Am. J. Hum. Genet. 62:159–170, 1998

Linkage Disequilibrium and Gene Mapping: An Empirical Least-Squares Approach

Laura C. Lazzeroni*

Department of Statistics, Stanford University, Stanford, CA

Summary

This paper proposes a novel approach for fine-scale mapping of disease genes that is based on the well-known linkage-disequilibrium parameter δ. Using a very simple, very general model, I show how δ can be interpreted in terms of identity-by-descent probabilities. The value of δ follows a piecewise curve along the chromosome, with the maximum occurring at the disease locus where the two pieces intersect. A semiparametric, multilocus approach is used to fit this nonlinear regression curve in order to estimate the gene location. Using the bootstrap to empirically estimate much of the probability model from the data avoids the need for many detailed population assumptions. One advantage of the approach is its use of the observed covariance structure of the data, which can be highly informative as to the gene location. I illustrate the method on the cystic fibrosis data of Kerem et al.

Introduction

Marker loci near a disease gene are often observed to be in linkage disequilibrium with the disease; that is, the relative frequencies of marker alleles in affected individuals differ from those in the general population. Linkage disequilibrium occurs because each new disease-predisposing mutation originally appears on a single chromosome. Individuals who inherit a disease mutation are likely to also inherit the alleles of the original chromosome, at neighboring marker loci. As generations pass, recombination or mutation can disrupt the joint transmission of disease mutation and marker allele. Because recombination with the disease gene happens less often for nearby marker loci, markers in the immediate vicinity of the gene should remain in greater disequilibrium than more distant marker loci.

This paper proposes a new multilocus method for fine-scale genetic mapping using linkage-disequilibrium data. The method is intended for genes that, through family-based linkage analysis, have already been mapped to a particular interval of a chromosome. In theory, disequilibrium mapping can further shorten that interval, reducing the remaining cost and effort needed to clone the gene (Jorde 1995). Because each observation represents several historical meioses, linkage-disequilibrium data can provide more opportunities for recombination on a short interval than are provided by pedigree data. This effectively increases the sample size and should make it easier to construct fine-scale maps. On the other hand, the choice of the probability model used for linkage-disequilibrium mapping is less clear than it is for linkage analysis. For linkage, the model can be greatly simplified by conditioning on the known pedigree structure. In contrast, the distant relationships on which linkage-disequilibrium mapping implicitly depends are unobserved.

One approach for linkage disequilibrium is to use methods, such as the transmission/disequilibrium test (TDT), that condition inference on parental genotypes. Although conditioning controls against possible confounding by population admixture, it also ignores information contained in the parental genotype distribution. TDT methods are especially appropriate when one is screening large numbers of loci or is seeking a conclusive verdict about the simultaneous presence of linkage and linkage disequilibrium (Lazzeroni and Lange, in press). Once linkage has been established, it can be preferable to ignore the issue of confounding and to use all information available for localizing the gene.

Previously, fine-scale linkage-disequilibrium mapping has followed one of three general strategies. The simplest approach is to examine each locus individually. The gene is mapped relative to a locus that provides strong evidence of linkage disequilibrium. For a review of several disequilibrium measures used for this purpose, see the work of Devlin and Risch (1995). A second approach,

Received July 3, 1997; accepted for publication November 11, 1997; electronically published January 16, 1998.

Address for correspondence and reprints: Dr. Laura C. Lazzeroni, HRP Redwood Building, T101D, Stanford University, Stanford, CA 94305-5405. E-mail: laura@osiris.stanford.edu

*Present affiliation: Division of Biostatistics and Department of Genetics, Stanford University.

0002-9297/98/6201-0024$02.00

haplotype analysis, uses multilocus data to identify part of the chromosome where many affected individuals share a common haplotype. Haplotype analysis often relies more on the geneticist's intuition and less on formally defined procedures. In practice, geneticists often use both single-locus analysis and haplotype analysis to search for a gene (e.g., see Hästbacka et al. 1994; Goddard et al. 1996). Recently, several authors have explored a third approach, in which likelihood analysis is applied to multilocus data. In contrast to haplotype analysis, likelihood analysis requires an explicit probability model for the data. Some likelihood methods (Ramsay et al. 1993; Hill and Weir 1994; Kaplan et al. 1995; Xiong and Guo 1997) have formalized aspects of haplotype analysis, with the goal of determining the relative order of the disease gene and a pair of marker loci. Other methods (Terwilliger 1995; Devlin et al. 1996; Xiong and Guo 1997) depend on likelihoods for single-locus linkage-disequilibrium data, combining these likelihoods by multiplication for larger numbers of loci.

The present paper describes a somewhat different multilocus strategy. It takes the view that the pattern of disequilibrium along the chromosome is more informative than the amount of disequilibrium. From this perspective, linkage-disequilibrium mapping is a regression or curve-fitting problem. The semiparametric method used to estimate the regression curve and corresponding gene location includes a novel application of the bootstrap, which replaces many detailed population assumptions.

Below, the section "A Model for Disequilibrium" develops a general expression for the well-known parameter δ in terms of identity-by-descent probabilities. This interpretation makes explicit the requisite underlying independence assumptions implied by stricter population-genetics models. Because of the simplicity of the derivation, it is possible to give very general conditions under which it holds. With this expression, δ can be written as a piecewise function of the position of the marker locus on the chromosome. The maximum value of δ occurs at the disease locus, where the two pieces of the curve intersect. Many different sets of detailed population assumptions lead to this same general pattern of linkage disequilibrium.

Subsequently, the section "Statistical Methods" tells how to estimate the gene location by use of an empirical least-squares strategy. The regression model above describes the value of δ as a function of the position of the marker locus on the chromosome. However, the model says nothing about the sampling properties of estimates of δ at a set of loci. Using the bootstrap to explore the joint sampling distribution of these statistics, one can obtain (1) a transformation that normalizes their distributions, (2) adjustments that remove possible bias, and (3) an estimate of their covariance structure. The bootstrap results are used to fit the proposed curve to the transformed, adjusted estimates, in order to estimate the gene location. Kerem et al.'s (1989) cystic fibrosis (CF) data illustrate the statistical methods and give insight into the nature of linkage-disequilibrium data.

A Model for Disequilibrium

Classifying chromosomes according to "disease" (D) or "normal" (N) status is the first step in developing a linkage-disequilibrium model. Of course, if the gene has not yet been cloned, it is impossible to know with certainty whether a disease mutation is present. Instead, this paper defines chromosomes that segregate with individual disease status to be "disease" chromosomes and defines "normal" chromosomes analogously. For a recessive disease, for example, both chromosomes from an affected individual are considered to be "disease" chromosomes. Both untransmitted chromosomes from that person's unaffected parents are considered to be "normal" chromosomes. In some cases, it is possible to refine this classification scheme, discarding ambiguous cases in order to enrich the "disease" chromosomes for the presence of ancestral mutations, relative to the "normal" chromosomes. Although chromosomes are classified in terms of observable status rather than in terms of the actual presence of a mutation, for convenience the quotation marks will be omitted. Because of heterogeneity, incomplete penetrance, and other factors, normal chromosomes, as defined here, can actually carry a disease mutation while disease chromosomes might not.

The model presented here depends on δ, a well-known linkage-disequilibrium measure first introduced in this context by Bengtsson and Thomson (1981). Another name for δ that has appeared in the literature is "P_{excess}" (Lehesjoki et al. 1993). Suppose that $P(A \mid D) \geqslant P(A \mid N)$, where $P(A \mid D)$ is the probability that a disease chromosome carries marker allele A and where $P(A \mid N)$ is the probability that a normal chromosome carries A. For such an allele, the parameter δ is defined to be

$$\delta = \frac{P(A \mid D) - P(A \mid N)}{1 - P(A \mid N)} .$$

By definition, $0 \leqslant \delta \leqslant 1$, with larger values indicating greater disequilibrium. Devlin and Risch (1995) point out a major advantage that δ has over most other measures of linkage disequilibrium. Because δ is a function of the two conditional probabilities, $P(A \mid D)$ and $P(A \mid N)$, it can be estimated from so-called case-control data consisting of separate samples of disease and normal chromosomes.

Completing the definition of δ for multiallelic loci requires that A be specified more precisely. In the fol-

lowing subsection, "An Interpretation of δ"), A represents a set of distinct alleles. This set has a higher probability of being represented on a disease chromosome than on a normal chromosome. Let a be the complementary composite allele consisting of the remaining alleles at the same locus. Thus, $P(a \mid D) = 1 - P(A \mid D)$, and $P(a \mid N) = 1 - P(A \mid N)$. A useful, equivalent definition of δ is

$$\delta = \frac{P(a \mid N) - P(a \mid D)}{P(a \mid N)} = 1 - \frac{P(a \mid D)}{P(a \mid N)} .$$

An Interpretation of δ

A general expression for δ can be written in terms of identity-by-descent probabilities. Suppose that disease-predisposing mutations were introduced into the population on one or more ancestral chromosomes. By assumption, many disease chromosomes today are descended, at the disease locus, from one of these ancestral founder chromosomes. For a given marker locus, let A be the composite allele consisting of all alleles present at the marker locus on at least one founder chromosome. Let a be the complementary composite allele consisting of the remaining alleles at that locus.

I now extend the definition of identity by descent at a single locus, in order to apply it to the interval between two loci. In this setting, when the three following conditions are met, a present-day chromosome and a founder chromosome are said to be identical by descent along the interval between the marker and the disease locus. (1) The interval on the former chromosome is a direct descendant of the same interval on the latter chromosome. (2) There has been no recombination anywhere on the interval in any intervening generation. (3) There has been no mutation at either of the two end loci in any intervening generation.

Let $P(I \mid D)$ and $P(I \mid N)$ be the probability of identity by descent along the mutation-marker interval, for disease and normal chromosomes, respectively. All such chromosomes carry a copy of allele A at the marker locus, by definition. If it is assumed that all remaining chromosomes have the same probability, say $P(a \mid \text{not } I)$, of carrying marker allele a, it follows that

$$P(a \mid D) = P(a \mid \text{not } I)[1 - P(I \mid D)]$$

and

$$P(a \mid N) = P(a \mid \text{not } I)[1 - P(I \mid N)] .$$

Consequently,

$$\delta = \frac{P(a \mid N) - P(a \mid D)}{P(a \mid N)} = \frac{P(I \mid D) - P(I \mid N)}{1 - P(I \mid N)} .$$

Note that δ depends on the marker locus only through the interval-identity-by-descent probabilities and not through the allele probabilities at the marker locus. For a rare disease, $P(I \mid N)$ is negligible and δ is equivalent to the probability that a disease chromosome is identical by descent, along the interval between the marker and the disease locus, to one of the founder chromosomes. When $P(I \mid N)$ is not negligible, application of the Taylor series expansion, $1/(1-x) = 1 + x + \ldots + x^n + O(x^{n+1})$ for $|x| < 1$, yields the approximation

$$\delta = \frac{P(I \mid D) - P(I \mid N)}{1 - P(I \mid N)} \approx [P(I \mid D) - P(I \mid N)] \times \left[1 + \sum_{j=1}^{J} P(I \mid N)^j\right] .$$

Let $P(M \mid D)$ and $P(M \mid N)$ be the probability of identity by descent for the mutation among disease and normal chromosomes, respectively. Let $P(I \mid M)$ be the conditional identity-by-descent probability for the interval, given identity by descent for the mutation, under the further assumption of conditional independence of disease status. It follows that $P(I \mid D) = P(M \mid D)P(I \mid M)$ and that $P(I \mid N) = P(M \mid N)P(I \mid M)$. After substitution into δ, these yield

$$\delta \approx P(I \mid M)[P(M \mid D) - P(M \mid N)] \times \left[1 + \sum_{j=1}^{J} P(M \mid N)^j P(I \mid M)^j\right] ; \quad (1)$$

that is, δ can be approximated by a low-degree polynomial function of $P(I \mid M)$.

The δ Curve

Here standard approximations are used to express $P(I \mid M)$ as a function of the physical distance between the marker locus and the disease mutation. Combining this result with equation (1) yields a regression curve describing the pattern of δ relative to the mutation site.

First suppose that the disease mutation appeared on a single founder chromosome G generations ago. Let τ be the probability that no mutation occurred at the

marker locus in the intervening G generations. Let θ be the recombination fraction between the marker locus and the disease mutation. Then,

$$P(I \mid M) = \tau(1-\theta)^G \approx \tau + \tau \sum_{j=1}^{J} \binom{G}{j} (-\theta)^j \ , \quad (2)$$

where

$$\binom{G}{j} = \frac{G \times (G-1) \times \ldots \times (G-j+1)}{j \times (j-1) \times \ldots \times 1} \ .$$

The binomial approximation in equation (2) is accurate up to an error of the order of θ^{J+1}. Under stricter assumptions, equation (2) implies that $\delta = P(M \mid D)(1-\theta)^G$, which is equivalent to the expression given by Lehesjoki et al. (1993). Suppose now that there are multiple founder chromosomes. Let π_k be the conditional probability that a chromosome inherits the mutation from the kth founder, given that it inherits some founder mutation. Let G_k be the number of generations since the kth founder, and let τ_k be the probability of no marker mutation in G_k generations. Then,

$$P(I \mid M) = \sum_k \pi_k \tau_k (1-\theta)^{G_k} \approx \sum_k \pi_k \tau_k + \sum_k \pi_k \tau_k \sum_{j=1}^{J} \binom{G_k}{j} (-\theta)^j \ . \quad (3)$$

Next, let y be the distance between the marker locus and the mutation site, measured in Morgans. Simple Taylor-series expansions of most mapping functions in the genetics literature (Ott 1991) have the general form

$$\theta \approx y + \sum_{j=2}^{J} b_j y^j \ , \quad (4)$$

where b_j is a constant for $j = 2, \ldots, J$. For example, under Morgan's mapping function, in which there is complete positive chiasma interference, $\theta = y$. Under Haldane's mapping function, in which there is no chiasma interference,

$$\theta = \frac{1 - \exp(-2y)}{2} \approx y - y^2 \ .$$

The implicit assumption here—that θ is a monotonically increasing function of genetic distance—could fail only if there were very strong negative interference. In humans, interference is generally thought to be positive.

Last, let x be the physical location of the marker locus, and let μ be the site of the disease-predisposing mutations. An implicit assumption underlying virtually all linkage-disequilibrium mapping is that genetic and physical distance are roughly proportional on the region spanned by the observed marker loci; that is,

$$y \approx b\,|x - \mu| \quad (5)$$

for some constant b.

When equations (1), (3), (4), and (5) are combined by substitution, δ can at last be approximated by

$$g(x) = \sum_{j=0}^{J} \beta_j |x - \mu|^j \ . \quad (6)$$

To avoid unneeded detail, $g(x)$ is written as a generic polynomial in $|x - \mu|$. Each coefficient β_j, $j = 0, \ldots, J$, is an unknown constant incorporating various terms in the derivation. Along the chromosome, this approximation of δ follows a piecewise polynomial curve that is symmetric about the mutation site, where the two pieces intersect. A small value of J gives a reasonable degree of accuracy, because of the small amount of error introduced at each step of the approximation. In view of the monotonic nature of each approximated function, δ is a decreasing function of $|x - \mu|$. Thus, $g(x)$ should also be a decreasing function of $|x - \mu|$ on the range requiring an accurate approximation.

What if the disease gene lies near a hotspot of increased recombination activity? In that case, the region of increased activity has a greater genetic length than would be suggested by its physical length, violating the proportionality assumption. Within such a region, the curve descends more quickly than would otherwise be the case, altering the coefficients—and, possibly, the degree—of the best-fitting polynomial approximation. A region of decreased activity leads to similar distortion. Of course, it is unlikely that the distortion on one side of the gene will mirror that on the other. Thus, I propose an asymmetric model in which the curve

$$g(x) = \begin{cases} \beta_0 + \sum_{j=1}^{J} \beta_{1j} |x - \mu|^j, & \text{if } x \leqslant \mu \\ \beta_0 + \sum_{j=1}^{J} \beta_{2j} |x - \mu|^j, & \text{if } x > \mu \end{cases} \quad (7)$$

is composed of two distinct polynomials that intersect at the mutation site. As before, β_0 and β_{ij} are unknown constants for $i = 1, 2$ and $j = 1, \ldots, J$, and $g(x)$ is constrained to decrease as $|x - \mu|$ increases on each side of μ. Asymmetry has a historical interpretation as well as a biological one. It can reflect the widespread propagation of individual recombination events that occurred during the early history of a mutation.

Implications of the Model

The model above is quite general. It does not depend on a formal stochastic model for the development of the population. Instead, the derivation proceeds backward

Table 1

CF Data for 23 Marker Loci

i	x_i	$\hat{\delta}_i$	P$f(\hat{\delta}_i)$ ± SE
1	0	.419	.722 ± .063
2	9	.674	.511 ± .144
3	24.8	.246	.844 ± .086
4	524.8	.465	.687 ± .100
5	534.8	.520	.644 ± .096
6	554.8	.431	.713 ± .053
7	569.8	.415	.725 ± .054
8	594.8	.586	.589 ± .099
9	614.8	.763	.421 ± .063
10	619.8	.771	.413 ± .062
11	654.8	.767	.417 ± .062
12	684.8	.786	.397 ± .076
13	709.8	.779	.404 ± .069
14	744.8	.786	.397 ± .076
15	779.8	.796	.385 ± .072
16	859.8	.693	.493 ± .056
17	869.8	.701	.485 ± .055
18	889.8	.551	.619 ± .131
19	899.8	.784	.398 ± .068
20	949.8	.667	.517 ± .224
21	1,599.8	.273	.826 ± .048
22	1,669.8	.340	.779 ± .059
23	1,769.8	.316	.796 ± .055

in time, using only a few conditional independence assumptions. Specific population-genetics models typically include stricter assumptions, such as random mating, which lead to the same type of conditional independence. The simplicity of the approach taken here makes it possible to simultaneously cover a number of complications, including multiple alleles, multiple founders, disease heterogeneity, normal carriers, mutation, and chiasma interference. The asymmetric curve further extends the model, allowing the intensity of recombination activity to vary with physical location.

The piecewise curve has some general implications for linkage-disequilibrium mapping. Consider the simple symmetric model with a single founder and $J = 1$. In that case,

$$\delta \approx \tau[P(M \mid D) - P(M \mid N)](1 - Gb|x - \mu|) \ .$$

Thus, $\beta_0 = \tau\,[P(M \mid D) - P(M \mid N)]$, $\beta_1 = -\beta_0 Gb$, and $|x - \mu| \approx (\delta - \beta_0)/\beta_1$. To estimate the distance $|x - \mu|$ from the value of δ at a single locus, the values of τ, $P\,(M \mid D)$, $P\,(M \mid N)$, G, and b are needed. Inaccurate assumptions or imprecise external estimates of these population quantities carry over to the estimated distance. Consequently, single-locus linkage-disequilibrium mapping is rarely effective for precise gene localization. Estimates of δ at multiple loci can yield estimates of β_0 and β_1. However, these alone will be insufficient to determine underlying population values, such as $P\,(M \mid D)$. Finding such values requires assumptions about other aspects of the population. Fortunately, this is not a problem when one is estimating the gene location. In fact, generic parameters, such as β_0, absorb the impact of incorrectly specified population quantities, suggesting that multilocus disequilibrium mapping can be robust to inaccurate population assumptions.

Statistical Methods

To estimate the gene location, one can fit the piecewise curve described above to estimates of δ at a set of L marker loci. The data consist of the haplotypes of one sample of disease chromosomes and another sample of normal chromosomes. The haplotypes should span the region where the gene is believed to lie. Although the method is easily modified to handle data collected separately at each locus, the ability to incorporate the additional information in haplotype data is a major advantage of this approach. A map of the marker locations, usually a work in progress at the time, is also needed. This paper addresses estimation with biallelic markers only.

Let δ_i be the disequilibrium parameter at locus i. To estimate δ_i, let $\hat{a}_i$ be the allele, at locus i, that is relatively more frequent in the normal sample than it is in the disease sample. Thus, $\hat{a}_i$ estimates which allele is associated with normal status in the population. Among the disease chromosomes typed at i, let p_i be the proportion with allele $\hat{a}_i$. Define r_i similarly for the normal chromosomes. The obvious estimate of δ_i is $\hat{\delta}_i = 1 - p_i/r_i$. If $\hat{a}_i$ is undefined because the allele frequencies are the same in both samples, $\hat{\delta}_i = 0$.

The CF data originally published in *Science* (Kerem et al. 1989) serve to illustrate the statistical methods. Table 3 of the *Science* paper presents the haplotypes of 94 disease chromosomes from a sample of affected individuals, as well as the haplotypes of 92 untransmitted normal chromosomes from their unaffected parents. Complete haplotypes include 23 RFLP loci spanning 1,770 kb. Ninety chromosomes used in the analysis have incomplete haplotypes.

Table 1 in the present paper shows the markers developed and mapped during the search for the CF gene and reported in the *Science* paper. The locations are standardized so that x_i is the number of kilobases between the ith locus and the first locus, metD. The CF gene, as shown on the *Science* map, extends from $x = 790$ to $x = 990$ (rounded to the nearest 10 kb) and includes loci 16–20. The CF mutation Δ508, present in 67% of the affected sample, lies between $x = 880$ and $x = 885$, between the 17th and 18th loci. In theory, linkage-disequilibrium mapping should point to this location.

The CF disequilibrium estimates (table 1 and fig. 1) follow a decidedly irregular pattern. Of the eight loci with estimates greater than .76, seven are clustered together just outside the gene. The remaining such locus

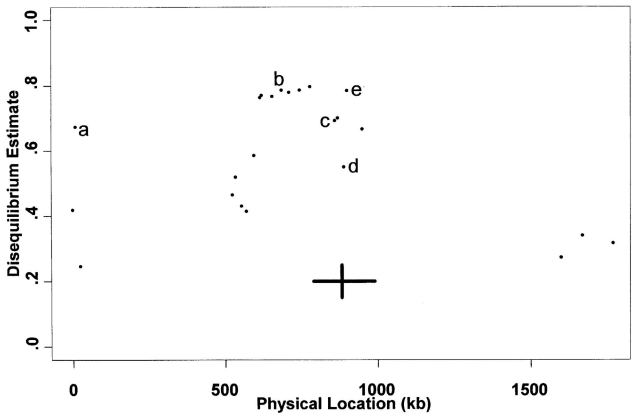

Figure 1 Original $\hat{\delta}_i$s for CF data. Loci 9–15 (*b*) and 19 (*e*) have high values and are highly correlated. Locus 18 (*d*) within the gene has a much lower value. Loci 16 and 17 (*c*) are intermediate. The estimate at locus 2 (*a*) is unusually high, given its distance from the CF gene, which is denoted by the horizontal bar. The vertical crosspiece represents the location of the Δ508 mutation.

is within the gene, very near Δ508. Four other loci inside the gene have lower estimates, ranging from .55 to .70. In contrast, $\hat{\delta}_2 = .67$ is as high, although that locus is nearly 800 kb away from the gene. Its immediate neighbors to either side show much less linkage disequilibrium.

The Bootstrap

To accurately estimate the gene location, it is necessary to understand the statistical properties of the $\hat{\delta}_i$s and their relationship to the true unknown δ_is. A parametric model for the joint sampling distribution of the statistics would require a number of additional assumptions. An alternative is to use the empirical distribution based on the observed data. The bootstrap is a statistical tool that, by resampling the observed data, simulates the empirical distribution of a statistic (Efron 1979).

The bootstrap distribution is obtained by randomly sampling *B* separate bootstrap data sets, as follows. Let *m* and *n* be the numbers of disease and normal chromosomes, respectively, in the original data. To create each bootstrap data set, *m* "bootstrap" disease chromosomes are drawn randomly with replacement from the *m* disease chromosomes in the original data. Similarly, *n* "bootstrap" normal chromosomes are drawn randomly with replacement from the *n* normal chromosomes in the original data. Each chromosome is sampled as a unit including all the typed alleles in its haplotype and any missing data.

On the basis of the chromosomes in the *b*th bootstrap data set, a bootstrap estimate $\hat{\delta}_i^*(b)$ is computed for each locus *i*, in the same way as $\hat{\delta}_i$ was computed on the basis of the original data. Let $\hat{a}_i^*(b)$ be the allele that is more frequent among the normal chromosomes than the disease chromosomes of the *b*th bootstrap data set. Let $p_i^*(b)$ and $r_i^*(b)$ be the proportions of allele $\hat{a}_i^*(b)$ among the disease and the normal chromosomes, respectively, in the *b*th data set. The *b*th bootstrap estimate at locus *i* is $\hat{\delta}_i^*(b) = 1 - p_i^*(b)/r_i^*(b)$. As in the original data, $\hat{\delta}_i^*(b) = 0$ if the frequencies in the disease and normal samples of the *b*th data set are the same.

Each bootstrap data set produces one set of estimates, $\hat{\delta}_1^*(b), \ldots, \hat{\delta}_L^*(b)$, on the basis of the same set of resampled chromosomes. The variability of these bootstrap estimates tells us about the variability of $\hat{\delta}_1, \ldots, \hat{\delta}_L$. Because each bootstrap estimate is based on a sample from the original data, it behaves as a statistical estimate of the disequilibrium present in the original data; that is, each $\hat{\delta}_i^*(b)$ is an estimate of $\hat{\delta}_i$, just as $\hat{\delta}_i$ is an estimate of δ_i. By replicating the resampling process many times, one can observe the random behavior of $\hat{\delta}_i^*$ with respect to $\hat{\delta}_i$. This bootstrap distribution provides an empirical estimate of the statistical behavior of $\hat{\delta}_i$ with respect to δ_i.

The following details are worth noting: (1) Because of the resampling of missing values, the number of chromosomes typed at a locus in a given bootstrap data set can differ from the number typed in the original data. (2) By chance, all resampled chromosomes in both samples of the *b*th bootstrap data set might share the same allele at locus *i;* in that case, $\hat{\delta}_i^*(b)$ is undefined. (3) The sampling distribution of $\hat{\delta}_i$ depends partly on the probability that $\hat{a}_i$, the disease-associated allele in the original data, is not the same as a_i, the disease-associated allele in the population. The bootstrap replicates this feature, since the disease-associated allele in the *b*th bootstrap sample $\hat{a}_i^*(b)$ can, in turn, differ from $\hat{a}_i$. For the CF data, this occurs at the 3rd locus, for 13% of the bootstrap data sets, and at the 22nd locus, for 7% of the bootstrap data sets. Five other loci behave similarly, at rates no greater than 2%.

Normalizing Transformation

The generalized least-squares procedure used to estimate the regression curve and corresponding mutation site works best if the data come from a symmetric distribution, such as the normal distribution (Seber and Wild 1989). The bootstrap distribution can be used to check how close the observed $\hat{\delta}_i$s are to normality. For the CF data, I resampled $B = 2,000$ bootstrap data sets and calculated the skewness of the resulting 2,000 bootstrap estimates, at each locus. The bootstrap estimates at all loci except one are negatively skewed.

An appropriate transformation of $\hat{\delta}_i$ can reduce the apparent asymmetry. The same transformation is used for all loci. For CF, I considered transformations of the form $f(\delta) = \delta^\gamma$ and $f(\delta) = (1 - \delta)^\gamma$, for various values of γ. I used each candidate to transform the bootstrap es-

timates and then recalculated the skewness at each locus. Table 2 summarizes the results. Although no perfect transformation eliminates the asymmetry at all loci, the selected transformation $f(\delta) = (1 - \delta)^{.6}$ reduces the maximum absolute skewness at any locus, from .88 to .37. It also sets the median and mean skewness close to zero. Applying this transformation to the original estimates yields the values in table 1.

Does this transformation change the form of the model? If δ is a polynomial in $|x - \mu|$, say $g'(x)$, then $(1 - \delta)^{.6} \approx 1 - .6g'(x) - .12g'(x)^2 + ...$ Thus, $(1 - \delta)^{.6}$ can still be approximated by $g(x)$, where $g(x)$ is another polynomial in $|x - \mu|$. Because $(1 - \delta)^{.6}$ is a decreasing function of δ, $g(x)$ now increases as $|x - \mu|$ increases, and lower values correspond to greater disequilibrium. Except for this change in direction, either the symmetric model in equation (6) or the asymmetric model in equation (7) should still hold. In principle, the same is true for any transformation $f(\delta)$ that is a smooth, finite, monotonic function on the [0,1] interval.

It is possible that the transformed estimate $f(\hat{\delta}_i)$ is a biased estimate of $f(\delta_i)$. The bootstrap provides an estimate of the bias, which can then be removed; for details, see the Appendix. Let $\hat{f}_i$ be the bias-corrected version of the transformed estimate $f(\hat{\delta}_i)$.

Table 2

Summary of Bootstrap Skewness for 23 Loci

	Skewness			
Transformation	Minimum	Median	Mean	Maximum
δ	−.88	−.29	−.36	.08
$\delta^{1.3}$	−.56	−.22	−.18	.41
$\delta^{1.4}$	−.47	−.19	−.12	.51
$\delta^{1.5}$	−.39	−.17	−.07	.61
$\delta^{1.6}$	−.32	−.13	−.02	.71
$(1 - \delta)^{.4}$	−.91	−.11	−.14	.20
$(1 - \delta)^{.5}$	−.41	−.05	−.04	.29
$(1 - \delta)^{.6}$	−.22	−.01	.05	.37
$(1 - \delta)^{.7}$	−.18	.08	.13	.46

Note.—The top row shows values for the 23 loci as estimated by the bootstrap for the untransformed estimate; the remaining rows show the same quantities for eight possible transformations.

Covariance

The primary motivation for using the bootstrap is to estimate the covariance matrix of the transformed disequilibrium estimates. The bootstrap-estimated covariance matrix plays an important role both in efficiently estimating the gene location and in constructing a confidence interval to describe the accuracy of that estimate.

Because the transformed estimates at different loci are based on the same set of chromosomes, they are clearly correlated. In fact, the feasibility of linkage-disequilibrium mapping may depend on the presence of such correlation. Contrary to a common misconception, correlated data can sometimes yield better estimates than are provided by uncorrelated data; in linear regression, for example, positive correlation increases the variance of the intercept estimate but reduces the variance of the slope estimate. In fact, under some circumstances, two perfectly correlated observations are sufficient for the slope of a line to be estimated with complete certainty. For linkage-disequilibrium mapping, the consequences of correlated data are less clear.

The covariance structure of a set of disequilibrium estimates probably reflects several historical factors, including variable mutation rates, population admixture, and multiple founder haplotypes. The variances of the estimates at individual loci depend on the allele frequencies in the disease chromosomes and in the normal chromosomes. Correlations occur between pairs of estimates if there is linkage disequilibrium between the two marker loci in either group. Both CF samples, in fact, exhibit very high levels of linkage disequilibrium between some markers. In any given setting, it is difficult to design a population model that generates a covariance structure similar to the true one. The empirical covariance matrix of the transformed estimates is a readily available alternative that can be estimated from the transformed bootstrap estimates; for details, see the Appendix.

Table 1 shows the bootstrap-estimated standard error (SE) values of the transformed CF estimates. The large range, .048 to .224, shows that $f(\delta_i)$ has been much better estimated at some loci than at others. Ten pairwise correlations are at least .9, and 53 are greater than .5. These strongly correlated pairs typically consist of either neighboring markers or markers that are both close to the disease gene. However, there are several exceptions to this anticipated pattern.

Loci 9–15 and 19 form an interesting example. The estimated disequilibrium at these eight loci is greater than that at all other loci. The pairwise correlations among the eight are all also high, ranging from .68 to 1.00. In contrast, the estimate at the 18th locus, which shows much less disequilibrium, is negatively correlated (between −.19 and −.09) with the eight previous estimates. The estimates at the 16th and 17th loci are moderately correlated (between .52 and .65) with all nine previous estimates. A possible explanation may be that, as suggested by Kerem et al. (1989), two distinct haplotypes are segregating with Δ508. The two haplotypes proposed by those authors agree at loci 9–15, differ at loci 16–18, and agree again at locus 19.

Estimating the Gene Location

The gene location μ can now be estimated by use of generalized least squares (GLS), to find the best-fitting curve under each plausible model of the general class

Table 3

Results for the CF Data

Model	$Q(\hat{\mu},\hat{\beta})$	p[a]	$Q(\hat{\mu},\hat{\beta})+2p$	$\hat{\mu}$
Symmetric:				
$J=1$	40.18	3	46.18	790.0
$J=2$	39.51	4	47.51	796.0
$J=3$	39.51	5	49.51	796.0
Asymmetric:				
$J=1$	37.82	4	45.82	892.6
$J=2$	35.12	6	47.12	899.8
$J=3$	35.03	8	51.03	899.8

NOTE.—For these data, AIC chooses the asymmetric model with $J = 1$.

[a] For symmetric models, $p = J + 2$; for asymmetric models, $p = 2 + 2J$.

suggested above in the section "A Model for Disequilibrium." The final model, selected by comparison of the GLS results, yields an estimate and confidence interval for μ. The following are the underlying assumptions: For locus $i = 1, ..., L$, the transformed, bias-corrected estimate $\hat{f}_i$ is an unbiased estimate of $f(\delta_i)$. The value of $f(\delta_i)$ can be approximated by $g(x_i)$, where $g(x)$ is given by either the symmetric curve

$$g(x) = \sum_{j=0}^{J} \beta_j |x-\mu|^j$$

or the asymmetric curve

$$g(x) = \begin{cases} \beta_0 + \sum_{j=1}^{J} \beta_{1j} |x-\mu|^j, & \text{if } x \leqslant \mu \\ \beta_0 + \sum_{j=1}^{J} \beta_{2j} |x-\mu|^j, & \text{if } x > \mu \end{cases}.$$

In addition, μ satisfies the range constraint $c_L \leqslant \mu \leqslant c_U$ for constants c_L and c_U, chosen on the basis of prior information about the gene location. The defaults $c_L = x_1$ and $c_U = x_L$ keep the location estimate within the span of the observed markers. For another constant, c_M, $g(x)$ satisfies the monotonicity constraint on each side of the mutation. If $f(\delta)$ is a decreasing (increasing) function of δ, $g(x)$ is an increasing (decreasing) function of $|x-\mu|$, on the range $|x-\mu| \leqslant c_M$. By default, $c_M = x_L - x_1$, forcing the curve on each side of μ to be monotonic across the observed range of the markers for any possible μ. Last, the covariance matrix of $\hat{f}_i$ is consistently estimated by V, the smoothed bootstrap covariance matrix described in the Appendix. With these assumptions, the choice of symmetry or asymmetry and the degree J of the approximation specify a particular model of this class.

The GLS estimate of the curve $g(x)$ for a given model is defined as follows (e.g., see Seber and Wild 1989): Let $\boldsymbol{\beta} = \{\beta_0, ..., \beta_J\}$ for a symmetric model; for an asymmetric model, let $\boldsymbol{\beta} = \{\beta_0, \beta_{11} ... \beta_{1J}, \beta_{21}, ..., \beta_{2J}\}$. To emphasize the dependence on the unknown parameters μ and $\boldsymbol{\beta}$ let $g_i(\mu,\beta) = g(x_i)$. Let $\mathbf{D}(\mu,\boldsymbol{\beta})$ be a vector of length L with the ith element equal to the difference $\hat{f}_i - g_i(\mu,\beta)$. Last, let the quadratic form $Q(\mu,\boldsymbol{\beta}) = \mathbf{D}(\mu,\boldsymbol{\beta})^T \mathbf{V}^{-1} \mathbf{D}(\mu,\boldsymbol{\beta})$. The GLS estimates $\hat{\mu}$ and $\hat{\boldsymbol{\beta}}$ are the values of μ and $\boldsymbol{\beta}$ that, among those values satisfying the range and monotonicity constraints, minimize $Q(\mu,\boldsymbol{\beta})$. The GLS criterion takes the estimated covariance structure into account in deciding which curve best fits the observed $\hat{f}_i$s. For example, GLS fits the curve more closely to estimates with smaller variances. It is now useful to add a normality assumption justified, at least approximately, by the transformation and the substantial number of observations underlying each estimate. Under the assumption that $\hat{f}_i$ is multivariate normal with mean $g_i(\mu,\boldsymbol{\beta})$ and covariance matrix $\mathbf{V}$, the log likelihood evaluated at μ and $\boldsymbol{\beta}$ is $-Q(\mu,\boldsymbol{\beta})/2$. In that case, the GLS estimates $\hat{\mu}$ and $\hat{\boldsymbol{\beta}}$ are also maximum-likelihood estimates.

Within the likelihood framework, Akaike's (1974) information criterion (AIC), a model-selection technique, can be used to select a specific curve under which to estimate the mutation site. Adding parameters, even meaningless ones, almost always increases the log likelihood and can never decrease it. However, such overparameterization can increase statistical variability, yielding estimates worse than those produced by more-parsimonious models. To avoid overfitting, AIC requires each additional parameter to increase the log-likelihood by at least one. In this setting, AIC is equivalent to minimizing $Q(\hat{\mu},\hat{\boldsymbol{\beta}}) + 2p$, where p is the number of parameters in the model.

For CF, AIC chooses the asymmetric piecewise-linear

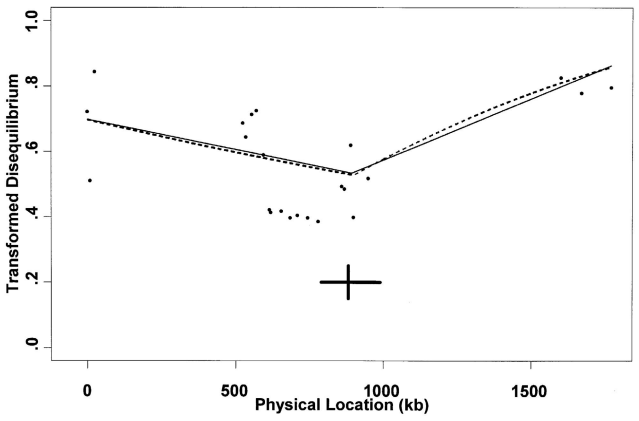

Figure 2 Fitted curves. The unbroken line is the fitted curve under the final first-order asymmetric model, and the broken line is the fit under the second-order asymmetric model. The points are the observed, transformed disequilibrium estimates. As can be seen, the greatest deviation between the two curves occurs within the large gap between the 20th and 21st markers. Both curves place the gene location very close to the vertical dashed line showing the location of Δ508. The horizontal crosspiece represents the entire CF gene.

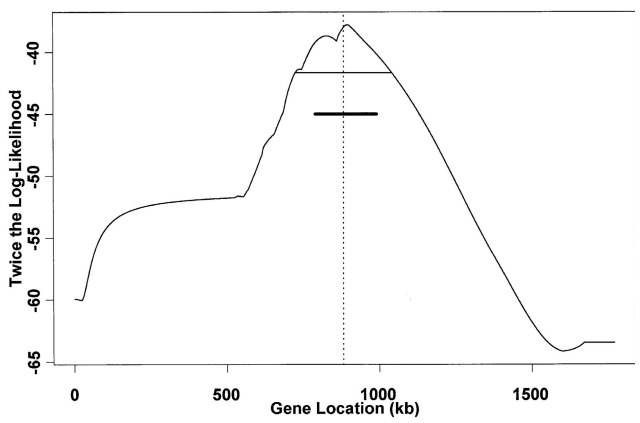

Figure 3 Twice the log likelihood under the final model. Equivalently, this is the least-squares curve, pictured upside down. The vertical line (which in this case is broken) and the lower, thicker horizontal crosspiece (represent, respectively, the location of Δ508 and the approximate extent of the CF gene. The upper, thinner horizontal line intersects the log-likelihood curve at the critical values for the .05 significance test. All values of μ for which the curve is above this line are included in the 95% confidence interval (724.1, 1,039.9).

model (table 3). This model places the estimated gene location at $\hat{\mu} = 892.6$, only ~10 kb away from Δ508 (fig. 2). Confidence intervals for μ are obtained by inverting a likelihood-ratio test, as described in the Appendix. For the CF data, the 95%, 90%, and 80% confidence intervals for the gene location are (724.1, 1,039.9), (760.1, 1,002.7), and (782.6, 963.3), respectively. Figure 3 shows the first of these intervals superimposed on the log-likelihood curve. An alternative strategy is to start with a predetermined length and find the corresponding interval and its achieved confidence level. For example, the 50-kb interval (872.0, 922.0) achieves a 54% confidence level and does, in fact, include the location of Δ508.

Undoubtedly, the right-hand side of the curve, with only five marker loci, is less well estimated than the left-hand side. In particular, because of the 650-kb gap between the 20th and 21st markers, there is almost no evidence with regard to the curvature on the right. Although it is almost impossible to distinguish between the second-order asymmetric curve (fig. 2) and the first-order model, the greatest deviation occurs within this uninformative gap. Because of the monotonicity constraint, increasing the degree of the polynomial sometimes fails to decrease $Q(\hat{\mu}, \hat{\boldsymbol{\beta}})$. For CF, the curve under the first-order model is the same regardless of whether the constraint is imposed. For higher-order models, the constraint excludes curves with implausible interpretations, such as those that place the gene intersection at a point of maximum transformed disequilibrium.

It is revealing to refit the data while modifying aspects of the original analysis. Table 3 shows the effect that the selected model has on the estimated location. The estimates under asymmetric models are always near Δ508, falling between 892.6 and 899.8. The estimates under symmetric models are near the beginning of the CF gene, between 790.0 and 796.0. If the correlation structure is ignored as if the $\hat{f}_i$s were independent observations, then asymmetric models do produce estimates similar to those of symmetric models (fig. 4). When only the diagonal of the bootstrap covariance matrix is used, the estimates under the asymmetric model are 781.0, for $J = 1$, and 779.8, for $J = 2$. Previous approaches that are equivalent to a symmetric, independence model obtained similar estimates (Terwilliger 1995; Devlin et al. 1996; Xiong and Guo 1997). For CF, the inclusion of asymmetry in the final model allows the regression procedure to better exploit the correlation structure. The fitted curve (fig. 2) can then roughly parallel the eight highly correlated estimates at loci 9–15 and 19.

In light of the CF results, some researchers might choose to consider only asymmetric curves. Convincing prior evidence that symmetry must hold is generally unlikely, yet it may not always be possible to detect asymmetry on the basis of the data; on the other hand, the impact that this choice has on the location estimate might be less when that is the case. An alternative might be to lower the criterion for moving from symmetry to asymmetry, although setting an appropriate level for that criterion would take some experience.

In contrast, the transformation choice does not affect the gene-location estimate very much. The other trans-

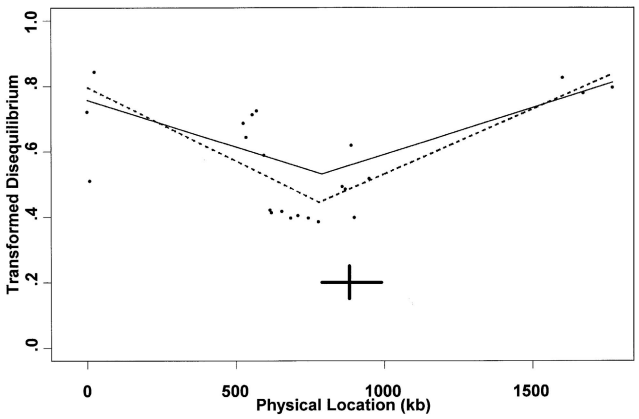

Figure 4 Alternative fitted curves. The unbroken line is the fitted curve under the first-order symmetric model, and the broken line is the fitted curve under the first-order asymmetric model and with independence being assumed. Both curves place the estimate near the front of the horizontal crosspiece representing the extent of the CF gene, well to the right of Δ508, which is shown by the vertical broken line.

formations that were considered, including all those in table 2, estimate the gene location to be near Δ508, between the 18th and 19th loci, under the asymmetric model with $J = 1$ or $J = 2$. The transformation does alter the confidence intervals, making them more conservative than intervals computed without a transformation. Because the transformation improves the fit of the normal approximation, this change probably reflects a more accurate, if less optimistic, assessment of the information in the data.

Future Improvements

Undoubtedly, there are many possible refinements of the methods described in this paper. Before mentioning a few possibilities, I will review the three basic steps of the estimation procedure. In the first step, the disequilibrium parameter δ_i is estimated for each locus. The second step uses the bootstrap to explore the joint sampling distribution of the estimates. The resulting information is used to better implement the third step: fitting the disequilibrium curve to the data.

For multiallelic marker loci, estimation of δ_i becomes more complicated. In particular, it is necessary to decide which alleles are associated with the disease. Even if it were possible, exhaustive enumeration of the marker alleles in the disease-mutation founders might not be the best approach for mapping the gene location. That approach would leave δ_i undefined at loci where every allele appeared on at least one founder chromosome. Instead, a better rule might attempt to identify only the alleles present in the founders representing the majority of today's cases. This paper does not propose a specific rule for deciding which alleles are disease associated at multiallelic loci. However, the impact of any such rule can clearly be assessed by including it within the bootstrap procedure.

Ideally, the resampling strategy of the bootstrap replicates the data-collection strategy. If family data are used, the chromosomes from each family should be resampled as a single unit. Unfortunately, family information for the CF data was unavailable. With such information, interesting comparisons could be made between the empirical distributions—and the location estimates—under alternative resampling schemes.

Currently, the method does not allow for disease mutations at multiple sites. However, the model could be generalized to express $g(x)$ as a weighted average of separate curves, each with a distinct peak representing a single mutation site. If the observed marker loci covered the region well enough, it might be possible to estimate such a model. More often, it would be difficult to distinguish the peaks, especially when they are close to each other. Thus, it would be useful to explore the behavior of the single-site estimate when multiple mutation sites are, in fact, present.

Likelihood theory, as used in this analysis, encounters certain limitations. For example, the likelihood-based confidence intervals act as if the selected model is the correct one, although there is no guarantee that the AIC choice is optimal, let alone correct. Although such intervals are standard in practice, they can be somewhat narrower than is justified by the available information. A second level of bootstrapping, which would replicate the entire estimation process, might produce a more accurate confidence interval for μ. Unfortunately, implementing this strategy is less straightforward than it sounds, since there are several possible approaches, all of which dramatically increase the required amount of computation. In the meantime, likelihood-based intervals are likely to achieve a reasonable degree of accuracy.

Discussion

The semiparametric linkage-disequilibrium mapping method proposed in this paper pays as much attention to the data as possible. The parametric part of the model describes only the trend of disequilibrium about the gene location. The bootstrap replaces additional explicit modeling that would otherwise be required for the joint sampling distribution of the disequilibrium estimates. This strategy reduces the risk of overparameterization and eliminates the need to depend on imprecise assumptions or external estimates of specific population quantities. On the basis of the fit to the data, a specific disequilibrium curve is chosen from among a class of plausible alternatives. As a consequence, the final model for a given data set will depend on the amount of data and on the observed marker spacing, as well as on the pattern of linkage disequilibrium in the population.

I have shown that varied sets of population assumptions lead to similar patterns of linkage disequilibrium. In fact, linkage-disequilibrium models proposed elsewhere also implicitly define a regression curve. For any given method, the space of permissible regression curves depends on which population quantities are externally specified. Because the underlying curves have similar shapes, it is not surprising that a variety of approaches obtain similar gene-location estimates. For CF, the empirical least-squares approach works particularly well. The combined use of asymmetry and the covariance structure substantially improves the location estimate in this case.

There is a growing consensus that linkage-disequilibrium mapping is a useful tool for gene localization. The CF example certainly reinforces that view, for this and other methods. In particular, multilocus methods can exploit the regression curve and the covariance structure to obtain information that is inherently different from the information contained in a single disequilibrium value. However, the details of the CF data show that real data can be very unlike ideal mathematical models.

Thus, only experience can establish the general feasibility of linkage-disequilibrium mapping.

Acknowledgments

This research was supported in part by U.S. Public Health Service grant GM53275 and by National Science Foundation grant DMS-9510516.

Appendix

Bias Correction

The bias of a parameter estimate is the amount by which the expectation of the estimate differs from the parameter. Thus, the bias of $f(\hat{\delta}_i)$ is $\mathrm{E}[f(\hat{\delta}_i)] - f(\delta_i)$. The bootstrap can be used to estimate and remove bias from $f(\hat{\delta}_i)$ (Efron and Tibshirani 1993).

To estimate the bias, the B bootstrap data sets are combined to form a single data set. For locus i, let $\hat{a}_i^{**}$ be the allele that, in the combined data set, appears more frequently among the normal chromosomes than among the disease chromosomes. Let p_i^{**} be the proportion of disease chromosomes in the combined data set that carry allele $\hat{a}_i^{**}$, and define r_i^{**} similarly for the normal chromosomes; then, the estimate for the combined data set is $\hat{\delta}_i^{**} = 1 - (p_i^{**}/r_i^{**})$. For a given transformation f, let the mean of the transformed bootstrap estimates at locus i be

$$\bar{f}_i^{*} = \frac{\sum_{b=1}^{B} f[\hat{\delta}_i^{*}(b)]}{B} .$$

Then, the bootstrap estimate of the bias of $f(\hat{\delta}_i)$ is $\bar{f}_i^{*} - f(\hat{\delta}_i^{**})$. For the transformed CF estimates, the bootstrap detects little bias. To remove bias, the estimated bias is subtracted from $f(\hat{\delta}_i)$, to give the bias-corrected estimate $\hat{f}_i = f(\hat{\delta}_i) - [\bar{f}^{*} - f(\hat{\delta}_i^{**})]$. Consequently, $\hat{f}_i$ is an approximately unbiased estimate of $f(\delta_i)$.

Covariance Matrix

For L loci, the $L \times L$ covariance matrix is computed as follows: The element in the ith row and jth column of the matrix is the sample covariance of the transformed bootstrap estimates $f(\hat{\delta}_i^{*})$ and $f(\hat{\delta}_j^{*})$, given by

$$\frac{\sum_{b=1}^{B} \left\{ f[\hat{\delta}_i^{*}(b)] - \bar{f}_i^{*} \right\} \left\{ f[\hat{\delta}_j^{*}(b)] - \bar{f}_j^{*} \right\}}{B - 1} .$$

The ith diagonal element of this matrix is the bootstrap estimate of variance for the transformed estimate at the ith locus.

As a technical note, the covariance matrix used in GLS must be positive definite (Seber and Wild 1989). Given the large number of estimated covariances, it is not too surprising that, if unmodified, the bootstrap covariance matrix for the CF data is not positive definite. Multiplying the off-diagonal elements by $1 - d$, where d is a small positive number, smooths the correlation structure, making the matrix positive definite. Because very high correlations can put strong constraints on the fitted line, it is better to risk underestimation of the correlations, in order to avoid overestimation of them. In effect, the smoothing parameter d relaxes the correlation structure in case some low-probability events, inconsistent with the estimated covariance, have been missed in the sample. A rule of thumb is to set $d \approx 2 - 2(1 - .5^{1/k})^{1/(m+n)}$, where k is the number of pairwise correlations above .98 in absolute value. For the CF data, $k = 3$ and $d = .02$ were chosen. In general, d should be kept small, subject to the following sensitivity diagnostic: For CF, the final model was refitted for several values of d, to evaluate its effect. For the range $.002 \leqslant d \leqslant .2$, the estimated gene location stayed between 890 and 900, indicating a lack of sensitivity to this choice.

GLS Optimization

All programming was done in the statistical language Splus. The routine "nlminb" was used to minimize the quadratic form. Splus functions are available from the author.

Although $Q(\mu,\boldsymbol{\beta})$ is multimodal, it has a unique minimum within each interval $x_i \leqslant \mu \leqslant x_{i+1}$. Thus, the global minimum is easily found by comparison of the local minima obtained from separate optimization procedures conducted within each interval inside the range constraint. Within an interval, absolute value signs in $Q(\mu,\boldsymbol{\beta})$ can be dropped, and the model becomes a constrained linear model. If an interval is such that few loci lie to one side, there may not be unique estimates of μ and $\boldsymbol{\beta}$, but the minimum value of $Q(\mu,\boldsymbol{\beta})$ inside the interval can still be found. If that value is the global minimum, then the estimate of μ should be placed inside the interval but should not be given an exact location. When $\mu = x_i$ or $\mu = x_{i+1}$, some partial derivatives are not continuous if the model is parameterized in terms of μ. However, the partial derivatives are continuous if the model is reparameterized in terms of z, where

$$\mu = \frac{x_i + x_{i+1}}{2} + \sin(z)\frac{x_{i+1} - x_i}{2} .$$

Evaluating the least-squares, or likelihood, surface for μ requires a second computational procedure. For each location at which the likelihood surface is evaluated, optimization is carried out over $\boldsymbol{\beta}$, with the value of μ held fixed.

Penalty functions are used to enforce the monotonicity

constraint. When the algorithm attempts to move to a set of parameter estimates for which the corresponding curve does not satisfy monotonicity, adding a penalty function to $Q(\mu,\boldsymbol{\beta})$ forces the algorithm to half-step until the constraint is satisfied. When the algorithm is currently near a constraint boundary and attempts to cross it, the numerical vector and matrix used to determine the search direction are modified. This allows the algorithm to search, if ncessary, along contraint boundaries.

Confidence Intervals

Inverting the following likelihood-ratio test of the gene location yields a confidence interval for μ under the final model. Let the null hypothesis be that μ_0, say, is the true mutation site. Let $\hat{\boldsymbol{\beta}}(\mu)$ be the value of $\boldsymbol{\beta}$ that minimizes $Q(\mu,\boldsymbol{\beta})$ when the value of μ is fixed. Let $\chi_1^2(1-\alpha)$ be the $(1-\alpha)100$th percentile of a χ^2 distribution with 1 df. The likelihood-ratio test rejects the null hypothesis $\mu = \mu_0$ at significance level α if

$$Q[\mu_0, \hat{\boldsymbol{\beta}}(\mu_0)] - Q\left(\hat{\mu}, \hat{\boldsymbol{\beta}}\right) \geqslant \chi_1^2(1-\alpha) \ .$$

A $(1-\alpha)100\%$ confidence interval consists of the μ values not rejected by the likelihood-ratio test at significance level α. This is obtained from the least-squares surface evaluated on a dense grid of points over the range of permissible mutation sites. The resulting interval is (μ_L, μ_U), where

$$\mu_L = \max \ \mu{:}Q\left[\mu', \hat{\boldsymbol{\beta}}(\mu')\right] \geqslant \left\{Q\left[\hat{\mu}, \hat{\boldsymbol{\beta}}(\hat{\mu})\right] + \chi_1^2(1-\alpha)\right\}$$

for all $\mu' \leqslant \mu$ and

$$\mu_U = \min \ \mu{:}Q\left[\mu', \hat{\boldsymbol{\beta}}(\mu')\right] \geqslant \left\{Q\left[\hat{\mu}, \hat{\boldsymbol{\beta}}(\hat{\mu})\right] + \chi_1^2(1-\alpha)\right\}$$

for all $\mu' \geqslant \mu$. The endpoints are the locations at which the likelihood surface rises above the critical value of the likelihood-ratio test. The resulting interval includes all μ accepted by the test (see fig. 2).

References

Akaike H (1974) A new look at statistical model identification. IEEE Trans Automatic Control AU-19:716–723

Bengtsson BO, Thomson G (1981) Measuring the strength of associations between HLA antigens and diseases. Tissue Antigens 18:356–363

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

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

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman & Hall, New York

Goddard KAB, Yu C-E, Oshima J, Miki T, Nakura J, Piussan C, Martin GM, et al (1996) Toward localization of the Werner syndrome gene by linkage disequilibrium and ancestral haplotyping: lessons learned from analysis of 35 chromosome 8p11.1-21.1 markers. Am J Hum Genet 58: 1286–1302

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 78:1073–1087

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

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

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

Kerem B, 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

Lazzeroni LC, Lange K. A conditional inference framework for extending the transmission/disequilibrium test. Hum Hered (in press)

Lehesjoki AE, 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

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

Ramsay M, Williamson R, Estivill X, Wainwright BJ, Ho MF, Halford S, Kere J, et al (1993) Haplotype analysis to determine the position of a mutation among closely linked DNA markers. Hum Mol Genet 2:1007–1014

Seber GAF, Wild CJ (1989) Nonlinear regression. John Wiley, New York

Terwilliger JD (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56: 777–787

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