1.0
0	0	13 (20.6%)	Yes	Microsatellite	.95	.25 cM	.999	1.00	1.0
5×10^{-7}	5×10^{-7}	13 (20.6%)	Yes	Biallelic	.95	.24 cM	.85	.996	1.0
2×10^{-3}	2×10^{-3}	13 (20.6%)	Yes	Microsatellite	.94	.30 cM	.98	1.00	1.0
1×10^{-3}	2×10^{-3}	13 (20.6%)	Yes	Microsatellite	.95	.30 cM	.996	1.00	1.1
1×10^{-3}	5×10^{-4}	13 (20.6%)	Yes	Microsatellite	.95	.25 cM	.99	1.00	.9

[a] See footnote to table 1.

quite small, 0.16–0.3 cM. They are substantially larger for the conditional-coalescent model, 0.35–0.68 cM. Similarly, the proportion of location estimates that are in the correct marker interval is different under the two models. For the 0.5-cM microsatellite map, the location estimates are in the correct marker interval virtually 100% of the time under independence, versus 86%–96% of the time under the conditional-coalescent model. The biallelic marker map that we used has much smaller marker intervals, 0.2-cM, and in that case the location estimates are in the correct marker interval 85%–90% of the time under the independence model, versus 62%–71% of the time under the conditional-coalescent model. In nearly all cases, the estimated location is no more than one marker interval off the correct interval. When the mutation rate either is correctly specified in the model or is extremely low, the estimate of τ^{-1} taken at the maximum-likelihood estimate of location has no apparent bias. We also consider the case when mutation is incorporated into the model but the mutation rate is misspecified by a factor of 2, either too high or too low; this misspecification does not cause problems in the mapping results, but it does result in a bias in the estimate of τ^{-1} taken at the maximum-likelihood estimate of location. As would be expected since presence of mutation tends to decrease haplotype sharing, if the assumed mutation rate is too high the estimate of τ^{-1} is biased upward, and if the assumed mutation rate is too low the estimate of τ^{-1} is biased downward. These results suggest that microsatellite markers or other markers with a high mutation rate can be successfully used in LD mapping studies even if the mutation rate is slightly misspecified; however, they also suggest that estimation of τ^{-1} (or of time to the most recent common ancestor of the sample) is not reliable, even in the case of a relatively recent variant, unless the mutation rate either is correctly specified in the model or is extremely low. We did not try to estimate mutation rate in these simulations, but that is also a possibility; however, the true mutation rate is likely to vary across loci, so, unless the rate could be well estimated for each locus, there would still be mutation-rate misspecification.

In table 5 we give results of the mapping simulations when location, ancestral haplotype, τ^{-1}, and p are all unknown. Because of computational limitations, these simulations are limited to the case of biallelic loci. We consider only the conditional-coalescent model, since the results under the independence model are always more favorable. For each case, 100 realizations are generated. Since this number is small, we report standard errors for the coverage, for the proportion of location estimates in the correct interval, and for the proportion ≤1 interval off. For the case of 50 ancestral haplotypes and 13 random nonancestral haplotypes, we can compare the first line of table 5 with the third line of table 4, to see that there is little difference in the results when the ancestral haplotype is known versus when it is estimated. This gives us some confidence that the other results in tables 3 and 4 may extrapolate readily to the case of unknown ancestral haplotype. In table 5 we also consider the case in which the sample consists of descendant haplotypes of two distinct ancestral haplotypes. In each case, we simulated 50 individuals under the coalescent model conditional on a time of 100 generations to the most recent common ancestor and simulated 13 individuals from another conditional-coalescent process, independent of the first, conditional on a time of 100 generations (line 2 of table 5), 133 generations (line 3 of table 5), or 80 generations (line 4 of table 5) to the most recent common ancestor. The DHS method used is the version in which only one ancestral haplotype is estimated and haplotypes not descended from it are assumed to be randomly drawn and independent; nonetheless, the method still seems to perform quite well in this case.

Table 4

Performance of DHS for Fine Mapping with Mutations and Nonancestral Haplotypes, Where True Value of τ^{-1} = 1, Conditional-Coalescent Model Holds, Ancestral Haplotype Is Known, and Location and τ^{-1} Are Unknown, 50 Ancestral Haplotypes

Mutation Rate					95% CI for Location		Proportion of Location Estimates		
True	Assumed	No. (%) of Nonancestral Haplotypes	p Estimated?	Map[a]	Coverage	Median Length	In Correct Interval	≤1 Interval off Correct Interval	Mean Estimate of τ^{-1}
0	0	0	No	Biallelic	.95	.36 cM	.71	.97	1.0
0	0	0	No	Microsatellite	.93	.35 cM	.96	1.00	1.0
0	0	13 (20.6%)	Yes	Biallelic	.97	.52 cM	.62	.96	1.0
0	0	13 (20.6%)	Yes	Microsatellite	.94	.45 cM	.94	1.00	1.0
5×10^{-7}	5×10^{-7}	13 (20.6%)	Yes	Biallelic	.96	.52 cM	.64	.96	1.0
2×10^{-3}	2×10^{-3}	13 (20.6%)	Yes	Microsatellite	.94	.68 cM	.86	1.00	1.0
1×10^{-3}	2×10^{-3}	13 (20.6%)	Yes	Microsatellite	.95	.65 cM	.89	1.00	1.1
1×10^{-3}	5×10^{-4}	13 (20.6%)	Yes	Microsatellite	.95	.53 cM	.91	.999	.9

[a] See footnote to table 1.

There is some suggestion of undercoverage for the last case in table 5, but the standard error is high. The lengths of the CIs and proportions maximizing in the correct interval and no more than one interval off are very similar for the cases of one and two ancestral haplotypes.

Analysis of EPM1 Data Set of Virtaneva et al. (1996)

Virtaneva et al. (1996) report refinement of the location of the EPM1 gene to a 175-kb interval in distal 21q, on the basis of data that include 88 five-locus haplotypes spanning a 900-kb region (D21S1885-D21S2040-D21S1259-D21S1912-PFKL) sampled from affected individuals in Finland. Pennacchio et al. (1996) report cloning of the EPM1 gene, which is found to be between D21S2040 and D21S1259, ~30 kb from D21S2040. We apply the DHS method to the data set of Virtaneva et al. (1996), to do multipoint mapping of the EPM1 gene and to estimate the parameter τ^{-1}, which is the expected genetic distance out to which the ancestral haplotype is preserved, or, equivalently, 1/(time in generations to the ancestral haplotype).

Figure 1 shows the log-likelihood curve for multipoint LD mapping when the heterogeneity parameter p is estimated and the mutation rate m is assumed to be 5×10^{-4}. The likelihood maximizes in the correct marker interval, between D21S2040 and D21S1259. The 95% CI based on independence (fig. 1, *dotted horizontal bar*) does not contain the true gene (*unbroken vertical bar*), whereas the 95% CI based on the conditional-coalescent model (*unbroken horizontal bar*) does. This suggests that the independence model does not fit the data. The estimated heterogeneity parameter is not significantly different from zero, so we are led to fit the model with $p = 0$ and $m = 5 \times 10^{-4}$. This log-likelihood plot is shown in figure 2. In that case, both 95% CIs contain the true gene location, and again the likelihood maximizes in the correct marker interval, this time even closer to the gene. For comparison, the method of Terwilliger (1995), as applied by Virtaneva et al. (1996), maximizes in the wrong interval, although a 95% CI obtained by inverting the likelihood-ratio test does contain the true gene location. On the basis of the information in figure 3 of Xiong and Guo (1997), their method maximizes ~0.40 cM from D21S1885, a position identical to our estimate when $p = 0$ and $m = 5 \times 10^{-4}$, and a 99% CI obtained by inverting the likelihood-ratio test appears to contain the true gene location, but a 95% CI obtained by the same method would not contain the true gene location.

In the data set of Virtaneva et al. (1996), 65 (74%) of the 88 haplotypes are identical, so there has been very little decay of haplotype sharing across the data set. For the problem of accurately estimating τ^{-1}, it would be advantageous to have longer haplotypes, so that more decay of sharing could be observed. (This is not necessary for the mapping problem, in which it is the comparison of log-likelihoods across locations that is relevant, rather than the estimated value of τ^{-1}.) Furthermore, note that this data set involves microsatellite markers. Our simulations indicate that, when the mutation rate is high, such as when microsatellites are used, estimation of τ^{-1} can be sensitive to assumptions about mutation rate, whereas LD mapping is very robust to such assumptions. For these reasons, the data set of Virtaneva et al. (1996) does not provide a lot of information for estimation of τ^{-1}, although it is very useful for mapping. When the mutation rate is assumed to be $m = 5 \times 10^{-4}$, we estimate τ^{-1} to be 2.9 cM, with a 95% CI of 0.48–5.40 cM, under the conditional-coalescent model. This gives an estimate of 34 generations to the most recent common ancestor of the sample, with a 95% CI of 19–207 generations, under the conditional-coalescent model. If m is assumed to be 0, then the corresponding estimates are 2.1 cM, with a 95% CI of

0.35–3.9 cM, under the conditional-coalescent model, or, equivalently, 47 generations, with a CI of 26–286 generations. For the same data set, Xiong and Guo (1997) report an estimate of ~70 generations for the age of the mutation, which is consistent with our CI. In fact, Xiong and Guo's (1997) method, like ours, does not actually estimate the age of the mutation in this case but, rather, estimates the time to the most recent common ancestor of the sample, which will, in general, be less than the actual age of the mutation. The population of Finland, from which the sampled haplotypes are drawn, is believed to be largely descended from a small founder population whose expansion began 2,000–2,500 years ago (Nevanlinna 1972; Norio 1981; de la Chapelle 1993). This suggests a rough guess of 100 generations for the age of the mutation, which is consistent with our estimate and CI for the time to the most recent common ancestor of the sample.

Analysis of CF Data Set of Kerem et al. (1989)

Kerem et al. (1989) identify the Δ508 mutation at the CF gene on chromosome 7, which is believed to be responsible for >70% of CF cases among whites. Data published by Kerem et al. (1989) include 94 CF haplotypes, 63 of which contain the Δ508 mutation, and 92 normal haplotypes. Each haplotype consists of 23 biallelic markers within a 2-Mb region covering the gene. Because normal haplotypes and not just allele frequencies are available, we apply a Markov-chain model to calculate the likelihood contribution of nonancestral portions of haplotypes. All physical distances are converted to genetic distances by use of the equivalence 1 cM ≈ 1 Mb. Figure 3 gives the resulting log-likelihood curve when p is estimated. Assumed mutation rates of 0 and 5×10^{-7} give indistinguishable results. The estimated ancestral haplotype is the same across all choices of location of the variant, as was also true for the data set of Virtaneva et al. (1996). In this example, both the 95% CI based on independence and that based on the conditional-coalescent model contain the true gene location. At the estimated location, the heterogeneity parameter is .21, with an estimated standard error of .05, indicating significant heterogeneity. A likelihood-ratio test comparing the likelihood maximized over location, ancestral haplotype, and τ^{-1} when $p = 0$, versus the likelihood maximized over the these parameters plus p also indicates that the model with p set equal to 0 demonstrates very severe model misfit (results not shown). A similar conclusion was reached by Devlin et al. (1996).

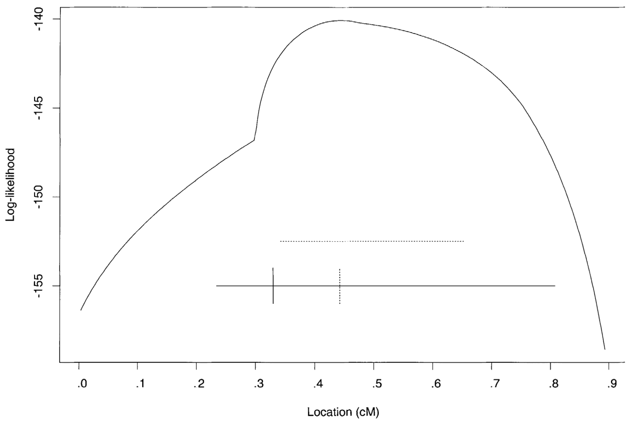

Figure 1 Log-likelihood versus location, for EPM1 data set of Virtaneva et al. (1996), with the likelihood maximized over ancestral haplotype, τ^{-1} and heterogeneity parameter p, and with the mutation rate fixed at 5×10^{-4}. Location is given as distance from D21S1885. The unbroken vertical line is the true variant location, the dotted vertical line is the estimated 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.

In comparison with our results, the method of Terwilliger (1995) maximizes at position 0.77 on our plot, but on the basis of his figure 4, a 99% CI obtained by inverting the likelihood-ratio test would not contain the true variant location, as has been pointed out by Devlin et al. (1996). The method of Devlin et al. (1996) maximizes at 0.81, with a 99% search interval containing the variant location, although a 95% search interval would not contain the variant. (Note that Devlin et al.

Table 5

Performance of DHS for Fine Mapping in the Presence of Multiple Ancestral Haplotypes; Location, Ancestral Haplotype, τ^{-1}, and p Are All Unknown, and Conditional-Coalescent Model Holds, .2-cM Map of Biallelic Markers with Allele Frequencies of 70% and 30%

	No. (%) of Haplotypes					95% CI for Location		Proportion (SD) of Location Estimates	
m	From Ancestor 1	From Ancestor 2	Random Nonancestral	True τ_1^{-1}	True τ_2^{-1}	Coverage (SD)	Median Length	In Correct Interval	≤1 Interval off Correct Interval
5×10^{-7}	50	0	13 (20.6%)	1	…	.98 (.01)	.53 cM	.6 (.05)	.95 (.02)
5×10^{-7}	50	13 (20.6%)	0	1	1	.95 (.02)	.48 cM	.5 (.05)	.97 (.02)
5×10^{-7}	50	13 (20.6%)	0	1	.75	.94 (.02)	.53 cM	.5 (.05)	.93 (.03)
5×10^{-7}	50	13 (20.6%)	0	1	1.25	.90 (.03)	.44 cM	.6 (.05)	.94 (.02)

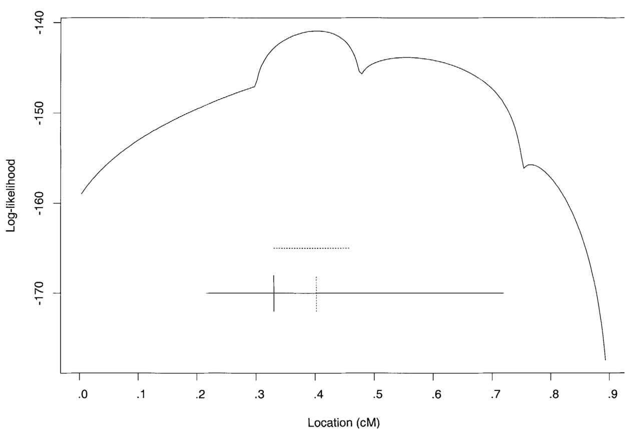

Figure 2 Log-likelihood versus location, for EPM1 data set of Virtaneva et al. (1996), with the likelihood maximized over ancestral haplotype and τ^{-1}, with heterogeneity parameter fixed at 0 and mutation rate fixed at 5×10^{-4}. Location is given as distance from D21S1885. The unbroken vertical line is the true variant location, the dotted vertical line is the estimated 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.

do not claim 95% coverage for their 95% search interval.) The method of Xiong and Guo (1997) maximizes at 0.8 cM, but, on the basis of their figure 1, the drop in log-likelihood from the maximum of their curve to the true gene location is ~100, which excludes the true gene location, $P \ll 10^{-10}$. Lazzeroni (1998) obtains several possible estimates for the gene location, with the asymmetric piecewise linear version giving an estimated location of 0.89, with a 95% CI containing the true location.

For estimation of LD around the Δ508 mutation, or, equivalently, the time to the most recent common ancestor of the Δ508 mutation in the sample, we make use of the additional information of which of the haplotypes actually contain this mutation (63 of the 94 haplotypes in the sample of affected individuals contain the mutation). We estimate τ^{-1} to be 0.95 cM, with a 95% CI of 0.44–1.46 cM under the conditional-coalescent assumption and a 95% CI of 0.70–1.21 cM under independence. This τ^{-1} would correspond to a time of 105 generations, with a CI of 69–225 generations under the conditional-coalescent assumption and a CI of 83–144 generations under independence. Devlin et al. (1996) obtain an estimate of 96 generations, which is in close agreement with our estimate.

Discussion

We have developed and implemented a new method for assessment of LD—the DHS method—which is designed to use multilocus haplotypes containing a particular variant, rather than to operate on pairs of loci. Dependence among loci within a haplotype is explicitly modeled, and dependence due to population structure is taken into account with a quasi-score estimating equation that uses approximate correlations from a conditional-coalescent model. Loci may be multiallelic, and the method allows for multiple origin of variants, chance sharing of alleles, mutations, ambiguous haplotype information, and missing data. This framework can be applied either to the problem of assessment and comparison of levels of LD around different variants or to the problem of multipoint fine-scale LD mapping.

If the ancestral haplotype were known, then our estimate of LD would be equivalent to 1/(age of the variant). When the ancestral haplotype is estimated, our estimate of LD is better interpreted as being 1/(time to the most recent common ancestor of the variant-containing haplotypes in the sample). The time to the most recent common ancestor of the variant-containing haplotypes in the sample will generally be less than the age of the variant, even if the entire population of variant-containing haplotypes is sampled. A similar interpretation holds for the times estimated by Devlin et al. (1996) and Xiong and Guo (1997), although they do not explicitly draw this distinction between age of the variant and time to the most recent common ancestor.

Simulation results show that the method performs very well, providing low bias, small CIs, and accurate coverage for estimation of τ^{-1} and for LD mapping. Although we estimate ancestral haplotype as a parameter in the model, simulations indicate that this does not introduce much bias into the estimation of the LD mea-

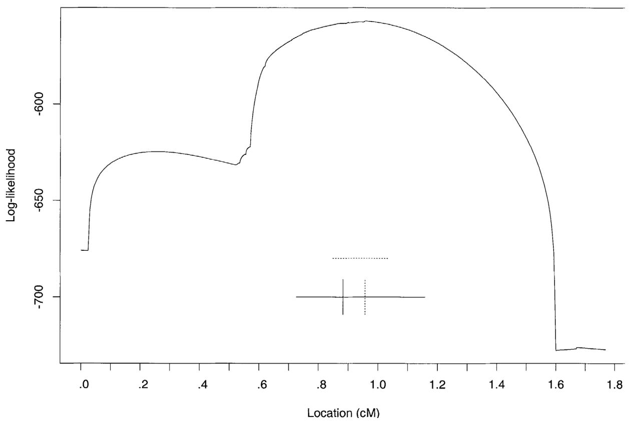

Figure 3 Log-likelihood versus location, for CF data set of Kerem et al. (1989), with the likelihood maximized over ancestral haplotype, τ^{-1}, and heterogeneity parameter p. Location is given as distance from D21S1885. The unbroken vertical line is the true variant location, the dotted vertical line is the estimated 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.

sure τ^{-1}, even for small sample sizes and with some haplotypes nonancestral. We also consider the effect that the presence of descendants of multiple ancestral haplotypes in the sample has on LD mapping. The mapping procedure still works extremely well. (It is not clear what the "correct" value of τ^{-1} should be in this case, since, presumably, each ancestral haplotype has its own value, so we did not study the effect that multiple ancestral haplotypes have on the estimation of τ^{-1}.)

For the mapping problem, the simulations show that the conditional-coalescent CI is much larger than that under independence, and yet each has the correct coverage for its particular case. Not surprisingly, the CI based on independence gives severe undercoverage in the conditional-coalescent case (results not shown). This indicates that it is important to take into account population structure, but the difficulty is that the appropriate model is not known. The coalescent model applies when the population size remains constant over time; thus it is reasonable to expect that CIs under this model will be conservative when there is population expansion, a case that could be regarded as intermediate between the coalescent and star-shaped/independence model. Such population expansion seems likely to be the case for humans.

When mutation is taken into account in the model, slight misspecification of the mutation rate has little adverse effect on mapping. A case that we did not consider in our simulations—but that could have a greater effect—would be the case when one marker near the variant has a mutation rate much higher than those of all the others. One possible approach to detection of such a scenario would be to perform the analysis while leaving out one marker at a time, and to see if there is a noticeable change in the results. Estimation of τ^{-1} is more sensitive to mutation rate than is mapping. When the mutation rate is high and is misspecified in the model, it leads to biased estimation of τ^{-1}. Thus, microsatellite markers whose mutation rates are not well known may still be very useful for LD mapping, but they are not ideal if one is most interested in estimating either τ^{-1} or the time to the most recent common ancestor.

The model that we have implemented is based on the assumption of no interference. In Appendix A, we describe how to include interference in the model. We argue that the assumption of no interference even though interference is known to exist is of little consequence here. This is so because the genetic distances at which LD around a variant would be detected are small enough that multiple crossover events occurring there would be extremely unlikely, even under no interference.

When control haplotypes are available, we apply a simple Markov model to describe nonancestral haplotypes. When only allele frequencies are available, there is no alternative but to assume an independence model for alleles in nonancestral haplotypes. In practice, this seems to have relatively little effect on estimates of τ^{-1} but seems to have a greater effect on the likelihood. Thus, it is probably more important to have control haplotypes for the mapping problem than for the problem of the assessment of LD around a variant.

The framework that we have described can, in principle, be extended to apply to individuals of known relationship, in which case the correlation between score functions for different pairs of observations would differ. In that case, the quasi-likelihood estimates would differ from the maximum-likelihood estimates under independence. This would be appropriate for application to disequilibrium mapping in small isolated populations for which relationships among individuals are known.

Acknowledgments

The support of National Institutes of Health (National Institute of Diabetes and Digestive and Kidney Diseases) grant 1 R01 DK55889 (to M.S.M.) and Department of Education grant P 200 A70512 (to A.S.) is gratefully acknowledged.

Appendix A

Distributions of Length and Breakpoints of Ancestral Segment, under Various Crossover Models

Stationary renewal–process models.—Assume that the occurrence of crossovers along a single chromatid strand follows a stationary renewal–process model in the genetic-distance metric with mean interarrival time $\mu_I = 1$ (by the definition of genetic distance) and for which the interarrival density $f_I(x)$ is defined. Examples include the no-interference model of Haldane, for which the interarrival density is $f_I(x) = e^{-x}$, and the stationary gamma model of Foss et al. (1993), McPeek and Speed (1995), and Zhao et al. (1995), for which the interarrival density is $f_I(x) = \Sigma_{k=1}^{\infty} 1/2^k (2\gamma)^{k\gamma} x^{k\gamma-1} e^{-2\gamma x}/(k\gamma - 1)!$, with the former being a special case of the latter, with $\gamma = 1$. For a stationary renewal process, the density of the length of the interval containing the variant (which we identify with the origin) is $xf_I(x)/\mu_I = xf_I(x)$ (e.g., see Cox and Isham 1980, pp. 7–8). It follows that, for $a,b > 0$,

$$P\{\text{no events in } (-a,b) \text{ in one meiosis}\}$$

$$= \int_{x \geqslant a+b} (x - a - b) f_I(x) dx \ ,$$

because the left edge of an interval of length l containing 0 is uniform on $(-l,0)$. Thus,

$$P\{\text{no events in } (-a,b) \text{ in } \tau \text{ meiosis}\} = \left[\int_{x \geqslant a+b} (x-a-b) f_I(x) dx \right]^{\tau},$$

which we denote as $g(a+b,\tau)$. Then the joint density of the left and right edges of the ancestral interval containing 0 after τ generations is $\partial^2 g(a+b,\tau)/\partial a \partial b = \partial^2 g(l,\tau)/\partial l^2$. In the case of no interference, this gives independent exponential(τ) breakpoints. Then the density of the length of the segment containing the variant after τ generations is $l\partial^2 g(l,\tau)/\partial l^2 = \tau l g(l,\tau-2)[(\tau-1)G_I(l)^2 + g(l,1)f_I(l)]$, where

$$G_I(l) = \int_l^{\infty} f_I(x) dx \ .$$

In the case of no interference, this gives a gamma(2,τ) density for the length of the segment containing the variant after τ generations.

Complete interference.—In the case of complete interference, we have P{no events in $(-a,b)$ in one meiosis} $= 1-a-b$. Then, applying calculations analogous to those used above, we determine that the density of the length of the segment containing the variant after τ generations is $f(l) = \tau(\tau-1)(1-l)^{\tau-2}l$ (i.e., the length is beta[$\tau-1$,2] distributed).

Appendix B

Reconstruction of the Ancestral Haplotype

In practice, to assess the magnitude of LD around a variant, we maximize the likelihood over the ancestral haplotype, using a branch-and-bound algorithm. This procedure works as follows: start with the two loci flanking the variant, and consider all possible ancestral haplotypes for those two loci. Any ancestral haplotype for the entire panel of markers will reduce to one of these two-locus haplotypes when restricted to just these two markers. Suppose that we can give upper and lower bounds on the maximized likelihood for any ancestral haplotype, for the full panel, that reduces to a particular two-locus haplotype when haplotypes are restricted to the given marker pair. If the upper bound for one two-locus haplotype is lower than the lower bound for another, then we need not give further consideration to any ancestral haplotypes that give the former when haplotypes are restricted to the two-marker set. We eliminate all such two-locus haplotypes that will lead to suboptimal ancestral haplotypes for the full set. For the remaining two-locus haplotypes, we add the nearest marker and consider all possible three-locus haplotypes containing the two-locus haplotypes kept at the previous step. Again, we give upper and lower bounds on the maximized likelihood for any ancestral haplotype, for the full panel, that reduces to a particular three-locus haplotype when restricted to the given marker trio. Then we eliminate any three-locus haplotypes that will lead to suboptimal ancestral haplotypes for the full set. We proceed stepwise in this fashion, until we obtain the maximum-likelihood estimates of ancestral haplotype and τ^{-1}. Note that a lower bound as described above is easily obtained by picking a representative ancestral haplotype, for the full panel, that reduces to the particular k-marker haplotype under consideration. For each marker outside the k-marker haplotype, the predominant allele in the sample at that marker may be chosen for the representative haplotype. Determination of an upper bound is more involved and will be described elsewhere. As part of the determination of the upper bound, we maximize the likelihood for each current k-marker haplotype.

Consider estimation of τ^{-1}. If the set of markers spans an extremely long distance, then, at some point, there will be no more sharing of an ancestral haplotype; however, an estimated ancestral haplotype containing markers from one of the haplotypes in the data will always exhibit some sharing. Thus, if reconstruction of the ancestral haplotype continues too far away from the variant, upward bias in the estimated τ^{-1} will be introduced. To guard against this, we do not reconstruct the ancestral haplotype beyond the markers at which the expected number of haplotypes still sharing the ancestral haplotype, conditional on the data and the parameter estimates, drops to either <5 or <25% of the sample size, whichever is larger. When the likelihood is maximized for each k-marker haplotype, the Baum algorithm automatically provides the conditional expected number of individuals still sharing at the ends of the haplotype. If this drops to either <5 or <25% of the sample size, reconstruction is terminated. On the other hand, for the multilocus mapping problem, it is necessary to include the same data in the likelihood for each possible location, and, in that case, we are generally not concerned with bias in the estimation of τ^{-1}. Thus, in the mapping problem, we do not impose any threshold for stopping the reconstruction of the ancestral haplotype.

Appendix C

Joint Distribution of Breakpoints of Ancestral Segments of Two Related Haplotypes

Under the assumption of no interference, right and left breakpoints are independent, so consider only the right breakpoints. The two haplotypes are assumed to be descended from the most recent common ancestor of the sample, with the most recent common ancestor of the pair occurring at time t_t generations after the most recent common ancestor of the sample. Thus, at time t_t generations after the most recent common ancestor of the sample, the two haplotypes have the same right breakpoint B_t, which is exponential(t_t) distributed. After that point, their recombinational histories are assumed to be independent for time t_a. During that time, let B_1 be the location of the nearest crossover, to the right of the variant, occurring on haplotype 1, and let B_2 be the location of the nearest crossover, to the right of the variant, occurring on haplotype 2. B_1 and B_2 are independent exponential(t_a). The resulting right breakpoint for haplotype 1 is $R_1 = \min(B_1,B_t)$, and the resulting right breakpoint for haplotype 2 is $R_2 = \min(B_2,B_t)$. We use the following three properties of the exponential distribution: (*a*) If X_i is exponential(a_i), with the X_i's independent, $i = 1,\ldots,n$, then $P\{X_j < X_i \text{ for all } i \neq j\}$ for a given j in $1,\ldots,n$ is $a_j/\Sigma_i a_i$. (*b*) Conditional on the event $\{X_j < X_i \text{ for all } i \neq j\}$ for a given j in $1,\ldots,n$, the distribution of X_j is exponential($\Sigma_i a_i$). (*c*) Conditional on $\{X_i > x\}$, X_i has the distribution of an exponential(a_i) random variable plus x. Applying the properties, we find that $P\{R_1 = R_2\} = P\{B_t \leqslant B_1 \text{ and } B_t \leqslant B_2\} = t_t/(t_t + 2t_a)$. Conditional on the event $\{B_t \leqslant B_1 \text{ and } B_t \leqslant B_2\}$, the distribution of B_t is exponential($t_t + 2t_a$). With probability $2t_a/(t_t + 2t_a)$, the right breakpoints of the two haplotypes will be distinct (i.e., the event $\{B_t > B_1 \text{ or } B_t > B_2\}$ will occur). Conditional on this event, the distribution of $\min(R_1,R_2)$ is exponential($t_t + 2t_a$), and $\max(R_1,R_2) = \min(R_1,R_2) + E$, where E is exponential(τ) independent of $\min(R_1,R_2)$.

Appendix D

Correlation between Ancestral-Segment Lengths, for Haplotypes Related by a Conditional-Coalescent Model

First, we give the density of the *j*th coalescent time, conditional on time α to the most recent common ancestor, where j is counted by proceeding backward in time. Here, all times are on the coalescent scale of $2N_e$ generations, where N_e is generally taken to be 10^4; thus, $\alpha = 5 \times 10^{-5}\tau$. The conditional density of the *j*th coalescent time is

$$f_j(s) = (-1)\sum_{i=n-j}^{n-1}\prod_{\substack{k=n-j\\k\neq i}}^{n-1}[(i+1)i/2-(k+1)k/2]^{-1}e^{-s(i+1)i/2}$$
$$\times\sum_{h=1}^{n-j-1}\prod_{\substack{l=1\\l\neq h}}^{n-j-1}[(h+1)h/2-(l+1)l/2]^{-1}e^{(\alpha-s)(h+1)h/2}$$
$$\div\sum_{i'=1}^{n-1}\prod_{\substack{j'=1\\j'\neq i'}}^{n-1}[(i'+1)i'/2-(j'+1)j'/2]^{-1}e^{-\alpha(i'+1)i'/2}\ ,$$

where n is the sample size. Then the correlation between ancestral-segment lengths for a randomly chosen pair of haplotypes from a coalescent model with sample size n, conditional on time α to the most recent common ancestor can be computed as

$$\sum_{j=1}^{n-1}\frac{2(n+1)}{(n-1)(n-j+1)(n-j+2)}\int_0^{\alpha}f_j(s)\frac{\alpha-s}{\alpha+s}ds\ .$$

When α approaches 0, we obtain correlation

$$\sum_{k=1}^{n-2}\Bigg\{2(n-2)!(n+1)$$
$$\div[(n-1)(n-k+1)(n-k+2)$$
$$\times(n-k)(k-1)!(n-k-2)!]$$
$$\times\sum_{i=1}^{\infty}(-1)^i\binom{n+i-1}{n-k}^{-1}\Bigg\}\ .$$

This approximation is useful when τ is small relative to 2×10^4.

References

Baum LE (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3:1–8

Bennett JH (1954) On the theory of random mating. Ann Eugenics 184:311–317

Cox DR, Isham V (1980) Point processes. Chapman & Hall, New York

de la Chapelle A (1993) Disease gene mapping in isolated human populations: the example of Finland. J Med Genet 30:857–865

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc [B] 39:1–38

Devlin B, Risch N, Roeder K (1996) Disequilibrium mapping: composite likelihood for pairwise disequilibrium. Genomics 36:1–16

Foss E, Lande R, Stahl FW, Steinberg CM (1993) Chiasma interference as a function of genetic distance. Genetics 133: 681–691

Graham J, Thompson EA (1998) Disequilibrium likelihoods for fine-scale mapping of a rare allele. Am J Hum Genet 63: 1517–1530

Kaplan NL, Hill WG, Weir BS (1995) Likelihood methods for locating disease genes in nonequilibrium populations. Am J Hum Genet 56:18–32

Kerem B-S, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245: 1073–1080

Kingman JFC (1982) The coalescent. Stochastic Proc Appl 13: 235–248

Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–1188

Lazzeroni LC (1998) Linkage disequilibrium and gene mapping: an empirical least-squares approach. Am J Hum Genet 62:159–170

Lindsay BG (1988) Composite likelihood methods. Contemp Math 80:221–239

McCullagh P, Nelder JA (1989) Generalized linear models, 2d ed. Chapman & Hall, New York

McPeek MS, Speed TP (1995) Modeling interference in genetic recombination. Genetics 139:1031–1044

Nevanlinna HR (1972) The Finnish population structure: a genetic and genealogical study. Hereditas 73:195–236

Norio R (1981) Diseases of Finland and Scandinavia. In Rotschild H (ed) Biocultural aspects of disease. Academic Press, New York, pp 359–415

Pennacchio LA, Lehesjoki A-E, Stone NE, Willour VL, Virtaneva K, Miao J, D'Amato E, et al (1996) Mutations in the gene encoding cystatin B in progressive myoclonus epilepsy (EPM1). Science 271:1731–1734

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77: 257–285

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

Slatkin M (1972) On treating the chromosome as the unit of selection. Genetics 72: 157–168

Terwilliger JD (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

Virtaneva K, Miao J, Träskelin A-L, Stone N, Warrington JA, Weissenbach J, Myers RM, et al (1996) Progressive myoclonus epilepsy EPM1 locus maps to a 175-kb interval in distal 21q. Am J Hum Genet 58:1247–1253

Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika 61:439–447

Whittemore AS (1996) Genome scanning for linkage: an overview. Am J Hum Genet 59:704–716

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

Zhao H, Speed TP, McPeek MS (1995) Statistical analysis of crossover interference using the chi-square model. Genetics 139:1045–1056