Theoretical Population Biology **61**, 153–161 (2002)
doi:10.1006/tpbi.2001.1562, available online at http://www.idealibrary.com on IDEAL®

Likelihood-Based Disequilibrium Mapping for Two-Marker Haplotype Data

Chad Garner[1] and Montgomery Slatkin

Department of Integrative Biology, University of California, Berkeley, California 94720-3140

Received January 16, 2001

We report a theory that gives the sampling distribution of two-marker haplotypes that are linked to a rare disease mutation. The sampling distribution is generated with successive Monte Carlo realizations of the coalescence of the disease mutation having recombination and marker mutation events placed along the lineage. Given a sample of mutation-bearing, two-marker haplotypes, the maximum likelihood estimate of the location of the disease mutation can be calculated from the generated sampling distribution, provided that one knows enough about the population history in order to model it. The two-marker likelihood method is compared to a single-marker likelihood and a composite likelihood. The two-marker maximum likelihood gives smaller confidence intervals for the location of the disease locus than a comparable single-marker maximum likelihood. The composite likelihood can give biased results and the bias increases as the extent of linkage disequilibrium on mutation-bearing chromosomes decreases. Haplotype configurations exist for which the composite likelihood will fail to place the disease locus in the correct marker interval.

Key Words: **linkage disequilibrium mapping; maximum likelihood; haplotype; disease; coalescence; composite likelihood.**

INTRODUCTION

The term linkage disequilibrium (LD) is used to describe a non-random association of alleles between loci that is the result of shared population history. When a mutation occurs, the new allele is in complete LD with every other site on the chromosome. With the meioses of each following generation genetic recombination breaks down the ancestral haplotype, reducing the LD between the mutation and linked alleles. The extent of LD between the mutation and linked sites in the present population is a function of the recombination rate between them and of the population history. In the simplest model, the age of a mutation, the size of the population in which it exists, and the rate at which the population has grown can provide an estimate of the recombination rate between the mutation and a linked marker given a present-day sample of haplotypes.

The general method of localizing a disease mutation based on LD has been called linkage disequilibrium mapping. We propose a likelihood-based method for two-marker linkage disequilibrium mapping that is an extension of the single-marker method of Rannala and Slatkin (1998). The method can be used to compute the maximum likelihood (ML) location of a disease locus given a configuration of two-marker disease-bearing haplotypes and a vector of coalescence times. A fundamental assumption of the method is that the sample of chromosomes is known to carry the disease mutation, limiting the application of the method to fully penetrant diseases. The greatest advantage of the two-marker approach over the single-marker approach is that it allows for interval mapping of the disease mutation. We compare the performance of the two-marker likelihood method to a single-marker likelihood and a composite

[1] To whom correspondence should be addressed. Fax: (510-)643-6264. E-mail: cgarner@socrates.berkeley.edu.

0040-5809/02 $35.00

likelihood (CL) method. We show that using two-marker haplotype data can reduce the confidence interval of the ML location of a disease mutation. Examples are given of when the location of the disease locus cannot be accurately determined with the marginal data and we illustrate some of the pitfalls of a CL approach.

EXISTING METHODS FOR LD MAPPING

The first reported application of linkage disequilibrium mapping that used a statistical model for the population history was by Hästbacka *et al.* (1992) in mapping the diastrophic dysplasia (DTD) mutation in the Finnish population. Hästbacka *et al.* (1992) used Luria–Delbrück theory, originally developed for the estimation of mutation rates in bacterial colonies, as a model for the exponential growth of the Finnish population in order to estimate the location of the DTD mutation. The frequency of the disease mutation would have been small in the history of the population and subject to the influence of drift, not accommodated for in the Luria–Delbrück model. Kaplan *et al.* (1995) and later Kaplan and Weir (1995) showed that a simple deterministic model for LD results in smaller confidence intervals (CI) than a stochastic model incorporating the evolutionary history. Pritchard and Feldman (1996) further described the importance of modeling the variance due to population genetic influences in the context of allele age estimation.

The existing theoretical approaches for likelihood-based LD mapping have a common purpose, that is, to predict the recombination rate between a mutant allele, M, and a marker or markers from the configuration of haplotypes in a sample of M-bearing chromosomes. The approaches differ in two primary ways: how the population history of the mutant allele is modeled, or more specifically, how the coalescence times are generated, and how the likelihood of the data is computed. The likelihood is based on a probability model for recombination and mutation. Kaplan *et al.* (1995) provided the first likelihood-based approach that modeled the population history of the disease locus. Their likelihood-based method modeled the number of copies of the mutant allele in the population with a discrete-time branching process. They proposed a rejection sampling method in which simulated population histories are rejected if they are not consistent with the total number of current-day copies of the mutant allele, given the expected population frequency. From this distribution they could estimate the probability of the observed configuration of haplotypes. Rannala and Slatkin (1998) and Graham and Thompson (1998) both proposed methods that first generate the coalescent ancestry of the sample of disease alleles and then place recombination and mutation events along the lineage according to a probability model. Rannala and Slatkin (1998) used a continuous-time birth death process for generating the vector of coalescence times while Graham and Thompson (1998) use a continuous-time Moran model. The differences in how the methods generate the ancestry of the mutant allele are minor with the larger differences being in how the methods account for recombination and mutation in the likelihood. Rannala and Slatkin (1998) used continuous-time Markov chain theory to compute the probabilities of transitions between haplotype states given the length of a lineage and generated configurations of haplotypes at each coalescent event moving forward in time. They calculated the likelihood from the simulated conditional distribution of current-day haplotype configurations. Graham and Thompson (1998) defined a set of sampled disease haplotypes that descend from a meiosis at which a recombination event occurred without subsequent recombination events as a "recombinant class." The likelihood is found from the distribution of recombinant classes.

Xiong and Guo (1997) and Devlin *et al.* (1996) have proposed approximations to the likelihood. Xiong and Guo (1997) calculate the expectation of the likelihood of the multinomial haplotype data using a Taylor expansion and then use this expectation to calculate the full likelihood including the population parameters. Devlin *et al.* (1996) used simulation to show that the negative logarithm of their LD statistic was approximately distributed as a gamma variant when subjected to population genetic and sampling affects and proposed a method to approximate the likelihood using the first two moments of the distribution.

Extension of a single-marker likelihood to multiple markers is not straightforward. The problem quickly becomes computationally intractable; for n diallelic markers, there are 2^n possible haplotypes, each having n-1 intervals where a recombination event could occur. Two-marker likelihood models were given as extensions to the single-marker models by Kaplan *et al.* (1995) and Graham and Thompson (1998). Each of the proposed models carry assumptions; Kaplan *et al.* (1995) assumed no recombination interference and Graham and Thompson (1998) assumed linkage equilibrium between the markers. Both of these assumptions seem unlikely in the context of LD mapping. Graham and Thompson's (1998) two-marker likelihood is a simple extension of the single-locus recombinant class model when the disease

locus lies between the markers but is rather more complicated when the disease locus lies outside the interval.

In order to avoid the computational difficulties associated with the likelihood of multiple-marker haplotype data, Devlin *et al.* (1996) and Xiong and Guo (1997) have suggested using CL approaches. The CL is constructed from a set of conditional or marginal events for which one can write the log likelihoods, $L_i(\phi)$, $i = 1$ to n. The composite log likelihood is $CL(\phi) = \sum_i L_i(\phi)$. In linkage disequilibrium mapping, the CL is the sum of the likelihoods computed for the individual markers. CL does not require that the conditional or marginal log likelihoods be independent; however, if the terms are correlated the asymptotic theory of likelihood ratios does not apply. Twice the log of the composite likelihood ratio is not distributed as chi-square and there is no natural definition of a CI. Considering a model with two, diallelic, markers, the haplotypes contribute the observations in a 2×2 contingency table; however, the single-marker model predicts only marginal counts from this table. The marginal values of the table provide no information about the interdependence between the markers. Because the terms of the CL are combined under an assumption of independence, the evidence can be overstated when the terms are highly correlated because the information from the full likelihood is less than the sum of the marginal values. For example, if two marker loci were in complete linkage disequilibrium, one of the markers would provide all of the information about linkage to a third locus; however, the CL combines the evidence from both markers, essentially leading to a doubling of the available evidence. A detailed description of CL theory is given by Lindsay (1988).

THE TWO-MARKER LIKELIHOOD

The two-marker likelihood is an extension to the model developed by Rannala and Slatkin (1998). The data are the sample counts of the two-marker haplotypes in a collection of M-bearing chromosomes (the haplotype configuration), where M is a unique, non-recurrent mutation having occurred at time t_1 generations in the past. We consider a model with diallelic markers; the two marker loci A and B are tightly linked and have alleles A_1 and A_2 and B_1 and B_2, respectively. Linkage equilibrium or disequilibrium can exist between the markers on chromosomes in the population that do not carry the disease mutation. The frequency of allele A_1 among non-mutant chromosomes is p and the frequency of B_1 is q. The mutation rate from allele i to allele j at marker A is v_{ij} and from allele i to allele j at marker B is μ_{ij}. The four possible two-locus haplotypes A_1B_1, A_1B_2, A_2B_1, and A_2B_2 are denoted 1 to 4, respectively, and occur with frequencies Q_{ij}, where the subscripts indicate which allele is present at markers A and B, respectively. The frequencies of the marker alleles (and haplotypes) among the non-mutant chromosomes are assumed to have remained constant since M first arose in the population. The rate of recombination between the markers is c_1, and between marker A and M is c_2. We assume that M is at a low enough frequency in the population so that the frequency of individuals that are homozygous for the disease mutation is negligible; consequently M-bearing chromosomes are assumed only to recombine with non-mutant chromosomes. There are three haplotype orderings that are possible: (1) M–A–B, (2) A–M–B, and (3) A–B–M. We assume that the recombination rates, c_1 and c_2, are low enough that map distances are additive. The distance between marker B and M, c_3, can be specified with c_1 and c_2 for each of the three possible haplotype orderings: for order 1, $c_3 = c_2 + c_1$; for order 2, $c_3 = c_1 - c_2$; and for order 3, $c_3 = c_2 - c_1$.

The model of the process that generated the observed configuration of M-bearing haplotypes has two components: (i) the genealogical process, and (ii) the process of recombination between and mutation of the two markers. In terms of a continuous-time Markov chain, the two components are: (i) the independent exponential holding times in the successive states, and (ii) the embedded Markov chain which describes the sequence of states that are visited. The genealogical process describes the distribution of intra-allelic coalescence times of the sampled M-bearing chromosomes. Our method computes the likelihood given a vector of coalescence times.

The probability of a change in the two-marker haplotype on an M-bearing chromosome is a function of the population haplotype frequencies and the recombination and mutation rates. Table I gives the infinitesimal transition probabilities for the three possible marker and disease locus orderings. The four by four matrix of infinitesimal transition probabilities, **V**, can be found from the six general expressions for each ordering given in the table. Two transition probabilities have zero probability under order 2 because it is impossible for them to occur with a single mutation or recombination event. The probabilities of transition between mutation-bearing haplotypes i and j during the lineage of length t, $P_{ij}(t)$, can be found by solving the system of Kolmogorov forward equations given by the elements in matrix **V**. A description of the Kolmogorov forward equations can be found in any general text on continuous-time Markov chain processes; for example, see Ross (1996).

TABLE I

Transition Probabilities for the Three Possible Orderings of the Disease Mutation and Markers A and B

Transition	Order 1 (M–A–B)	Order 2 (A–M–B)	Order 3 (A–B–M)
$A_iB_i \rightarrow A_iB_j$	$\mu_{ij}+\sum_j c_1 Q_{\cdot j}+c_2 Q_{ij}$	$\mu_{ij}+\sum_j (c_1-c_2)\, Q_{\cdot j}$	$\mu_{ij}+(c_2-c_1)\, Q_{ij}$
$A_iB_i \rightarrow A_jB_i$	$v_{ij}+c_2 Q_{ji}$	$v_{ij}+\sum_j c_2 Q_{j\cdot}$	$v_{ij}+\sum_j c_1 Q_{j\cdot}+(c_2-c_1)\, Q_{ji}$
$A_iB_i \rightarrow A_jB_j$	$c_2 Q_{jj}$	0	$(c_2-c_1)\, Q_{ij}$
$A_iB_j \rightarrow A_iB_i$	$\mu_{ji}+\sum_i c_1 Q_{\cdot i}+c_2 Q_{ii}$	$\mu_{ji}+\sum_i (c_1-c_2)\, Q_{\cdot i}$	$\mu_{ji}+(c_2-c_1)\, Q_{ii}$
$A_iB_j \rightarrow A_jB_i$	$c_2 Q_{ji}$	0	$(c_2-c_1)\, Q_{ji}$
$A_iB_j \rightarrow A_jB_j$	$v_{ij}+c_1 Q_{jj}$	$v_{ij}+\sum_j c_2 Q_{j\cdot}$	$v_{ij}+\sum_j c_1 Q_{j\cdot}+(c_2-c_1)\, Q_{jj}$

Note. The 12 transition probabilities can be found from the six general forms shown.

The process of recombination and mutation is modeled moving forward in time. During the waiting time between the lth and $(l-1)$th coalescence events, there are $l-1$ independent lineages, each undergoing the transition process. At the lth coalescent event, one of the $l-1$ lineages existing in the genealogy is selected at random and duplicated. The configuration of haplotypes in the sample immediately after the $(l-1)$th lineage is specified by the numbers of each haplotype type, $Y_{h_{l-1}}$, where h denotes the haplotype, defined as 1 through 4 above. During the waiting time, t_l, between the $(l-1)$th and the lth coalescent events some of the haplotypes will be replaced by other forms. Denoting the number of haplotypes being replaced by A_1B_1, A_1B_2, A_2B_1, and A_2B_2 haplotypes as j, k, r, and s, respectively, the probability of the observed number of replacements is given by the multinomial equations,

$$P(k, r, s, Y_{1_{l-1}}-k-r-s \mid Y_{1_{l-1}}) = \frac{Y_{1_{l-1}}!}{k!\, r!\, s!\, (Y_{1_{l-1}}-k-r-s)!} (P_{12}(t_l))^k (P_{13}(t_l))^r (P_{14}(t_l))^s \times (1-(P_{12}(t_l)+P_{13}(t_l)+P_{14}(t_l)))^{Y_{1_{l-1}}-k-r-s} \quad (1)$$

$$P(j, r, s, Y_{2_{l-1}}-j-r-s \mid Y_{2_{l-1}}) = \frac{Y_{2_{l-1}}!}{j!\, r!\, s!\, (Y_{2_{l-1}}-j-r-s)!} (P_{21}(t_l))^j (P_{23}(t_l))^r (P_{24}(t_l))^s \times (1-(P_{21}(t_l)+P_{23}(t_l)+P_{24}(t_l)))^{Y_{2_{l-1}}-j-r-s} \quad (2)$$

$$P(j, k, s, Y_{3_{l-1}}-j-k-s \mid Y_{3_{l-1}}) = \frac{Y_{3_{l-1}}!}{j!\, k!\, s!\, (Y_{3_{l-1}}-j-k-s)!} (P_{31}(t_l))^j (P_{32}(t_l))^k (P_{34}(t_l))^s \times (1-(P_{31}(t_l)+P_{32}(t_l)+P_{34}(t_l)))^{Y_{3_{l-1}}-j-k-s} \quad (3)$$

$$P(j, k, r, Y_{4_{l-1}}-j-k-r \mid Y_{4_{l-1}}) = \frac{Y_{4_{l-1}}!}{j!\, k!\, r!\, (Y_{4_{l-1}}-j-k-r)!} (P_{41}(t_l))^j (P_{42}(t_l))^k (P_{43}(t_l))^r \times (1-(P_{12}(t_l)+P_{13}(t_l)+P_{14}(t_l)))^{Y_{4_{l-1}}-j-k-r} \quad (4)$$

The number of each haplotype immediately before the lth coalescent event is $Y'_{1_{l-1}} = Y_{1_{l-1}} + \sum_{l-1} j$, $Y'_{2_{l-1}} = Y_{2_{l-1}} + \sum_{l-1} k$, $Y'_{3_{l-1}} = Y_{3_{l-1}} + \sum_{l-1} r$, and $Y'_{4_{l-1}} = Y_{4_{l-1}} + \sum_{l-1} s$. The probabilities of each of the $Y'_{h_{l-1}}$ are given by the product of Eqs. (1) to (4) summed of all possible values j, k, r, and s that are consistent with $Y_{h_{l-1}}$ and $Y'_{h_{l-1}}$.

Rannala and Slatkin (1998) showed the probability of a current-day configuration of mutation-bearing haplotypes, Y_0, for their single-marker model and noted that the exact evaluation of the probability was intractable for sample sizes greater than 10 chromosomes; therefore, they proposed a Monte Carlo estimator for the probability. We have used a Monte Carlo algorithm similar to Rannala and Slatkin's (1998) to compute the likelihood for the two-marker haplotype case. We define a vectors of parameters: $\boldsymbol{\theta} = \{\mathbf{V}, Q, t_1, n, \mathbf{w}\}$, where t_1 is the age of the mutation, n is the total number of M-bearing haplotypes in the sample, and $\mathbf{w}$ is a vector of demographic parameters used to generate the coalescence times. In the birth–death model described in Rannala and Slatkin (1998), $\mathbf{w}$ would include f, the fraction of the population sampled, and ξ, a parameter for the combined effects of population growth and selection. The Monte Carlo estimator of $P(Y_0 \mid \boldsymbol{\theta})$ is obtained as

$$P(Y_0 \mid \theta) \approx \frac{1}{R} \sum_{k=1}^{R} P(Y_0 \mid \tilde{\mathbf{Y}}_n(k), \tilde{\mathbf{t}}(k); \mathbf{V}),$$

where the sum is evaluated over R Monte Carlo realizations of the random variables $\tilde{\mathbf{Y}}_n(k)$ and $\tilde{\mathbf{t}}(k)$. The vector of random variables $\tilde{\mathbf{t}}(k)$ is the vector of coalescence times. $\tilde{\mathbf{Y}}_n(k)$ are simulated by generating random variables j, k, r, and s from the multinomial distributions given in (1) through (4), conditional on the configuration at the previous step. The probability is computed by simulating up to the configuration $\tilde{\mathbf{Y}}'_n(k)$ and taking the sum of all possible multinomial probabilities that are consistent with $\tilde{\mathbf{Y}}_n(k)$.

TABLE II

Description of 10 Data Sets Analyzed Using the Two-Marker (2-mrkr) and Single-Marker (A and B) Methods and the Resulting Confidence Intervals

	Marginal Frequencies		Recombination Rates		Haplotype Frequencies			Sample Haplotype Configuration				Confidence Intervals		
Data set	p_A	p_B	c_1	c_2	Q_{11}	Q_{12}	Q_{21}	Y_{11}	Y_{12}	Y_{21}	Y_{22}	2-mrkr	A	B
1	0.05	0.05	0.0005	0.0005	0.0025	0.0475	0.0475	46	2	0	2	0.002	0.0023	0.0032
2	0.10	0.10	0.0005	0.0005	0.01	0.09	0.09	46	2	0	2	0.0019	0.0024	0.0034
3	0.50	0.50	0.0005	0.0005	0.25	0.25	0.25	47	2	0	1	0.0025	0.0033	0.0055
4	0.05	0.50	0.0005	0.0005	0.025	0.025	0.475	47	1	1	1	0.0019	0.0023	0.0044
5	0.50	0.05	0.0005	0.0005	0.025	0.475	0.025	46	3	0	1	0.0022	0.0033	0.0032
6	0.05	0.05	0.001	0.001	0.0025	0.0475	0.0475	42	4	0	4	0.0029	0.0032	0.0048
7	0.05	0.05	0.0005	0.001	0.0025	0.0475	0.0475	44	2	0	4	0.0028	0.0032	0.004
8	0.05	0.05	0.001	0.0005	0.0025	0.0475	0.0475	44	4	0	2	0.0019	0.0022	0.004
9	0.05	0.50	0.002	0.0005	0.025	0.025	0.475	44	4	1	1	0.0019	0.0023	0.0077
10	0.50	0.50	0.0005	0.0005	0	0.50	0.50	47	2	1	0	0.0025	0.0032	0.0044

The likelihood is conditional on the recombination rate between the markers, c_1. The method depends on the markers being tightly linked and it may not be possible to get a very accurate estimate of the distance between them. Figure 1 shows the likelihood curves computed using the same data but with five different values of c_1. The sample consisted of 30 disease-bearing chromosomes: 26 A_1B_1, 2 A_1B_2, 0 A_2B_1, and 2 A_2B_2 haplotypes having non-disease population frequencies of 0.05, 0.15, 0.15, and 0.65, respectively. Coalescent times were generated using parameter values corresponding to the DTD mutation in Finland (see Rannala and Slatkin, 1998 for details). The true recombination rate between the markers was 0.001. The ML location of the disease mutation was at a recombination rate of 0.0013 from the proximal marker. A large bias in the location estimate was observed only when the estimate of c_1 was considerably less than the true value. From this simple investigation, the method appears to be robust to small errors in the c_1 parameter; however, a more extensive study would be required to explore this aspect of the method in detail.

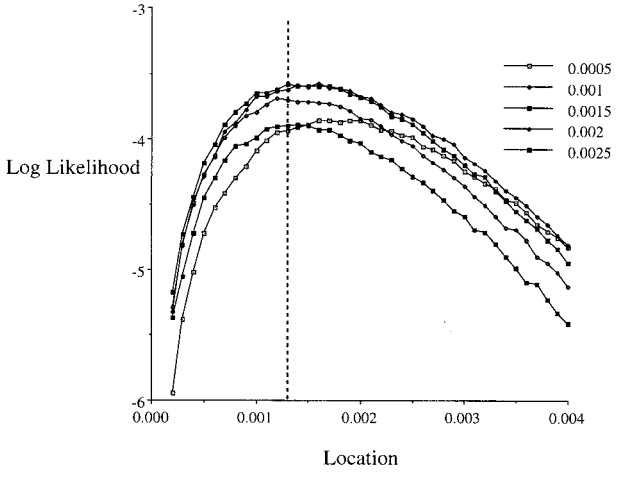

FIG. 1. Two-marker likelihoods computed for five different values of c_1. The location of marker A on the plot is at 0.000 with the location of marker B being at the various values of c_1 to left of marker A. The maximum likelihood location of the disease mutation is $c_2 = 0.0013$ and is indicated by the dashed line.

COMPARISON WITH A SINGLE-MARKER LIKELIHOOD

Haplotype configurations were generated by first simulating a vector of coalescence times based on a set of demographic parameters and the age of the disease mutation using the birth–death model described by Rannala and Slatkin (1998). The transition process was then modeled moving forward in time with the probabilities $P_{ij}(t)$ being dependent on the specified recombination rates and haplotype frequencies. The probability of mutation at the diallelic markers was assumed to be 0. The age of the disease mutation was assumed to be 2000 years ($t_1 = 100$) and to have a frequency of 0.8%. The demographic parameters used to generate the coalescence times were set to correspond to the DTD mutation in the Finnish population; $f = 1.125 \times 10 - 3$ and $\xi = 0.085$. Table II shows the CIs for the two-marker and the single-marker ML recombination rate estimates with 10 different sets of data. Likelihoods were computed from 50,000 Monte Carlo realizations. The sample size was 50 M-bearing chromosomes for all data sets and the disease mutation was assumed to have originally occurred on a haplotype having A_1 and B_1 alleles. The order of the loci was M–A–B for all of the data.

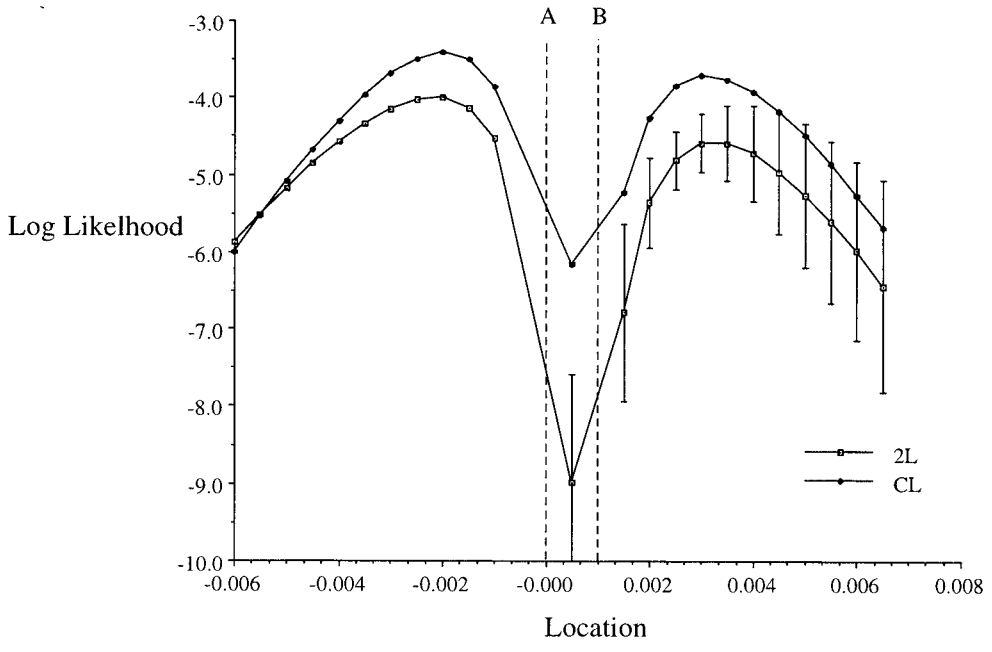

FIG. 2. Plot of the means and standard deviations of two-marker (2L) and composite (CL) likelihoods calculated from 20 replicates of the same data set. The dashed lines give the locations of markers A and B. The maximum likelihood location of the disease mutation is at a recombination distance of 0.002 to the left of marker A.

The obvious advantage of two-marker LD mapping over a single-marker approach is that the information from two markers allows for interval mapping. This advantage comes at the costs of needing haplotype data for two markers, increased computational time, and the having to know the recombination rate between the two markers. Comparing the CIs under the two methods assessed the amount of information that is gained by using the two-marker ML over the single-marker ML, ignoring the interval information. The two-marker likelihood described here was compared to the single-marker likelihood method of Rannala and Slatkin (1998).

The CIs for the single-marker and the two-marker ML increase as recombination between the markers and the disease mutation increases. The two-marker CI was less than the single-marker CI for all data sets. Data sets 1–5 show that the frequency of the disease-associated marker allele affects the size of the difference between the CIs of the two methods. The difference in the CIs increases as the frequency of the disease-associated marker allele increases. Data sets 6–9 show the effect of the distance to the disease mutation on the size of the single- and two-marker CIs. For low-frequency alleles, the difference in the CIs of the two methods is small; however, the difference can be relatively large when the disease-associated allele frequency is higher. In general, the size of the single-marker CI increases in relation to the two-marker CI as the marker allele frequency and distance to the disease mutation increase.

COMPARISON TO A COMPOSITE LIKELIHOOD

Figure 2 shows the two-marker and composite likelihood plots for a data set having 26 A_1B_1, 2 A_1B_2, 1 A_2B_1, and 1 A_2B_2 mutation-bearing chromosomes with frequencies of 0.10, 0.30, 0.30, and 0.30, respectively, in the non-disease population. The age of the mutation and demographics of the population were assumed to be those for the DTD mutation in Finland as described above. The two-locus and composite likelihoods of the data were computed 20 times, each from a different random seed and 50,000 Monte Carlo replicates, at each location shown and the mean and standard deviation were calculated. Figure 2 shows that there is almost no variation in the two-locus likelihoods around the maximum; however, there is considerable variation among the replicates in the intervals were the disease mutation does not exist. There is very little variation in the composite likelihood replicates regardless of the distance from the maximum.

TABLE III

Haplotype and Allele Frequencies for Nine Models Used to Generate Haplotype Configurations

Model	Q_{11}	Q_{12}	Q_{21}	Q_{22}	p	q	D'
1	0.0025	0.0475	0.0475	0.9025	0.05	0.05	0
2	0.05	0.0	0.0	0.95	0.05	0.05	1
3	0.0125	0.0375	0.2375	0.7125	0.05	0.25	0
4	0.05	0.0	0.20	0.75	0.05	0.25	1
5	0.0125	0.2375	0.0375	0.7125	0.25	0.05	0
6	0.05	0.20	0.00	0.75	0.25	0.05	1
7	0.06	0.24	0.14	0.56	0.30	0.20	0
8	0.20	0.10	0.00	0.70	0.30	0.20	1
9	0.01	0.29	0.19	0.51	0.30	0.20	0.83

Data were simulated as described in the previous section with the following exceptions. Two-marker haplotype data were simulated under two generating models defined by the order of the loci. For order I, the location of the disease mutation was simulated to be in the interval between markers A and B, and for order II, the mutation was simulated to be outside of the markers and proximal to marker A. For both orders the data were simulated under one of nine possible population haplotype frequency models with varying linkage disequilibrium (models 1–9, Table III). The recombination rate between markers A and B, c_1, was 0.0005. We have used a CL of the single-marker likelihood model described by Rannala and Slatkin (1998). Likelihoods were computed at 24 locations (intervals of 0.01cM) across a 0.25-cM region spanning the markers. The three intervals, outside of marker A, between markers A and B, and outside of marker B, are denoted 1, 2, and 3, respectively.

Table IV summarizes the comparison between the ML and CL location estimates. There were 780 replicates analyzed under order I and 368 under order II. The large difference in the number of replicates analyzed under the two models is due to the larger amount of time required to compute the likelihoods for haplotype configurations generated under order II. Twenty-two replicates were discarded because the maximum and composite likelihood estimates were at the edge of the test range. The Pearson correlation coefficients between the maximum and composite likelihood location estimates of the disease mutation were 0.89 and 0.83 for orders I and II, respectively. The composite and maximum likelihood locations were in different intervals for 64 replicates under order I. In all cases, the maximum likelihood location of the disease mutation was in interval 2 and the composite likelihood location was in interval 3. This discrepancy was associated with uninformative markers having low population allele frequencies. In 19 of the replicates of order II, the composite and maximum likelihood locations were in different intervals.

Figure 3 shows the two-locus and composite likelihood plots for a replicate in which the maximum likelihood location of the disease mutation was in interval 2 and the composite likelihood location was in interval 3. The data were a configuration of 89 A_1B_1, 4 A_1B_2, 6 A_2B_1, and 1 A_2B_2 disease-bearing haplotypes, with the population frequencies of the non-disease haplotypes being 0.05, 0.0, 0.0, and 0.95, respectively. Because there are no A_1B_2 and A_2B_1 haplotypes in the population, a transition from a MA_1B_1 haplotype to a MA_2B_1 haplotype requires two recombination events to occur. Similarly, a transition from an A_1B_1M haplotype to an A_1B_2M haplotype requires two recombination events. However, a transition from an A_1MB_1 haplotype to an A_1MB_2 or A_2MB_1 haplotype requires only a single recombination event. There are four A_1B_2 and six A_2B_1 M-bearing haplotypes in the sample so that, given the parameters of the model, the most likely order is A–M–B. The haplotype frequency information is not found in the marginal, single-locus data so the composite likelihood does not take the population frequency of the A_1B_2 and A_2B_1 haplotypes into account and the composite likelihood location for the disease locus lies in interval 3.

TABLE IV

Comparison between Two-Marker Maximum and Composite Likelihoods for Data Generated under Two-Marker and Disease Locus Orderings

Order	n	Correlation		Locations (mean/SD)		Different	Δ interval		
		likelihood	location	Max L	Comp L	location	0	1	2
I	780	0.96	0.89	13.4 (0.71)	13.5 (0.76)	10%	716	64	0
II	368	0.88	0.83	7.0 (3.10)	6.29 (3.79)	61%	346	17	2

Note. The Δ interval columns show the number of times that the composite likelihood places the disease locus one or two intervals away from the maximum likelihood location.

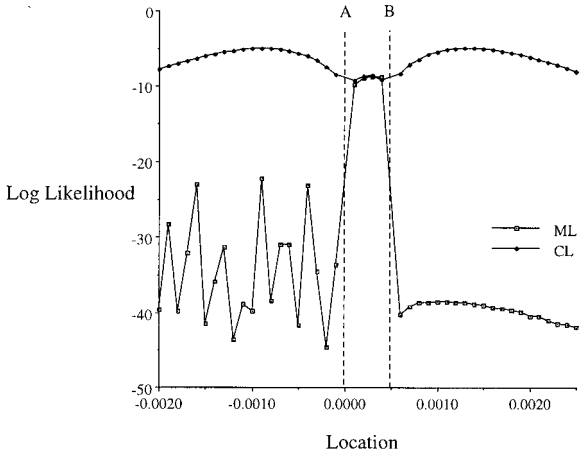

FIG. 3. Two-marker (2L) and composite (CL) likelihoods across a recombination distance of 0.005. The dashed lines give the locations of markers A and B. A recombination rate of 0.0005 separates the two markers. The maximum likelihood location of the disease mutation is between the markers and the maximum composite likelihood location lies outside of the markers.

The nine generating models (models 1–9, Table III) were coded as an independent class variable with nine levels and a generalized linear model was used to test whether the generating model had a significant effect on the distribution of differences in the ML and CL location estimates. A significant effect was observed for order I but not for order II. The ML and CL locations were more similar under models 5 through 9; these models are characterized by lower disease-associated allele frequencies for marker A (the marker proximal to the disease mutation). For all of the cases in which ML and CL locations were different for order I, at least one of the markers was not informative (had no recombinants in the sample).

The relationship between the amount of LD between the markers on the sampled M-bearing chromosomes and the difference between the ML and CL location estimates was assessed by linear regression. LD was measured as the difference between the M-bearing A_1B_1 haplotype frequency calculated from the sample and the product of the A_1 and B_1 allele frequencies in the population. Samples of M-bearing chromosomes with lower LD tended to show larger differences between the ML and CL location estimates under order II. In general, as the degree of LD decreased a bias in the CL location estimate toward larger distances from the proximal marker was observed. No significant relationship was observed under order I, this is likely due to the lower simulated recombination rates under the model, resulting in higher LD among the M-bearing chromosomes.

CONCLUSIONS

We present a two-marker, likelihood-based method for LD mapping. The two-marker approach is an improvement over the single-marker likelihood because the information provided by the second marker determines the position of the disease locus in relation to the markers (i.e., it allows for interval mapping) and the location estimate has smaller confidence intervals. Our method is an improvement over existing two-marker methods for several reasons. Our method does not assume linkage equilibrium between the markers and the disease locus or between the markers. The maximum likelihood estimate of the location of a disease locus is computed from the sample of disease-bearing haplotype data and the location can be either outside of the markers or between them. Because our method uses the full haplotype data, all of the available information is being used. The coalescent ancestry of the disease allele is generated independent of the likelihood calculation so that the vector of coalescence times can be generated under any demographic model. Our two-marker method cannot be extended to an arbitrary number of loci; however, it can accommodate multiallelic markers with non-negligible mutation rates.

For the two-marker method one needs two polymorphic sites in a small genomic region, estimates of the population haplotype frequencies, and a good estimate of the genetic distance between the markers. These costs can be outweighed by a substantial reduction in the candidate interval for the disease mutation. For markers having ideal characteristics for linkage disequilibrium mapping (e.g., low associated allele frequency, small genetic distance to disease mutation) the two-marker maximum likelihood had a confidence interval that was 0.03 cM or 30 kb smaller on average than the single-marker confidence interval. As markers become less ideally suited to mapping, the difference between the single-marker and two-marker confidence intervals becomes greater.

We show that the composite likelihood can be biased toward increased estimates of the distance between the disease mutation and the markers and that the bias increases as the LD between the markers on the M-bearing chromosome decreases. The correlation between the composite and maximum likelihood estimates was higher when the disease locus was between the marker loci; when the genetic distances were smaller and LD was greater among the M-bearing chromosomes. Because the composite likelihood uses information only from the marginal values it provides no information regarding how the data fit the specified marker map. The effect of

using only the marginal information can be a false inference of the location of the disease locus or ambiguity in its location. When there are no single-marker recombinants observed in the sample, the single-marker likelihood surface is determined by the marker allele frequencies; non-informative marker loci can lead to bias in the composite likelihood but presumably they would not be included in the analysis. Markers that are not informative for single-marker or composite likelihood analysis do provide information in the two-marker maximum likelihood. The single-marker and two-marker maximum likelihoods have statistically interpretable confidence intervals while the composite likelihood does not.

This report shows that two-marker haplotype data provide more information for linkage disequilibrium mapping than single markers and the composite statistic used here; however, there are questions that remain to be investigated. If one wishes to map a disease locus using two-marker haplotype data, which pairs of markers should one use? One could use all marker pairs but clearly some pairs will be better suited to LD mapping than others given the population history of the disease locus and the markers. We have tested the method in ideal circumstances (a rare, fully penetrant disease mutation) and the utility of this method and others like it for mapping mutations involved in disease of a more complex nature needs to be investigated.

A computer program for carrying out the likelihood calculations described in this paper has been written in the C language and is available from the corresponding author (C.G.).

ACKNOWLEDGMENT

This work was supported by the National Institutes of Health (NIH) under Grant GM-40282.

REFERENCES

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

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

Hastbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A., and Lander, E. 1992. Linkage disequilibrium mapping in isolated founder populations: Diastrophic dysplasia in Finland, *Nat. Gen.* **2**, 204–211.

Kaplan, N., Hill, W., and Weir, B. 1995. Likelihood methods for locating disease genes in nonequilibrium populations, *Am. J. Hum. Genet.* **56**, 18–32.

Kaplan, N. L., and Weir, B. S. 1995. Are moment bounds on the recombination fraction between a marker and a disease locus too good to be true? Allelic association mapping revisited for simple genetic diseases in the Finnish population, *Am. J. Hum. Genet.* **57**, 1486–1498.

Lindsay, B. G. 1988. Composite likelihood methods, *Contemp. Math.* **80**, 221–239.

Pritchard, J. K., and Feldman, M. W. 1996. Genetic data and the African origin of humans, *Science* **274** 1548.

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

Ross, S. M. 1996. "Stochastic Processes," Wiley, New York.

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