Genetic Epidemiology 19(Suppl 1):S71-S77 (2000)

Methods for Multipoint Disease Mapping Using Linkage Disequilibrium

Bruce Rannala[1]* and Montgomery Slatkin[2]

[1]*Department of Ecology and Evolution, State University of New York, Stony Brook, New York*
[2]*Department of Integrative Biology, University of California, Berkeley, California*

A disease-associated mutation arises on a single chromosome such that alleles at linked markers are initially in complete linkage disequilibrium (LD) with the mutation. LD can be used as a tool for high-resolution mapping of the position of a disease mutation relative to a set of linked marker loci. When more than two linked marker loci are considered, developing a maximum likelihood approach is a challenging mathematical problem. To reduce the complexity, approximate and composite likelihood (CL) methods have been developed for multipoint LD mapping that use simplified models of population history, or of recombination, that ignore some of the statistical dependence among disease chromosomes and among marker loci. We describe the relationship among several composite likelihood methods for multipoint LD mapping, and suggest an alternative CL method that takes better account of the statistical dependence among marker loci. Genet. Epidemiol. 19(Suppl 1):S71-S77, 2000. © 2000 Wiley-Liss, Inc.

Key words: linkage disequilibrium; multipoint gene mapping; composite likelihood; mutation

INTRODUCTION

Family-based linkage methods for mapping disease mutations have relatively low resolution; even when numerous extended families are available, the data are usually informative about recombination at a few hundred meioses at most [Boehnke, 1994]. Thus, linkage methods typically provide estimates of map distance at a resolution of no more than 1 cM. Positional cloning requires a much narrower candidate region: alternative methods for high-resolution mapping of disease mutations are needed. One of the most promising methods, linkage disequilibrium (LD) mapping [Lander and Botstein, 1986], takes advantage of the fact that a mutation arises on a single chromosome and is

*Correspondence to: Dr. Bruce Rannala, Department of Medical Genetics, 8-39 Medical Sciences Bldg., University of Alberta, Edmonton, Alberta T6G 2H7 Canada. E-mail: brannala@ualberta.ca

initially in complete LD with alleles at nearby marker loci. LD between the disease mutation and a marker allele decays by recombination over time at a rate determined by the map distance of the marker from the disease mutation. Because the genealogy underlying a population sample of chromosomes that bear a disease mutation may be very large (much larger than a family pedigree), there is the opportunity for thousands of informative meiotic events to occur. LD mapping can therefore provide much higher resolution than linkage mapping, narrowing the candidate region of a mutation down to 0.0001 cM (roughly 100 kb), or less. Recently, maximum likelihood (ML) methods have been developed for LD mapping using one, or two, linked diallelic markers [Kaplan et al., 1995; Rannala and Slatkin, 1998]. Multipoint ML methods for LD mapping, on the other hand, present a number of technical difficulties that will likely be overcome only by improved computer programs that implement numerical techniques to evaluate likelihoods. The parametric multipoint LD mapping methods developed thus far rely on approximations that allow rapid computation of the likelihood [Terwilliger, 1995; Xiong and Guo, 1997; Graham and Thompson, 1998; Service et al., 1999]. Several use a composite likelihood (CL) to approximate the exact likelihood (EL). Here we review existing CL methods, clarify their assumptions, and suggest an alternative CL method.

METHODS

Consider a sample of n chromosomes carrying a disease-associated allele, D. For each chromosome, L linked marker loci are typed. Let X_{ij} be the allele observed at the jth marker locus on chromosome i, define $\mathbf{X}_i$ to be a vector of the alleles observed at all L markers on chromosome i (i.e., the multilocus haplotype of chromosome i), define $\mathbf{X}_j$ to be a vector of the alleles observed over all n chromosomes at a specific marker locus j, and let $\mathbf{X} = \{X_{ij}\}$ be a matrix of the haplotypes of the n sampled chromosomes. Several parameters are needed to model the population of disease chromosomes. Let t be the time, in the past, when the mutation D arose, let X_0 be the multilocus haplotype on which the disease mutation first arose, define α_i to be the map position of marker locus i relative to marker locus 1, where marker locus 1 is defined to be the locus closest to the telomere, and let $\boldsymbol{\alpha} = \{\alpha_i\}$ be a vector of the map distances of the L - 1 remaining markers from marker 1. Let the distance of the disease mutation from marker 1 be α_M and let $\mathbf{p} = \{p_{jk}\}$ be a matrix of marker allele frequencies on normal chromosomes (those not carrying disease mutation D), where p_{jk} is the frequency of the kth allele at marker locus j. We will assume that the allele frequencies $\mathbf{p}$ have been constant since mutation D arose, but this assumption may be relaxed. Define $\boldsymbol{\beta}$ to be a vector of population demographic factors influencing the intraallelic genealogy (sampling fraction, population growth rate, etc.). These parameters will be irrelevant for the CL methods considered in this paper, but may have an important influence on the EL. In the case of a single marker locus, the second subscript will be dropped for the variables and parameters with double subscripts.

Exact Likelihood

The likelihood of the haplotypes observed among a sample of disease chromosomes depends on their underlying genealogical relationships, the rates of recombination among markers, and the time since the mutation arose. Since the genealogy of the chromosomes can never be known with certainty, it is necessary to integrate over all possible genealogies to obtain a marginal likelihood which does not depend on a specific

genealogy. If we define $\Omega = \{\alpha, \mathbf{p}, X_0, t\}$ to be a vector of the unknown (nuisance) parameters, the likelihood may be written as

$$\Pr(\mathbf{X} \mid \Omega, \beta; \alpha_M) = \int \Pr(\mathbf{X} \mid \Omega, \tau; \alpha_M) dF(\tau \mid \beta), \quad (1)$$

where τ jointly represents the genealogical tree and the coalescence times underlying the sampled disease chromosomes, the so-called intraallelic genealogy [Slatkin, 1996]. The above integral is of general Lebesgue-Stieltjes form, and evaluating it would involve an integration over n - 1 coalescence times and a sum over $n!(n - 1)!/2^{n-1}$ distinct labelled genealogies. In principle, this can be done numerically using computer intensive methods, e.g., Rannala and Slatkin [1998] developed a simple Monte Carlo integration method for numerically evaluating this likelihood given a single diallelic marker locus.

Composite Likelihoods

Although an EL calculation for arbitrary numbers of marker loci and alleles is possible in principle by using equation (1) and evaluating a high-dimensional integral, this turns out to be difficult in practice. To minimize computational problems, approximate methods have been developed for calculating the likelihood aimed at avoiding some of the difficulties posed by the above formula. Most of these approaches involve the use of a CL. The likelihood of a parameter Θ that completely determines the probability distribution of an m-dimensional discrete random variable $\mathbf{A} = \{A_1, A_2, \ldots A_m\}$ is given by the joint probability distribution $\Pr(\mathbf{A}|\Theta)$. If the variables are independent the joint probability distribution is simply the product of the marginal probabilities

$$\Pr(\mathbf{A} \mid \Theta) = \prod_{i=1}^{m} \Pr(A_i \mid \Theta). \quad (2)$$

The CL is the product of the marginal likelihoods, which is the EL if the variables are independent, but is otherwise only an approximation to the EL. Computing the CL is usually much easier than computing the EL because the marginal probabilities are often functions of fewer parameters and are more easily derived. Several different composite likelihoods have been proposed for LD mapping that ignore either the non-independence among chromosomes resulting from their common genealogical history, or the non-independence among marker loci arising because recombination acts on pairs of haplotypes rather than independent pairs of markers.

Type I composite likelihood

The first CL method we consider treats chromosomes as independent, thus ignoring the effects of shared genealogy. This CL is the EL if the genealogy is a star-genealogy with all disease-associated chromosomes in the sample first coalescing to a shared haplotype precisely at time t in the past [Rannala and Slatkin, 1998]. Terwilliger [1995] appears to have been the first to employ this CL by considering a model in which the frequency, q_i, of the ith marker allele on disease chromosomes is

$$q_i = p_i + \lambda(1 - p_i), \quad (3)$$

where p_i is the frequency of the ith marker on normal chromosomes. If mutation D first arose on a chromosome bearing marker allele i, then λ is the excess frequency of i on D-bearing chromosomes. It will be convenient to define a new variable, for use here and in later sections, $Y_i = \Sigma_j \Im(i, X_j)$, where $\Im(i, X_j) = 1$ if $X_j = i$, and 0 otherwise. Terwilliger [1995] proposed the likelihood function (4) below for estimating λ, given that the

mutation arose on a chromosome bearing marker allele i. We correct a typographical error in his paper and substitute Y and Z for X and Y to be consistent with our notation

$$L_i = C\prod_{j=1}^{m} q_j^{Y_j} r_j^{Z_j}, \tag{4}$$

where q_i is given by equation (3), $q_j = (1 - \lambda)p_j$ (for all $j \neq i$), Y_j is the number of disease chromosomes bearing allele j, Z_j is the number of normal chromosomes bearing allele j, and m is the number of distinct haplotypes in the total sample of normal and D-bearing chromosomes. For a rare disease, the frequency of normal chromsomes in the sample bearing allele j is $r_j = p_j$. The sample of normal chromosomes carries no information about λ and may be subsumed into the irrelevant constant C. The parameter λ, estimated by maximizing the above likelihood, does not allow one to directly estimate the rate of recombination per generation, θ, and instead provides a test of association (or disequilibrium) between mutation D and marker i (i.e., $\lambda = 0$ versus $\lambda > 0$). Terwilliger [1995] suggested approximating λ by $(1 - \theta)^t$ in the likelihood function to estimate θ.

It is possible to derive the expected value of λ as a function of θ and t, under an explicit model of recombination, and in this way obtain an approximate estimator of θ. Consider a chromosome j, descended from a single D-bearing chromosome that arose t generations ago. If D arose on a chromosome bearing allele i (in our notation $X_0 = i$), the probability that j carries marker allele i is

$$\begin{aligned} \Pr(X_j = i \mid X_0 = i, t, \mathbf{p}; \theta) &= (1-\theta)^t + [1-(1-\theta)^t]p_i, \\ &= (1-p_i)(1-\theta)^t + p_i. \end{aligned} \tag{5}$$

This is also the expectation of the frequency q_i of allele i on disease chromosomes after t generations because the expectation of a sum (the sum of the disease chromosomes bearing allele i, used in calculating the frequency of i) is equal to the sum of the expectations, regardless of whether the variables are independent. Thus, the expectation of the frequency of i does not depend on the precise model of population demography (the marginal probability depends only on the recombination process). The EL does depend on population demography (and evolutionary history). It is simple to solve for λ as a linear function of q_i

$$\lambda = \frac{q_i}{1-p_i} - \frac{p_i}{1-p_i}.$$

Substituting the expectation of q_i into the above equation, the expectation of λ is found to be $(1 - \theta)^t$ suggesting that Terwilliger's approximation equates λ with its expected value under a model of population recombination. The approximate likelihood derived by Terwilliger is actually a CL because the resulting likelihood may instead be written as

$$\Pr(\mathbf{X} \mid X_0, t, \mathbf{p}; \theta) = C\prod_{j=1}^{n} \Pr(X_j \mid X_0, t, \mathbf{p}; \theta). \tag{6}$$

This CL approximation recently has been used by others [e.g., Service et al., 1999]. Xiong and Guo [1997] arrive at the same result by using a first-order Taylor series approximation for the EL in a model with explicit population demography. Treating the joint probability distribution of the marker alleles over n chromosomes as the product of the marginal probability of the marker on each chromosome ignores the dependence among chromosomes due to shared genealogy.

A digression is needed to clarify the relationship of these formulae to those of Rannala and Slatkin [1998], and others, who instead write equation (5) as

$$\Pr(X_j = i \mid \theta, X_0 = i) = e^{-t\theta} + (1 - e^{-t\theta}) p_i .$$

Note that $(1 - \theta)^t$ can be rewritten as $\exp\{t \log(1 - \theta)\}$ and that, for small θ, the expression $t \log(1 - \theta)$ is approximated as $-t\theta$ (a first-order Taylor series approximation). The expression $\exp\{-t\theta\}$ then approximates $(1 - \theta)^t$ when θ is small. The same result may be obtained directly using a continuous time Markov process model of recombination [Rannala and Slatkin, 1998]. In our earlier paper, we noted that maximizing the above CL to estimate θ produces an estimator that is identical to the moment estimator

$$\hat{\theta} = \frac{1}{t}\{\log(1 - p_0) - \log(Y_0 - np_0) + \log(n)\}.$$

For most datasets, the CL presented above will provide confidence intervals (CIs) that are too narrow (Figure 1), because it assumes independence where it does not exist and exaggerates the amount of information available from the data. Xiong and Guo [1997] remark that for most datasets they examined, little difference is observed between first and second order approximations. This suggests a CL approach often may be sufficient.

Multiple marker alleles

One merit of the above CL approximation is that it produces a simple analytical expression that can be used to directly explore properties of the likelihood function, with the hope that the results obtained will generalize to the more complicated EL. One question that can be easily addressed is the effect that additional marker alleles have on estimation of the parameter θ. In particular, do we gain any information by explicitly including multiple alleles, or can the marker alleles not associated with the disease mutation simply be pooled? If no information is added by the additional allele counts then a sufficient statistic for estimating θ in the case of a single marker locus is the number of sampled chromosomes, Y_0, bearing the marker found on the chromosome on which the disease mutation arose, $X_0 = 0$ (see Casella and Berger [1990] for a discussion of sufficient statistics). In other words, the sample information from $X_j \neq X_0$ carries no

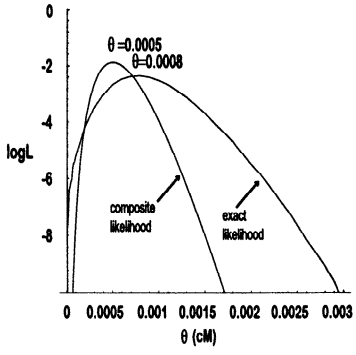

Fig. 1. Log-likelihood of a microsatellite marker in linkage disequilibrium with the diastrophic dysplasia mutation [Hastbacka et al., 1992], as a function of the recomination rate, θ [for details of data treatment, see Rannala and Slatkin, 1998]. Log-likelihoods are shown for both the CL method (equation (6)) and the EL method (obtained using our DMLE program). The difference between the CL and the EL is informative about how well the CL approximates the EL. In this case, the CL underestimates θ, in comparison to the EL, and provides a radius of curvature (confidence interval) that is too narrow.

additional information about θ. The probability of the marker allele counts in a sample of disease chromosomes, denoted as $\mathbf{Y} = \{Y_0, Y_1, \ldots Y_{m-1}\}$, using the CL approximation is

$$\Pr(\mathbf{Y} \mid \mathbf{p}, \theta, t) = \frac{n!}{Y_0! \cdots Y_m!} \prod_{i=1}^{m-1} \{p_i(1-e^{-\theta t})\}^{Y_i} \times \{e^{-\theta t} + p_0(1-e^{-\theta t})\}^{Y_0},$$

and the marginal probability of Y_0 may be written

$$\Pr(Y_0 \mid \mathbf{p}, \theta, t) = \frac{n!}{Y_0!(n-Y_0)!} \{(1-e^{-\theta t})(1-p_0)\}^{n-Y_0} \times \{e^{-\theta t} + (1-e^{-\theta t})p_0\}^{Y_0}.$$

If Y_0 is a sufficient statistic for θ, the conditional probability distribution of **Y** given Y_0 should be independent of θ. The conditional probability distribution is

$$\Pr(\mathbf{Y} \mid Y_0, \mathbf{p}, \theta, t) = \frac{\Pr(\mathbf{Y} \mid \mathbf{p}, \theta, t)}{\Pr(Y_0 \mid \mathbf{p}, \theta, t)},$$

$$= \frac{Y_1! \cdots Y_m!}{(n-Y_0)!} \times \frac{\prod_{i=1}^{m-1} p_i^{Y_i}}{(1-p_0)^{n-Y_0}},$$

which does not depend on θ. It is more difficult to prove that Y_0 is a sufficient statistic for estimating θ under the EL, but we conjecture that this is the case. McPeek and Strahs [1999] suggested that existing likelihood methods are too simplistic because they consider only a diallelic locus. However, if X_i is a sufficent statistic for θ then no information is lost by simply pooling the remaining (non-ancestral) alleles into a single class and using a diallelic method. This is common practice in gene mapping studies, and many disease mutations have been mapped by this approach.

Type II composite likelihood

The second CL method we consider ignores the dependence of recombination events among linked markers. For example, if two markers *a* and *b* are both centromeric to the disease mutation, *D*, with marker *a* closer to the mutation than marker *b*, any recombination events that occur in the interval *D*-*a* will also occur in the interval *D*-*b*. Recombination events between the markers and the disease locus are therefore not independent. The type II CL approach ignores this dependence among marker loci and multiplies the marginal probabilities of markers *a* and *b* to obtain the joint probability of the *a*-*b* haplotype. Terwilliger (1995) employed both a type I and a type II CL approximation in deriving a multipoint CL method using

$$\Pr(\mathbf{X} \mid \boldsymbol{\alpha}, X_0, t, \mathbf{p}; \alpha_M) = \prod_{i=1}^{n} \prod_{j=1}^{L} \Pr(X_{ij} \mid X_0, t, \mathbf{p}; \theta_j). \quad (7)$$

where $\theta_j = |\alpha_j - \alpha_M|$. This greatly simplifies the problem, i.e., the only probability required is that for the single chromosome, single marker, haplotype (equation (5)). An alternative CL that takes account the dependence among marker loci, but not the genealogical dependence among chromosomes, uses

$$\Pr(\mathbf{X} \mid \boldsymbol{\alpha}, X_0, t, \mathbf{p}; \alpha_M) = \prod_{i=1}^{n} \Pr(\mathbf{X}_{i.} \mid \boldsymbol{\alpha}, X_0, t, \mathbf{p}; \alpha_M). \quad (8)$$

An exact expression can be obtained for $\Pr(\mathbf{X}_{i.} \mid \boldsymbol{\alpha}, X_0, t, \mathbf{p}; \alpha_M)$ in the special case that the marker alleles at each locus are in linkage equilibrium on normal chromosomes, but it is quite complex and will be given elsewhere.

DISCUSSION

Formulating an exact likelihood for multipoint LD mapping is a difficult mathematical problem. In this paper, we have described several approximate methods for single marker and multipoint LD mapping that make use of composite likelihood (CL) approximations. We show that for a single marker locus, if the genealogical relationship among chromosomes is ignored (a type I CL), all available information about the map position, θ, of the disease mutation relative to the marker is contained in the number of disease chromosomes that carry the marker allele found on the chromosome on which the disease-associated mutation first arose (this is a sufficient statistic). We also show that in the case of a single locus, the maximum-CL and method of moments estimators of θ are identical. The use of CLs undoubtedly involves a trade-off of statistical accuracy and efficiency for mathematical simplicity and rapid computability. Additional studies are needed comparing the statistical performance of various CL, and EL, methods to determine how serious a cost is incurred by this trade-off. Although CLs offer a simple, approximate, approach for high-resolution LD gene mapping, in our opinion it is still worthwhile to pursue exact likelihood approaches despite the additional mathematical and computational challenges. Exact methods will ultimately provide the most accurate and efficient techniques for high-resolution LD mapping, although probably not the fastest.

ACKNOWLEDGMENTS

This work was supported by NIH grant HG01988. DMLE software is available on the World Wide Web at http://allele.bio.sunysb.edu.

REFERENCES

Boehnke M. 1994. Limits of resolution of genetic linkage studies. Am J Hum Genet 55:379-390.

Casella G, Berger RL. 1990. Statistical inference. Belmont: Duxbury Press.

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.

Hastbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E. 1992. Linkage disequilibrium mapping in isolated founder populations. Nat Genet 2:204-211.

Lander ES, Botstein D. 1986. Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harb Symp Quant Biol 51:49-62.

McPeek MS, Strahs A. 1999. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet 65:858-875.

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

Service SK, Temple Lang DW, Freimer NB, Sandkuijl LA. 1999. Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. Am J Hum Genet 64:1728-1738.

Slatkin M. 1996. Gene genealogies within mutant allelic classes. Genetics 143:579-587.

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.

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