Am. J. Hum. Genet. 60:1513–1531, 1997

Fine-Scale Genetic Mapping Based on Linkage Disequilibrium: Theory and Applications

Momiao Xiong and Sun-Wei Guo

Division of Epidemiology, University of Minnesota, Minneapolis

Summary

Linkage-disequilibrium mapping (LDM) recently has been hailed as a powerful statistical method for fine-scale mapping of disease genes. After reviewing its historical background and methodological development, we present a general, mathematical, and conceptually coherent framework for LDM that incorporates multilocus and multiallelic markers and mutational processes at the marker and disease loci. With this framework, we address several issues relevant to fine-scale mapping and propose some efficient computational methods for LDM. We implement various LDM methods that incorporate population growth, recurrent mutation, and marker mutations, on the basis of a general framework. We demonstrate these methods by applying them to published data on cystic fibrosis, Huntington disease, Friedreich ataxia, and progressive myoclonus epilepsy. Since the genes responsible for these diseases all have been cloned, we can evaluate the performance of our methods and can compare ours with that of other methods. Using the proposed methods, we successfully and accurately predicted the locations of genes responsible for these diseases, on the basis of published data only.

Introduction

The recent successes of positional cloning have been instrumental in elucidating the genetic mechanisms underlying many human diseases. In essence, positional cloning seeks to identify disease genes on the basis of their chromosomal locations, in the absence of information on the underlying biological defect (Collins 1992). It is now well known that meiotic event–based linkage analysis needs huge (sometimes too huge to be realistic) sample sizes for fine-scale mapping of disease genes (Lange et al. 1985; Bodmer 1986; Boehnke 1994). Linkage disequilibrium (LD) recently has emerged as a very promising tool for fine-scale genetic mapping.

LD (Lewontin and Kojima 1960), or, more precisely, *gametic phase disequilibrium* (Crow and Kimura 1970), or *gametic disequilibrium* for short, refers to the nonrandom association of alleles at different loci into *gametes*. It should be pointed out that the nonrandom association of alleles also could arise for *unlinked* loci (Turner 1971; Sinnock and Sing 1972; Smouse and Neel 1977; Weir and Cockerham 1989). The discovery of LD dates back to 1909, when Weinberg (1909) noted that, in a random-mating population, the alleles at two loci approach a random association only asymptotically. Shortly thereafter, Jennings (1917) and Robbins (1918) described the actual mode of approach to equilibrium frequencies, for the two-locus model. Since then, the population genetics of LD has been studied extensively. It is now generally understood that many factors, such as selection, admixture, finite population size, migration and mutation, coancestry, genetic hitchhiking, and growing population size, can affect LD (e.g., see Kojima and Schaffer 1967, Hill and Robertson 1968, Karlin 1969, Ohta and Kimura 1969, Weir et al. 1972, Nei and Li 1973, Hill 1976, Thomson 1977, Hedrick 1980, and Slatkin 1994).

LD mapping (LDM) is based on the following phenomenon (Hästbacka et al. 1992; Jorde 1995; Kaplan et al. 1995). When a chromosome carrying a disease allele is first introduced into a population as a result of either mutation or migration, the mutant allele is on a chromosome with a unique set of marker alleles (i.e., the haplotype). As the chromosome is propagated in the following generations, the length of the characteristic haplotype decreases monotonely and stochastically, with each generation. As a result of recombination, markers in the immediate vicinity of the disease locus are more likely to remain in the same strand than those farther away. Since the number of recombinations that accumulate through many generations is far greater than that observed in or inferred from any pedigree-based linkage study, the mapping resolution achieved through the analysis of LD patterns is much higher than that of linkage studies. Thus, it is possible to map genes at a scale finer than 1 cM by the identification of markers that are in strong LD with the disease allele. The so-called fineness of the map depends on how many generations have passed since the introduction of the mutation.

Received April 5, 1996; accepted for publication April 1, 1997.

Address for correspondence and reprints: Dr. Sun-Wei Guo, Institute of Human Genetics and Department of Epidemiology, School of Public Health, University of Minnesota, 1300 South Second Street, Suite 300, Minneapolis, MN 55454-1015. E-mail: swguo@med.umn.edu

0002-9297/97/6006-0031$02.00

LDM can be complicated by many factors. Mutations at marker loci and recurrent mutations at the disease locus can obscure the LD patterns observed in the neighborhood of the disease locus. Other factors, such as drift, selection, population stratification or admixture, the unknown age of the mutant allele, and nonrandom sampling, also can create difficulties in LDM.

Although Fisher (1947) had inferred decades ago the locus order for the rhesus factor, on the basis of gametic frequencies, the application of fine-scale mapping based on LD is fairly recent, compared with traditional linkage analysis. This is probably because the need for fine-scale mapping becomes pressing only when coarse-scale mapping becomes routine. Furthermore, unlike linkage analysis, the methodological development of LDM also requires profound knowledge of population genetics.

Bodmer (1986) appears to be the first to have advocated the use of LD for fine-scale mapping of a human population. Lander and Botstein (1986) proposed the use of LDM for recent genetic isolates, in lieu of the use of linkage analysis based on family data. Although some researchers argued that LD could not be used for fine-scale mapping (Weir 1989; Hill and Weir 1994), remarkable successes in fine-scale mapping based on LD quickly dispelled this view (Cox 1989; Snell et al. 1989; Theilmann et al. 1989; Hästbacka et al. 1992, 1994; MacDonald et al. 1992; Huntington's Disease Collaborative Research Group 1993; A. Chakravarti, personal communication).

These successes led gene mappers to embrace LDM as a promising tool for fine-mapping and to develop better theoretical methods. For example, Terwilliger (1995) proposed a likelihood method for LDM, on the basis of one or more marker loci, without assuming the evolutionary history of the population. In contrast, Kaplan et al. (1995) used a Poisson branching process to model a growing population. By simulating the evolutionary history of the population, they provided estimates for the location of the disease gene, on the basis of a likelihood function. This likelihood approach provides a more reliable estimate of confidence limits for the recombination fraction than does the Luria-Delbrück-type model used by Hästbacka et al. (1992). The method also can evaluate the order of a disease locus and two marker loci. On the basis of a similar model, Kaplan and Weir (1995) investigated the effects of mutation, at either the marker or the disease locus, on the upper boundaries of the recombination-fraction estimate. Their results showed that their approach is superior to the method based on the Luria-Delbrück-type model.

However, the approach proposed by Kaplan et al. (1995) is not without its shortcomings. It is difficult for simulation methods (SIM) to provide solutions to statistical inference problems, such as properties of estimators and sample-size requirements, which are important for the practical use of the method. SIM also is difficult for practitioners to use. Furthermore, there is an added problem of sampling variations due to simulation, which may demand a large number of replicates.

There are many other unresolved issues in LDM. Is the assumption of exponential expansion of the population, as made by Hästbacka et al. (1992) and Kaplan et al. (1995), or any assumption about population growth, indispensable for LDM? Under what circumstances can apparently nonassociated marker alleles be lumped into one group? In the neighborhood of the disease locus, why do some markers show strong LD whereas others do not? How can frequencies of alleles associated with the disease be lower than those in the normal population?

Without an appropriate framework, it is difficult to answer these questions. It will be difficult to use LDM to finely map disease genes, in the face of factors such as marker mutation, recurrent mutations at the disease locus, and unknown population growth rate. In fact, for some recently developed methods for LDM, fine-scale gene mapping for diseases like Huntington disease (HD) and Friedreich ataxia (FA) still poses a challenge (Kaplan et al. 1995) and raises the question of how useful the LDM methods are (Jorde 1995). Indeed, if Kaplan et al. (1995) are correct in their suspicion that LDM only works for some simple monogenic diseases, then its utility would be very limited.

In this paper, we present a general, mathematical, and conceptually coherent framework for LDM that incorporates multilocus and multiallelic markers and mutational processes at the marker and disease loci. Under this framework, the issues raised above can be resolved readily. The framework still assumes a homogeneous population, but it is not limited to an exponentially growing population. We show that our framework encompasses several existing LDM methods as special cases.

We also propose some efficient computational methods for LDM. We then demonstrate these methods by applying them to data published prior to cloning of the genes for cystic fibrosis (CF), HD, FA, and progressive myoclonus epilepsy (EPM1). The genes for these diseases all have been cloned. Thus, the exact locations of these genes are known, and these data provide a useful benchmark for the evaluation and comparison of various LDM methods, including ours.

We demonstrate that our proposed methods perform remarkably well for these data. Thus, we believe that the utility and scope of LDM, if carried out appropriately, is wider than previously thought. Finally, we provide some general considerations for LDM and describe areas for further research.

The Likelihood Function for LDM

Consider a disease locus with two alleles, a disease allele, *d,* and a normal allele, *n.* At the linked marker

locus, there are m alleles M_i ($i = 1, \ldots, m$). The recombination fraction between the two loci is assumed to be θ. Following Kaplan et al. (1995), let k_n and k_d be sample sizes from the normal and disease chromosomes, respectively. Also, let p_{i_n} and p_{i_d} ($i = 1, \ldots, m$) be the marker allele frequencies for allele M_i, for the normal and disease chromosomes, respectively. Note that $\Sigma_{i=1}^{m} p_{i_n} = 1$ and $\Sigma_{i=1}^{m} p_{i_d} = 1$. For relatively young diseases, marker allele frequencies in normal chromosomes will be assumed to be constant over time, but, within the disease population, the frequencies will be assumed to change over time. Therefore, frequencies $p_{i_d}(t)$ ($i = 1, \ldots, m$) are time dependent. For notational convenience, we suppress t. Here, time is measured in generations, with $t = G$ being the generation from which the samples are taken. For simplicity, we assume that generations are nonoverlapping. More methods of estimation of the age of the mutant allele have been developed (S.-W. Guo and M. Xiong, unpublished data).

With the random union of gametes, replacement from the disease population, and random sampling, the conditional probability of obtaining the sample, given the marker allele frequencies $P(t) = [p_{1_d}, \ldots, p_{m_d}]^T$, follows the multinomial distribution

$$f[k_{1_d}, \ldots, k_{m_d} | P(t)] = \frac{k_d!}{\prod_{i=1}^{m} k_{i_d}!} \prod_{i=1}^{m} p_{i_d}^{k_{i_d}}, \qquad (1)$$

where k_{i_d} is the observed number of disease chromosomes carrying allele M_i ($i = 1, \ldots, m$).

Marker-frequency changes between generations are governed by a Wright-Fisher population-genetics model. Evolutionary forces, such as random drift, mutation, and recombination, will cause marker frequency p_{i_d} to change stochastically. Therefore, frequency p_{i_d} at any generation t is a *random variable*. Taking the expectation of equation (1) over $P(t)$, we obtain the unconditional sampling distribution

$$f(k_{1_d}, \ldots, k_{m_d}) = \frac{k_d!}{\prod_{i=1}^{m} k_{i_d}!} \mathrm{E}\left(\prod_{i=1}^{m} p_{i_d}^{k_{i_d}} \right) \qquad (2)$$

(Hill and Weir 1994). In general, $\mathrm{E}(p_{i_d})$ is a function of θ. Therefore, $f(k_{1_d}, \ldots, k_{m_d})$ is the likelihood function of θ. Ignoring the constant term, we define the likelihood function $l(\theta)$ as

$$l(\theta) = \mathrm{E}\left(\prod_{i=1}^{m} p_{i_d}^{k_{i_d}} \right). \qquad (3)$$

To obtain the maximum-likelihood estimate of θ, we need to evaluate the likelihood function $l(\theta)$. It should be noted that the simple form of the above likelihood function is deceptive. Since the expectation is taken over $P(t)$, which contains all the evolutionary history of p_{i_d} prior to t, the likelihood function is actually very difficult to evaluate. Kaplan et al. (1995) approached the problem by simulation. That is, they simulated the evolutionary history, given a set of population and genetic parameters, and let

$$\mathrm{E}\left(\prod_{i=1}^{m} p_{i_d}^{k_{i_d}} \right) \approx \frac{1}{J} \sum_{j=1}^{J} \prod_{i=1}^{m} p_{i_d}^{k_{i_d}}(j), \qquad (4)$$

where $p_{i_d}(j)$ is the jth simulated realization of random variable p_{i_d}, in J realizations.

This is a standard maneuver of approximating an expectation by a sample mean, by use of the Monte Carlo method (Hammersley and Handscomb 1964). With a given population model (e.g., the Poisson branching process), the likelihood can be evaluated approximately for any given θ.

However, there are several problems associated with this approximation. First, although in principle any degree of accuracy of the approximation can be achieved at the cost of more computation time, the number of replicates needed for a desired accuracy is hard to determine a priori for a specific problem, since, in general, it depends on various factors. Second, as a result of the Monte Carlo approach, the estimate of θ is subject to variations in the Monte Carlo sampling, in addition to statistical uncertainty. Similarly, the boundaries computed by the simulation also are subject to variations in the Monte Carlo sampling. Third, the simulation is subject to several constraints imposed by the data. For example, the simulated evolutionary history that gives rise to values for the total number of disease chromosomes in the population has to be close to the estimated value. In addition, there is a nonnegligible chance that one or more alleles at the marker locus could reach fixation or extinction in simulation. This problem may be more acute for biallelic markers. In reality, of course, we would not have used the nonpolymorphic marker in the first place. Thus, the SIM of Kaplan et al. (1995), which is basically a rejection sampling scheme, may not be entirely realistic or computationally efficient.

Here we present a computationally economical approximation, which allows us to consider more complex genetic models and provides more insight into LD between the marker and the disease loci. Let $\mu_i(t) = \mathrm{E}(p_{i_d})$ ($i = 1, \ldots, m$), $\mu(t) = [\mu_1(t), \ldots, \mu_{m-1}(t)]^T$, $h[p_{1_d}(t), \ldots, p_{m_d}(t)] = \prod_{i=1}^{m} p_{i_d}^{k_{i_d}}$, and the Hessian matrix of $h[p_{1_d}(t), \ldots, p_{m_d}(t)]$ be

$$H(t) = \left(\frac{\partial^2 h}{\partial p_{i_d} \partial p_{j_d}} \bigg|_{P(t)=\mu(t)} \right)_{(m-1)\times(m-1)}.$$

Noting that

$$\begin{aligned} &\mathrm{E}\{[P_0(t) - \mu(t)]^T H(t)[P_0(t) - \mu(t)]\} \\ &\quad = tr[H(t)D(t)] - \mu(t)^T H(t)\mu(t) , \end{aligned} \tag{5}$$

where $D(t) = \mathrm{E}[P_0(t)P_0^T(t)]$, $P_0(t) = (p_{1_d}, \ldots, p_{m-1_d})^T$, and *tr* denotes the trace of the matrix, we obtain the first-order approximation (FOA) to the likelihood function

$$l(\theta) \approx \prod_{i=1}^{m} {\mu_i}^{k_{i_d}} \tag{6}$$

and the second-order approximation to the likelihood function

$$l(\theta) \approx \prod_{i=1}^{m} {\mu_i}^{k_{i_d}} + {}^1\!/_2\{tr[H(t)D(t)] - \mu^T(t)H(t)\mu(t)\} . \tag{7}$$

When the marker has only two alleles—that is, when $m = 2$—equation (7) becomes

$$l(\theta) \approx {\mu_1}^{k_{1_d}}(1 - \mu_1)^{k_{2_d}} + {}^1\!/_2 H(t)[E(p_{1_d}^2) - \mu_1^2] ,$$

where

$$\begin{aligned} H(t) = {} & k_{1_d}(k_{1_d} - 1){\mu_1}^{k_{1_d}-2}(1 - \mu_1)^{k_{2_d}} \\ & - 2k_{1_d}k_{2_d}{\mu_1}^{k_{1_d}-1}(1 - \mu_1)^{k_{2_d}-1} \\ & + k_{2_d}(k_{2_d} - 1){\mu_1}^{k_{1_d}}(1 - \mu_1)^{k_{2_d}-2} . \end{aligned}$$

We note that the above approximations hold in form *regardless* of the population-genetics model considered.

FOA- and Second-Order Approximation

To evaluate the approximate likelihood functions (6) and (7), it is necessary to calculate the first and second moments of the marker frequencies p_{i_d} ($i = 1, \ldots, m$). The marker frequency p_{i_d} is a random variable subject to evolutionary forces, such as recombination, mutation, and migration. In order to compute the first two moments, we need to specify a population-genetics model for marker frequencies.

For simplicity, we assume that there is no substructure in the population and that there is random mating in the population. As Kaplan et al. (1995) pointed out, although there may be a selective advantage for carriers, for practical purposes all carrier individuals can be assumed to be selectively equivalent. It is easy to see that this assumption is reasonable for a recessive disease. For a dominant disease, the assumption of selective equivalence also may be reasonable for late-onset (postreproductive age) diseases.

Since microsatellite markers usually have high mutation rates (from $\sim 10^{-3}$ to $\sim 10^{-5}$) (Weber and Wong 1993), their use may obscure the LD patterns. Thus, it is appropriate to consider the mutation at the marker locus, in LDM. We assume that the mutations at microsatellite loci occur according to a stepwise mutation model (SMM) (M. Xiong and S.-W. Guo, unpublished data). The SMM stipulates that the repeat number changes by a few, as a result of mutation (Ohta and Kimura 1973; Shriver et al. 1993; Valdes et al. 1993; M. Xiong and S.-W. Guo, unpublished data). For ease of exposition, we consider a one-step SMM in which there is only one repeat change in the event of a mutation. Extension to a multistep SMM is straightforward but may be more complicated.

We consider multiple alleles for the microsatellite markers and assume that allele M_i, indexed according to the number of repeats, can mutate to the next-larger allelic state, M_{i+1} (i.e., expansion), with probability u, and to the next-smaller allelic state, M_{i-1} (i.e., contraction), with probability v. Let M_1 denote the allele with the smallest number of repeats and M_m denote the allele with the largest number of repeats. We assume that allele M_1 can mutate only to allele M_2 and that allele M_m can mutate only to allele M_{m-1}. For diallelic loci, let u be the forward mutation rate for allele M_1 mutating to M_2 and v be the backward mutation rate.

We also assume that disease is due to mutations of a normal allele to a disease allele. Backward mutation is assumed to be negligible. Let γ_d be a disease-allele mutation rate and p_d be the disease-allele frequency. In generation t, suppose that there are $X_i(t)$ disease chromosomes carrying marker allele M_i, and $X_T(t) = \Sigma_{i=1}^m X_i(t)$ total disease chromosomes in the population.

We consider a two-locus Wright-Fisher model for mutation, recombination, and random genetic drift. The joint evolutionary process of the marker allele frequency p_{i_d} can be approximated by use of a diffusion process (see Appendix A).

It can be shown (see Appendix B) that the first two moments of p_{i_d} satisfy the following ordinary differential equations:

$$\frac{d\mathrm{E}[p_{i_d}(t)]}{dt} = \mathrm{E}[g_i(t)] , \; i = 1, \ldots, m ; \tag{8}$$

$$\begin{aligned} \frac{d\mathrm{E}[p_{i_d}(t)p_{j_d}(t)]}{dt} = {} & -\mathrm{E}\left[\frac{p_{i_d}(t)p_{j_d}(t)}{X_T(t)}\right] \\ & + \mathrm{E}[g_i(t)p_{j_d}(t)] + \mathrm{E}[g_j(t)p_{i_d}(t)] , \\ & i \neq j ; \end{aligned} \tag{9}$$

and

$$\begin{aligned} \frac{d\mathrm{E}[p_{i_d}^2(t)]}{dt} = \mathrm{E}\left\{\frac{p_{i_d}(t)[1 - p_{i_d}(t)]}{X_T(t)}\right\} + 2\mathrm{E}[g_i(t)p_{i_d}(t)] , \\ i = 1, \ldots, m , \end{aligned} \tag{10}$$

where $g_i(t)$ is defined in Appendix A. Note that equation (8) can be rewritten in a matrix form as follows:

$$\frac{d\mu(t)}{dt} = A\mu(t) + B , \tag{11}$$

where A is a matrix that depends on θ, disease-allele frequency, recurrent-mutation rate, and marker mutation rates (see Appendix C) and where $B = (b_1, \ldots, b_m)^T$ with $b_1 = (1 - u)\alpha p_{1_n} + v\alpha p_{2_n}$, $b_i = u\alpha p_{i-1_n} + [1 - (u + v)]\alpha p_{i_n} + v\alpha p_{i+1_n}$, in which $i = 2, \ldots, m - 1$, and $b_m = u\alpha p_{m-1_n} + (1 - v)\alpha p_{m_n}$, where α is a function of θ, disease-allele frequency, recurrent-mutation rate, and marker mutation rates (see Appendix C).

Solving equation (11) for $\mu(t)$ (see Appendix C) yields

$$\mu(t) = e^{At}\mu(0) + A^{-1}(e^{At} - I)B , \tag{12}$$

where $\mu(0) = [p_{1_d}(0), \ldots, p_{m_d}(0)]^T$ is a vector of the initial values for p_{i_d}, I is an identity matrix, and $\exp(At)$ denotes an exponential matrix defined by $e^{At} = I + \Sigma_{k=1}^{\infty} (At)^k/k!$. Equation (12) provides a nice explanation of the dynamics of marker allele distribution in the disease population. The expected marker allele frequencies at generation t is a function of two components: the first is the initial distribution of marker alleles and its evolution through cumulative recombination and mutation, and the second involves the evolution of marker allele frequencies in the normal population, as a function of time, recombination, and mutation. Thus, as t increases, the expected marker allele frequency in the disease population approaches that in the normal population, that is, eventual equilibrium.

To see this more clearly, we assumed that initially there is complete LD between the marker and the disease loci—that is, $p_{1_d}(0) = 1$, $p_{j_d}(0) = 0$, $j = 1, \ldots, m$, and $j \neq 1$—and that there is no mutation at either the marker locus or the disease locus (i.e., $u = v = \gamma_d = 0$). Then, equation (12) can be simplified to $E(p_{1_d}) = e^{-\theta t} + (1 - e^{-\theta t})p_{1_n}$ and

$$\begin{aligned} E(p_{j_d}) &= (1 - e^{-\theta t})p_{j_n} , \\ j &= 1, \ldots, m \text{ and } j \neq 1 . \end{aligned} \tag{13}$$

We point out that the result obtained by Cox et al. (1989) is a special case of equation (13).

It is interesting to note that the first moments of p_{i_d} can be computed regardless of how the disease population or the normal population changes with time. This feature has an important implication: If we have little knowledge of how a population of interest changes with time, we just may use the FOA to the likelihood of equation (3) for fine-mapping purposes. The computation of second moments is similar to that of first moments and is outlined in Appendix C.

Extensions to Multiple Marker Loci

The above approach can be extended to include multiple marker loci. For ease of exposition, we only discuss the extension to two-locus haplotype data and the composite likelihood for multilocus LDM based on multilocus nonhaplotype data. Extensions to multilocus haplotype data are straightforward but more complicated.

Two-Locus Haplotype Data

For two-locus haplotype data, there are three possible orderings—marker$_1$–disease–marker$_2$, marker$_1$–marker$_2$–disease, and disease–marker$_1$–marker$_2$. We only discuss the case of marker$_1$–disease–marker$_2$, since the other two cases can be dealt with in a similar fashion.

We denote p_{ij_d} as the conditional frequency of haplotype C_i–C_j in disease chromosomes. Let θ_k be the θ between the disease locus and the kth ($k = 1, 2$) marker.

By use of a similar argument as that used for one marker locus, the evolutionary process of the marker frequency p_{ij_d} ($i = 1, \ldots, m$ and $j = 1, \ldots, m$) also can be approximated by use of a diffusion process (see Appendix D). It can be shown that the expectation of the haplotype frequency in the disease population, p_{ij_d}, satisfies

$$\frac{dE(p_{ij_d})}{dt} = E[g_{ij}(t)] , \tag{14}$$

where $g_{ij}(t)$ is defined in Appendix D. If there is no mutation at either the marker locus or the disease locus—that is, $u = v = \gamma_d = 0$—then equation (14) reduces to

$$\begin{aligned} \frac{dE(p_{ij_d})}{dt} &= -(\theta_1 + \theta_2)E(p_{ij_d}) + \theta_1 p_{i_{\cdot n}} E(p_{\cdot j_d}) \\ &\quad + \theta_2 p_{\cdot j_n} E(p_{i_{\cdot d}}), \\ i &= 1, \ldots, m_1 \text{ and } j = 1, \ldots, m_{2\cdot} , \end{aligned} \tag{15}$$

where the dot subscript indicates summation over all values of the corresponding index. Solving the above equations for $E(p_{ij_d})$ yields

$$\begin{aligned} E(p_{ij_d}) &= [p_{ij_d}(0) - \beta_1 - \beta_2 - p_{i_{\cdot n}} p_{\cdot j_n}]e^{-(\theta_1 + \theta_2)t} \\ &\quad + \beta_1 e^{-\theta_1 t} + \beta_2 e^{-\theta_2} + p_{i_{\cdot n}} p_{\cdot j_n} , \end{aligned}$$

where $\beta_1 = p_{\cdot j_n}[p_{i_{\cdot d}}(0) - p_{i_{\cdot n}}]$, $\beta_2 = p_{i_{\cdot n}}[p_{\cdot j_d}(0) - p_{\cdot j_n}]$, and $p_{ij_d}(0)$ is a set of initial values of the conditional haplotype frequencies.

The second moment of marker frequencies also can

be derived. In particular, if there is no mutation at either the marker locus or the disease locus—that is, $u = v = \gamma_d = 0$—it can be shown that

$$\frac{d\mathrm{E}(p_{ij_d}^2)}{dt} = -\left[\frac{1}{X_T(t)} + 2\theta_1 + 2\theta_2\right]\mathrm{E}(p_{ij_d}^2) + \frac{1}{X_T(t)}\mathrm{E}(p_{ij_d}) + 2\theta_1 p_{i,_n}\mathrm{E}(p_{\cdot j_d}p_{ij_d}) + 2\theta_2 p_{\cdot j_n}\mathrm{E}(p_{i,_d}p_{ij_d}) \quad (16)$$

and that

$$\frac{d\mathrm{E}(p_{ij_d}p_{kl_d})}{dt} = -\left[\frac{1}{X_T(t)} + 2\theta_1 + 2\theta_2\right]\mathrm{E}(p_{ij_d}p_{kl_d}) + \theta_1[p_{i,_n}\mathrm{E}(p_{\cdot j_d}p_{kl_d}) + p_{k,_n}\mathrm{E}(p_{\cdot l_d}p_{ij_d})] + \theta_2[p_{\cdot j_n}\mathrm{E}(p_{i,_d}p_{kl_d}) + p_{\cdot l_n}\mathrm{E}(p_{k,_d}p_{ij_d})] . \quad (17)$$

Multilocus Nonhaplotype Data

Whereas multilocus haplotype data may be difficult to obtain in some cases, single-locus data can be relatively easier to obtain for multiple loci. Analogous to the location score in multipoint-linkage analysis (Ott 1991), we also can compute the location score for multipoint LDM, using Haldane's (1919) map function,

$$\theta = {}^1\!/_2(1 - e^{-2l}) . \quad (18)$$

Suppose that $k + 1$ markers are located at chromosomes that are in accordance with the order C_0, C_1, . . . , C_k. Let l_i denote the map distance between markers C_i and C_{i-1} ($i = 1, \ldots, k$). Let x denote the distance between the disease locus and marker C_0. Then, from equation (18), θ_j, between marker C_j ($1 \leq j \leq k$) and the disease locus, is given by

$$\theta_j = {}^1\!/_2(1 - e^{-2|x-\sum_{i=1}^{j} l_i|}) . \quad (19)$$

We define the likelihood function L_j of θ_j as $L_j = \prod_{i=1}^{m} p_{i_d}^{k_{i_d}(j)}$, where p_{i,j_d} denotes the frequency of allele M_i at C_j, in the disease population, and $k_{i_d}(j)$ denotes the observed number of allele M_i at C_j, sampled from the disease population. Then, the logarithm of the overall likelihood function L across all markers is defined as

$$l = \sum_{j=1}^{k} \log L_j . \quad (20)$$

Let $\mu_i(j) = \mathrm{E}[p_{i_d}(j)]$. From the previous discussion, when mutations can be ignored,

$$\mu_i(j) = p_{ij_d}(0)e^{-\theta_j t} + (1 - e^{-\theta_j t})p_{i,j_n} , \quad (21)$$

where $p_{ij_d}(0)$ is an initial value of the frequency of allele M_i at C_j and where p_{i,j_n} is the frequency of the allele M_i at C_j, in the normal population. Thus, the FOA to l is given by $l_a = \sum_{j=1}^{k}\sum_{i=1}^{m} k_{i_d}(j)\log \mu_i(j)$.

Similarly, we can determine the second-order approximation to l. Because the extension of previous results is straightforward, we omit details.

It should be pointed out that, strictly speaking, equation (20) is not a likelihood, because it implicitly assumes that marker frequencies at different loci are independent. For markers that are closely linked, this clearly is not true. Without knowing the exact dependencies in marker frequencies among the markers, equation (20) is at least an FOA to the true, yet unknown, likelihood. For this reason, we will call the likelihood expressed in equation (20) the "composite likelihood."

Some Implications of the Proposed Model

We point out two immediate implications of our proposed model: First, most investigators have concentrated on the simplest cases, in which there are two types of alleles at the marker locus—associated and nonassociated alleles. This may be reasonable if there is a single ancestral mutation in the population. However, if there are multiple disease mutations or multiple founders carrying different mutations, then focusing on the simplest case no longer may be sufficient. One way to deal with this situation is to specify initial values for $p_{i_d}(0)$, where $i = 1, \ldots, m$. Of course, these values usually are unknown. However, since all disease alleles are assumed to be selectively neutral, the marker frequencies in the current population may be an approximation to the frequencies at the time the mutation(s) was introduced. Suppose that there are r alleles with disease mutations, indexed by $i_{1_d}, \ldots, i_{r_d}$. Let $\hat{p}_{i_{1_d}}, \ldots, \hat{p}_{i_{r_d}}$ be the observed marker frequencies within the disease population. Let $\hat{p}_0 = \sum_{j=1}^{r} \hat{p}_{i_{j_d}}$. Then, we may simply specify $p_{i_{j_d}}(0)$ as $p_{i_{j_d}}(0) = \hat{p}_{i_{j_d}}/\hat{p}_0$, where $j = 1, \ldots, r$, and, for other alleles, let their initial values be 0. After specifying initial values $p_{i_d}(0)$, we obtain, by solving equation (8) for $\mathrm{E}(p_{i_d})$,

$$\mathrm{E}(p_{i_d}) = p_{i_d}(0)e^{-\theta t} + (1 - e^{-\theta t})p_{i_n} , \quad i = 1, \ldots, m . \quad (22)$$

That is, the current frequency of the associated allele consists of two parts: one is the attenuation of the initial frequency (owing to recombination) and the gradual attainment to the frequency of the same allele in the normal population.

Second, in the disease population, the frequency of the associated marker allele usually is assumed to be higher than in the normal population. Both Kaplan et al. (1995) and Terwilliger (1995) build this assumption

into their models. Indeed, in most cases this assumption is true. This assumption also is sensible because, if it was observed to be otherwise, the marker would not be identified as in LD with the disease locus. However, in many practical situations, it is often the case that, in a region that is supposedly linked to the disease locus, some markers show strong LD with the disease locus whereas others do not, despite the fact that they all may be linked to the disease. One can find such examples in an FA data set considered by Pandolfo et al. (1990).

We offer three explanations of why this may happen sometimes. The first is that the inequality $E(p_{1_d}) > p_{1_n}$, where allele 1 is associated with the disease, is *stochastic* in nature. It may be violated in some observed samples. The second is that there may be early recombinations between the marker locus and the disease locus or that recurrent mutations may have occurred in the past. If either of these events happens, then it is possible that $p_{1_d}(0) < p_{1_n}$, which implies that $E(p_{1_d}) = p_{1_n} + e^{-\theta t}[p_{1_d}(0) - p_{1_n}] < p_{1_n}$.

A third explanation is that there may be unequal mutation rates at the marker locus. If this happens, the frequency of a marker allele associated with the disease-allele mutations is no longer required to be higher in the disease population than in the normal population. To see this, suppose that there are two alleles at the marker locus. For the sake of argument, suppose also that there is no recurrent mutation and no backward mutation (which is equivalent to $\gamma_d = \nu = 0$ but $u > 0$). Suppose further that the mutant disease allele initially is in complete LD with M_1. Thus, $p_{1_d}(0) = 1$ and $p_{2_d}(0) = 0$. In this situation, equation (8) is reduced to

$$\frac{dE(p_{1_d})}{dt} = -[\alpha + u(1 - \alpha)]E(p_{1_d}) + (1 - u)\alpha p_{1_n} .$$

When this equation is solved for $E(p_{1_d})$,

$$E(p_{1_d}) = e^{-[\alpha+u(1-\alpha)]t} + \frac{(1 - u)\alpha}{u + (1 - u)\alpha} p_{1_n}[1 - e^{-[\alpha+u(1-\alpha)]t}] . \quad (23)$$

It is clear that in this case $[(1 - u)\alpha]/[u + (1 - u)\alpha]p_{1_n} < p_{1_n}$. Thus, for a t that is large enough, it is possible that $E(p_{1_d}) < p_{1_n}$. Intuitively, when marker allele M_1, associated with the disease allele, mutates to marker allele M_2, both mutation and recombination will reduce marker frequency p_{1_d}. Reduction of $E(p_{1_d})$ owing to recombination will have lower boundary p_{1_n}, but reduction owing to mutation will not be restricted by p_{1_n}.

Connections among Existing LDM Methods

On the basis of the results we have obtained so far, it is possible to relate some existing LDM methods. Rewriting equation (13), we have $1 - E(p_{1_d}) = (1 - e^{-\theta t})(1 - p_{1_n}) \approx 1 - e^{-\theta t}$, provided $p_{1_n} \approx 0$. Now, if we replace $E(p_{i_d})$ by its sample estimate, $\hat{p}_{1_d}$, obtained from the *current* population, we have $1 - \hat{p}_{1_d} \approx 1 - e^{-\theta t}$, which was used by Hästbacka et al. (1992) as one way to estimate θ. Obviously, this estimate is very crude if p_{1_n} is nonnegligible. It is somewhat surprising that the same formula can be derived *without* the assumption of exponential population growth.

The method proposed by Terwilliger (1995) also is a special case of our FOA to the likelihood. To see this, we define λ, using Terwilliger's notation, to satisfy $q_1 = p_1 + \lambda(1 - p_1)$ and $q_i = p_i - \lambda p_i$ $(i \neq 1)$, where q_1 and q_i (i.e., p_{1_d} and p_{i_d} in our notation), are the conditional frequencies of the putative ancestral allele and of other nonancestral alleles, respectively, in the disease chromosomes, and where p_1 and p_i $(i \neq 1)$ are the population frequencies of the progenitor allele and of other alleles, respectively, which are approximately equal to p_{1_n} and p_{i_n} in our notation (assuming that the disease is rare). Denoting $r_i = p_{i_n}$, Terwilliger (1995) proposed the following likelihood function for θ:

$$L = \prod_{j=1}^{m} q_j^{k_{j_d}} r_j^{k_{i_n}} .$$

If we let $\lambda = e^{-\theta t}$, then

$$q_1 \approx p_{1_n} + e^{-\theta t}(1 - p_{1_n}) = e^{-\theta t} + (1 - e^{-\theta t})p_{1_n} ;$$

$$q_i \approx (1 - e^{-\theta t})p_{i_n} , \; i \neq 1 ,$$

which is exactly our FOA to the likelihood in equation (3), in the absence of marker mutation and recurrent mutation and when there is initially complete LD. Terwilliger did point out that λ should be roughly proportional to $(1 - \theta)^t$ (Terwilliger 1995, p. 780), which equals $e^{-\theta t}$ when θ is small, just as we showed above.

Terwilliger (1995) introduced an additional parameter, α, which can be thought of as the proportion of disease chromosomes that are identical, by descent from a common founder chromosome (p. 780). In this case, $\lambda = \alpha(1 - \theta)^t \approx \alpha e^{-\theta t}$. Thus,

$$q_1 \approx \alpha e^{-\theta t} + (1 - \alpha e^{-\theta t})p_{1_n} ;$$

$$q_i \approx (1 - \alpha e^{-\theta t})p_{i_n} , \; i \neq 1 . \quad (24)$$

To incorporate this heterogeneity, we let $q_1(0) < 1$; that is, there is an incomplete LD initially. Replacing q_1 with $E(p_{1_d})$ in equation (22), we have

$$q_1 \approx e^{-\theta t}q_1(0) + (1 - e^{-\theta t})p_{1_n} ;$$

$$q_i \approx e^{-\theta t}q_i(0) + (1 - e^{-\theta t})p_{i_n} , \; i \neq 1 , \quad (25)$$

which is somewhat different from equation (24). It also can be shown that our result is different from equation (24), even if there are mutations at the marker locus and/or the disease locus. We note that equation (25) has a very nice interpretation. The current pool of disease chromosomes comes from two sources: one is descended from the common ancestral chromosome that underwent no recombination between the marker locus and the disease locus, and the other is descended from normal chromosomes that recombined with the disease chromosomes. We also note that the likelihood derived by Terwilliger (1995) can be embedded in our composite likelihood, which is an approximation. Since $q_i(0)$ has a much clearer meaning in our model and because the derivation of equation (25) was based on a dynamic population-genetics model, we expect that our method should perform better.

Numerical Examples

To illustrate our proposed LDM methods, we applied them to four genetic diseases, CF, HD, FA, and EPM1, for which the genes all have been cloned. Since the physical distance between the disease loci and their surrounding markers now are known, LD data published prior to cloning provides an opportunity to evaluate the performance of our methods and to compare our methods with that of others.

We chose the CF data set because it has been well analyzed by different researchers and can serve as a yardstick for comparison. The HD and FA data were chosen because the LD patterns for these two diseases were quite complicated, and no LDM method has been shown to be satisfactory.

Throughout our analysis, we used the empirical conversion rate of 1 cM $\approx$ 1,000 kb. We used the FOA and the second-order approximation, assuming a constant population size (SCP) and assuming an exponentially growing population (SEG). However, since the FOA works remarkably well, we used the SEG and SCP only for the CF example. When applicable, the results were compared with those obtained by the SIM of Kaplan et al. (1995), the Luria-Delbrück-type method (LDT) used by Hästbacka et al. (1992, 1994), and the method of Terwilliger (1995).

CF

The CF gene was cloned in 1989. The most common mutation, $\Delta F508$, accounts for >70% of Caucasian CF cases and was identified in a region flanked by markers 10-1x.6 (*Hae*III) and T6/20 (Kerem et al. 1989). Following Kaplan et al. (1995), we assumed that the CF mutation occurred ~200 generations ago ($G = 200$).

For the SEG model, following Kaplan et al. (1995), we assumed that the current number of disease chromosomes is $X_T(G) = 2 \times 10^7$ and, hence, that the population growth rate is $\lambda = 0.078$.

Table 1 summarizes the results. It can be seen that the results using FOA and SEG are almost identical in most cases and are in broad agreement with the true distance. SCP tends to overestimate the distance, whereas LDT tends to underestimate. For markers within 80 kb from the CF locus, however, LDT gives slightly better estimates. Table 2 shows the largest, the smallest, and the average absolute estimation errors of the four methods, for 19 markers. It can be seen that, for this data set, the accuracy of the estimations by FOA and by SEG is almost identical and is fairly satisfactory. The accuracy of the estimation by SCP is compatible with that of LDT but has a higher variation.

The nearly identical results obtained by FOA and SEG suggest that the assumption of exponential population growth is not critical to the accuracy of the estimation. With our proposed framework, the population size only affects the variance and covariance of allele frequencies in the diffusion process. Inappropriately specified population size, however, may affect the accuracy of the Taylor expansion. For this example, the FOA is good enough, and little is gained by the use of the second-order approximation. LDT, in general, is not as good as our two likelihood methods, although it is quite accurate when the markers are very close ($\leq$70 kb) to the CF locus. The formulation of Hästbacka et al. (1992) for estimation of θ involves the marker allele frequency in the disease population only but does not involve the marker allele frequency in the normal population, and, hence, it loses some information. Therefore, the accuracy of LDT may not be very satisfactory if the markers used are not very close to the disease locus.

It also is interesting to compare the support intervals obtained by use of the four methods. Following customary methods, we established support intervals for θ by decreasing the log likelihood by 2 units from its maximum value. For this example, the proportions of upper-support boundaries that are smaller than the actual distance are 16%, 10%, 10%, and 78% for FOA, SEG, SCP, and LDT, respectively. Since the second-order approximation more closely resembles the curvature of the true likelihood, it is not surprising that the support boundaries obtained by either SEG or SCP are better than those obtained by FOA. The boundaries obtained by FOA are not as good as those obtained by the second-order approximation, but they are reasonable. However, the upper boundaries obtained by LDT are somewhat disappointing. Similar conclusions were reached by Kaplan et al. (1995) and by Kaplan and Weir (1995).

We also used the multilocus composite likelihood, on the basis of information on the genetic distance among 23 markers (fig. 1). It can be seen that the composite likelihood reached its peak at 0.8 cM (or 800 kb) from

Table 1

Estimates of Genetic Distance between the CF Locus and Various Marker Loci, by Four LDM Methods

Marker	Actual Distance (kb)	Estimated Distance (kb), by[a]			
		FOA	SEG	SCP	LDT
E6	350	360 [180–740]	350 [110–510]	620 [480–990]	130 [120–150]
E7	340	340 [170–710]	340 [110–490]	580 [460–990]	130 [120–160]
pH131	320	350 [230–530]	350 [190–610]	480 [170–640]	240 [210–280]
W3D1.4	305	370 [240–560]	370 [190–650]	520 [310–690]	240 [210–280]
XV2C	280	220 [100–450]	210 [70–560]	380 [150–650]	110 [90–130]
*Hinc*II	260	80 [30–160]	75 [50–180]	140 [50–230]	60 [50–70]
*Bgl*II	240	90 [40–180]	90 [60–220]	150 [60–240]	74 [70–90]
KM19	220	100 [50–190]	100 [30–230]	160 [70–240]	80 [70–90]
E2.6	190	90 [30–220]	90 [20–270]	180 [60–300]	60 [50–80]
H2.8A	165	110 [50–210]	110 [70–260]	190 [20–290]	90 [80–100]
E4.1	130	130 [50–270]	120 [30–340]	220 [10–370]	70 [60–80]
J44	95	80 [30–180]	80 [10–230]	150 [50–260]	50 [50–60]
*Acc*I	15	140 [80–240]	140 [60–310]	470 [230–730]	120 [110–140]
*Hae*III	5	130 [70–230]	130 [50–310]	210 [50–290]	120 [110–140]
T6/20	15	250 [10–670]	40 [20–100]	110 [50–160]	70 [60–80]
H1.3	25	80 [30–180]	80 [50–120]	140 [50–230]	60 [50–70]
CE1.0	75	290 [50–1,000]	240 [10–490]	140 [30–260]	23 [20–30]
J3.11	660	730 [430–1,660]	740 [340–1,000]	1,310 [320–2,000]	280 [250–330]
J29	760	670 [400–1,260]	670 [330–890]	440 [310–860]	290 [250–340]

[a] The numbers in brackets are the estimated lower and upper support boundaries. In all calculations, a generation time of 200 and a conversion rate of 1 cM ≈ 1,000 kb were assumed. Data were taken from the study by Kerem et al. (1989).

marker metD (*Ban*I), as compared with the actual physical distance of ~875 kb. Thus, the error is only ~75 kb. This agreement suggests that the composite likelihood gives a more reliable estimation of the disease locus than the use of individual markers. Terwilliger (1995) applied his method to the same data set, yielding an estimate of 770 kb. Thus, in this sense, our method gives a somewhat more accurate estimate of the CF-gene location than that of Terwilliger.

We also investigated the impact of the choice of population-growth models by using SCP for the estimation of the location of the disease gene for marker E6. We found that the model assuming a large, constant population size is approximately equivalent to the model assuming an exponential growth (data not shown).

Table 2

Errors in the Estimation of the Location of the CF Gene, by Different LDM Methods

Method	Largest Error (kb)	Smallest Error (kb)	Average Error (kb)
FOA	240	0	90
SEG	240	0	90
SCP	650	10	170
LDT	470	45	160

One common opinion holds that LDM can be applied only to genetic diseases without recurrent mutations (e.g., Kaplan et al. 1995). Without recurrent mutations and with the barring of marker mutations, there is usually a predominant ancestral marker allele with a higher frequency in the disease population. However, this fre-

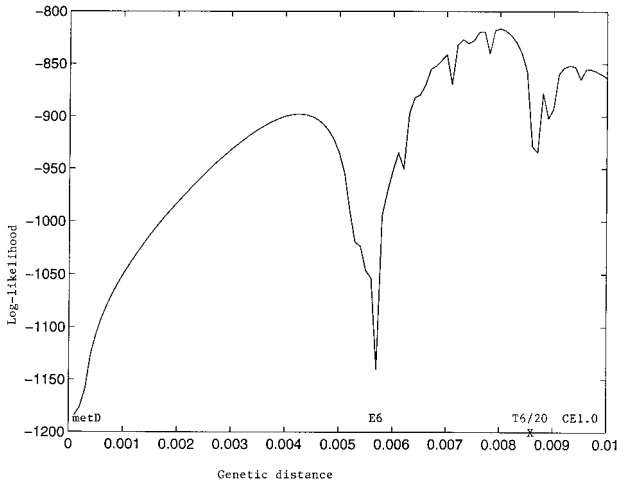

Figure 1 Composite log likelihood for estimation of the location of the CF locus, on the basis of 19 markers from E6 to J29. Marker metD (*Ban*I) is used as a reference point. The true location of the gene is marked by an "X."

Table 3

Estimates of Genetic Distance bertween the CF Locus and Various Marker Loci, by the FOA and SIM Methods, for the Finnish-Population Data

Marker	Actual Distance (kb)	Estimated Distance (kb), by[a]	
		FOA	SIM[b]
XV2C	280	170 [30–510]	300 [?–900]
KM19	220	370 [140–850]	600 [?–1,400]
Mp6d-9	130	250 [80–630]	400 [?–1,110]
G2	~70	220 [25–950]	400 [?–1,700]
J3.11	660	1,080 [540–2,490]	... [?–>2,000]

[a] The numbers in brackets are the estimated lower and upper support boundaries.

[b] Estimates are from the article by Kaplan and Weir (1995).

quency differential and its magnitude are determined by the distance between the marker and the disease locus. Markers *close* to the disease locus tend to have a predominant allele associated with the disease. This may not be true for markers farther away from the disease locus.

Kaplan et al. (1995) estimated, by using SIM, the distances between the CF gene and markers XV2C and KM19. They used data collected from several European populations and assumed that 200 generations was the age of the Δ*F*508 mutation in all the populations. These data sets may not be appropriate for the comparison of different LDM methods, because it is very likely that the age of the Δ*F*508 mutation is different in different populations. The likelihood that the same 3-bp deletion occurred more than once in different populations is much smaller than the likelihood that the Δ*F*508 mutation was introduced, by gene flow, at different times. Kaplan and Weir (1995) selected 5 of 11 markers (XV2C, KM19, Mp6d-9, G2, and J3.11) in the Finnish population to demonstrate their method. Using the same data set, we can compare our method with theirs, assuming 100 generations as the age of the CF disease mutation in the Finnish population (table 3). It can be seen that, in general, SIM considerably overestimates the distances. For markers, such as J3.11, that are not very close to the CF locus, SIM even failed to give a sensible estimation of the CF-gene location. We point out that estimates obtained by our method can be improved considerably if the age of mutation is estimated simultaneously, rather than fixed.

HD

In 1983, the gene responsible for HD was mapped to chromosome 4, by use of linkage analysis (Gusella et al. 1983). Haplotype analysis using multiallelic markers indicated that a 500-kb segment between D4S180 and D4S182 is the most likely site of the mutation (MacDonald et al. 1991). Subsequent work by the Huntington Disease Collaborative Research Group (1993) identified in this region a large gene, IT15, spanning ~210 kb, with an expandable unstable trinucleotide repeat, which is responsible for HD.

In the published HD data (MacDonald et al. 1991), marker allele frequencies have several patterns. There seem to be multiple ancestral haplotypes, but no single haplotype is predominant. Some markers show strong allelic associations with HD, but they are interspersed with intervening markers that show no association. Some markers that are linked to HD do not show any LD at all.

Following Kaplan et al. (1995), we assumed the age of the HD mutation to be $G = 200$ generations. This number agrees broadly with our estimate based on marker data (S.-W. Guo and M. Xiong, unpublished data). Because HD is a dominant disease and affects ~1/10,000 people of European descent, the frequency of the disease chromosomes is ~1/20,000.

It is now known that IT15, with an expandable unstable trinucleotide repeat, lies within the region between D4S180 and D4S182 or is 240 kb, 110 kb, and 250 kb away from D4S180, D4S95, and D4S182, respectively (D. A. Tagle, personal communication). Both D4S95/*Acc*I and D4S95/*Mbo*I show strong LD with the HD locus, but a nearby marker (*Taq*I) does not. Assuming no mutation at either the marker locus or the disease locus, our method placed the HD gene to be ~260 kb and ~290kb away from D4S95/*Mbo*I and D4S95/*Acc*I, respectively, which are ~150 kb and ~180 kb from the true location.

MacDonald et al. (1991) noted that the most common haplotypes on HD chromosomes differ in their D4S95/*Taq*I alleles. One factor that causes the lack of a predominant allele in the HD chromosomes could be the mutation at marker loci. Such a mutation process would decrease the frequency of the progenitor allele and increase the frequency of the other allele, in HD chromosomes. To examine this scenario, we estimated the distance between the marker D4S95/*Taq*I and the HD locus and the marker mutation rates. The mutation rate was estimated to be $\sim 2 \times 10^{-3}$, and the distance was ~330 kb, as compared with the true distance of 110 kb. When this model was extended to D4S180/*Bam*HI, D4S180/*Xmn*I, and D4S182/*Eco*T23, the mutation-rate estimates were within the range of $0–3.0 \times 10^{-3}$ (table 4). Although marker mutation is a factor, recurrent mutations at the CAG repeat in the HD locus may be a more plausible explanation for the lack of a predominant allele.

We considered a model that incorporated the marker mutation and the recurrent disease mutations. Three parameters, θ, the mutation rates at the marker loci, and

Table 4

Estimates of Genetic Distance between the HD Locus and Various Marker Loci, by FOA with and without Recurrent Mutations

Marker	Mutation Rate at Marker Locus[a]	Mutation Rate at HD Locus[b]	Estimated Distance (kb)	Error (kb)	Lower Boundary (kb)	Upper Boundary (kb)
D4S180/*Bam*HI	3.0	0	320	80	3	3,390
	3.0	50.0	220	20	4	3,290
*Xmn*I	0	0	1,518	~1,278	...	...
	0	40.0	240	~0	4	990
D4S95/*Mbo*I	0	0	260	150	130	480
	0	3.0	200	90	80	440
*Taq*1	2.0	0	330	220	40	910
	2.0	10.0	220	110	40	990
*Acc*I	0	0	290	180	90	>10,000
	0	1.4	260	150	60	1,030
D4S182/*Eco*T23	0	0	500	240	260	990
	0	12.0	260	10	50	850

NOTE.—The searching-grid sizes of θ and the mutation rates at the marker locus and the HD locus were 10^{-5}, 10^{-5}, and 10^{-9}, respectively.
[a] $\times 10^{-3}$.
[b] $\times 10^{-8}$.

the mutation rate at the disease locus, were incorporated, and their corresponding estimates, by use of FOA, also are listed in table 4. The estimated recurrent-mutation rates vary from marker to marker. At some loci, for example D4S95/*Mbo*I and D4S95/*Acc*I, the estimated mutation rate γ_d is small, suggesting that the effect of mutation on these markers is negligible. It also can be seen that, after the incorporation of marker mutations and recurrent mutations at the disease locus, the accuracy of the location estimates improved substantially. The overall average error of the estimation, by use of the model with mutations at both the marker and disease loci, is 89 kb, which is almost as accurate as our reanalysis of the CF data.

It may seem a bit strange that the estimate of the mutation rate at the disease locus varies from marker to marker. We point out that this is perfectly reasonable, since all marker data are subject to sampling errors. In fact, the magnitude of the estimated mutation rates (10^{-3} for the markers and from $\sim 10^{-8}$ to $\sim 10^{-9}$ for the HD locus) seems to be reasonable.

The composite likelihood involving D4S180/*Bam*HI, D4S95/*Mbo*I, and D4S182/*Eco*T23 peaked at the point ~250 kb away from the marker D4S180, as compared with the actual distance of ~240 kb (fig. 2). The error of the estimation is only ~10 kb!

FA

The cloning of the FA gene, called "X25," was reported early last year (Campuzano et al. 1996). Five exons of X25 were found to be spread over 40 kb. There are two point mutations, T→G in exon 3 and A→G in exon 4, but an unstable GAA trinucleotide expansion in the first X25 intron appears to be the predominant mutation site (Campuzano et al. 1996).

The FA gene, mapped to chromosome 9 in 1988 (Chamberlain et al. 1988), was found to be tightly linked to D9S15 and D9S5 (Fujita et al. 1990). In addition, LD analysis suggested that the FA gene was located within a 1-cM region bounded by these two tightly linked markers. Fujita et al. (1990) estimated that the θs between the FA gene and D9S15 and between the FA gene and D9S5 are 0.5 cM and 0 cM, respectively. Using the data in Fujita et al. (1990), Kaplan et al. (1995) applied SIM, hoping to finely map the gene. However, they got results no better than those of Fujita et al. (1990).

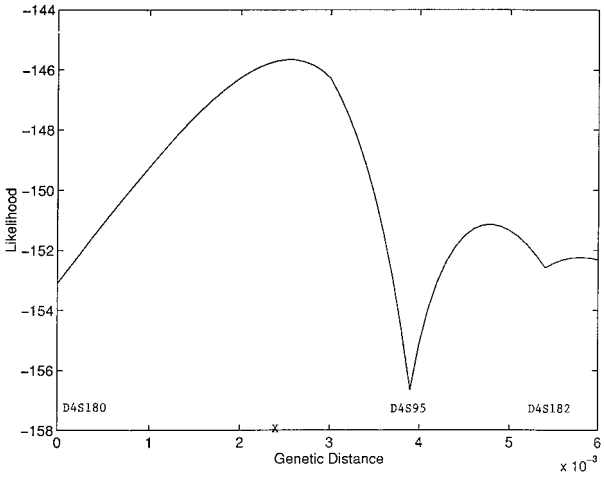

Figure 2 Composite log likelihood for estimation of the location of the HD locus, across markers D4S180, D4S127, D4S95, and D4S182. The true location of the gene is marked by an "X."

Table 5

Estimates of Genetic Distance and the Lower and Upper Boundaries, between the FA Locus and Two Marker Loci, by Different Methods

Marker	Mutation Rate at Marker Locus	Mutation Rate at FA Locus	Estimated Distance (kb)	Lower Boundary (kb)	Upper Boundary (kb)
D9S15	0	0	620	400	1,000
D9S5	2.5×10^{-3}	4.0×10^{-5}	480	220	1,220

Since there is no information on the age of the FA gene, Kaplan et al. (1995) assumed the age to be $G = 200$. Using the same data, we took a different approach, estimating simultaneously the age of the FA mutation and the location. By maximizing the composite likelihood based on D9S15 and D9S5 over the age of the FA mutation and θ, we estimated the age to be ~180 generations, for the Italian population (Pandolfo et al. 1990). In the discussion below, we use this figure and assume the frequency of the FA gene to be $\sqrt{1/50{,}000} = .0045$.

Fujita et al. (1990) found that D9S15, a six-allele microsatellite marker, is in strong LD with the FA locus. Kaplan et al. (1995) did not report their estimate of θ for this marker but only reported an upper boundary for θ of ~2 cM, which they admitted was too large to be useful. Here we assume a six-allele model with no mutation at the marker loci and consider the allele A2, the most common in the disease chromosomes, as the putative ancestral allele. With this model, the distance between D9S15 and the FA gene is estimated to be 620 kb (table 5), which is ~50 kb away from the true location (Campuzano et al. 1996). Note also that our upper boundary is only half that of Kaplan et al. (1995).

D9S5 is a bit problematic because no single allele has a predominant frequency in the FA population. We suspect that there may have been an early recombination between the marker and the disease locus, after the disease mutation occurred or that there may have been recurrent mutations. Therefore, we incorporated mutations at both loci into our model and designated the allele with the highest frequency in the disease sample as the common ancestral allele. The resultant estimation precisely placed the FA gene in the first X25 intron, where there is an unstable GAA trinucleotide expansion (table 5). These estimations suggest the order of D9S15–D9S5–FA, which agrees with the actual locations of these markers and the FA gene.

We also used the two-locus composite likelihood with the fixed mutation rates 0, 2.5×10^{-3}, and 4×10^{-5} at D9S15, D9S5, and the FA locus, respectively, for which the mutation rates were estimated from previous analyses (table 5). This yielded the distance of 690 kb between D9S15 and the FA gene, which again placed the FA gene 20 kb away from F8101, that is, exactly in an exon of X25.

EPM1

The EPM1 gene was mapped to chromosome 21q22.3 by use of linkage analysis and was narrowed further to a 0.6-cM region around markers D21S25 and PFKL, by use of LD (Lehesjoki et al. 1993). Recently, the EPM1 gene was cloned and was found to be 2.5 kb in length and ~30 kb away from marker D21S2040 (Pennacchio et al. 1996). The EPM1 gene consists of three small exons. The first base-pair mutation (G→C) and the second (G→C) were found at the last nucleotide of intron 1 and at amino acid position 68 of the cystatin B gene, respectively.

We assumed, as did Lehesjoki et al. (1993), the age of the disease mutation to be 100 generations and estimated that EPM1 is ~350 kb away from marker D21S25 (support interval 150 kb–750 kb; see table 6). The true location of the EPM1 gene now is known to be ~393 kb away from D21S25, which is remarkably close to our prediction.

On the basis of the marker-distance information that recently has become available (Stone et al. 1996), we applied our methods to data for markers PFKL and D21S25, published in Lehesjoki et al. (1993). We found that the age of the disease mutation is approximately $\hat{t} = 74$ generations and that the EPM1 gene is 610 kb away from PFKL. The error of our estimate is only ~30 kb.

Recently, Virtaneva et al. (1996) generated new data at D21S1885, D21S2040, D21S1259, D21S1912, and PFKL. Using this data set, we calculated the composite likelihood for these markers (fig. 3). Again, the age of mutation is ~70 generations, and the distance between D21S1885 and the EPM1 locus is estimated to be 370 kb, which is only 40 kb away from the true location (fig. 3).

The results of the likelihood-based multipoint LD analysis, according to Terwilliger (1995), placed the disease gene in the region between D21S1259 and D21S1912 and estimated the EPM1 gene to be 80 kb away from D21S1259 (Virtaneva et al. 1996). The error (220 kb) of their estimate is almost six times higher than

Table 6

Estimates of Genetic Distance between the EPM1 Locus and Two Marker Loci, by Different Methods

	Estimated Distance (kb), by[a]			
Marker	FOA	SIM[b]	LDT	Modified LDT
PFKL	360 [210–570]	500 [?–100]	280 [230–360]	360 [?–490]
D21S25	350 [150–750]	600 [?–1,300]	140 [110–180]	350 [?–480]

[a] The numbers in brackets are the estimated lower and upper support boundaries.
[b] Data and estimates from tables 2 and 3 in the article by Kaplan and Weir (1995).

ours. This demonstrates again that our method provides a more accurate estimate than that of Terwilliger (1995).

Discussion

To make efficient inferences in LDM, it is necessary to base the inference on the maximum-likelihood principle, which requires an explicit expression for the expectation of the conditional likelihood function, that is, the unconditional likelihood. The unconditional likelihood is deceptively simple in form, but it can be very difficult to evaluate, even in the single-marker case. In contrast to SIM, proposed by Kaplan et al. (1995), we have approximated the likelihood using the Taylor expansion. The approximations require the computation of the first and second moments of the marker allele frequency in the disease population. The first moments of the allele frequencies can be derived regardless of the population model considered. The derivation of the second moments, however, does require the specification of a population model. Through derivation of the first and second moments of the marker allele frequencies in the disease-causing chromosomes, we have presented a general, mathematical, and conceptually coherent framework for LDM, which incorporates multilocus and multiallelic markers and mutational processes, at both the marker and disease loci. This framework provides many new insights into the patterns of LD and the mathematical links between seemingly unrelated methods for LDM.

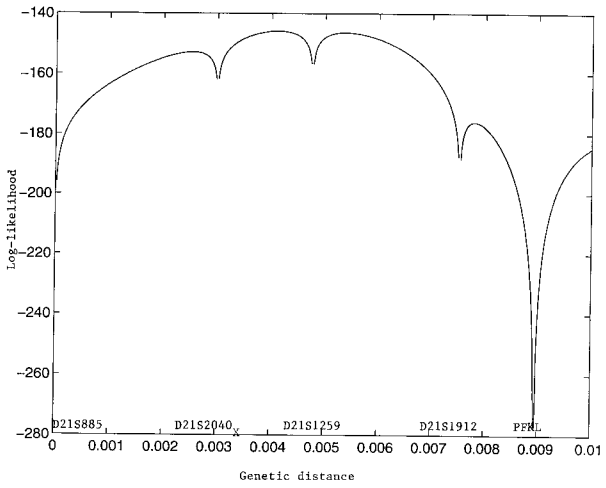

Figure 3 Composite log likelihood for estimation of the location of the EPM1 locus, on the basis of markers D21S1885, D21S2040, D21S1259, D21S1912, and PFKL. The true location of the gene is marked by an "X."

The methods for LDM can be classified roughly into two groups. One is simple disequilibrium mapping (Weir 1989; Jorde et al. 1994; Devlin and Risch 1995), which is based solely on the magnitude of the disequilibrium measures. The other group is what we called "model based," which is represented by the work of Hästbacka et al. (1992), Hill and Weir (1994), Kaplan et al. (1995), Kaplan and Weir (1995), Risch et al. (1995), and Terwilliger (1995). The latter group can be distinguished further, depending on whether one imposes a population model (e.g., an exponentially growing population).

Like most population-genetics models of LD, Hill and Weir's (1994) model assumes a constant effective population size N_e. With that model, θ unfortunately is confounded with an unknown N_e. This makes it difficult to estimate θ. Moreover, the model has the problem that once the allele frequencies of disease-causing chromosomes reach the state of equilibrium, all information about θ, generated by LD, will be lost. The major contribution of Hästbacka et al. (1992) was to consider the nonequilibrium (i.e., a rapid-growing population) situation of a so-called young and isolated population. In this kind of model, all information on recombination events accumulated throughout the entire history of the population is manifested by LD. As a result, θ is confounded only with the age of the disease mutation, which sometimes can be estimated approximately through other sources. In fact, when multilocus data are used and interlocus genetic distances are known, the composite likelihood can be used to estimate simultaneously the age of the mutation and the location of the disease locus.

Kaplan et al. (1995) recognized that one does not

need to model the evolutionary history of the whole population. Instead, one can model only the dynamics of the disease-causing chromosomes. Since the disease of interest is usually rare, the proportion of disease-causing chromosomes in the entire population is typically very small. Once information on the marker allele frequencies of the normal chromosomes is gathered, all information on θ is in the disease-causing chromosomes. However, this is true only when the disease under study is rare.

Does this mean that we always have to know or to assume the growth rate of a population, for LDM? This question is important, since determination of the growth rate for a particular population for the last 20 or more generations can be difficult, despite the fact that most human populations have expanded considerably in the last century. Our results challenge this notion, on two grounds. First, the results derived with the assumption of an exponentially growing population, obtained by Hästbacka et al. (1992) and Lehesjoki et al. (1993), also can be derived with our framework without any assumptions about population growth. In fact, the equations for the estimation of θ, proposed by the two groups, were derived without respect to growth rate. Second, our numerical results suggest that FOA likelihood function (6) performs remarkably well. As we pointed out before, the FOA is valid regardless of which population model is used.

The framework that we proposed also has broadened the scope of LDM. Several methods assume that the frequency of the associated allele in disease-causing chromosomes always should be higher than that in the normal chromosomes. Terwilliger's (1995) method implicitly assumes that this is the case (i.e., $\lambda \leqslant 0$). The assumption that $P_{\text{excess}} \geqslant 0$, made by Lehesjoki et al. (1993), also explicitly assumes so. Kaplan et al. (1995) noted that, in the case of FA and HD, some markers show LD with the disease locus, but for these markers, the allele frequencies in the samples of disease-causing chromosomes are lower than those in the normal sample. Kaplan et al. (1995) and Kaplan and Weir (1995) thought that these observations were not consistent with their evolutionary theory. Assuming that sampling error can be ignored, however, we know from the above discussions that this phenomenon can be accommodated within our model, owing either to random drift (since the inequality is stochastic in nature) or to mutations at the marker locus.

Kaplan et al. (1995), Kaplan and Weir (1995), and we found that the upper boundaries estimated by the LDT method were too restrictive and missed the true location of the disease locus in almost 80% of cases. This clearly is unacceptable. We also found, however, that support intervals estimated by SIM were too conservative to be useful.

One potential source of inaccuracy in SIM is the simulation itself. By necessity, SIM generates a prespecified number of replicates, according to some parameters and to population-dynamics models. Because of their Monte Carlo nature, sampling variations are introduced into the parameter estimate, in addition to noise in the data and to intrinsic statistical variations in the estimation.

For HD and FA, for which no single marker allele has a predominantly high frequency in disease chromosomes, SIM and other methods do not work at all. It should be noted that the analysis of HD and FA data was based on data collected from large continental populations whose histories are not well understood. It is likely that there are multiple disease-causing mutations on different alleles. For this class of so-called multimutant diseases, a single allele with a predominantly high frequency among disease chromosomes may not exist. Mutations at marker loci also can cause the same problem. To deal with these possibilities, we incorporated mutations at both marker and disease loci. For the same data sets used by Kaplan et al. (1995), our method mapped the HD gene with remarkable accuracy: the average error of the estimation was only ~89 kb. On the basis of limited published data, we predicted, prior to cloning, that the FA gene is ~690 kb away from D9S15, which is exactly the location of the FA gene. We are convinced that, given the right population and data, it is technically feasible to fine-map disease genes by use of LDM.

On the basis of our experiences with LDM, using published data, we offer some general considerations for the fine-scale mapping of disease genes. First and foremost, it is important to understand the disease and the population. Is the disease rare in the population? This question should be examined carefully before an LDM analysis is launched. If the disease is heterogeneous, it may be a good idea to select one specific subtype of the disease, for LDM. It also may be ideal to have a genetically isolated population for LDM, with the additional requirements that the disease mutation (not necessarily the population) is old enough for recombination to narrow the region of disequilibrium but not so old as either to reach linkage equilibrium or to accumulate many new mutations. Second, it is useful to know the locations of the markers to be saturated, in the region of interest. If we know the interlocus distances among the markers, we can use the composite likelihood and can extract information on the disease locus, from multiple markers. Third, it also is worthwhile to place the markers carefully. For example, assigning markers approximately equally to both sides of the disease locus would allow more accurate localization of the disease locus. This can be done, for example, by the even placement of markers in the region of interest. Fourth, it may be efficient to saturate the region of interest with

markers, in two steps. At the first step, the region would be saturated with markers spaced at ~500 kb apart. Once a narrower region is identified, the region would be saturated with markers spaced at ~60–100 kb. Owing to the inherent limitations, a map that is too dense may be a waste.

Throughout this article, we have used a one-step SMM to describe the mutation process at microsatellite loci. Although the model is simple and seems to work well, it may not work well in all cases. If this is true, a multistep SMM should be used.

Although allelic heterogeneity can be handled in LDM by the introduction of recurrent mutations, locus heterogeneity may be more difficult to deal with. Also, the assumption of the constant allele frequency in the normal population may not hold when the mutation rate at the marker is very high and the age of the disease mutation is old. Population substructure, incomplete penetrance, phenocopies, and nonrarity of the disease also can pose problems. Thus, there is room for improvement for LDM methodology.

Acknowledgments

This research was supported by the National Institutes of Health grants R29-GM52205 and R01-GM56515. The authors thank Dr. Aravinda Chakravarti, who drew their attention to Fisher's (1947) paper on the Rhesus factor and who kindly supplied his unpublished data. They also thank Dr. Michael Boehnke, Dr. James T. Elder, Ms. Robin Hemenway, and two anonymous reviewers, for their helpful comments on earlier versions of this paper, which helped improve the presentation of this paper.

Appendix A

Let N_G be the current ($t = G$) size of the normal population, with an exponential growth rate ρ. The amount of mutation at the disease locus in each generation depends on the γ_d as well as on N_G.

There are three ways to obtain the disease chromosomes carrying marker allele M_i, in generation $t + 1$:

1. The disease chromosomes carrying M_i in generation t do not recombine with other chromosomes during the time period $(t, t + 1)$.
2. Disease chromosomes recombine with the normal chromosomes carrying the marker allele M_i.
3. Mutations occur on normal chromosomes carrying the marker allele M_i.

Given $X_i(t)$, the number of disease chromosomes carrying marker allele M_i in generation t, if mutation at the marker locus is ignored, then

$$X_i(t+1) = (1-\theta)X_i(t) + [\theta X_T(t) + \gamma_d 2N_G(1+\rho)^{-(G-t)}]p_{i_n},$$

where p_{i_n} is the frequency of the allele M_i in the normal population. It is easy to see that $X_T(t + 1) = X_T(t) + \gamma_d 2N_G(1 + \rho)^{-(G-t)}$. Recall that $p_{i_d}(t)$ is the frequency of the marker allele M_i in the disease population. Let $p^*_{i_d}(t + 1)$ be the frequency of the allele M_i in the disease population, after recombination and mutation, during the time period $(t, t + 1)$. Furthermore, let p_d be the disease-allele frequency, that is, $p_d = X_T(t)/2N(t)$, where $N(t)$ is the size of the population in generation t. Assume that p_d is constant over time. Then $p^*_{i_d}(t + 1)$ is given by

$$\begin{aligned} p^*_{i_d}(t+1) &= \frac{(1-\theta)X_i(t) + [\theta X_T(t) + \gamma_d 2N_G(1+\rho)^{-(G-t)}]p_{i_n}}{X_T(t) + \gamma_d 2N_G(1+\rho)^{-(G-t)}} \\ &\approx (1-\theta)p_{i_d}(t)\frac{1}{1+\frac{\gamma_d}{p_d}} + \alpha p_{i_n} \\ &\approx (1-\alpha)p_{i_d}(t) + \alpha p_{i_n}, \end{aligned} \tag{A1}$$

where $\alpha = \theta + \gamma_d/p_d$.

Under the one-step SMM, marker allele M_i can mutate to the next-larger allelic state M_{i+1}, with probability u, and to the next-smaller allelic state M_{i-1}, with probability v. Clearly, M_1 can mutate only to the allelic state M_2, and M_m can mutate only to the allelic state M_{m-1}. Given $p^*_{i_d}(t + 1)$, after meiosis the frequency $p_{i_d}(t + 1)$ at the $(t + 1)$th generation has a multinomial distribution with parameters

$$\begin{aligned} \pi_i(t) = {} & [1-(u+v)]p^*_{i_d}(t+1) \\ & + up^*_{i-1_d}(t+1) + vp^*_{i+1_d}(t+1) \end{aligned} \tag{A2}$$

(Ohta and Kimura 1973; M. Xiong and S.-W. Guo, unpublished data).

It follows from equations (A1) and (A2) that

$$\begin{aligned} g_i(t) &= \mathrm{E}[p_{i_d}(t+1) - p_{i_d}(t)\,|\,P(t)] \\ &\approx -[\alpha + (u+v)(1-\alpha)]p_{i_d} \\ &\quad + u(1-\alpha)p_{i-1_d} + v(1-\alpha)p_{i+1_d} + u\alpha p_{i-1_n} \\ &\quad + [1-(u+v)]\alpha p_{i_n} + v\alpha p_{i+1_n}, \\ &\qquad i = 2, \ldots, m-1, \\ g_1(t) &= \mathrm{E}[p_{1_d}(t+1) - p_{1_d}(t)\,|\,P(t)] \\ &\approx -[\alpha + u(1-\alpha)]p_{1_d} + v(1-\alpha)p_{2_d} \\ &\quad + (1-u)\alpha p_{1_n} + v\alpha p_{2_n}, \\ g_m(t) &= \mathrm{E}[p_{m_d}(t+1) - p_{m_d}(t)\,|\,P(t)] \\ &\approx -[\alpha + v(1-\alpha)]p_{m_d} + u(1-\alpha)p_{m-1_d} \\ &\quad + (1-v)\alpha p_{m_n} + u\alpha p_{m-1_n}, \end{aligned}$$

and

$$w_{ij}(t) = \mathrm{E}\{[p_{i_d}(t+1) - p_{i_d}(t)][p_{j_d}(t+1) - p_{j_d}(t)] \mid P(t)\} \approx \frac{p_{i_d}(t)[\delta_{ij} - p_{j_d}(t)]}{X_T(t)} .$$

Therefore, the joint evolutionary process $p_{i_d}(t)$ ($i = 1, \ldots, m$) at the disease and marker loci can be approximated by a diffusion process with a generator given by

$$L = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{P_{i_d}(t)[\delta_{ij} - p_{j_d}(t)]}{X_T(t)}\frac{\partial^2}{\partial p_{i_d}\partial p_{j_d}} + \sum_{i=1}^{m} g_i(t)\frac{\partial}{\partial p_{i_d}}$$

(Revuz and Yor 1994).

Appendix B

Let f be a function of $p_{i_d}(i = 1, \ldots, m)$. By the Hille-Yosida theorem (Ethier and Kurtz 1986), we have $d\mathrm{E}(f)/dt = \mathrm{E}[L(f)]$, where L is the generator of the diffusion process. In particular, if $f = p_{i_d}(t)$, then $\partial^2 f/\partial p_{i_d}\partial p_{j_d} = 0$ and $\partial f/\partial p_{i_d} = 1$. Thus, $d\mathrm{E}[p_{i_d}(t)]/dt = \mathrm{E}[g_i(t)]$, where $i = 1, \ldots, m$. Similarly, if $f = p_{i_d}(t)p_{j_d}(t)$, then $\partial^2 f/\partial p_{i_d}\partial p_{j_d} = 1$ and $\partial f/\partial p_{i_d} = p_{j_d}$, and, hence,

$$\frac{d\mathrm{E}[p_{i_d}(t)p_{j_d}(t)]}{dt} = -\mathrm{E}\left[\frac{p_{i_d}(t)p_{j_d}(t)}{X_T(t)}\right] + \mathrm{E}[g_i(t)p_{j_d}(t)] + \mathrm{E}[g_j(t)p_{i_d}(t)] .$$

Clearly, $\partial^2 p_{i_d}^2/\partial p_{i_d}^2 = 2$ and $\partial p_{i_d}^2/\partial p_{i_d} = 2p_{i_d}$. By the same argument, we obtain

$$\frac{d\mathrm{E}[p_{i_d}^2(t)]}{dt} = \mathrm{E}\left[\frac{p_{i_d}(t)(1 - p_{i_d}(t)}{X_T(t)}\right] + 2\mathrm{E}[g_i(t)p_{i_d}(t)] .$$

Appendix C

The matrix A has the following elements:

$$\begin{aligned}
a_{11} &= -[\alpha + u(1-\alpha)] \\
a_{12} &= v(1-\alpha) \\
a_{1j} &= 0;\ j = 3, \ldots, m \\
a_{i,i-1} &= u(1-\alpha) \\
a_{ii} &= -[\alpha + (u+v)(1-\alpha)] \\
a_{i,i+1} &= v(1-\alpha);\ i = 2, \ldots, m \\
a_{i,j} &= 0;\ j \neq i-1, i, i+1 \\
a_{m-1,m} &= u(1-\alpha) \\
a_{mm} &= -[\alpha + v(1-\alpha)] \\
a_{jm} &= 0;\ j = 1, \ldots, m-2
\end{aligned}$$

It is easy to see that

$$\frac{de^{-At}}{dt} = -Ae^{-At}$$

and that

$$\int_0^t e^{-As}ds = tI + \sum_{k=1}^{\infty}\frac{(-A)^k t^{k+1}}{(k+1)!} = -A^{-1}[e^{-At} - I] .$$

Thus, we have

$$\frac{d[e^{-At}\mu(t)]}{dt} = e^{-At}B . \qquad \text{(C1)}$$

When both sides of equation (C1) are integrated,

$$e^{-At}\mu(t) - \mu(0) = -A^{-1}(e^{-At} - I)B . \qquad \text{(C2)}$$

Thus, it follows from equation (C2) that $\mu(t) = e^{At}\mu(0) + A^{-1}(e^{At} - I)B$.

To apply the second-order approximation, it is necessary to compute the second moments of the marker allele frequencies, which depends on (1) the recurrent-mutation rate γ_d at the disease locus, (2) the mutation process at the marker locus, and (3) the population-growth model. There are an infinite number of choices for all of these variables. Here, we only consider some simple, yet reasonably realistic, models.

For ease of exposition, we consider a two-allele marker. Let ρ be the rate of population expansion. Then, $X_T(t) = 2N_d e^{\rho(t-G)}$ ($0 \leq t \leq G$). Equation (10) can be rewritten as

$$\frac{d\mathrm{E}(p_{1_d}^2)}{dt} = -\left[\frac{e^{-\rho(t-G)}}{2N_d} + a_1\right]\mathrm{E}(p_{1_d}^2) + \left(\frac{e^{-\rho(t-G)}}{2N_d} + a_2\right)\mathrm{E}(p_{1_d}) , \qquad \text{(C3)}$$

where $a_1 = 2[\alpha + u(1-\alpha) + v(1-\alpha)]$ and $a_2 = 2v(1-\alpha) + 2(1-u)\alpha p_{1_n} + 2v\alpha p_{2_n}$. When $\mathrm{E}(p_{1_d})$ is substituted into equation (C3),

$$\mathrm{E}(p_{1_d}^2) = p_{1_d}(0)e^{-a_1 t + (e^{\rho G}/2N_d\rho)(e^{-\rho t} - 1)} + e^{-a_1 t + (e^{-\rho(t-G)}/2N_d\rho)}\int_0^t h(s)ds , \qquad \text{(C4)}$$

where

$$
\begin{aligned}
h(s) = {} & \frac{p_{1_d}(0) + \frac{b}{\lambda}}{2N_d} e^{(a_1+\lambda-\rho)s+\rho G-(1/2N_d\rho)e^{-\rho(s-G)}} \\
& - \frac{b}{2\lambda N_d} e^{(a_1-\rho)s+\rho G-(1/2N_d\rho)e^{-\rho(s-G)}} \\
& + a_2\left[p_{1_d}(0) + \frac{b}{\lambda}\right] e^{(a_1+\lambda)s-(1/2N_d\rho)e^{-\rho(s-G)}} \qquad \text{(C5)} \\
& - \frac{a_2 b}{\lambda} e^{a_1 s-(1/2N_d\rho)e^{-\rho(s-G)}} \\
\lambda = {} & -[\alpha + u(1-\alpha) + v(1-\alpha)] \\
b = {} & (1-u)\alpha p_{1_n} + v\alpha p_{2_n} + v(1-\alpha)
\end{aligned}
$$

As long as we know the population growth rate ρ, the expectations of the second moments of p_{i_d} can be expressed as a function of u, v, θ, t, and γ_d. The case of multiallelic markers can be considered similarly if an SMM is assumed.

Appendix D

$$
\begin{aligned}
L = {} & \frac{1}{2}\sum_i\sum_j\sum_k\sum_l a_{ijkl}(t)\frac{\partial^2}{\partial p_{ij_d}\partial p_{kl_d}} \qquad \text{(D1)} \\
& + \sum_i\sum_j g_{ij}\frac{\partial}{\partial p_{ij_d}},
\end{aligned}
$$

where

$$
\begin{aligned}
a_{ijkl}(t) = {} & \frac{p_{ij_d}(t)(\delta_{ik}\delta_{jl} - p_{kl_d})}{X_T(t)}, \\
g_{ij}(t) = {} & -\alpha p_{ij_d} - (u+v)(1-\alpha)p_{ij_d} \\
& + [1-(u+v)]\left(\theta_1 p_{i_n}p_{\cdot j_d} + \theta_2 p_{\cdot j_n}p_{i_d} + \frac{\gamma_d}{p_d}p_{ij_n}\right) \\
& + u\left[(1-\alpha)p_{i-1,j_d} + \theta_1 p_{i-1_n}p_{\cdot j_d}\right. \\
& \qquad \left. + \theta_2 p_{\cdot j_n}p_{i-1_d} + \frac{\gamma_d}{p_d}p_{i-1,j_n}\right] \\
& + v\left[(1-\alpha)p_{i+1,j_d} + \theta_1 p_{i+1_n}p_{\cdot j_d}\right. \\
& \qquad \left. + \theta_2 p_{\cdot j_n}p_{i+1_d} + \frac{\gamma_d}{p_d}p_{i+1,j_n}\right] \\
& + u\left[(1-\alpha)p_{ij-1_d} + \theta_1 p_{i_n}p_{\cdot j-1_d}\right. \\
& \qquad \left. + \theta_2 p_{\cdot j-1_n}p_{i_d} + \frac{\gamma_d}{p_d}p_{ij-1_n}\right] \\
& + v\left[(1-\alpha)p_{ij+1_d} + \theta_1 p_{i_n}p_{\cdot j+1_d}\right. \\
& \qquad \left. + \theta_2 p_{\cdot j+1_n}p_{i_d} + \frac{\gamma_d}{p_d}p_{ij+1_n}\right],
\end{aligned}
$$

and

$$
\alpha = \theta_1 + \theta_2 + \frac{\gamma_d}{p_d}, \qquad \text{(D2)}
$$

and, for $i = 1, m_1$, $j = 1$, and m_2, where m_i is the number of alleles at locus i, we need to consider corresponding boundaries for u and v. Using the Hille-Yosida theorem, we obtain a system of differential equations with regard to the expectations of marker frequencies in a disease population:

$$
\frac{d\mathrm{E}(p_{ij_d})}{dt} = \mathrm{E}[g_{ij}(t)] \qquad \text{(D3)}
$$

From equation (14) it follows that

$$
\begin{aligned}
\mathrm{E}(p_{i_d}) &= [p_{i_d}(0) - p_{i_n}]e^{-\theta_1 t} + p_{i_n} \\
\mathrm{E}(p_{\cdot j_d}) &= [p_{\cdot j_d}(0) - p_{\cdot j_n}]e^{-\theta_2 t} + p_{\cdot j_n}.
\end{aligned} \qquad \text{(D4)}
$$

Substituting $\mathrm{E}(p_{i_d})$ and $\mathrm{E}(p_{\cdot j_d})$ from equation set (D4) into equation (15), we obtain

$$
\begin{aligned}
\frac{d\mathrm{E}(p_{ij_d})}{dt} = {} & \lambda\mathrm{E}(p_{ij_d}) + a_1 e^{-\theta_1 t} \\
& + a_2 e^{-\theta_2 t} + (\theta_1+\theta_2)p_{i\cdot n}p_{\cdot j_n},
\end{aligned}
$$

where $\lambda = -(\theta_1+\theta_2)$, $a_1 = \theta_2 p_{\cdot j_n}[p_{i_d}(0) - p_{i_n}]$, and $a_2 = \theta_1 p_{i_n}[p_{\cdot j_d}(0) - p_{\cdot j_n}]$. Thus,

$$
\begin{aligned}
\frac{d[e^{-\lambda t}\mathrm{E}(p_{ij_d})]}{dt} &= -\lambda e^{-\lambda t}\mathrm{E}(p_{ij_d}) + e^{-\lambda t}\frac{d\mathrm{E}(p_{ij_d})}{dt} \\
&= -\lambda e^{-\lambda t}\mathrm{E}(p_{ij_d}) + e^{-\lambda t}[\lambda\mathrm{E}(p_{ij_d}) \\
&\quad + a_1 e^{-\theta_1 t} + a_2 e^{-\theta_2 t} + (\theta_1+\theta_2)p_{i_n}p_{\cdot j_n}] \qquad \text{(D5)} \\
&= a_1 e^{-(\lambda+\theta_1)t} + a_2 e^{-(\lambda+\theta_2)t} \\
&\quad + (\theta_1+\theta_2)p_{i_n}p_{\cdot j_n}e^{-\lambda t}.
\end{aligned}
$$

When both sides of equation (D5) are integrated,

$$e^{-\lambda t}\mathrm{E}(p_{ij_d}) - p_{ij_d}(0) = -\frac{a_1}{\lambda + \theta_1}[e^{-(\lambda+\theta_1)t} - 1] - \frac{a_2}{\lambda + \theta_2}[e^{-(\lambda+\theta_2)t} - 1] - \frac{(\theta_1 + \theta_2)p_{i_{\cdot}n}p_{\cdot j_n}}{\lambda}(e^{-\lambda t} - 1)\,. \quad \text{(D6)}$$

After some algebra is performed, it follows from equation (D6) that

$$\begin{aligned}\mathrm{E}(p_{ij_d}) &= e^{\lambda t}p_{ij_d}(0) - \frac{a_1}{\lambda + \theta_1}(e^{-\theta_1 t} - e^{-\lambda t}) \\ &\quad - \frac{a_2}{\lambda + \theta_2}(e^{-\theta_2 t} - e^{\lambda t}) \\ &\quad - \frac{(\theta_1 + \theta_2)p_{i_{\cdot}n}p_{\cdot j_n}}{\lambda}(1 - e^{\lambda t}) \\ &= [p_{ij_d}(0) - \beta_1 - \beta_2]e^{-(\theta_1+\theta_2)t} \\ &\quad + \beta_1 e^{-\theta_1 t} + \beta_2 e^{-\theta_2 t} + p_{i_{\cdot}n}p_{\cdot j_n}\,,\end{aligned}$$

where $\beta_1 = p_{\cdot j_n}[p_{i_{\cdot}d}(0) - p_{i_{\cdot}n}]$ and $\beta_2 = p_{i_{\cdot}n}[p_{\cdot j_d}(0) - p_{\cdot j_n}]$.

References

Bodmer WF (1986) Human genetics: the molecular challenge. Cold Spring Harbor Symp Quant Biol 51:1–13

Boehnke WF (1994) Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. Am J Hum Genet 55:379–390

Chamberlain S, Shaw J, Rowland A, Wallis J, South S, Nakamura Y, von Gabain A, et al (1988) Mapping of mutation causing Friedreich's ataxia to human chromosome 9. Nature 334:248–249.

Collins FS (1992) Positional cloning: let's not call it reverse anymore. Nat Genet 1:3–6

Compuzano V, Montermini L, Molto MD, Pianese L, Cossee M, Cavalcanti F, Monticeli A, et al. (1996) Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science 271:1423–1427

Cox TK, Kerem B, Rommens J, Iannuzzi MC, Drumm M, Collins FS, Dean M, et al (1989) Mapping of the cystic fibrosis gene using putative ancestral recombinants. Am J Hum Genet Suppl 45:A136

Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper & Row, New York

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

Ethier S, Kurtz TG (1986) Markov processes: characterization and convergence. Wiley, New York

Fisher RA (1947) The rhesus factor: a study in scientific method. Am Sci 35:95–103

Fujita R, Hanauer A, Sirugo G, Heilig R, Mandel JL (1990) Additional polymorphisms at marker loci D9S5 and D9S15 generate extended haplotypes in linkage disequilibrium in Friedreich ataxia. Proc Natl Acad Sci USA 87:1796–1800

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, et al (1983) A polymorphic DNA marker genetically linked to Huntington's Disease. Nature 306:234–238

Haldane JBS (1919) The combination of linkage values and the calculation of distance between the loci of linked factors. J Genet 8:299–309

Hammersley JM, Handscomb DC (1964) Monte Carlo methods. Methuen & Co, London

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–211

Hästbacka J, de la Chapelle A, Mahtani MM, Clines G, Reeve-Daly MP, Daly M, Hamilton BA, et al (1994) The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell 78:1078–1087

Hedrick PW (1980) Hitchhiking: a comparison of linkage and partial selfing. Genetics 94:791–808

Hill WG (1976) Non-random association of neutral linked genes in finite populations. In: Karlin S, Nevo E (eds) Population genetics and ecology. Academic Press, New York, pp 339–376

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

Hill WG, Weir BS (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. Am J Hum Genet 54:705–714

Huntington Disease Collaborative Research Group, The (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72:971–983

Jennings HS (1917) The numerical results of diverse systems of breeding, with respect to two pairs of characters, linked or independent, with special relation to the effects of linkage. Genetics 2:97–154

Jorde LB (1995) Linkage disequilibrium as a gene-mapping tool. Am J Hum Genet 56:11–14

Jorde LB, Watkins WS, Carlson M, Groden J, Albertsen H, Thliveris A, Leppert M (1994) Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. Am J Hum Genet 54:884–898

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

Kaplan NL, Weir BS (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

Karlin S (1969) Equilibrium behavior of population genetics models with nonrandom mating. Gordon & Breach, London

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

Kojima K, Schaffer HE (1967) Survival processes of linked mutant genes. Evolution 21:518–531

Lander ES, Botstein D (1986) Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harbor Symp Quant Biol S1:49–62

Lange K, Kunkel L, Aldridge J, Latt SA (1985) Accurate and superaccurate gene mapping. Am J Hum Genet 37:853–867

Lehesjoki A-E, Koskiniemi M, Norio R, Tirrito S, Sistonen P, Lander E, de la Chapelle A (1993) Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping. Hum Mol Genet 2:1229–1234

Lewontin RC, Kojima K (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14:450–472

MacDonald ME, Lin C, Srinidhi L, Bates G, Altherr M, Whaley WL, Lehrach H, et al (1991) Complex patterns of linkage disequilibrium in the Huntington disease region. Am J Hum Genet 49:723–734

MacDonald ME, Novelletto A, Lin C, Tagle D, Barnes G, Bates G, Taylor S, et al (1992) The Huntington's disease candidate region exhibits many different haplotypes. Nat Genet 1:99–103

Nei M, Li W-H (1973) Linkage disequilibrium in subdivided populations. Genetics 75:213–219

Ohta T, Kimura M (1969) Linkage disequilibrium due to random genetic drift. Genet Res 13:47–55

——— (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res 22:201–204

Ott J (1991) Analysis of human genetic linkage, rev ed. Johns Hopkins University Press, Baltimore

Pandolfo M, Sirugo G, Antonelli A, Weitnauer L, Ferretti L, Leone M, Dones I, et al (1990) Friedreich ataxia in Italian families: genetic homogeneity and linkage disequilibrium with the marker loci D9S5 and D9S15. Am J Hum Genet 47:228–235

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

Revuz D, Yor M (1994) Continuous Martingales and Brownian motion. Springer-Verlag, New York

Risch N, de Leon D, Ozelius LJ, Kramer P, Almasy L, Singer B, Fahn S, et al (1995) Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nat Genet 9:152–159

Robbins RB (1918) Some applications of mathematics to breeding problems. III. Genetics 3:375–389

Shriver MD, Lin L, Chakraborty R, Boerwinkle E (1993) VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics 134:983–993

Sinnock P, Sing CF (1972) Analysis of multilocus genetic systems in Tecumseh, Michigan. II. Consideration of the correlation between nonalleles in gametes. Am J Hum Genet 24: 393–415

Slatkin M (1994) Linkage disequilibrium in growing and stable populations. Genetics 137:331–336

Smouse PE, Neel JV (1977) Multivariate analysis of gametic disequilibrium in the Yanomama. Genetics 85:733–752

Snell RG, Lazarou L, Youngman S, Quarrell OWJ, Wasmuth JJ, Shaw DJ, et al (1989) Linkage disequilibrium in Huntington's disease: an improved localization for the gene. J Med Genet 26:673–675

Stone NE, Fan JB, Willour V, Pennacchio LA, Warrington JA, Hu A, de la Chapelle A, et al (1996) Construction of a 750-kb bacterial clone contig and restriction map in the region of human chromosome 21 containing the progressive myoclonus epilepsy gene. Genome Res 6:218–225

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

Theilmann J, Kanani S, Shiang R, Robbins C, Quarrell O, Huggins M, Hedrick A, et al (1989) Non-random association between alleles detected at D4S95 and D4S98 and the Huntington's disease gene. J Med Genet 26:676–681

Thomson G (1977) The effect of a selected locus on linked neutral loci. Genetics 85:753–788

Turner JR (1971) Selection and stability in the complex polymorphism of *Moraba scurra*. Evolution 26:334–343

Valdes AM, Slatkin M, Freimer N (1993) Allele frequencies at microsatellite loci: the stepwise model revisited. Genetics 133:737–749

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

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128

Weinberg W (1909) Über Vererbungsgesetze beim Menschen. Z Abst V Vererb 1:277–330

Weir BS (1989) Locating the cystic fibrosis gene on the basis of linkage disequilibrium with markers? In: Elston RC, Spence MA, Hodge SE, MacCluer JW (eds) Multipoint mapping and linkage based on affected pedigree members: Genetic Analysis Workshop 6. AR Liss, New York, pp 81–86

Weir BS, Allard RW, Kahler AL (1972) Analysis of complex allozyme polymorphisms in a barley population. Genetics 72:505–523

Weir BS, Cockerham CC (1989) Complete characterization of disequilibrium at two loci. In: Feldman MW (ed) Mathematical evolutionary theory. Princeton University Press, Princeton, pp. 86–110