Genetic Epidemiology 17:237–252 (1999)

Fieller's Theorem and Linkage Disequilibrium Mapping

Heather J. Cordell* and Robert C. Elston

Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio

Linkage disequilibrium mapping exploits the fact that at genetic markers close enough to a disease locus on a particular chromosome, we expect to find an association between the disease and marker alleles. Furthermore, the magnitude of the association is expected to follow a unimodal curve when plotted against location, with the peak at the disease location. In practice, for real data, we usually see deviations from such a curve due to other influences such as evolutionary variability, mutation, and selection. Here we propose fitting a quadratic curve to data of this nature, estimating the location of the disease locus by the point at which the curve is maximum. A key feature of our method is the use of transformations of both location and disequilibrium, so that departures from a unimodal curve are incorporated by fitting the curve not to the original location and disequilibrium values but to the transformed values. In addition, we estimate the covariances between the disequilibrium values at linked loci using either a multinomial approximation or a bootstrap procedure. The location estimate from our method is the ratio of two quantities that, in large samples, are normally distributed, and so we use Fieller's theorem to obtain a confidence interval for the disease gene location. We successfully apply our method to data from several published studies in which the true disease gene location is known. Genet. Epidemiol. 17:237–252, 1999.

Key words: allelic association; mutation; confidence interval

*Correspondence to: Dr. Heather J. Cordell, Department of Epidemiology and Biostatistics, Case Western Reserve University, Rammelkamp Center for Education and Research, MetroHealth Medical Center, 2500 MetroHealth Drive, Cleveland, OH 44109. E-mail: cordell@darwin.cwru.edu

Contract grant sponsor: National Institute of General Medical Sciences; Contract grant number: U.S. Public Health Service grant GM28356.

Received for publication 13 October 1998; revision accepted 18 March 1999

INTRODUCTION

Fine-scale mapping of disease genes is impractical using traditional recombination-based linkage methods. This is because the sample sizes required to map a disease gene to within a few kilobases (kb) using traditional methods are prohibitively large [Lange et al., 1985]. For this reason, interest has recently focused on linkage disequilibrium methods as a promising alternative approach. Linkage disequilibrium is the phenomenon of non-random association between alleles at different loci on the same chromosome, caused by the fact that each disease-predisposing mutation occurred in the chromosome of a specific ancestor of the current population. Individuals inheriting the disease mutation are also likely to inherit the ancestral alleles in neighboring regions of that chromosome.

The extent of linkage disequilibrium depends on how many generations have passed since the introduction of the disease mutation. Since the probability of recombination is a monotonic function of genetic distance, one might expect the degree of disequilibrium observed across a chromosome to follow a unimodal curve with a peak at the true location of the disease mutation. In practice, this does not generally happen, owing to a number of complicating factors. Mutations at marker loci and subsequent mutations at the disease locus can obscure the associations between the marker and disease alleles. If the original mutation occurs on an ancestral haplotype of common alleles, there will be little or no association to begin with. Disease mutations may be introduced by more than one ancestor (on chromosomes with different haplotypes) and, for complex diseases, may exist at more than one disease locus. Other factors such as genetic drift, selection, and population admixture can lead to a pattern of linkage disequilibrium that is far from clear, with markers showing the strongest disease associations not necessarily being those closest to the disease locus.

Two main approaches have been used in the development of linkage disequilibrium methods for fine mapping disease loci. The first, which we shall term "model-based," 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], Terwilliger [1995], Xiong and Guo [1997], Rannala and Slatkin [1998] and Graham and Thompson [1998]. In this approach, a specific population genetics model for the introduction and propogation of the disease mutation is assumed, although the results obtained may be robust to some of the model assumptions. We use here an alternative approach that is similar to that described by Lazzeroni [1998], and may be described as a "model-free" or "semi-parametric" method for linkage disequilibrium mapping. This approach considers the mapping of the disease mutation to be a statistical problem suited to a regression approach in which we fit a curve to the pattern of disequilibrium across a chromosome. Both the model-based and model-free approaches may be subdivided according to whether they incorporate information from each locus individually, or whether they use multilocus haplotype data, which may be more informative but is often more difficult to obtain. In the case of the model-free methods, although the pattern of disequilibrium across a chromosome will be affected by the underlying population genetics model, we may replace many detailed population assumptions by an estimate of the variances and covariances of the disequilibrium measures obtained from the data. A novel feature of the method we propose here is the additional use of transformations of both location and disequilibrium measure, so that departures from a unimodal curve are incorporated by fitting the curve not to the

original location and disequilibrium values but to some transformed values, *x*(location) and *y*(disequilibrium).

Several measures have been proposed for quantifying the degree of disequilibrium between alleles at a single marker locus and a disease. Devlin and Risch [1995] discuss some of the most popular disequilibrium measures currently in use. Most rely on counting frequencies of disease associated and non-associated alleles in cases and controls (see Table I). Rather than using genuine individuals as matched or unmatched controls, a popular approach in recent years has been to use the non-transmitted parental alleles as a "family-based" matched control [Spielman et al., 1993; Thomson, 1995]. This has the advantage of removing associations solely due to population stratification in a case/control study.

In our study, we have considered the six measures given in Table II. As we shall see, the measure used can have quite a substantial effect on the final estimated location of the disease mutation. For this reason, it is important to choose the measure that will give the most accurate results. One of the most popular measures in the literature has been the etiologic fraction, often denoted, as here, by δ [Devlin and Risch, 1995; Lazzeroni, 1998] or by p_{excess} [Lehesjoki et al., 1993]. Devlin and Risch [1995] used simulation to show that the measure δ is generally the most directly related to the recombination fraction between marker and disease loci. In their simulations, the measure D' also performed well, whereas Q, Δ, and *dif* were less reliable. They did not consider log *OR*, the logarithm of the odds ratio, which is the only measure of association that is free of the effects of the marginal distributions in Table I [Edwards, 1963]. We have investigated a slightly different strategy for choosing an appropriate measure, namely to use the measure that shows the greatest rank correlation between magnitude and genetic distance when the measure is calculated for pairs of marker loci. For two marker loci *A* and *B*, a disequilibrium measure is calculated by arranging the data as in Table I, but with the case and control categories replaced by disease associated (b) and non-associated ($\bar{b}$) alleles at marker B. This gives us a measure of association between the alleles at the two marker loci, rather than between a single marker and the disease. For all measures except δ, the measure will be symmetrical in A and B. The measure may be calculated for all possible pairs of marker loci, and the correlation or rank correlation between the magnitude of the measure and the inter-marker distance provides an indication of which measure shows the strongest relationship to genetic distance in the region under consideration.

METHODS

Suppose we have haplotype data for cases and controls at a set of *m* linked markers surrounding a putative disease gene region. For the purposes of this analysis, the "control" data may either derive from a genuine sample of control individuals, or may consist of the untransmitted haplotypes obtained from the parents of the affected individuals. Let

TABLE I. Frequencies of Disease Associated (a) and Non-Associated ($\bar{a}$) Alleles at a Marker Locus

Marker allele	Case	Control	
a allele	n_1	n_3	$n_1 + n_3$
$\bar{a}$ allele	n_2	n_4	$n_2 + n_4$
	$n_1 + n_2$	$n_3 + n_4$	N

TABLE II. Disequilibrium Measures Often Used for Fine-Scale Mapping*

Symbol	Formula
δ	$\frac{n_1 n_4 - n_2 n_3}{n_4(n_1 + n_2)}$
Q	$\frac{n_1 n_4 - n_2 n_3}{n_1 n_4 + n_2 n_3}$
Δ	$\frac{n_1 n_4 - n_2 n_3}{[(n_1 + n_2)(n_3 + n_4)(n_1 + n_3)(n_2 + n_4)]^{0.5}}$
D'	$\frac{n_1 n_4 - n_2 n_3}{(n_1 + n_2)(n_2 + n_4)}$
dif	$\frac{n_1 n_4 - n_2 n_3}{(n_1 + n_2)(n_3 + n_4)}$
log *OR*	$\log \frac{n_1 n_4}{n_2 n_3}$

*Formulae are given in terms of the counts in Table I

l_i be the location of the ith marker locus, measured with respect to some fixed origin, and d_i be some measure of association (disequilibrium) between the disease and alleles at marker i. The basis of our method is to find transformations $x(l_i)$ and $y(d_i)$ such that we can fit a quadratic curve to $y(d_i)$ against $x(l_i)$, i.e., we fit

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

or, in matrix notation, $y = X\beta$, where the coefficients β are to be estimated. This can be fitted using weighted least squares so that

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y,$$

where V is the $m \times m$ variance/covariance matrix of the y_i, which can be estimated using a bootstrap approach [Lazzeroni, 1998] or a multinomial distribution approximation, as described in the following subsection.

Note that $\hat{\beta}$ is equivalent to the maximum likelihood estimate if $y \sim \text{MVN}(X\beta, V)$. We can, therefore, obtain $\hat{\beta}$ by maximizing the log likelihood

$$-\frac{1}{2}\log|V|^{0.5} - \frac{1}{2}(y - X\beta)^T V^{-1}(y - X\beta)$$

over β and unknown parameters in $y(d_i)$, $x(l_i)$, $z(h_i)$. The disease location is estimated where the quadratic curve is maximum, i.e., at the point where $x(l) = -\hat{\beta}_1/(2\hat{\beta}_2)$. A confidence interval for this value of x may then be obtained using Fieller's theorem, as described below.

Fieller's Theorem

Given an estimate of location $-\hat{\beta}_1/(2\hat{\beta}_2)$ from the maximum likelihood procedure, we can use Fieller's theorem [Fieller, 1932] to construct a confidence interval

for the true location. Fieller's theorem gives a method of obtaining a confidence interval for the ratio of two means, $\lambda = \mu_a/\mu_b$, given that the sample means $(a,b) \sim$ bivariate normal $(\mu_a, \mu_b, \sigma_a^2, \sigma_b^2, \sigma_{ab})$. Consider the distribution of $a - \lambda b$. This quantity is normally distributed with expectation 0 and variance $\sigma_a^2 + \lambda^2\sigma_b^2 - 2\lambda\sigma_{ab}$. Assume we can estimate $\mathrm{Var}(a - \lambda b)$ by $(v_a + \lambda^2 v_b - 2\lambda v_{ab})s^2$ where v_a, v_b, and v_{ab} are known constants and s^2 is estimated, independent of a and b, with υ d.f. Then we obtain a $1 - \alpha$ confidence interval for λ from

$$P\left\{\frac{(a-\lambda b)^2}{(v_a+\lambda^2 v_b-2\lambda v_{ab})s^2} \le t^2_{\upsilon,\alpha/2}\right\} = 1-\alpha$$

$$\text{i.e., setting } t = t_{\upsilon,\alpha/2},\ P\{(a-\lambda b)^2 - t^2(v_a+\lambda^2 v_b - 2\lambda v_{ab})s^2 \le 0\} = 1-\alpha$$

$$\text{i.e., } P\{(b^2 - t^2 v_b s^2)\lambda^2 - 2(ab + t^2 v_{ab} s^2)\lambda + a^2 - t^2 v_a s^2 \le 0\} = 1-\alpha.$$

Therefore a confidence interval for λ is given by the set

$$\{\lambda:\ (b^2 + t^2 v_b s^2)\lambda^2 - 2(ab + t^2 v_{ab} s^2)\lambda + a^2 - t^2 v_a s^2 \le 0\}.$$

For our data, we let $a = -\hat{\beta}_1$ and $b = 2\hat{\beta}_2$. The variance estimates $v_a s^2$, $v_b s^2$, $v_{ab} s^2$ are obtained from the Hessian matrix. The roots of the parabola in λ give the end points of the interval, provided $(b^2 - t^2 v_b s^2) > 0$. An interesting feature of this set, therefore, is that depending on the coefficient $(b^2 - t^2 v_b s^2)$ and the roots of the parabola, the confidence region may be either an interval, the complement of an interval, or the whole real line. Note that if $(b^2 - t^2 v_b s^2) < 0$, this implies that $b^2/(v_b s^2) < t^2$, i.e., the coefficient b is itself not significantly different from 0.

Estimation of V

The variance/covariance matrix V can be estimated using a bootstrap approach as described in Lazzeroni [1998]. This implementation implicitly assumes independence of the case and control haplotypes, since we sample from each independently at each replicate. If the case and control haplotypes are not independent, such as when haplotypes from the parents of the affected individuals are used as controls, it would make more sense to sample the nuclear families, so that the matched case and "control" are both included, or not included, in any given bootstrap sample. To be used in the maximization of the likelihood, the covariance matrix V must be positive definite. This may be achieved [Lazzeroni, 1998] by multiplying the off-diagonal elements of V by a smoothing factor $1 - \varepsilon$, where ε is a small positive number.

As an alternative to using the bootstrap, the variance/covariance matrix V may be estimated using a multinomial distribution approximation. This has the advantage of being less computationally intensive than the bootstrap. Suppose the data at a single marker locus can be arranged as in Table I, where a is the allele associated with disease and $\bar{a}$ the allele not associated with disease at this marker. Now consider arranging the data in a joint table according to what happens at two marker loci, as in Table III. Here, n_{11}, for example, is the number of disease haplotypes with allele a at locus 1 and allele b at locus

2, where b is the allele associated with disease and $\bar{b}$ the allele not associated with disease at locus 2. Assuming cases and controls are independent, the counts in the eight cells of Table III are distributed as two independent multinomials, one for the case cells and one for the control cells. If the disease is monogenic and the only cause of familial aggregation is due to this one gene (no environmental causes), then we also get two independent multinomials in the situation where untransmitted haplotypes from the parents of the affected individuals are used as controls. Otherwise, there will be dependencies between the multinomials for the case group and the control group, although it may be that these dependencies are negligible.

The calculation of V using the multinomial approximation requires a calculation of the variances and covariances of the disequilibrium measures (d_i,d_j) for $i, j = 1,2...m$. The (i,j)th element of the matrix V, which is the variance/covariance matrix for the y_i, may then be written

$$V_{ij} = \frac{\partial y}{\partial d_i}\frac{\partial y}{\partial d_j}\ \mathrm{Cov}(d_i, d_j)$$

where the variance/covariance matrix with elements $\mathrm{Cov}(d_i,d_j)$ may be calculated as D^TSD, where S is the 8×8 variance/covariance matrix of the 8 cells of Table III, and D is the 8×2 matrix of derivatives corresponding to the transformation from cell counts to a disequilibrium measure.

The estimation of V, either by the bootstrap or the multinomial approximation, assumes that haplotype data are available rather than just allele counts at individual loci. If haplotype data are not available, we may set V equal to the identity matrix, which makes our method equivalent to fitting a quadratic curve by ordinary least squares, except for the transformations. Alternatively, we may use the genotype data from the individual loci to estimate haplotype frequencies using the Expectation-Maximization (EM) algorithm or the method of gene counting [Slatkin and Excoffier, 1995; MacLean and Morton, 1985].

Transformations

A feature of our method is the use of transformations of the location and disequilibrium parameters. Broadly speaking, they fall into two categories: fixed transformations, which are carried out before maximization of the likelihood, and variable transformations, which depend on parameters that are estimated during the maximization procedure.

TABLE III. Frequencies of Disease Associated (a, b) and Non-Associated ($\bar{a}$, $\bar{b}$) Alleles at Two Marker Loci

Alleles at				
Marker 1	Marker 2	Case	Control	
a	b	n_{11}	n_{33}	
a	$\bar{b}$	n_{12}	n_{34}	
$\bar{a}$	b	n_{21}	n_{43}	
$\bar{a}$	$\bar{b}$	n_{22}	n_{44}	
		N_d	N_n	N

Disequilibrium Transformations

We considered four possible fixed transformations for y, namely $y(d) = d$, $y(d) = \log d$, $y(d) = d^{\alpha}$, and $y(d) = -(1 - d)^{\alpha}$. The first two require no parameter estimation. The other two were suggested by Lazzeroni [1998], who proposes estimating the parameter α to minimize the skewness (which may be estimated at each marker locus using the bootstrap procedure).

We also considered a variable transformation of the form

$$y(d) = \begin{cases} (d^p - 1)/p & p > 0 \\ \log d & p = 0, \end{cases}$$

where $p \geq 0$ is to be estimated in the likelihood maximization. This transformation corresponds to that proposed by Box and Cox [1964] and also tends to remove skewness.

Location Transformations

For transforming the location, we used the generalized modulus power transformation (GEMPT) [George and Elston, 1988]. This transformation is of the form

$$x(l) = \frac{\operatorname{sgn}(l - \hat{l})[(|\, l - \hat{l}\,| + 1)^{\kappa} - 1]}{\kappa}$$

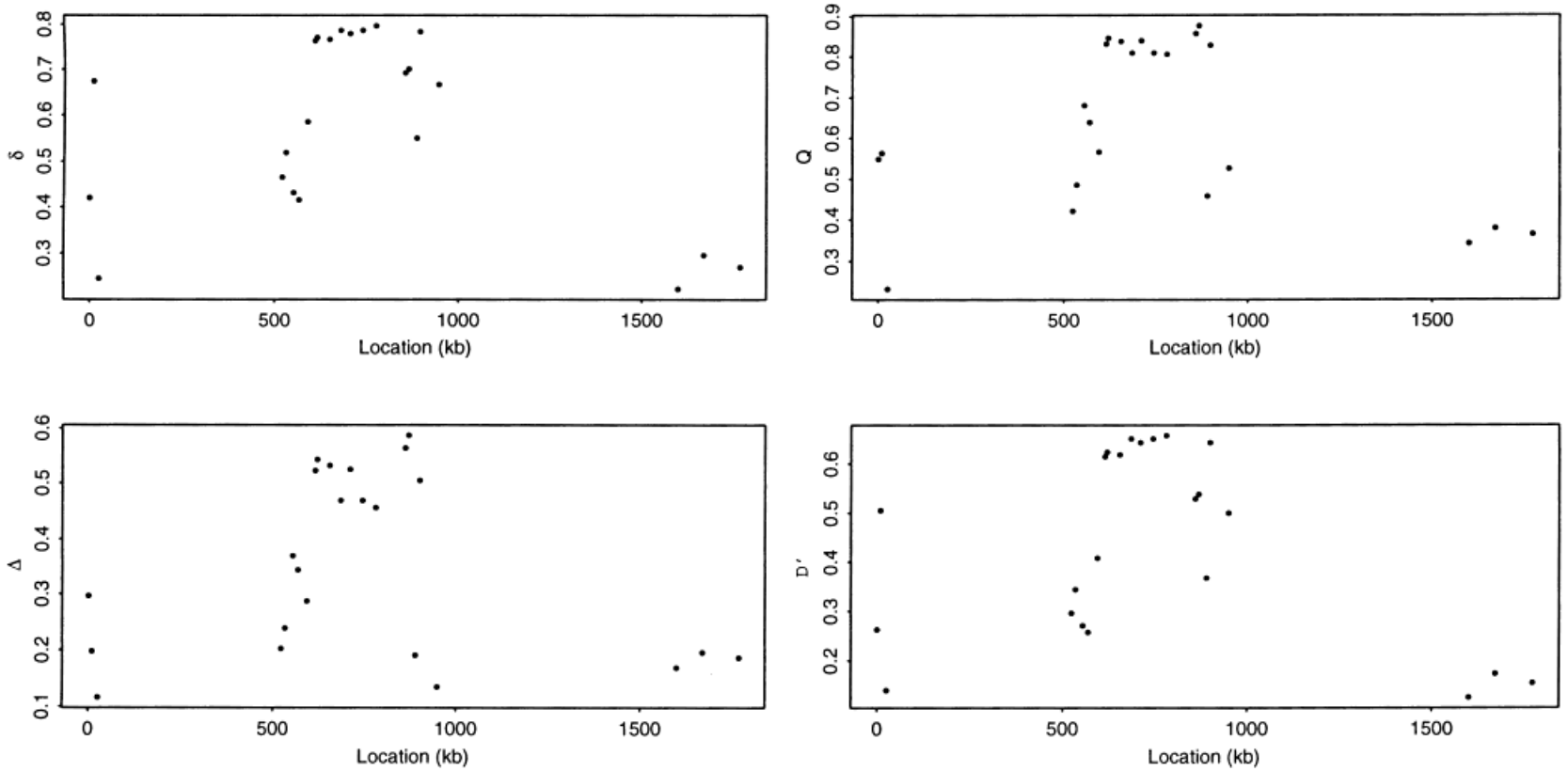
where κ is to be estimated in the maximization. This transformation is an extension of that proposed by John and Draper [1980], which tends to remove kurtosis. It relies on a shift parameter $\dot{l}$ that may be thought of as an initial estimate of location. This may be chosen as the average of the marker locations $\Sigma_i l_i/m$, or as the location l_i where d_i is maximum. More generally, we considered using an iterative procedure where the entire maximization was carried out iteratively with the parameter $\dot{l}$ updated to the current best estimate of location $x^{-1}(\hat{\beta}_1/(2\hat{\beta}_2))$ at the end of each iteration. This generally gave convergence of $\dot{l}$ within 10 iterations.

RESULTS

The methods outlined above were applied to data for four diseases in which the true location of the disease mutation is known: cystic fibrosis, Huntingdon's disease, diastrophic displasia, and progressive myoclonus epilepsy. All maximization was carried out using the FORTRAN subroutine MAXFUN, which is part of the program package SAGE [1994].

Cystic Fibrosis (CF)

Cystic fibrosis (CF) is an autosomal recessive disorder, with the most common disease mutation accounting for more than 70% of Caucasian disease cases. The CF gene was cloned in 1989 [Kerem et al., 1989]. The data of Kerem et al. [1989] include haplotypes of 94 disease chromosomes from a sample of affected individuals, together with 92 untransmitted normal chromosomes from the unaffected parents. Haplotypes are given at 23 diallelic marker loci spanning 1,770 kb around the CF mutation.


δ
0.3 0.4 0.5 0.6 0.7 0.8
0 500 1000 1500
Location (kb)
Q
0.3 0.4 0.5 0.6 0.7 0.8 0.9
0 500 1000 1500
Location (kb)
Δ
0.1 0.2 0.3 0.4 0.5 0.6
0 500 1000 1500
Location (kb)
D'
0.2 0.3 0.4 0.5 0.6
0 500 1000 1500
Location (kb)

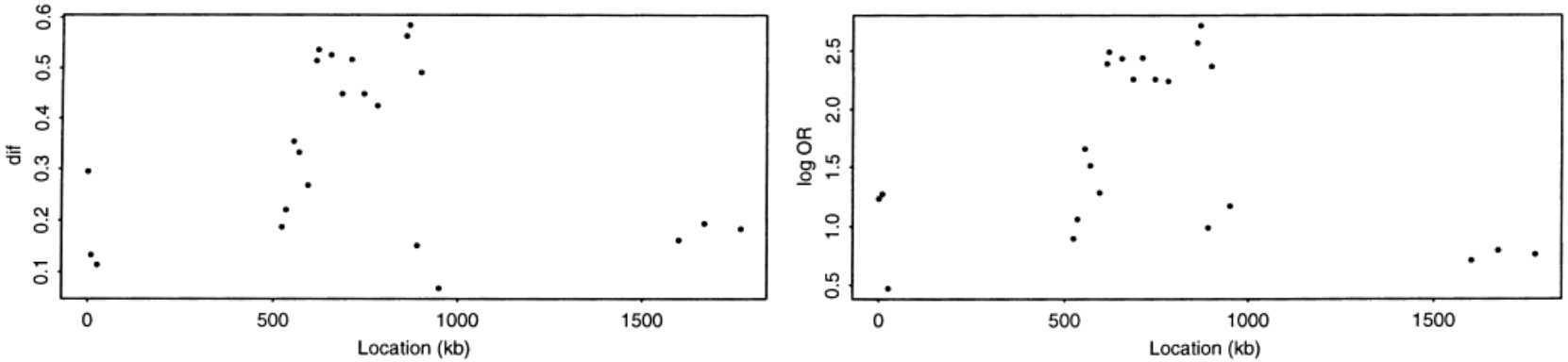


Fig. 1. Plots of disequilibrium against location for the cystic fibrosis data, using the six measures from Table II.

TABLE IV. Correlations Between Disequilibrium Measures and Distances Between the Marker Loci, in the Cystic Fibrosis Data

Measure	Correlation	Rank correlation
δ	–0.63	–0.67
Q	–0.68	–0.71
Δ	–0.61	–0.71
D'	–0.66	–0.70
dif	–0.61	–0.70
log *OR*	–0.55	–0.71

We applied our methods as outlined here to the problem of localizing and gaining a confidence interval for location of the CF gene. For the disequilibrium measure *d*, we considered the six possible measures given in Table II. Graphs of measure against location for the 23 markers in the CF data are given in Figure 1. Although these are broadly similar for the different disequilibrium measures, there are some differences in the patterns, particularly at the far left (0 kb) region. Each measure was then calculated for each pair of marker loci, in order to determine which disequilibrium measures show the clos-

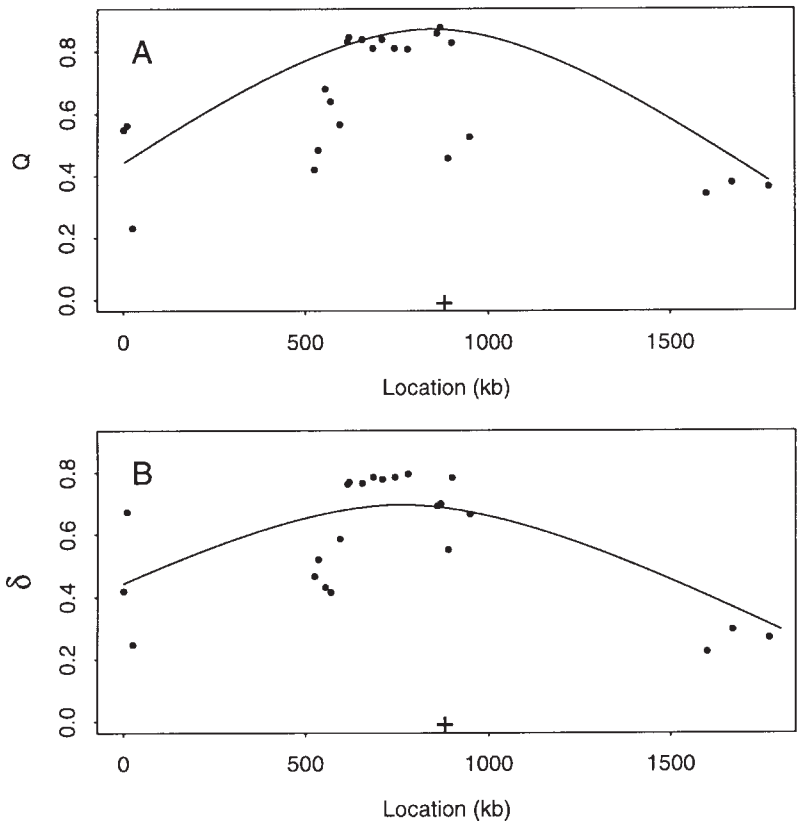


Fig. 2. Disequilibrium values and fitted curves for the cystic fibrosis data, using the measures *Q* and δ. The true location of the disease mutation is given by +.

est relationship to inter-locus distance in these data. Table IV gives the Pearson correlation and rank correlation coefficients between distance and each of these measures, for all pairs of marker loci. We see that the rank correlations are similar for all measures, but the (non-rank) correlations show somewhat more variability, with Yule's *Q* showing the strongest, and log *OR* the weakest, relationship between disequilibrium measure and distance. The etiologic fraction, δ, does not seem to perform quite as well as the other measures in terms of showing correlation between magnitude and distance. It would seem that Yule's *Q* is, therefore, the most appropriate measure to use to analyse these data. However, since many other investigators have used δ, and it is of interest to compare the performance of the different measures, we present results from using all six disequilibrium measures.

Figure 2A shows the results of fitting our model for $d = Q$. We took the fixed identity transformation $y = d$ and the iterative GEMPT transformation for $x(l)$. Problems were encountered when using a variable transformation for $y(d)$, suggesting that there is a limited amount of information in the data and it may be unwise to attempt to estimate too many parameters. The variance/covariance matrix V was estimated using the bootstrap and the off-diagonal elements of V were multiplied by $1 - \varepsilon$ where $1 - \varepsilon$ was taken, in common with Lazzeroni [1998], to equal 0.98. We obtain a location estimate of 848 kb with a confidence interval of (770, 942). This estimate is very satisfactory, lying only about 35 kb from the true location.

Figure 2B shows the results of fitting our model for $d = \delta$, the etiologic fraction. Here we again took the fixed identity transformation $y = d$, and used the iterative GEMPT transformation for $x(l)$. The final estimate of location was at 762 kb with a 95% confidence interval for location of (654, 856). This is somewhat disappointing as the true location of the CF mutation (shown by a "+") is at 880–885 kb. However, by examining the data we should perhaps not be too surprised, as the data themselves do not look particularly quadratic, particularly in the 0 kb region. Several investigators have used fixed transformations of δ, e.g., $y = -\log \delta$ [Devlin et al., 1996] or $y = (1 - \delta)^{0.6}$ [Lazzeroni, 1998] in their analyses. We, therefore, incorporated these transformations into our analyses, but they did not appear to make a large difference to the location estimate or to the confidence interval.

Table V shows the results for all 6 disequilibrium measures. We see that measures *Q*, Δ, and log *OR* perform well, while δ, D', and *dif* are less accurate. There seems to be a correspondence between the performance of a measure and its rank correlation between magnitude and distance in the marker data (Table IV), suggesting that the measure with the most negative rank correlation may indeed be the most appropriate one to use.

TABLE V. Final Location and Confidence Interval Results for Different Disequilibrium Measures, for the Cystic Fibrosis Data

Measure	Estimated location	95% Confidence interval
δ	762	(654, 856)
Q	848	(770, 942)
Δ	850	(683, 1,126)
D'	750	(573, 867)
dif	775	(542, 1,766)
log *OR*	885	(703, 1,096)

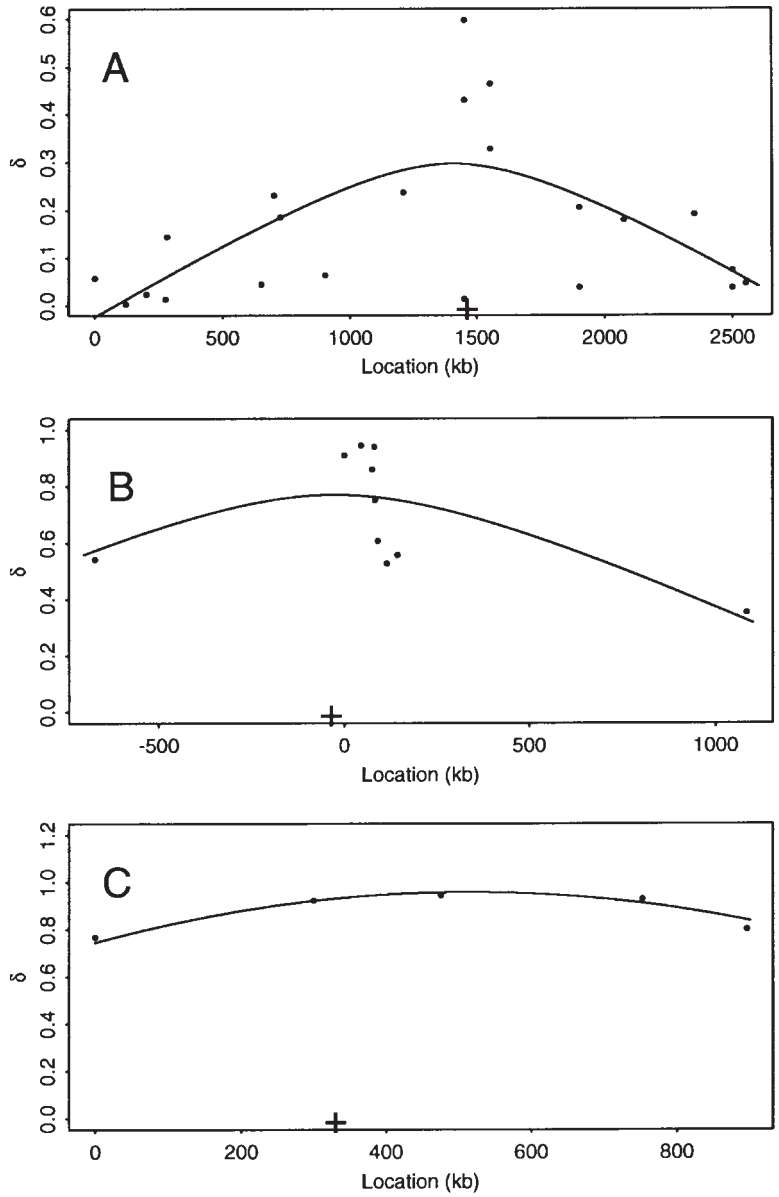


Fig. 3. Disequilibrium measure δ and fitted curves, plotted against location, for Huntington's disease (**A**), diastrophic displasia (**B**), and progressive myoclonus epilepsy (**C**). The true locations of the disease mutations are given by +.

So far, all results quoted were obtained using a bootstrap estimate of *V*. We found similar results using the multinomial distribution approximation, except when ε was very small (in which case *V* was not always positive definite). This suggests that the bootstrap estimate may be more useful and more robust than the multinomial estimate.

It is of interest to compare our results with those analyses of the same data by

Terwilliger [1995], Devlin et al. [1996], Xiong and Guo [1997], and Lazzeroni [1998]. The location of the CF gene was estimated by Terwilliger [1995] to be at 770 kb, about 110 kb away from the true location. Devlin et al. [1996] estimate the location at about 60 kb from the true location, and Xiong and Guo [1997] at about 75 kb from the true location. Therefore, our location estimates using the measure Q perform quite favourably in comparison. The location estimated by Lazzeroni [1998] was only 10 kb from the true location, but the data analyzed by Lazzeroni [1998] was later found to contain an error (Lazzeroni, personal communication). Using the method of Lazzeroni [1998], we estimated the location to be at 900 kb, that is, 15–20 kb from the true location, with a 95% confidence interval of (749, 1056). Our method does not, therefore, appear to locate the disease mutation quite as accurately as that of Lazzeroni [1998], but our confidence interval is somewhat narrower. Since this analysis is of a single data set, it is not possible to assess the true coverage properties of the confidence interval obtained; comparison of further data sets and simulation will be required to compare the coverage properties of intervals obtained using the two methods.

Huntington's Disease (HD)

The HD gene, identified by the Huntington Disease Collaborative Research Group [1993], is known to span a 210-kb region approximately 110 kb from *D4S95*, 240 kb from *D4S180*, and 250 kb from *D4S182* [Xiong and Guo, 1997]. The data published by MacDonald et al. [1991] gives counts of alleles on HD and normal chromosomes at 27 markers in this region. Since five of these markers were separated from the others by a gap of indeterminate length, we used our methods to analyze the data for the remaining 22 markers (from *D4S168* to *D4S10*). Counts were given at individual markers rather than haplotypes across the chromosome, so it was not possible to calculate disequilibrium measures between markers or to estimate the variance/covariance matrix V from these data. We therefore set V to equal the identity matrix. The data and the fitted curve for δ are shown in Figure 3A. The true disease mutation lies approximately 1,460–1,660 kb from the first marker (*D4S168*) (although note we had some difficulty in resolving the inter-marker distances given by Xiong and Guo [1997] with those in MacDonald et al. [1991]). Our estimates of the disease location for the six measures δ, Q, Δ, D', *dif*, and log *OR* are 1,406, 1,383, 1,396, 1,410, 1,432, and 1,378 kb. No confidence intervals were obtained for these data since the parabola constructed using Fieller's theorem for 95% confidence had no real roots. We see that the measures δ, D', and *dif* perform best in these data, giving location estimates 30–50 kb from the true location, which compares favorably with an average error of 89 kb from Xiong and Guo [1997].

Diastrophic dysplasia (DTD)

We also reanalyzed the data of Hästbacka et al. [1994], which consist of counts of alleles on DTD and normal chromosomes at 11 markers on chromosome 5q. These data have been analyzed by Devlin et al. [1996], who successfully located the DTD gene in approximately the correct region, 70 kb proximal to *CSF1R*. Since counts were given at individual markers rather than haplotypes across the chromosome, it was again not possible to calculate disequilibrium measures between markers or to estimate the variance/covariance matrix V, and so we set V equal to the identity matrix. In common with Devlin et al. [1996], we excluded *D5S413* because the distance between it and other loci could not be estimated. The exact locations of the two outermost markers *D5S372* and *RPS14* were also unknown; we used estimates based on the recombination estimates of Hästbacka et al. [1994] and Devlin et al. [1996].

The data and the fitted curve for δ are shown in Figure 3B. The true disease mutation lies approximately 35 kb proximal to *BT1*, or at the location –35 kb on this scale. Our estimates of the disease location for the six measures δ, *Q*, Δ, *D′*, *dif*, and log *OR* are –25, –221, –122, 16, –124, and –93 kb. The only 95% confidence interval obtainable was for log *OR*, which gave a rather wide interval of (–1,289, 156). We see quite a bit of variability in the location estimates, with δ performing best giving an estimate only 10 kb from the true location. The measure *D′* also performs quite well, giving an error of 51 kb. The errors for the other measures are somewhat larger, although quite comparable with the results of Devlin et al. [1996] whose location estimate lay approximately 100 kb from the true location. For these data we have a problem in that most of the data are clustered together in the center of the location axis, with very little information in the tails. This is not ideal for a curve-fitting type method such as proposed here, and is likely to be the cause of the rather erratic location estimates we obtain. It is interesting to note that, in common with our results for Huntington's disease, the measures that perform best are δ and *D′*, which are those proposed by Devlin and Risch [1995] to be most related to genetic distance.

Progressive Myoclonus Epilepsy (EPM1)

We also applied our methods to the data of Virtaneva et al. [1996], which consist of haplotypes on disease and normal chromosomes for 5 loci surrounding the EPM1 locus, which lies 30 kb away from marker *D21S2040* [Pennacchio et al., 1996], or 330 kb from *D21S885* [Xiong and Guo, 1997]. The iterative GEMPT did not converge for measures δ and *D′*, so results for these measures were calculated without this transformation. The rank correlations of disequilibrium with distance for the pairs of marker loci for the six measures δ, *Q*, Δ, *D′*, *dif*, and log *OR* were –0.68, –0.65, –0.77, –0.52, –0.70, –0.65, suggesting that Δ is the most appropriate measure for these data. However, the estimated locations using these six measures were at 513, 641, 787, 489, 899, and 633 kb from *D21S885*, as opposed to the true value of 330 kb. Confidence intervals for the six measures were (290, 614), (–∞, ∞), (667, ∞), (310, 588), (723, ∞), and (536, 1,372). As for HD and DTD, the measures that perform best are δ and *D′*, but even these have quite substantial errors. The errors are slightly less than the 220 kb error in Terwilliger [1995] but much greater than the 40 kb error of Xiong and Guo [1997], both of whom also analyzed these data. Figure 3C shows the data and fitted curve for δ. We see that with only five data points, we do not really have enough information for a curve-fitting method; the curve fits well but basically just repeats the data, giving a location estimate close to that location with the highest disequilibrium. Since this is quite far from the true location, particularly for measures Δ and *dif*, we find correspondingly poor estimates using our method. This would be true even if a piecewise polynomial were fitted as by Lazzeroni [1998]. It appears that, given data on just a few loci, model-based methods such as that of Xiong and Guo [1997] may be more robust for accurate localization of disease mutations.

DISCUSSION

High-resolution localization of disease loci is an important step before embarking on positional cloning, which may be infeasible in regions extending more than a few hundred kb. With this aim, linkage disequilibrium methods provide a promising tool for fine mapping disease genes.

In our analyses, the regression approach proposed here performed well when applied to data at a reasonable number of marker loci, but poorly when the number of marker loci was small. This is not unexpected for a method that, although population and disease model free, requires estimation of several regression and transformation parameters. In particular, in order for the method to work well, one would anticipate that a

sufficiently large number of loci scattered at roughly equal intervals over the location space would be required. Also, for fitting a quadratic curve, it would be reasonable to demand that the marker map extends sufficiently far from the true disease location to where the magnitude of the observed disequilibrium approaches zero.

Comparison of our method on real data with various previously proposed methods [Terwilliger, 1995; Devlin et al., 1996; Xiong and Guo, 1997; Lazzeroni, 1998] leads to the somewhat unsatisfactory conclusion that no one method or class of methods performs best in all situations. It does appear that the measures δ and D', proposed by Devlin and Risch [1995] as being most generally useful for linkage disequilibrium mapping, are indeed the most useful in many situations, although this is not invariably true, e.g., for the cystic fibrosis data the measures Q, Δ and log OR performed better. We speculate that the best measure of disequilibrium to use is most often the one that is most highly correlated with distance in the region being studied, but this was not the case for EPM1. It would be interesting to study more disease loci for which haplotype data are available, over a large enough region, to determine whether this would indeed be more often than not the best strategy.

The number of data sets currently available where the true location of the disease mutation is known is quite small, and so clearly further theoretical work and simulation studies will be required to investigate the properties and power of the various different approaches. In this paper, we have concentrated on fitting a quadratic curve, which provides a convenient method for obtaining a confidence interval for the location via Fieller's theorem. Other curve-fitting approaches such as the piecewise polynomial curve proposed by Lazzeroni [1998], or fitting a unimodal curve using non-parametric methods, may prove to be equally useful and warrant further investigation.

Proven successes with linkage disequilibrium methods have so far been confined to monogenic disorders. While there is some evidence that linkage disequilibrium may exist for complex diseases such as type 1 diabetes [Copeman et al., 1995; Merriman et al., 1997], this is likely to be population dependent and may be much diluted by the presence of allelic as well as locus heterogeneity. The introduction of disease mutations by more than one ancestor, at different loci and on different genetic backgrounds, makes the existence of associations between disease and marker loci far from guaranteed. Nevertheless when linkage disequilibrium does exist, it will provide a valuable tool for the localization and identification of genes involved in both complex and monogenic disorders.

ACKNOWLEDGMENTS

This work was supported in part by U.S. Public Health Service grant GM28356 from the National Institute of General Medical Sciences. We thank Anna-Elina Lehesjoki for sharing the EPM1 data. Some of the results of this paper were obtained by using SAGE, which is supported by a U.S. Public Health Service Resource Grant 1 P41 RR03655 from the National Center for Research Resources.

REFERENCES

Box GEP, Cox DR. 1964. An analysis of transformations (with Discussion). J R Stat Soc B 26:211–252.

Copeman JB, Cucca F, Hearne CM, Cornall RJ, Reed PW, Rønningen KS, Undlien DE, Nisticò L, Buzzetti R, Tosi R, Pociot F, Nerup J, Cornélis F, Barnett AH, Bain SC, Todd JA. 1995. Linkage disequilibrium mapping of a type 1 diabetes susceptibility gene (IDDM7) to chromosome 2q31-q33. Nature Genet 9:80–85.

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.

Edwards AWF. 1963. The measure of association in a 2×2 Table. J R Stat Soc A 126:109–114.
Fieller EC. 1932. The distribution of the index in a normal bivariate population. Biometrika 24:428–440.
George V, Elston RC. 1988. Generalized modulus power transforms. Commun Statist A 17:2933–2952.
Graham J, Thompson EA. 1998. Disequilibrium likelihoods for fine-scale mapping of a rare allele. Am J Hum Genet 63:1517–1530.
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. Nature Genet 2:204–211.
Hästbacka J, de la Chapelle A, Mahtani MM, Clines G, Reeve-Daly MP, Daly M, Hamilton BA, Kusumi K, Trivedi B, Weaver A, Coloma A, Lovett M, Buckler A, Kaitila I, Lander E. 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.
Huntingdon Disease Collaborative Research Group. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntingdon's disease chromosomes. Cell 72:971–983.
John JA, Draper NR. 1980. An alternative family of transformations. Appl Stat 29:190–197.
Kaplan NL, Weir BS. 1995. Are moment bounds on the recombination fraction between a marker and 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.
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 T, Chakravarti A, Buchwald M, Tsui L-C. 1989. Identification of the Cystic Fybrosis gene: Genetic analysis. Science 245:1073–1080.
Lange K, Kunkel L, Aldridge J, Latt SA. 1985. Accurate and superaccurate gene mapping. Am J Hum Genet 37:853–867.
Lazzeroni LC. 1998. Linkage disequilibrium and gene mapping: An empirical least squares approach. Am J Hum Genet 62:159–170.
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.
MacLean CJ, Morton NE. 1985. Estimation of Myriad Haplotype Frequencies. Genet Epidemiol 2:263–272.
MacDonald ME, Lin C, Srinidhi L, Bates G, Altherr M, Whaley WL, Lehrach H, Washmuth J, Gusella JF. 1991. Complex patterns of linkage disequilibrium in the Huntington disease region. Am J Hum Genet 49:723–734.
Merriman T, Twells R, Merriman M, Eaves I, Cox R, Cucca F, McKinney P, Shield J, Baum D, Bosi E, Pozzilli P, Nistico L, Buzzetti R, Joner G, Rønningen K, Thorsby E, Undlien D, Pociot F, Nerup J, Bain S, Barnett A, Todd J. 1997. Evidence by allelic association-dependent methods for a type 1 diabetes polygene (*IDDM6*) on chromosome 18q21. Hum Mol Genet 6:1003–1010.
Pennacchio LA, Lehesjoki A, Stone NE, Willour VL, Virtaneva K, Miao J, D'Amato E, Ramirez L, Faham M, Koskiniemi M, Warrington JA, Norio R, de la Chapelle A, Cox DR, Myers RM. 1996. Mutation as in the gene encoding cystatin B in progressive myocolonus epilepsy (EPM1). Science 271:1731–1734.
Rannala B, Slatkin M. 1998. Likelihood analysis of disequilibrium mapping, and related problems. Am J Hum Genet 62:459–473.
Risch N, de Leon D, Ozelius LJ, Kramer P, Almasy L, Singer B, Fahn S, Breakefield X, Bressman S. 1995. Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nature Genet 9:152–159.
SAGE. 1994. Statistical analysis for genetic epidemiology, Release 2.2. Computer program package available from the Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland.
Slatkin M, Excoffier L. 1995. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. Heredity 76:377–383.
Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission tests for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516.
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.
Thomson. 1995. Mapping disease genes: family-based association studies. Am J Hum Genet 57:487–498.
Virtaneva K, Miao J, Träskelin A-L, Stone N, Warrington JA, Weissenbach J, Myers RM, Cox DR, Sistonen P, de la Chapelle A, Lehesjoki A-E. 1996. Progressive myoclunus epilepsy EPM1 locus maps to a 175-kb interval in distal 21q. Am J Hum Genet 58:1247–1253.
Xiong M, Guo S-W. 1997. Fine-scale genetic mapping based on linkage disequilibrium: Theory and applications. Am J Hum Genet 57:487–498.