Am. J. Hum. Genet. 56:788–798, 1995

Multipoint Linkage Analysis Using Sib Pairs: An Interval Mapping Approach for Dichotomous Outcomes

Jane M. Olson

Department of Biomathematics, Roswell Park Cancer Institute, Buffalo

Summary

I propose an interval mapping approach suitable for a dichotomous outcome, with emphasis on samples of affected sib pairs. The method computes a lod score for each of a set of locations in the interval between two flanking markers and takes as its estimate of trait-locus location the maximum lod score in the interval, provided it exceeds the prespecified critical value. Use of the method depends on prior knowledge of the genetic model for the disease only through available estimates of recurrence risk to relatives of affected individuals. The method gives an unbiased estimate of location, provided the recurrence risks are correctly specified and provided the marker identity-by-descent probabilities are jointly, rather than individually, estimated. I also discuss use of the method for traits determined by two loci and give an approximation that has good power for a wide range of two-locus models.

Introduction

Tests of genetic linkage that utilize affected sib pairs provide an alternative to pedigree analysis when the genetic model underlying the disease is not known. Early methods that rely on sib-pair identity by descent (IBD) at the marker locus (Day and Simons 1976; de Vries et al. 1976; Green and Woodrow 1977; Suarez et al. 1978; Suarez and Hodge 1979) recently have been generalized to the substitution of marker identity by state for IBD relations (Lange 1986; Weeks and Lange 1988) and inclusion of all affected relatives in a pedigree (Weeks and Lange 1988) and multiple marker loci (Weeks and Lange 1992).

In this paper, I propose an interval mapping approach to linkage analysis of sib-pair data when the trait variable is dichotomous. Interval mapping has been previously proposed for mapping loci underlying quantitative traits in laboratory animals that have been subjected to breeding experiments (Lander and Botstein 1989; Haley and Knott 1992) and in humans (Fulker and Cardon 1994). Such methods increase the power of detecting linkage by utilizing information from two flanking markers rather than a single marker. For dichotomous traits, multipoint linkage methods are also more powerful than two-point methods (e.g., Lathrop et al. 1985).

Although the proposed method can include affected, unaffected, and discordant pairs, we emphasize its use for samples of affected sib pairs. A lod score is computed for each of a set of locations in the interval flanked by two markers. The maximum lod score in the interval, provided it exceeds the appropriate threshold, is taken as the estimate of the trait-locus location. Our approach takes advantage of multipoint IBD results obtained by Olson (in press). Although our approach is parametric, it depends on prior knowledge of the trait genetic model only through available estimates of the prevalence of the trait in the population and risks to relatives of affected probands.

The method may be extended to incorporate two trait loci. Some additional assumptions about the trait model are required, but we show that the method appears robust to violation of these assumptions and is preferable to the method that assumes a single trait locus when two loci are, in fact, present.

Methods and Results

Genetic Model

Trait Locus.—Consider a disease with a single underlying locus. For a disease locus with frequency p of the disease allele T, the prevalence K_p of the disease in the population is

$$K_p = p^2\delta_1 + 2p(1 - p)\delta_2 + (1 - p)^2\delta_3 ,$$

where δ_1, δ_2, and δ_3 are the probabilities of expressing the disease in individuals with trait genotypes *TT*, *Tt*, and *tt*, respectively. The additive (V_A) and dominance (V_D) genetic variances for the trait locus are given by

$$V_A = 2p(1 - p)[p(\delta_2 - \delta_1) + (1 - p)(\delta_3 - \delta_2)]^2$$

and

$$V_D = p^2(1 - p)^2(\delta_1 - 2\delta_2 + \delta_3)^2 ,$$

Received June 29, 1994; accepted for publication December 8, 1994.

Address for correspondence and reprints: Dr. Jane M. Olson, Department of Biomathematics, Roswell Park Cancer Institute, Elm and Carlton Streets, Buffalo, NY 14263-0001.

0002-9297/95/5603-0030$02.00

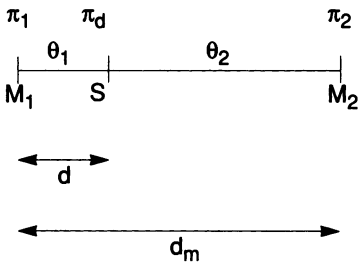

Figure 1 Configuration of loci in interval mapping

The population parameters K_p, V_A, and V_D completely specify the distribution of sib-pair IBD scores underlying the disease locus, conditional on the number of affected siblings (Suarez et al. 1978).

Marker Loci.—Let M_1 and M_2 be linked marker loci with known recombination fraction θ_m, and let the data available on these markers for the sample of sib pairs be denoted I_m. Let π_1 and π_2 denote the proportion of genes shared IBD by the members of a sib pair at M_1 and M_2, respectively. Now consider a locus S that lies between M_1 and M_2, so that the recombination fraction between M_1 and S is θ_1 and that between S and M_2 is θ_2. Let d_m be the map distance between M_1 and M_2, and let d denote the map distance between M_1 and S, so that d lies between 0 and d_m, inclusive. Figure 1 illustrates this configuration. A mapping function that describes the relationship between map distance and recombination fraction may be chosen at the user's discretion; in this paper, I use Haldane's (1919) mapping function in all computations.

Now, consider the quantity $\hat{f}_{ij} = P(\pi_1 = i/2, \pi_2 = j/2 \mid I_m)$, the estimated probability that a pair of siblings share i genes IBD at M_1 and j genes IBD at M_2, given the available marker information I_m and the recombination fraction θ_m. Also let $\hat{f}_{i\cdot} = \Sigma_j \hat{f}_{ij}$ and $\hat{f}_{\cdot j} = \Sigma_i \hat{f}_{ij}$, the estimated marginal IBD probabilities. (Computation of $\hat{f}_{ij}$ is discussed in a later section.) Olson (in press) showed that, under the assumptions of no crossover interference and equal male and female recombination fractions,

$$E(\pi_d \mid I_m) = \rho_0 + \rho_1 \hat{\pi}_1 + \rho_2 \hat{\pi}_2 , \quad (1)$$

where $\hat{\pi}_1 = \hat{f}_{1\cdot}/2 + \hat{f}_{2\cdot}$ and $\hat{\pi}_2 = \hat{f}_{\cdot 1}/2 + \hat{f}_{\cdot 2}$ are the estimated proportions of genes shared IBD by the sibs at the marker loci, ρ_0, ρ_1, and ρ_2 are parameters that depend on the recombination fractions (table 1), and π_d is the proportion of genes shared IBD for the locus at location d. Note that $\hat{\pi}_1$ and $\hat{\pi}_2$ are estimated jointly by using all available marker information. Fulker and Cardon (1994) employ a similar expression (substituting individual estimates of π_1 and π_2 for $\hat{\pi}_1$ and $\hat{\pi}_2$) in an interval mapping extension of the Haseman and Elston (1972) method.

Now let $\widehat{\pi_1\pi_2}$ equal $\frac{1}{4}\hat{f}_{11} + \frac{1}{2}\hat{f}_{12} + \frac{1}{2}\hat{f}_{21} + \hat{f}_{22}$. Olson (in press) further showed that

$$P(\pi_d = \tfrac{1}{2} \mid I_m) = \omega_0 + \omega_1(\hat{\pi}_1 + \hat{\pi}_2 - 2\widehat{\pi_1\pi_2}) + \omega_2 \hat{f}_{1\cdot} + \omega_3 \hat{f}_{\cdot 1} + \omega_4 \hat{f}_{11} . \quad (2)$$

Expressions for the regression parameters in equations (1) and (2) are reproduced in table 1 from Olson (in press). Expressions (1) and (2) contain the marker information necessary to consider linkage to a trait locus.

Likelihood of the Marker Data Conditional on Number of Affected Sibs in a Sib Pair

The distribution of the IBD state of a single marker conditional on the number of affected sibs in a sib pair was given by Suarez et al. (1978), and we extend that approach here. Using Bayes's rule,

$$P(I_m \mid A) = P(A \mid I_m)P(I_m)/P(A) ,$$

where A is the number of affected sibs in a sib pair. For a pair of markers flanking a trait locus at location d, assuming that the trait and marker loci have no pleiotropic effects on one another,

$$P(I_m \mid A) = \frac{P(I_m)}{P(A)} \sum_{\pi_d} P(A \mid \pi_d)P(\pi_d \mid I_m) . \quad (3)$$

As π_d takes only three possible values, $P(A \mid \pi_d)$ can be written as a linear function of π_d and $I(\pi_d = \frac{1}{2})$, the indicator that the sibs share exactly one gene IBD:

$$P(A \mid \pi_d) = \alpha + \beta\pi_d + \gamma I(\pi_d = \tfrac{1}{2}) ,$$

where α, β, and γ depend on K_p, V_A, and V_D and on the value of A. Table 2 gives expressions for α, β, and γ and for $P(A)$ when 0, 1, or 2 sibs are affected. Substituting this linear function into equation (3) and computing the summation, we obtain

$$P(I_m \mid A) = \frac{P(I_m)}{P(A)} [\alpha + \beta E(\pi_d \mid I_m) + \gamma P(\pi_d = \tfrac{1}{2} \mid I_m)] . \quad (4)$$

A similar expression for the likelihood in the case of known marker IBD states was given by Todorov (1992).

Expression (4) can be computed by using equations (1) and (2) for a set of known locations in the interval defined by the two markers. The likelihood ratio $P_d(I_m \mid A)/P_0(I_m \mid A)$, where $P_0(I_m \mid A) = P(I_m)$ is the probability of the marker data if the trait locus is unlinked to either marker locus, may also be computed. For a set of N independent sib pairs, the lod score assuming that d is the true location of the trait locus is therefore

Table 1

Regression Parameters in the Expressions for $E(\pi_2|I_m)$ and $P(\pi_2 = 1/2|I_m)$ in the Flanking Marker Case[a]

Parameter	Expression
ρ_0	$(1-\psi_1)(1-\psi_2)/\psi_m$
ρ_1	$-\psi_2(1-\psi_2)(1-2\psi_1)/\psi_m(1-\psi_m)$
ρ_2	$-\psi_1(1-\psi_1)(1-2\psi_2)/\psi_m(1-\psi_m)$
ω_0	$2\psi_1(1-\psi_1)\psi_2(1-\psi_2)/\psi_m^2$
ω_1	$2(1-2\psi_1)(1-2\psi_2)\psi_1(1-\psi_1)\psi_2(1-\psi_2)/\psi_m^2(1-\psi_m)^2$
ω_2	$(1-2\psi_1)^2\psi_2^2(1-\psi_2)^2/\psi_m^2(1-\psi_m)^2$
ω_3	$(1-2\psi_2)^2\psi_1^2(1-\psi_1)^2/\psi_m^2(1-\psi_m)^2$
ω_4	$(1-2\psi_1)^2(1-2\psi_2)^2\psi_1(1-\psi_1)\psi_2(1-\psi_2)/\psi_m^2(1-\psi_m)^2(1-2\psi_m+2\psi_m^2)$

SOURCE.—Olson (in press).
[a] $\psi_i = \theta_i^2 + (1-\theta_i)^2$; $i = 1, 2, m$.

$$Z_d = \sum_{i=1}^{N} \log_{10}[\alpha + \beta E(\pi_d|I_{mi}) + \gamma P(\pi_d = 1/2|I_{mi})] - \sum_{i=1}^{N} \log_{10} P(A_i) .$$

The last term on the right-hand side does not depend on the location of the trait locus and may be ignored when comparing log-likelihoods across the interval.

If K_p, V_A, and V_D are known, Z_d may be computed easily. If these quantities cannot be specified a priori with confidence, as is often the case in practice, it is necessary to estimate them. However, when the data set consists solely of affected sib pairs, all three parameters cannot be estimated using only the likelihood (4). One solution is to estimate α, β, and γ by using previously obtained recurrence-risk information, and I discuss this approach in a later section. First, however, we consider the problem of computing the $\hat{f}_{ij}$ and the expected lod score.

Estimation of Marker IBD Probabilities for Sib Pairs with No Parental Information

For the expected lod score computations presented in subsequent sections, we consider only the case of independent sib pairs for which no parental marker information is available.

Let π be the two-marker IBD state of the sib pair. Using Bayes's theorem,

$$\hat{f}_{ij} = P(\pi|I_m) = \frac{P(I_m|\pi)P(\pi)}{\sum_{\pi} P(I_m|\pi)P(\pi)} . \quad (5)$$

Expressions for $P(\pi)$ for each π are given in Olson (in press). Conditional on the marker joint IBD state, the markers are independent; that is,

$$P(I_m|\pi) = P(I_{m1}|\pi_1)P(I_{m2}|\pi_2) ,$$

where I_{m1} and I_{m2} represent the data for M_1 and M_2, respectively. To obtain this result, we rely on the facts that parents and offspring always share exactly one gene IBD and that noninbred parents share exactly zero genes IBD so that we need not include a summation over possible IBD states for parental and parent-offspring pairs. These facts also imply that the algorithm is valid when parental marker data are available. These individual marker terms may be obtained using quantities given in table 3 of Haseman and Elston (1972) by noting that, for example,

$$P(I_{m1}|\pi_1) = \frac{P(\pi_1|I_{m1})P(I_{m1})}{P(\pi_1)} .$$

Table 2

Coefficients of Regression of $P(A/\pi_d)$ on π_d and $I(\pi_d = 1/2)$

No. of Affected Siblings (A)	Regression Coefficients			
	α	β	γ	$P(A)$
2	K_p^2	$V_A + V_D$	$-V_D/2$	$K_p^2 + V_A/2 + V_D/4$
1	$2K_p(1-K_p)$	$-2(V_A + V_D)$	V_D	$2K_p(1-K_p) - V_A - V_D/2$
0	$(1-K_p)^2$	$V_A + V_D$	$-V_D/2$	$(1-K_p)^2 + V_A/2 + V_D/4$

Table 3

Maximum $E(Z_d)$ and Estimated Distance from M_1 ($\hat{d}$) for 25 Affected Sib Pairs, π Estimated Jointly and Individually (Additive Trait, K_p = .001)

Distance between Markers (cM)	No. of Marker Alleles		Location of Trait Locus							
			Center				One-fifth from M_1			
			Joint $\hat{\pi}$		Individual $\hat{\pi}$		Joint $\hat{\pi}$		Individual $\hat{\pi}$	
	M_1	M_2	max $E(Z_d)$	$\hat{d}$	max $E(Z_d)$	$\hat{d}$	max $E(Z_d)$	$\hat{d}$	max $E(Z_d)$	$\hat{d}$
5	2	2	1.09	2.5[a]	.83	2.5[a]	1.11	1.0[a]	.84	2.0
	6	6	2.27	2.5[a]	2.02	2.5[a]	2.35	1.0[a]	2.07	1.8
	2	6	1.85	2.5[a]	1.69	4.2	1.73	1.0[a]	1.52	3.8
	6	2					2.06	1.0[a]	1.92	.4
10	2	2	.86	5.0[a]	.72	5.0[a]	.91	2.0[a]	.75	3.2
	6	6	1.87	5.0[a]	1.75	5.0[a]	2.01	2.0[a]	1.88	2.8
	2	6	1.48	5.0[a]	1.38	7.4	1.34	2.0[a]	1.19	6.0
	6	2					1.82	2.0[a]	1.74	1.0
20	2	2	.57	10.0[a]	.52	10.0[a]	.67	4.0[a]	.61	5.0
	6	6	1.31	10.0[a]	1.28	10.0[a]	1.59	4.0[a]	1.54	4.2
	2	6	.99	10.0[a]	.96	12.6	.89	4.0[a]	.81	8.2
	6	2					1.46	4.0[a]	1.43	3.2
35	2	2	.31	17.5[a]	.30	17.5[a]	.46	7.0[a]	.44	7.7
	6	6	.75	17.5[a]	.74	17.5[a]	1.13	7.0[a]	1.12	7.0[a]
	2	6	.55	17.5[a]	.54	19.2	.54	7.0[a]	.52	9.8
	6	2					1.09	7.0[a]	1.08	6.3

[a] True trait-locus location.

The computation of the $\hat{f}_{ij}$ is therefore straightforward and requires little computer time.

The algorithm immediately extends to include information from multiple markers. Letting π more generally denote the multipoint IBD state of the sib pair, we may compute $P(\pi)$ for multiple markers using the Markov property for sib-pair IBD states (e.g., Olson, in press), giving

$$P(\pi) = P(\pi_1)P(\pi_2|\pi_1)P(\pi_3|\pi_2) \ldots P(\pi_K|\pi_{K-1})$$

for K markers. Similarly, $P(I_m|\pi)$ is a product of K contributions, one from each marker. Again, the algorithm is valid when parental marker information is included; the computational burden remains trivial.

For larger pedigree structures, including nuclear families with more than two offspring, computation of joint-marker IBD estimates is expected to be time consuming, particularly if one wishes to extend the methodology to include information from other relative pairs from a large pedigree and additional markers. For that reason, we will examine, in the sib-pair setting, the feasibility of substituting estimates of π_1 and π_2 obtained from each of the markers separately, that is, using table 3 of Haseman and Elston (1972).

Expected Lod Score

Consider a set of N independent sib pairs. Assume that data on two codominant markers are available for each sib pair but that no parental marker data are available. The expected value of the lod score Z_d is equal to

$$E(Z_d) = N \sum_g P(g|A)Z_d(g) ,$$

where g is in the set of possible sib-pair marker genotypes and $Z_d(g)$ is the lod score obtained for that genotype. Using Bayes's rule, we obtain

$$P(g|A) = \frac{P(A|g)P(g)}{P(A)} .$$

The quantities $P(A|g)$ and $P(A)$ can be computed as in the previous section. The quantity $P(g)$ is the prior probability of the sib-pair genotype and is given in Olson (1994) for the case of sib pairs with no parental marker information.

Three expected lod score functions for 140 affected sib pairs are shown in figure 2. In each plot, the distance between the two markers was 20 cM, the population prevalence K_p was .001, and the trait locus was additive ($\delta_1 = 1$, $\delta_2 = \frac{1}{2}$, $\delta_3 = 0$). Each marker has two equally frequent alleles. In each plot, four curves are shown: $E(Z_d)$ as discussed above, with π_1 and π_2 estimated jointly; $E(Z_d)$, with π_1 and π_2 estimated individually; $E(Z_d)$, using only information from M_1; and $E(Z_d)$, using only information from M_2. The latter two curves were computed by put-

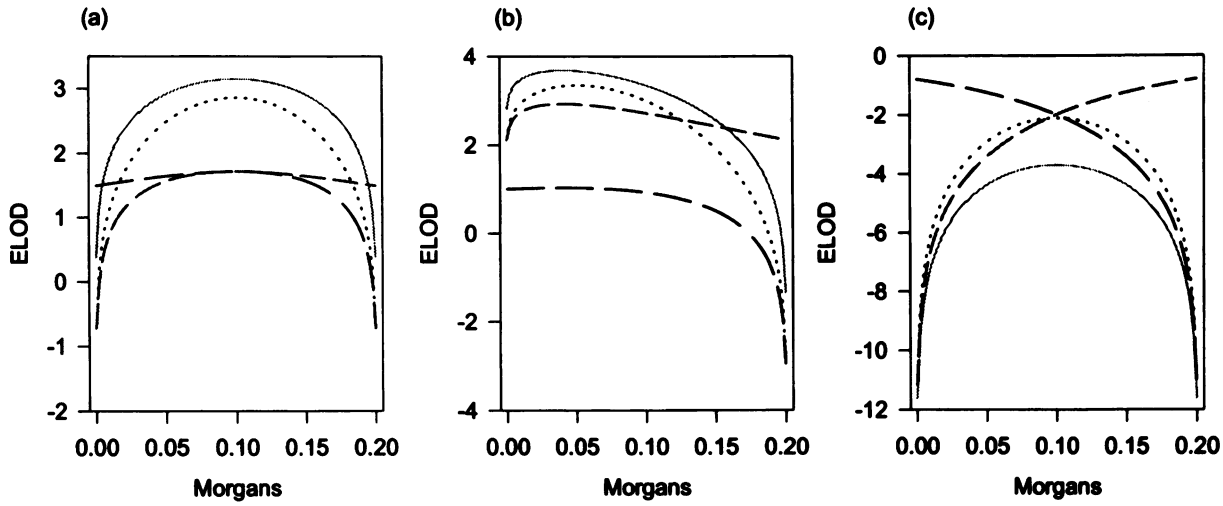

Figure 2 Expected lod score ($E[Z_d]$) curves for 140 affected sib pairs: two markers, joint estimation (——); two markers, individual estimation (---------), marker 1 only (– – – –); marker 2 only (— — —). For all figures, each marker has two equally frequent alleles; the trait model is additive, and the K_p = .001. *a*, Trait locus in center of interval (10 cM). *b*, Trait locus near marker 1 (4 cM). *c*, No linked trait.

ting $E(\pi_d | I_{mk}) = (1 - \psi) - (1 - 2\psi)E(\pi_k | I_{mk})$, $k = 1, 2$, where I_{mk} refers to the information from the kth marker locus. (Because the trait locus was additive, P ($\pi_d = \frac{1}{2} | I_{mk}$) is not needed here.)

Figure 2*a* shows a situation in which the trait locus lies halfway between the two markers. The curve computed by using the joint marker information (solid line) has the highest maximum $E(Z_d)$; $E(Z_d)$ computed by using two markers, individual estimates, also does well. Figure 2*b* shows a situation in which the trait locus lies 4 cM from M_1. Again, the curve computed by using the joint marker information has the highest maximum $E(Z_d)$. Figure 2*c* shows the case of no linked trait locus; here, joint marker information gives the most evidence against linkage.

These curves suggest that the information from two flanking markers has more power than a single marker; this finding parallels that of Fulker and Cardon (1994) and Olson (in press) for quantitative trait loci. Of interest is the comparison between jointly estimated and individually estimated π, and it is instructive to take a closer look at figure 2*b*. In addition to increasing power, the use of jointly estimated π is unbiased; that is, the maximum value of $E(Z_d)$ is at the correct location of 4 cM when π's are jointly estimated but at ~6 cM when individually estimated.

The expected bias generated by the individually estimated π is further illustrated in table 3, which gives the maximum $E(Z_d)$ and the location of $E(Z_d)$ for various marker distances, trait locations, and numbers of marker alleles for 25 affected sib pairs. In all cases, the trait locus is additive, and the population prevalence K_p equals .001. When the trait locus is in the center of the interval and markers are equally informative, both methods give the correct location. In all other cases, however, the maximum $E(Z_d)$ occurs at the correct location only when the π's are jointly estimated, with one exception occurring when marker informativity and distance between the markers are high.

To obtain information useful in study planning, we also computed $E(Z_d)$ at the true location for 25 sib pairs with either 0, 1, or 2 affected members, at various values of K_p and marker distance, and for dominant ($\delta_1 = 1$, $\delta_2 = 1$, $\delta_3 = 0$), additive ($\delta_1 = 1$, $\delta_2 = \frac{1}{2}$, $\delta_3 = 0$), and recessive ($\delta_1 = 1$, $\delta_2 = 0$, $\delta_3 = 0$) trait loci. All markers have two equally frequent alleles, and the trait locus is in the center of the interval; the results obtained therefore reflect close to the smallest $E(Z_d)$ in a genome mapping setting. The results are given in table 4. As expected, linkage information increases with decreasing marker distance and decreasing trait prevalence. Rare recessive diseases provide the most information.

As the likelihood function (4) is also valid for pairs with 0 or 1 affected members, we also included the contribution of these pairs to the expected lod score. With the exception of a common trait (K_p = .5), pairs with no affected members contribute virtually no linkage information. Pairs with one affected member contribute substantial information if the trait is dominant or a common recessive and a moderate amount of information if the trait is an uncommon or rare recessive. These results suggests that, if the trait locus is uncommon, the most powerful design samples only pairs with two affected, unless the trait locus is strongly dominant. On the other hand, if pairs with one affected member are available, adding them to the analysis will increase power. These findings are consistent with previous work (Blackwelder and Elston 1985).

Obtaining Values for K_p, V_A, and V_D

The likelihood (4) and its associated lod score may be computed directly if K_p, V_A, and V_D are known. Con-

Table 4

($E(Z_d)$) at True Trait-Locus Location for 25 Sib Pairs with 0, 1, or 2 Affected for Recessive, Additive, and Dominant Traits

K_p	Distance between Markers (cM)	Trait Model: Recessive — Two Affected Sibs	Recessive — One Affected Sib	Recessive — No Affected Sibs	Additive — Two Affected Sibs	Additive — One Affected Sib	Additive — No Affected Sibs	Dominant — Two Affected Sibs	Dominant — One Affected Sib	Dominant — No Affected Sibs
.5	5	.12	.92	.12	.04	.11	.04	.12	.92	.12
	10	.10	.73	.10	.03	.09	.03	.10	.73	.10
	20	.07	.48	.07	.02	.06	.02	.07	.48	.07
	35	.03	.26	.03	.01	.03	.01	.03	.26	.03
.1	5	1.29	.80	.00	.49	.11	.00	.72	.99	.00
	10	1.03	.63	.00	.40	.09	.00	.58	.79	.00
	20	.67	.40	.00	.26	.06	.00	.39	.53	.00
	35	.36	.21	.00	.15	.03	.00	.21	.29	.00
.01	5	3.79	.74	.00	1.00	.11	.00	1.05	1.00	.00
	10	2.96	.57	.00	.80	.09	.00	.84	.81	.00
	20	1.88	.35	.00	.53	.06	.00	.55	.53	.00
	35	.98	.18	.00	.29	.03	.00	.30	.30	.00
.001	5	5.48	.72	.00	1.09	.11	.00	1.09	1.00	.00
	10	4.22	.56	.00	.86	.09	.00	.87	.81	.00
	20	2.63	.34	.00	.57	.06	.00	.57	.53	.00
	35	1.35	.17	.00	.31	.03	.00	.31	.30	.00
.0001	5	6.23	.72	.00	1.10	.11	.00	1.10	1.00	.00
	10	4.76	.55	.00	.87	.09	.00	.87	.81	.00
	20	2.94	.33	.00	.57	.06	.00	.57	.53	.00
	35	1.49	.17	.00	.32	.03	.00	.32	.30	.00

NOTE.—Two markers with two equally frequent alleles; trait located in the center of the interval, π estimated jointly.

versely, estimation of these parameters is not possible if only affected individuals are sampled. Fortunately, for many traits, sufficient population information is available to give accurate estimates of these parameters. The parameter K_p is the prevalence of the trait in the population, and V_A and V_D may be obtained from information on incidence in siblings (K_s) and in parents/offspring (K_o) of affected probands (Suarez et al. 1976), sometimes referred to as recurrence risks. Specifically,

$$K_o = K_p + \frac{V_A}{2K_p},$$

and

$$K_s = K_p + \frac{2V_A + V_D}{4K_p},$$

so that

$$V_A = 2(K_o - K_p)K_p,$$

and

$$V_D = 4[(K_s - K_p)K_p - V_A/2].$$

If no dominance is present, then $K_o = K_s$, and

$$V_A = 2(K_s - K_p)K_p.$$

This formulation is particularly appropriate if unaffected and/or discordant pairs are present, as the values for K_p, V_A, and V_D may be used in the likelihood contributions from these pairs.

If only affected pairs are present, it is instructive to examine an alternative formulation. First, note that the expressions in table 2 are valid for relative pairs other than sib-pairs; the differing distributions of π_d for the different types of relative pairs give the probabilities

$$P(A_s = 2) = K_pK_s = \alpha + \beta/2 + \gamma/2,$$

$$P(A_o = 2) = K_pK_o = \alpha + \beta/2 + \gamma,$$

Table 5

$E(Z_d)$ at True Trait-Locus Location for 25 Sib Pairs with Two Affected as a Function of K_p, K_s, and K_o

K_p	K_s	K_o	Distance between Markers (cM) 5	10	20	35
.1	.2	.2	.25	.20	.14	.08
	.3	.3	.45	.36	.25	.14
	.4	.4	.58	.47	.31	.17
	.5	.5	.66	.54	.36	.20
	.3	.2	1.26	1.00	.65	.34
	.4	.3	1.19	.95	.62	.34
	.25	.2	.75	.60	.39	.21
.01	.05	.05	.66	.54	.36	.20
	.1	.1	.86	.69	.46	.25
	.2	.2	.97	.78	.52	.28
	.3	.3	1.01	.81	.53	.29
	.4	.4	1.03	.82	.54	.30
	.5	.5	1.05	.83	.55	.30
	.1	.05	2.59	2.05	1.32	.70
	.15	.05	3.62	2.83	1.80	.94
	.25	.08	3.83	2.99	1.90	.99
.001	.005	.005	.66	.54	.36	.20
	.01	.01	.86	.69	.46	.25
	.1	.1	1.07	.85	.56	.31
	.3	.3	1.09	.87	.57	.31
	.5	.5	1.09	.87	.57	.31
	.01	.005	2.59	2.05	1.32	.70
	.015	.005	3.62	2.83	1.80	.94

NOTE.—Two markers with two equally frequent alleles; trait located in the center of the interval; π estimated jointly.

and

$$P(A_g = 2) = K_p K_g = \alpha + \beta/4 + \gamma/2 ,$$

where K_g is the incidence in grandparents (or other second-degree relatives) of affected probands, and A_s, A_o, and A_g denote the number of affected individuals in the relative pair (s = sibling, o = parental, g = grandparental). Solving for α, β, and γ, we obtain

$$\alpha = K_p(2K_g - K_o) ,$$

$$\beta = 4K_p(K_s - K_g) ,$$

and

$$\gamma = 2K_p(K_o - K_s) .$$

Be careful to note that these values of α, β, and γ are valid only for the case of two affected relatives. Note that because K_p is a factor in all three parameters, it cancels in the likelihood and is not needed for lod score computation. This formulation is particularly useful when extended to the two-locus case, which I consider in the next section.

To summarize this subsection, given estimates of K_p, K_s, and K_o, the lod score and an estimate of trait-locus location may be obtained; in particular, prior knowledge of the trait allele frequencies and penetrances are not required. For samples of affected sib pairs, table 5 gives expected lod scores $E(Z_d)$ at the true trait-locus location for values of K_p, K_s, and K_o. The range of recurrence risks represented roughly covers the region consistent with the given K_p and a one-locus model (Suarez et al. 1976).

Extension to Two Trait Loci

Thus far, we have considered only the situation in which the genetic determinant of the trait is a single locus. For many traits, the possible presence of more than one trait locus complicates the analysis. Risch (1990*a*) gives formulas for the risk to relatives of affected probands for two-locus multiplicative, additive, and genetic heterogeneity models. These expressions may be used to adapt the present method to handle two trait loci.

To illustrate this, consider the genetic heterogeneity model. Risch (1990*a*) shows that the recurrence risk K_r to a relative of type r, is a function of the marginal disease prevalences K_{p1} and K_{p2} and the marginal recurrence risks K_{r1} and K_{r2} corresponding to trait loci 1 and 2, respectively. Specifically,

$$K_p K_r = 1 - 2(1 - K_{p1})(1 - K_{p2}) + (1 - 2K_{p1} + K_{p1}K_{r1})(1 - 2K_{p2} + K_{p2}K_{r2}) ,$$

where K_p is again the prevalence of trait in the population. Note that $K_p K_r$ is also the probability that both relatives in the pair are affected.

For a pair of affected sibs, assume data I_m are available on two linked markers and that one of the trait loci, say locus 1, lies between these two markers, and the other is not linked to either locus 1 or the markers. As before,

$$P(I_m|A) = \frac{P(I_m)}{P(A)} \sum_{\pi_1} P(A|\pi_1)P(\pi_1|I_m) ,$$

where π_1 is the IBD state at trait locus 1. Here, $P(A)$, the probability of an affected sib pair, is equal to $K_p K_s$, where K_s is the recurrence risk for siblings. To compute $P(A|\pi_1)$, the probability of an affected pair given the IBD state at locus 1 only, let w_{ij} be the penetrance of an individual with genotype g_i at locus 1 and g_j at locus 2. In a manner similar to Risch (1990*a*), we write

$$P(A|\pi_1) = \sum_i \sum_j \sum_k \sum_l w_{ij} w_{kl} P(g_i g_k|\pi_1) P(g_j g_l) .$$

In this summation, the subscripts i and j belong to the first individual, and k and l belong to the second; i and k belong

Table 6

Expressions for the Coefficients for Two-Locus Models for Affected Sib Pairs

Two-Locus Model[a]	α^*	β^*	γ^*
Multiplicative	$\alpha K_{p2}K_{s2}$	$\beta K_{p2}K_{s2}$	$\gamma K_{p2}K_{s2}$
Additive	$\alpha + K_{p2}K_{s2} + 2K_{p1}K_{p2}$	β	γ
Heterogeneity	$1 - 2(1 - K_{p1})(1 - K_{p2}) + (1 - K_{p1})(1 - 2K_{p2} + K_{p2}K_{s2}) + \alpha(1 - 2K_{p2} + K_{p2}K_{s2})$	$\beta(1 - 2K_{p2} + K_{p2}K_{s2})$	$\gamma(1 - 2K_{p2} + K_{p2}K_{s2})$

[a] From Risch (1990*a*).

to locus 1, and j and l belong to locus 2. For the genetic heterogeneity model, the penetrance w_{ij} is defined as $1 - (1 - x_i)(1 - y_j)$, where x_i and y_j are the marginal penetrances for the two trait loci. Substituting into the previous expression and simplifying, we get

$$P(A|\pi_1) = 1 - 2(1 - K_{p1})(1 - K_{p2}) + (1 - 2K_{p1})(1 - 2K_{p2} + K_{p2}K_{s2}) + (1 - 2K_{p2} + K_{p2}K_{s2})P(A_1|\pi_1) .$$

The expression $P(A_1|\pi_1)$ is the conditional probability that the sib pair is affected because of locus 1 only; it equals $\alpha + \beta\pi_1 + \gamma I(\pi_1 = 1/2)$ as in the one-locus case. As a result, $P(A|I_m)$ is also a linear function of $E(\pi_1|I_m)$ and $P(\pi_1 = 1/2|I_m)$ and may be written

$$P(A|I_m) = \alpha^* + \beta^* E(\pi_1|I_m) + \gamma^* P(\pi_1 = 1/2|I_m) ;$$

expressions for these coefficients are given in table 6. Results are similar for the additive and multiplicative models of Risch (1990*a*); expressions for α^*, β^*, and γ^* for these models are also given in table 6.

If the genetic model is known, then the coefficients are easily computed. If the genetic model is not known, the coefficients may not be computed without making some assumptions about the model. With some knowledge or assumptions about locus 2, the population recurrence-risk information may be used to gain estimates of the coefficients in the same manner as in the one locus case. This is easily done in the case of the multiplicative model. First note that

$$P(A|\pi_1) = [\alpha + \beta\pi_1 + \gamma I(\pi_1 = 1/2)]K_{p2}K_{r2} ,$$

for any type of relative pair r. As a result, it suffices to assume that K_{s2}/K_{g2} and K_{s2}/K_{p2} are known and equal c_g and c_o, respectively.

Then, following a program similar to that in the one-locus case, we can write

$$K_pK_s = \alpha^* + \beta^*/2 + \gamma^*/2 ,$$

$$c_oK_pK_o = \alpha^* + \beta^*/2 + \gamma^* ,$$

and

$$c_gK_pK_g = \alpha^* + \beta^*/4 + \gamma^*/2 .$$

Solving for the coefficients, we get

$$\alpha^* = K_p(2c_gK_g - c_oK_o) ,$$

$$\beta^* = 4K_p(K_s - c_gK_g) ,$$

and

$$\gamma^* = 2K_p(c_oK_o - K_s) .$$

In the absence of any knowledge of locus 2, we suggest putting $c_o = \sqrt{K_s/K_o}$ and $c_g = \sqrt{K_s/K_g}$; these choices put $K_{s2}/K_{g2} = K_{s1}/K_{g1}$ and $K_{s2}/K_{o2} = K_{s1}/K_{o1}$.

Table 7

Some One-Locus Models

No.	Type	p	δ_1	δ_2	δ_3	K_p	K_s	K_o	K_g
1	Additive	.01	1	1/2	0	.01	.2575	.2575	.1338
2	Dominant	.005	1	1	0	.01	.5043	.5038	.2569
3	Recessive	.1	1	0	0	.01	.3025	.1000	.0550
4	Additive	.1	1	1/2	0	.1	.3250	.3250	.2125
5	Dominant	.0513	1	1	0	.1	.5439	.5380	.3190
6	Recessive	.3162	1	0	0	.1	.4331	.3162	.2081

Table 8

Some Two-Locus Models

No.	Locus 1 (No.[a])	Locus 2 (No.[a])	Type	K_p	K_s	K_o	K_g
1	Additive (4)	Additive (4)	Multiplicative	.01	.1056	.1056	.0452
2	Dominant (5)	Dominant (5)	Multiplicative	.01	.2958	.2894	.1018
3	Recessive (6)	Recessive (6)	Multiplicative	.01	.1876	.1000	.0433
4	Dominant (5)	Recessive (6)	Multiplicative	.01	.2356	.1701	.0664
5	Additive (1)	Additive (1)	Heterogeneity	.0199	.2665	.2665	.1382
6	Dominant (2)	Dominant (2)	Heterogeneity	.0199	.5038	.5038	.2665
7	Recessive (3)	Recessive (3)	Heterogeneity	.0199	.3060	.1086	.0642
8	Dominant (2)	Recessive (3)	Heterogeneity	.0199	.4048	.3058	.1652
9	Dominant (5)	Recessive (3)	Heterogeneity	.109	.5311	.5099	.3094
10	Dominant (2)	Recessive (6)	Heterogeneity	.109	.4468	.3411	.2248

[a] From Table 7.

Unfortunately, the additive and heterogeneity models can not be so easily approximated; however, in practice, it may suffice to use the multiplicative model approximation for two-locus diseases. We examined the effect of using this approximation for several two-locus models. Table 7 gives parameter values for several one-locus models that were combined under either a multiplicative or genetic heterogeneity model to give the two-locus models in table 8. Table 9 gives the maximum $E(Z_d)$ and location estimates obtained when the true model, one-locus approximation, and the multiplicative two-locus approximation are used on a sample of 25 independent sib pairs. The parameters α^*, β^*, and γ^* for the true model were obtained from the parameters of their component loci. The parameters α^*, β^*, and γ^* for the two approximations used the true K_s, K_o, and K_g in the manner described above for the one-locus case and for the two-locus case. Each marker had two equally frequent alleles, and the markers were 20 cM apart; the trait locus was 4 cM from M_1.

The one-locus approximation does rather poorly in all cases, except when the contribution of one of the loci is much greater than that of the other (models 9 and 10). For the multiplicative models, parameter values obtained using the one-locus approximation generate negative likelihood

Table 9

Maximum $E(Z_d)$ and Estimated Distance in cM from M_1 ($\hat{d}$) for Two-Locus Models and their Approximations

				Approximation			
		True Model		One-Locus		Two-Locus	
Two-Locus Model	Given Marker IBD State at Locus	max $E(Z_d)$	$\hat{d}$	max $E(Z_d)$	$\hat{d}$	max $E(Z_d)$	$\hat{d}$
1	1	.30	4.0[a]	...[b]	...	.30	4.0[a]
2	1	.44	4.0[a]	...	...	.44	4.0[a]
3	1	.79	4.0[a]	...	...	.79	4.0[a]
4	1	.44	4.0[a]	...	...	.43	5.5
4	2	.79	4.0[a]	...	...	.78	2.4
5	1	.13	4.0[a]	−.04	8.4	.13	5.8
6	1	.14	4.0[a]	.00	8.2	.14	5.2
7	1	.50	4.0[a]	−.04	8.4	.44	6.4
8	1	.23	4.0[a]	−.03	8.2	.21	6.1
8	2	.28	4.0[a]	.03	8.4	.27	5.0
9	1	.37	4.0[a]	.37	5.0	.33	.0
9	2	.00	4.0[a]	−.35	9.8	−.09	9.4
10	1	.00	4.0[a]	−.54	9.8	−.15	9.4
10	2	.60	4.0[a]	.60	4.8	.53	.0

NOTE.—Twenty-five affected sib pairs; markers are 20 cM apart, with two equally frequent alleles; trait locus is at $d = 4$ cM.

[a] True trait-locus location.

[b] ... = Not a valid model.

contributions and are invalid. The two-locus approximation does well, however, even when the true model is one of heterogeneity. Some power is lost, and the location estimate may be inaccurate, but the method is nonetheless capable of giving an approximate location for the trait loci. In no case does the two-locus approximation result in an maximum $E(Z_d)$ that is <88% of that resulting from the true model, except for the rare loci in models 9 and 10. These results suggest that the two-locus approximation may be an effective screening tool when scanning the genome for linked markers. In addition, the recent work of Neuman and Rice (1992) gives values of recurrence risks compatible with several two-locus epistatic and heterogeneity models; this information will aid researchers in choosing between one- and two-locus models. An alternative approach might be to search simultaneously for two disease loci. Such an approach may increase the power of a linkage study (Schork et al. 1993; Knapp et al. 1994) and may be feasible for a genome scan in the affected sib-pair setting (Knapp et al. 1994).

Discussion

We describe an interval mapping approach to scanning the genome for markers linked to one or two loci underlying a dichotomous trait or disease. The method is parametric in the sense that it depends on prior knowledge of recurrence risks to relatives of affected individuals. On the other hand, prior knowledge of the trait allele frequencies and penetrances are not required, and thus the method retains some of the model-free properties of affected sib-pair methods. The method gives an unbiased estimate of trait-locus location, provided the recurrence risks are correctly specified and provided the marker IBD states are estimated jointly, rather than individually, as proposed by Fulker and Cardon (1994) for quantitative data. In addition, the jointly estimated marker IBD states are easy to compute, even for multiple markers, for sib pairs with or without parental marker information, provided the sib pairs are independent. Although I did not specifically examine the case in which parental marker information is available, power is likely to improve if additional family members are typed (e.g., Risch 1990*b*).

The method has good power for detecting linkage for rare diseases, particularly those displaying a strong dominance component. Some guidelines useful for study planning are given in the form of lod scores expected for traits consistent with a variety of one-locus models. A more detailed examination of power, particularly for two-locus trait models, and comparison to other mapping methods, such as full maximum likelihood and nonparametric methods, are needed to provide suitable guidelines for mapping complex diseases. Appropriate critical values also need to be chosen. A critical lod score value of 3.3 has been suggested for human data by Lander and Schork (1994), on the basis of work by Lander and Botstein (1989) in which the Ornstein-Uhlenbeck diffusion process is used to approximate the null distribution of linkage statistics. Further work is also needed to determine the most useful applications of the methods. It is anticipated that these methods will be most useful in a coarse, rather than fine, mapping context; in fine mapping, the lod score curve may be too flat to provide a sufficiently narrow confidence interval around location estimate. As a result, it is important that the coarse mapping methods provide as accurate a location estimate as possible.

I have emphasized the use of the method with samples of affected sib pairs, although the likelihood is also given for unaffected and discordant sib pairs for the one-locus case. Affected-unaffected pairs contribute substantial information about linkage primarily if the trait is strongly dominant. Inclusion of all available sib pairs is practical in the one-locus case, as estimates of K_p, V_A, and V_D may be used to construct the likelihoods of all types of sib pairs; in the two-locus case, recurrence-rate information is not sufficient to obtain estimates of all of the regression parameters required. In addition, if discordant pairs are included, the Haseman-Elston method may be applied directly and parameters estimated in a standard regression framework. This latter approach has the advantage of not requiring any prior trait knowledge; however, the estimates obtained in an interval mapping setting are likely to be biased (Fulker and Cardon 1994; Olson 1994*b*).

It is also possible, in the case of affected sibpairs, to estimate β and γ, given prior knowledge of α, by using standard likelihood methods. In the one-locus case, α is simply K_p^2, the square of the trait prevalence in the population and in many cases may be well specified. It is instructive, then, to examine the likelihood for an additive trait locus ($\gamma = 0$). When the trait is rare or uncommon, α is substantially smaller than β, and the likelihood as a function of β is very flat; it is very difficult to estimate β with good precision, without very large sample sizes. When the trait is common, the likelihood as a function of β has a much sharper maximum; in this case, simulations (not shown) suggest that both $\hat{\beta}$ and the maximum lod score are well estimated if the sample size is sufficiently large but greatly overinflated if the sample size is small.

Risch (1990*b*) gives a form of the likelihood in the single-marker case $P(I_m|A) = \Sigma_\pi P(\pi|A)P(I_m|\pi)$ and proposes estimation of the parameters $P(\pi|A)$, $\pi = 0, \frac{1}{2}, 1$, by using standard maximum-likelihood techniques. An extension of this method to the interval mapping framework may give an alternative form of the likelihood (4) that yields more stable parameter estimates. Further examination of parameter estimation using these likelihoods is needed.

The methods outlined in this paper rely on prior knowledge of trait recurrence-risk information. If such information is not available, a nonparametric approach may be taken instead when data consist solely of affected sib pairs. The quantity $E(\pi_d|I_m)$ may be used to con-

struct a modified Green and Woodrow (1977) statistic $T_d = (\bar{\hat{\pi}}_d - \frac{1}{2})/\sqrt{Var(\bar{\hat{\pi}}_d)}$, where $\bar{\hat{\pi}}_d$ is the mean of $\hat{\pi}_d$ for the sample of sib pairs. The statistic T_d may then be computed at values of d throughout the interval defined by the two markers. Unfortunately, T_d, although providing evidence for linkage, gives biased location estimates (not shown). The nonparametric approach does have the advantage of not depending on any prior knowledge of the trait model and may be suitable for a preliminary genome scan; detailed investigation of this alternative is left for future research. A related issue of great importance is the robustness of the methods to misspecification of recurrence risks and marker allele frequencies. With regard to the former, the results in table 9 and other preliminary work suggest that, while estimates may be biased, good power remains; much more work is needed to fully explore the robustness of the methods.

The method is easily extended to other types of relative pairs. If such pairs are independent, their contributions may be included by adding their terms to the log-likelihood. I leave the treatment of correlated relative pairs from an extended pedigree to future research. One of the problems with use of a multipoint method in large pedigrees is the computational burden involved in estimating the marker variables; this burden increases rapidly when more than two markers are considered. One challenge, therefore, is to develop efficient strategies for multipoint IBD computation for arbitrary pedigree structures. One final note is given regarding computation. Programs to implement the methods outlined in this paper and analogous methods for quantitative traits (Olson, in press) are in development and will be made available by the author at a future date.

Acknowledgments

I thank Barbara McKnight for helpful discussion and Robert Elston for continued support and encouragement. This work was supported by grant CA09168 from the National Institutes of Health and was undertaken while I was a Senior Fellow in the Department of Biostatistics at the University of Washington.

References

Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests from disease susceptibility loci. Genet Epidemiol 2:85–97

Day NE, Simons MJ (1976) Disease susceptibility genes—their identification by multiple case family studies. Tissue Antigens 8:109–119

de Vries RRP, Fat RFM, Lai A, Nijenhuis LE, van Rood JJJ (1976) HLA-linked genetic control of host response to *Mycobacterium leprae*. Lancet 2:1328–1330

Fulker DW, Cardon LR (1994) A sib-pair approach to interval mapping of quantitative trait loci. Am J Hum Genet 54:1092–1103

Green JR, Woodrow JC (1977) Sibling method for detecting HLA-linked genes in a disease. Tissue Antigens 9:31–35

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

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19

Knapp M, Seuchter SA, Baur MP (1994) Two-locus disease models with two marker loci: the power of affected-sib-pair tests. Am J Hum Genet 55:1030–1041

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:125–199

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

Lange K (1986) The affected sib-pair method using identity by state relations. Am J Hum Genet 39:148–150

Lathrop GM, Lalouel JM, Julier C, Ott J (1985) Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. Am J Hum Genet 37:482–498

Neuman RJ, Rice JP (1992) Two-locus models of disease. Genet Epidemiol 9:347–365

Olson JM (1994) Estimating two-point identity-by-descent probabilities for sibpairs with no parental information. Tech rep 132, Department of Biostatistics, University of Washington, Seattle

———. Robust multipoint linkage analysis: an extension of the Haseman-Elston method. Genet Epidemiol (in press)

Risch N (1990*a*) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228

——— (1990*b*) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242–253

Schork NJ, Boehnke M, Terwilliger JD, Ott J (1993) Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am J Hum Genet 53:1127–1136

Suarez BK, Hodge SE (1979) A simple method to detect linkage for rare recessive diseases: an application to juvenile diabetes. Clin Genet 15:126–136

Suarez BK, Reich T, Trost J (1976) Limits of the general two-allele single locus model with incomplete penetrance. Ann Hum Genet 40:231–243

Suarez BK, Rice JP, Reich T (1978) The generalized sib pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 42:87–94

Todorov A (1992) Robust methods for multipoint linkage analysis. PhD thesis, Louisiana State University, New Orleans

Weeks DE, Lange K (1988) The affected-pedigree-member method of linkage analysis. Am J Hum Genet 42:315–326

——— (1992) A multilocus extension of the affected-pedigree-member method of linkage analysis. Am J Hum Genet 50:859–868