Am. J. Hum. Genet. 51:773–784, 1992

Power of Sib-Pair and Sib-Trio Linkage Analysis with Assortative Mating and Multiple Disease Loci

William M. Sribney* and Michael Swift†'‡

*Department of Biostatistics, †Biological Sciences Research Center, and ‡Genetics Division, Department of Medicine, University of North Carolina, Chapel Hill

Summary

Sib-pair linkage analysis has been proposed for identifying genes that predispose to common diseases. We have shown that the presence of assortative mating and multiple disease-susceptibility loci (genetic heterogeneity) can increase the required sample size for affected-affected sib pairs several fold over the sample size required under random mating. We propose a new test statistic based on sib trios composed of either one unaffected and two affected siblings or one affected and two unaffected siblings. The sample-size requirements under assortative mating and multiple disease loci for these sib-trio statistics are much smaller, under most conditions, than the corresponding sample sizes for sib pairs. Study designs based on data from sib trios with one or two affected members are recommended whenever assortative mating and genetic heterogeneity are suspected.

Introduction

Linkage analysis using affected sib pairs has been proposed for identifying specific genes that predispose to common diseases. Sib-pair methods (Penrose 1935; Suarez et al. 1978; Blackwelder and Elston 1982, 1985; Suarez and Van Eerdewegh 1984; Amos et al. 1989) or other nonparametric methods (Weeks and Lange 1988; Amos and Elston 1989) are appealing for common diseases, since the mode of inheritance may be unknown and since sibships or partial families with one or more affected members may be easier to obtain than extended pedigrees. However, for any proposed linkage study of a common disorder, it is important to take into account the possibilities of assortative mating, multiple disease-susceptibility loci, incomplete penetrance, epistasis, and diagnostic uncertainties. The presence of one or more of these factors may substantially affect the power of linkage analysis (Cavalli-Sforza and King 1986; Goldin and Gershon 1988; Ploughman and Boehnke 1989; Risch 1990).

In evaluating the power of sib-pair linkage analysis, we considered the t_2 test statistic of Blackwelder and Elston (1985), which represents the mean number of marker alleles identical by descent (IBD) between the sibs. This test statistic was shown by Blackwelder and Elston (1985) to be more powerful than other proposed test statistics, under most penetrance and prevalence combinations for a single disease locus with random mating. Recently, Schaid and Nick (1990, 1991; also see Knapp 1991) showed, also for a single disease locus and random mating, that the power of the t_2 statistic was quite close to that of the asymptotically most powerful test.

We examined the power of the t_2 statistic for sib pairs with two affected members and for sib pairs with one affected member, when there is assortative mating, when there are several loci that predispose to the same disease, and when there is incomplete penetrance or epistasis. We also developed test statistics for sib trios with one, two, and three affected members, and we examined their power under the same conditions as those for sib pairs.

Received February 12, 1992; final revision received June 1, 1992.

Address for correspondence and reprints: Michael Swift, M.D., New York Medical College, Valhalla, NY 10595.

0002-9297/92/5104-0011$02.00

Assortative Mating and Observed Phenotypic Concordance

When the probability that one member of a mated pair has a specific phenotypic trait is dependent on

whether the other mate has the same trait, then we say that there is assortative mating relative to this trait. For a dichotomous phenotype classification ($x = 0$ if unaffected; and $x = 1$ if affected), the distribution of phenotypes among mated pairs can be described with the parametrization

$$w_{xy} = \frac{P(x,y)}{P(x)\,P(y)}, \tag{1}$$

where x is the phenotype of the female mate and y is the phenotype of male mate. For random mating, $w_{xy} \equiv 1$.

The w_{xy} are related to the prevalences of the affected phenotype among females who have mates, $v_f = P(x=1)$, and among males who have mates, $v_m = P(y=1)$:

$$v_f = P(1,0) + P(1,1) = w_{10}v_f(1 - v_m) + w_{11}v_f v_m\,;$$
$$v_m = P(0,1) + P(1,1) = w_{01}(1 - v_f)v_m + w_{11}v_f v_m\,.$$

In addition, there is the constraint $\Sigma_{xy}P(x, y) = 1$. Thus, the w_{xy} have only 1 df once the sex-specific prevalences (for those individuals who have mates) are determined. It is convenient to specify this df via the concordance rate c of the trait in mates of affected individuals:

$$\begin{aligned} c &= P(\text{mate of individual affected} \mid \text{individual affected}) \\ &= \frac{P(1,1)}{(v_f + v_m)/2} \\ &= \frac{2w_{11}v_f v_m}{v_f + v_m}\,. \end{aligned} \tag{2}$$

Thus, the concordance rate c, along with v_f and v_m, can be used to specify uniquely the phenotype distribution in mated pairs.

It is important to note that the "affected" or "disease" phenotype is defined by the observer; there is no way to know how it relates to the characteristics that actually determine the selection of mating partners in the population. There may be a "mate-selection" phenotype, related to the same genetic locus or loci as the disease phenotype, that is hidden to the observer but relative to which assortative mating occurs. Thus, the concordance rate for the disease phenotype may not reveal the true degree of assortative mating. The concordance rate for the disease phenotype may only represent a lower bound for the concordance of the full phenotype that is related to the genetic locus or loci associated with the disease phenotype. For example, although Palmour et al. (1991) found no evidence for positive assortative mating between alcoholics or alcoholics and psychiatrically ill persons in a Canadian population, they did observe a significant ($P < .002$) deviation from random mating between alcoholic males and nonalcoholic daughters of alcoholic fathers.

The concordance rates considered in this paper are similar to observed rates in clinical studies. Concordance rates (computed in accordance with eq. [2]) ranging from 29% to 64% for overall psychiatric disorder (Merikangas 1982*b*) and from 7% to 44% for alcoholism (Jacob and Bremer 1986) have been reported. Substantially elevated concordance rates for more narrowly defined disorders have also been observed; for example, Merikangas (1982*a*) reported that 24 (43%) of 56 married probands with primary affective disorder had spouses who were also diagnosed with primary affective disorder.

Statistical Model for Assortative Mating

To determine the power of linkage tests under assortative mating, the joint genotype distribution for mated pairs at the loci associated with the disease phenotype must be determined. This genotype distribution depends, in general, on the concordance rate, penetrance function, and prevalence of the mate-selection phenotype. For simplicity of presentation, we assume that these are the same for both the mate-selection phenotype and the disease phenotype. Thus, for the computation of the genotype distribution, we make the following assumptions:

1. There are k autosomal disease-susceptibility loci, which can act either independently or epistatically to predispose to a clinically indistinguishable disease phenotype. Each of these loci is unlinked to all the others.
2. Each of the k disease-susceptibility loci has two alleles: a disease allele A_i and a wild-type allele a_i, $i = 1,2, \ldots, k$. An individual's genotype at the k loci is represented by $s = (s_1, s_2, \ldots, s_k)$, where s_i is the genotype at the ith locus, $s_i = a_ia_i$, A_ia_i, or A_iA_i.
3. The disease phenotype has a dichotomous classification: $x = 0$ if unaffected, and $x = 1$ if affected.
4. The concordance rate, penetrance function, and prevalence are the same for both the mate-selection

phenotype and the disease phenotype. The penetrance function is denoted by $f(x|s) = P(x|s)$, where x is the individual's phenotype and s is the individual's genotype at the k disease-susceptibility loci.

5. The disease prevalence and penetrance function are the same for both males and females.

Genotype Distribution under Assortative Mating

Under random mating, the assumption that the genotype distribution is in equilibrium leads to a unique choice for the genotype distribution: the Hardy-Weinberg distribution. For the case of assortative mating when there is a single locus and complete dominant or recessive expression, the equilibrium genotype distribution is easily determined (Wright 1921; Crow and Felsenstein 1968). Other single-locus models have also been presented (O'Donald 1960; Karlin 1968; Karlin and Scudo 1969; Scudo and Karlin 1969; Stark 1977).

For multiple loci and for any form of the penetrance function, we determine the equilibrium genotype distribution under assortative mating by a computational method as follows: We start with an assortative-mating population with an arbitrary initial genotype distribution and then "breed" the population to produce the genotype distribution for the offspring of this population. We then impose the same degree of assortative mating on this second generation and compute the genotype distribution of their offspring. This process is continued until the genotype distribution reaches equilibrium.

Since we must evaluate the relationship between the phenotypes and genotypes of mated pairs, we consider the joint phenotype and genotype distribution of mated pairs, $P(x,s;y,t)$, where x and y are, respectively, the phenotypes of the female and male mates, and s and t are the genotypes of the female and male mates at the k disease-susceptibility loci. The assumption of assortative mating relative to the disease phenotype means that the joint genotype distribution for mated pairs is dependent only on the joint phenotype distribution. Thus, the genotypes of a mated pair conditional on their phenotypes are independent; i.e., $P(s,t|x,y) = P(s|x)P(t|y)$. Therefore,

$$\begin{aligned} P(x,s;y,t) &= P(s|x)P(t|y)P(x,y) \\ &= \frac{P(x,s)}{P(x)}\frac{P(y,t)}{P(y)}P(x,y) \\ &= \frac{P(x,y)}{P(x)P(y)}P(x|s)P(s)P(y|t)P(t) \\ &= w_{xy}f(x|s)g(s)f(y|t)g(t)\,, \end{aligned} \tag{3}$$

where w_{xy} is given by equation (1), $f(x|s) = P(x|s)$ is the penetrance function, and $g(s) = P(s)$ is the genotype distribution for an individual. As was noted earlier, the w_{xy} have only 1 df, which can be specified via the phenotypic concordance rate c of equation (2).

We now choose an arbitrary distribution for $g(s)$ in equation (3) and "breed" the population represented by the distribution $P(x,s;y,t)$. The genotype distribution $g^*(u)$ of the offspring from the matings described by $P(x,s;y,t)$ is given by

$$g^*(u) = \sum_{st} P(u|s,t)P(s,t)\,, \tag{4}$$

with

$$P(s,t) = \sum_{xy} P(x,s;y,t)\,. \tag{5}$$

Since the disease-susceptibility loci are assumed to be unlinked, the probability that a child has genotype u conditional on the parents' genotypes s and t is

$$P(u|s,t) = \prod_{i=1}^{k} P(u_i|s_i,t_i) = \prod_{i=1}^{k} \pi_{s_i t_i u_i}\,, \tag{6}$$

where $\pi_{s_i t_i u_i}$ is the usual genetic transition matrix for two alleles at one autosomal locus (Elston and Stewart 1971). Thus, equation (4) becomes

$$g^*(u) = \sum_{st}\left[\prod_{i=1}^{k} \pi_{s_i t_i u_i}\right]P(s,t)\,. \tag{7}$$

To compute the genotype distribution for successive generations, $g^*(\cdot)$ is substituted in place of $g(\cdot)$ in equation (3). The parameters w_{xy} must be recalculated, since they depend on the disease prevalence, which is altered by the new genotype distribution. The degree of assortative mating is kept fixed by holding c, the phenotypic concordance rate, constant. Equation (7) then gives the genotype distribution for the third generation. By iteration in this manner, the genotype distributions for successive generations are obtained, and this process is continued until the genotype frequencies reach equilibrium. For a specified disease prevalence, the equilibrium genotype distribution can be obtained by selecting, through an interpolation search procedure, the allele frequencies of the initial genotype distribution.

The uniqueness of the equilibrium genotype distribution has been shown for some models of assortative mating (Wright 1921; O'Donald 1960; Karlin 1968; Lange 1976). For each specification of our model of

assortative mating (i.e., number of disease loci, penetrance function, disease prevalence, and concordance rate), the uniqueness of the equilibrium genotype distribution can be demonstrated empirically by showing that it does not depend on the genotype distribution chosen for the first generation.

Test Statistics for Sib Pairs

In examining the power of sib pairs and sib trios to detect linkage, we make the following assumptions, in addition to those assumptions made in the model of assortative mating:

1. There is a polymorphic marker locus that is fully informative in each nuclear family in the study sample.
2. There is complete ascertainment: the probability that any sibship with at least one affected member will be included in the study sample is independent of the number of affected sibs.

We wish to test the hypothesis that one of the disease-susceptibility loci (the first of the k loci) is linked to the marker locus with recombination frequency $\theta < \frac{1}{2}$.

Since it is assumed that each family is fully informative at the marker locus (i.e., the father has genotype F1/F2, and the mother has genotype M1/M2), the distribution of marker alleles IBD for sib pairs can be calculated from the conditional probability of a child having marker genotype m (m = F1/M1, F1/M2, F2/M1, or F2/M2) and disease status x, given the parents' genotypes s and t at the disease-susceptibility loci. This probability is given by

$$P(x,m|s,t) = \sum_u P(x|u)P(m|u_1,s_1,t_1)P(u|s,t)\,, \quad (8)$$

where u takes on all the child's possible genotypes at the disease-susceptibility loci, and u_1, s_1, and t_1 represent genotypes at the first disease-susceptibility locus.

The probability $P(u|s,t)$ is given by equation (6). In a similar manner, $\psi_{s_1t_1u_1m} = P(m|u_1,s_1,t_1)$ can be expressed in terms of the recombination frequency θ by using the genetic transition matrix for two linked autosomal loci (Elston and Stewart 1971). Thus, equation (8) becomes

$$P(x,m|s,t) = \sum_u f(x|u)\psi_{s_1t_1u_1}\prod_{i=1}^{k}\pi_{s_1t_1u_1}\,. \quad (9)$$

Under the null hypothesis H_0 of no linkage between the marker locus and any of the k disease loci, the probabilities of observing either 0, 1, or 2 alleles IBD between two sibs are, respectively, $p_{00} = \frac{1}{4}$, $p_{01} = \frac{1}{2}$, and $p_{02} = \frac{1}{4}$. Under the alternative hypothesis H_1 of linkage between the marker locus and the first of the k disease loci, the probability p_j of observing j = 0, 1, or 2 alleles IBD between two sibs can be computed using equation (9) and the joint distribution $P(s,t)$ of the parents' genotypes that is given by equation (5). For a sib pair with disease status (x,y),

$$p_j \propto \sum_s\sum_t\sum_{mm'} P(x,m|s,t)P(y,m'|s,t)P(s,t)\,,$$

$$\text{3 } j \text{ alleles IBD}$$

where the sum over m and m' is restricted to the marker-locus genotypes that give exactly j alleles IBD; the constant of proportionality is determined by the constraint $p_0 + p_1 + p_2 = 1$.

If n affected-affected (1,1) sib pairs are observed and n_0, n_1, n_2 are the numbers of pairs with, respectively, 0, 1, 2 marker alleles IBD, then $(n_0,n_1,n_2) \sim$ trinomial $(n;\ p_0,p_1,p_2)$. The test statistic $T_{11} = b_0\hat{p}_0 + b_1\hat{p}_1 + b_2\hat{p}_2$, where $\hat{p}_j = n_j/n$ and $b_0 = 0$, $b_1 = 1$, and $b_2 = 2$, represents the mean number of marker alleles IBD. We choose to use the notation T_{11}, where the subscript denotes the (1,1) disease status of the pair, rather than the t_2 notation of Blackwelder and Elston (1985), to differentiate this statistic from the corresponding statistic T_{01} for unaffected-affected (0,1) sib pairs, which we also examine.

Test Statistics for Sib Trios

We consider test statistics for each of the sib trios with one or more affected members: (0,0,1), (0,1,1), and (1,1,1) trios. For each trio, the test statistic is based on the joint IBD distribution produced by all the possible unique combinations of marker alleles IBD in pairwise comparisons among the three sibs.

For sib trios with one or two affected members, the joint IBD distribution has seven possible states (table 1A). For trios consisting of three affected sibs, there are four unique states (table 1B).

We denote the probability of observing these IBD states by p_j, with $j = 1,2,\ldots,7$, for (0,0,1) or (0,1,1) trios, and $j = 1,2,3,4$ for (1,1,1) trios. The probabilities p_{0j} for these states under the null hypothesis H_0 of no linkage are given in table 1. Under the hypothesis H_1 of linkage, the probabilities of the IBD states can be computed by enumerating all the possible arrange-

Table 1

Possible Combinations of Marker Alleles IBD in Sib Trios

A. Trios with One or Two Affected Members: Trio Disease Status $(x,y_1,y_2) = (1,0,0)$ or $(x,y_1,y_2) = (0,1,1)$

State j	Markers IBD between Pairs (x,y_1)	(x,y_2)	(y_1,y_2)	Probability under H_0: p_{0j}	b_j[a]
1	2	2	2	1/16	0
2	2	1	1	1/4	0
3	2	0	0	1/8	0
4	1	1	0	1/8	0
5	1	0	1	1/4	1
6	1	1	2	1/8	1
7	0	0	2	1/16	2

B. Trios with Three Affected Members: Trio Disease Status $(x_1,x_2,x_3) = (1,1,1)$

State j	Markers IBD between Pairs (x_1,x_2)	(x_1,x_3)	(x_2,x_3)	Probability under H_0: p_{0j}	b_j[a]
1	2	2	2	1/16	2
2	2	1	1	3/8	1
3	2	0	0	3/16	0
4	1	1	0	3/8	0

[a] The test statistic for the trio is given by $\sum_j b_j\hat{p}_j$.

ments of marker alleles for each member of the trio with disease status (x,y,z):

$$p_j \propto \sum_s \sum_t \sum_{mm'm''} P(x,m|s,t)P(y,m'|s,t)P(z,m''|s,t)P(s,t)\,.$$

3 state j observed

If n trios each with disease status (0,1,1) are observed, and $n_1,n_2,\ldots,n_7$ are the numbers of trios for the seven possible IBD states, then $(n_1,n_2,\ldots,n_7) \sim$ multinomial$(n; p_1,p_2,\ldots,p_7)$. Following the form of the test statistic for sib pairs, we construct a test statistic for trios that is a linear combination of the estimators $\hat{p}_j = n_j/n$:

$$T_{011} = \sum_{j=1}^{7} b_j\hat{p}_j\,.$$

The b_j are chosen to optimize the power of the test statistic. By the Neyman-Pearson Lemma (Knapp 1991), the most powerful test for a specified linkage hypothesis H_1 (i.e., specified number of disease loci, recombination frequency, penetrance function, disease prevalence, and degree of assortative mating) is based on the ratio

$$\frac{P_{H1}(n_1,n_2,\ldots,n_7)}{P_{H0}(n_1,n_2,\ldots,n_7)} = \prod_{j=1}^{7}\left(\frac{p_j}{p_{0j}}\right)^{nj}.$$

Taking the logarithm and dividing by n gives

$$T^*_{011} = \sum_{j=1}^{7} b_j^*\hat{p}_j = \sum_{j=1}^{7}\left(\log\frac{p_j}{p_{0j}}\right)\hat{p}_j \qquad (10)$$

for the statistic that gives the most powerful test.

The optimal coefficients b_j^* vary only moderately with different linkage hypotheses; for example, the b_j^* for a dominant penetrance function are approximately the same as those for a recessive penetrance function. The selected values for the b_j are given in table 1A; this set of values gives the minimum, or near minimum, required sample size for a wide range of linkage hypotheses.

The statistic T_{001} for (0,0,1) trios is identical in form to T_{011}. Indeed, since both statistics have identical properties under the null hypothesis of no linkage, the statistics could be combined (by simply combining the observed numbers of IBD states: $n_1,n_2,\ldots,n_7$) for the purpose of testing the null hypothesis. However, while the probabilities p_j are the same for both trios

under the null hypothesis, they are different under the hypothesis of linkage. Thus, sample-size calculations for the two trios give different results, and so we consider them separately.

The statistic T_{111} for (1,1,1) trios is constructed in a similar manner:

$$T_{111} = \sum_{j=1}^{4} b_j \hat{p}_j ,$$

with b_j, as given in table 1B, chosen to give the minimum, or near minimum, required sample size for a range of linkage hypotheses, on the basis of comparison with the statistic for the Neyman-Pearson most powerful test.

Calculation of Required Sample Size

The mean and variance of the test statistics T_{11}, T_{01}, T_{011}, T_{001}, and T_{111} are $\mu = \sum_j b_j p_j$, and $\sigma^2 = \zeta^2/n$, respectively, where

$$\zeta^2 = \sum_j b_j^2 p_j (1 - p_j) - \sum_{i \neq j} b_i b_j p_i p_j .$$

With a normal approximation, the required sample size n for power $(1 - \beta)$ and significance level α (one-sided) is given by

$$n \geq \left(\frac{z_{1-\alpha}\zeta_0 + z_{1-\beta}\zeta_1}{\mu_0 - \mu_1} \right)^2$$

where μ_0 and ζ_0 are computed under the null hypothesis H_0 of no linkage, μ_1 and ζ_1 are computed under the hypothesis H_1 of linkage, and $z_{1-\alpha}$ is the $(1 - \alpha)$th quantile of the standard normal distribution.

Penetrance Functions

We first consider sample-size requirements for penetrance functions representing incomplete dominant expression. For these cases, when considering multiple disease loci, we assume that each of the disease-susceptibility loci acts independently to predispose to the disease phenotype. Table 2A gives the penetrance function for two loci acting independently, each with dominant expression and penetrance of .8.

To represent the possibility of sporadic cases, we set the penetrance for wild-type homozygotes to a positive value. Table 2B represents the situation in which sporadic events and two disease-susceptibility loci, each with dominant expression and penetrance of .8, are considered as three independent and equal causes of disease when the total disease prevalence is .05. Table 2C gives a penetrance function for two loci acting independently, each with recessive expression, and table 2D represents a penetrance function for two loci acting epistatically.

Table 2

Penetrance Functions for Two Disease-Susceptibility Loci

A. Two Loci Acting Independently with Incomplete Dominant Expression

Locus 1 Genotype	Penetrance Function Locus 2 Genotype		
	a_2a_2	A_2a_2	A_2A_2
a_1a_1	0	.8	.8
A_1a_1	.8	.96	.96
A_1A_1	.8	.96	.96

B. Sporadic Cases and Two Loci Acting Independently with Incomplete Dominant Expression[a]

Locus 1 Genotype	Penetrance Function Locus 2 Genotype		
	a_2a_2	A_2a_2	A_2A_2
a_1a_1	.0169	.8034	.8034
A_2a_2	.8034	.9607	.9607
A_1A_1	.8034	.9607	.9607

C. Two Loci Acting Independently with Incomplete Recessive Expression

Locus 1 Genotype	Penetrance Function Locus 2 Genotype		
	a_2a_2	A_2a_2	A_2A_2
a_1a_1	0	0	.8
A_1a_1	0	0	.8
A_1A_1	.8	.8	.96

D. Two Loci with Epistasis

Locus 1 Genotype	Penetrance Function Locus 2 Genotype		
	a_2a_2	A_2a_2	A_2A_2
a_1a_1	0	0	0
A_1a_1	0	.8	.8
A_1A_1	0	.8	.8

[a] The penetrance function represents the case in which sporadic events and two disease-susceptibility loci, each with dominant expression and penetrance .8, are considered as three independent and equal causes of disease when the total disease prevalence is .05.

Results

Tables 3–7 give, for sib pairs and sib trios, the sample sizes, based on the corresponding test statistic T_{11}, T_{01}, T_{111}, T_{011}, or T_{001}, that are needed to detect linkage between a disease-susceptibility locus and a single marker locus. These tables show the required sample size for a one-sided significance level of $\alpha = .001$ (roughly equivalent to a lod score of 3) and power $(1-\beta) = .90$. Tables 3–7 are derived using a recombination frequency $\theta = .02$ between the first disease locus and the marker locus and a disease prevalence of .05. Over the range of disease prevalence from .01 to .20, the sample-size requirements (data not shown) are fairly stable, and our findings about the relative power of the test statistics hold for this range of prevalence.

Tables 3–7 show the required sample sizes when mating is random and for varying degrees of assortative mating. In table 3, required sample sizes are given for one, two, and four disease loci, each with dominant expression. The penetrance function for the case of two disease loci is given in table 2A.

A situation in which sporadic cases of disease occur is illustrated in table 4 (penetrance function is given in table 2B). Table 5 gives sample-size requirements for different penetrance values when there are two disease-susceptibility loci acting independently, each with dominant expression. Table 6 gives an example of recessive expression (penetrance function is given in table 2C). An example of epistasis between two disease loci is shown in table 7 (penetrance function is given in table 2D).

Discussion

Assortative Mating

There are two ways in which assortative mating decreases the power of linkage tests based on affected siblings only: First, because affected-affected matings are more likely under assortative mating, there are

Table 3

Required Sample Sizes, by Number of Disease Loci, for Loci Acting Independently with Incomplete Dominant Expression[a]

No. of Disease Loci and Pair or Trio[b]	Required Sample Size[c]			
	Random Mating	Assortative Mating Concordance Rate c		
		.25	.50	.75
1:				
(1,1)	52	100	243	715
(1,1,1)	22	58	172	559
(0,1,1)	28	38	60	116
(0,0,1)	45	49	55	68
(0,1)	101	120	159	249
2:				
(1,1)	222	417	1,036	3,576
(1,1,1)	92	221	668	2,604
(0,1,1)	115	148	218	375
(0,0,1)	181	185	194	210
(0,1)	417	472	582	816
4:				
(1,1)	903	1,684	4,262	16,166
(1,1,1)	360	850	2,613	11,264
(0,1,1)	451	573	820	1,353
(0,0,1)	707	711	718	729
(0,1)	1,675	1,863	2,226	2,975

[a] Penetrance function for one disease locus is $f(A_1A_1) = f(A_1a_1) = .8$ and $f(a_1a_1) = 0$. Penetrance function for two loci is given in table 2A. Penetrance function for four loci is computed assuming four dominant loci acting independently each with penetrance .8. Disease prevalence is .05. Recombination frequency $\theta = .02$ between first disease locus and marker locus.

[b] Pair or trio is that on which the test statistic is based.

[c] For $\alpha = .001$ (one-sided) and power $(1-\beta) = .90$.

Table 4

Required Sample Sizes for Two Disease Loci Acting Independently with Incomplete Dominant Expression and Sporadic Cases[a]

Pair or Trio[b]	Required Sample Size[c]			
	Random Mating	Assortative Mating Concordance Rates c		
		.25	.50	.75
(1,1)	251	375	646	1,204
(1,1,1)	95	171	348	735
(0,1,1)	133	156	195	255
(0,0,1)	990	1,149	1,447	1,964
(0,1)	1,407	1,628	2,030	2,704

[a] Sporadic events and the two disease loci are three independent and equal causes of disease. Penetrance function is given in table 2B. Disease prevalence is .05. Recombination frequency $\theta = .02$ between first disease locus and marker locus

[b] On which test statistic is based.

[c] For $\alpha = .001$ (one-sided) and power $(1-\beta) = .90$.

Table 5

Required Sample Sizes, by Penetrance, for Two Disease Loci Acting Independently with Incomplete Dominant Expression[a]

Penetrance and Pair or Trio[b]	Required Sample Size[c]			
	Random Mating	Assortative Mating Concordance Rate c		
		.25	.50	.75
.3:				
(1,1)	330	401	521	695
(1,1,1)	167	218	307	440
(0,1,1)	489	565	687	853
(0,0,1)	4,219	4,409	4,696	5,062
(0,1)	6,860	7,281	7,927	8,766
.5:				
(1,1)	258	373	624	1,145
(1,1,1)	117	197	382	790
(0,1,1)	257	321	441	646
(0,0,1)	966	1,020	1,110	1,244
(0,1)	1,773	1,957	2,278	2,790
.7:				
(1,1)	230	397	864	2,367
(1,1,1)	98	210	551	1,719
(0,1,1)	150	194	284	474
(0,0,1)	307	320	343	382
(0,1)	651	736	900	1,221
.9:				
(1,1)	216	440	1,248	5,531
(1,1,1)	88	234	807	3,952
(0,1,1)	87	110	157	266
(0,0,1)	107	108	109	111
(0,1)	272	305	371	518

[a] Penetrance function is computed assuming two dominant loci acting independently, each with penetrance as given. Disease prevalence is .05. Recombination frequency $\theta = .02$ between first disease locus and marker locus.

[b] Pair or trio is that on which test statistic is based.

[c] For $\alpha = .001$ (one-sided) and power $(1-\beta) = .90$.

Table 6

Required Sample Sizes for Two Disease Loci Acting Independently with Recessive Expression[a]

Pair or Trio[b]	Required Sample Size			
	Random Mating	Assortative Mating Concordance Rates c		
		.25	.50	.75
(1,1)	111	178	324	643
(1,1,1)	73	182	467	1,158
(0,1,1)	90	105	131	176
(0,0,1)	353	361	371	382
(0,1)	618	656	716	802

[a] Penetrance function is given in table 2C. Disease prevalence is .05. Recombination frequency θ = .02 between first disease locus and marker locus.
[b] On which test statistic is based.
[c] For α = .001 (one-sided) and power $(1-\beta)$ = .90.

more families in which the mother and the father each carry a disease-susceptibility allele. Two children of such a mating can both be affected and yet not share the same (by descent) disease allele. Clearly, this reduces the power of tests based on the number of marker alleles shared among affected sibs.

Second, assortative mating causes an increase in the proportion of individuals with two or more disease alleles. For example, a single disease locus with dominant expression, penetrance of .8, and disease prevalence of .05 under random mating results in a population with .0010 of individuals homozygous for the disease allele. Under the same conditions, assortative mating with a concordance rate c = .5 gives a .0083 proportion of homozygotes, an eightfold increase.

In the case of two disease loci each with dominant expression and the penetrance function of table 2A and with disease prevalence of .05, random mating gives a proportion of .0015 for individuals with two or more disease alleles. Under assortative mating this proportion becomes .0121, also an eightfold increase. This increase, under assortative mating, in the proportion of mothers and fathers with multiple disease alleles also leads to an increased likelihood of different (by descent) disease alleles among affected siblings and to a concomitant loss of power for test statistics based on affected sibs only.

Required sample sizes for test statistics based on only affected sibs increase quite rapidly with increasing degrees of assortative mating. For the conditions of table 3, assortative mating with a phenotypic concordance rate c = .75 increases the required sample size for (1,1) sib pairs to roughly 16-fold over the required sample size under random mating. The re-

Table 7

Required Sample Sizes for Two Disease Loci with Epistasis[a]

Pair or Trio[b]	Required Sample Size[c]			
	Random Mating	Assortative Mating Concordance Rate c		
		.25	.50	.75
(1,1)	115	180	334	705
(1,1,1)	69	136	301	716
(0,1,1)	97	118	158	232
(0,0,1)	362	366	374	388
(0,1)	650	691	764	894

[a] Penetrance function is given in table 2D. Disease prevalence is .05. Recombination frequency θ = .02 between first disease locus and marker locus.
[b] On which test statistic is based.
[c] For α = .001 (one-sided) and power $(1-\beta)$ = .90.

quired sample size for (1,1,1) sib trios increases by over 25-fold. But required sample sizes for pairs and trios composed of both affected and unaffected sibs change to a much lesser degree, by factors of 4 or less.

As the number of disease loci increases (table 3), the required sample size increases for all types of sib pairs and trios. However, with higher degrees of assortative mating, the greater the difference between statistics based on affected sibs and those based on affected and unaffected sibs. For (1,1) pairs and (1,1,1) trios, the combined effect of assortative mating and multiple disease loci is greater than a multiplicative effect; for (0,1,1), (0,0,1), and (0,1) sibships, the combined effect of assortative mating and multiple disease loci is less than multiplicative.

For the conditions of table 3, the power of (0,1,1) trios is just a little better than that of twice as many (0,1) pairs. Thus, it appears that the greater power of (0,1,1) trios may be simply due to the fact that each (0,1,1) trio is composed of two (0,1) pairs and one (1,1) pair. However, when there is the possibility of sporadic disease (table 4), (0,1,1) trios are superior to twice as many independent (0,1) pairs plus an equal number of independent (1,1) pairs, even under random mating.

Sporadic Disease

Under the conditions of table 3, the test statistic based on (0,0,1) trios has the smallest required sample sizes for large degrees of assortative mating. But this assumes that there are no sporadic cases of disease. If sporadic cases of disease can occur, then the power of the statistic for (0,0,1) trios is dramatically lower (table 4). The statistics based on (1,1), (1,1,1), and (0,1,1) sibships are, however, quite robust with respect to the occurrence of sporadic cases. This is expected, since, with cases occurring sporadically, many of the (0,0,1) trios have the affected member being a sporadic case; on the other hand, a much smaller proportion of sibships with two or more affected members contain sporadic cases.

A close comparison of table 4 and the corresponding entries of table 3 for two loci reveals an apparent anomaly: Under assortative mating, the sample-size requirements for (1,1) and (1,1,1) sibships are less when there are sporadic cases than when there are no sporadic cases. This is due to different joint genotype distributions for parents in the two situations. With assortative mating, many families with affected children will have parents who are both affected. But when sporadic disease is possible, one member of an affected-affected mating will sometimes be a sporadic case. Hence, for the same degree of assortative mating, there will be more families in which only one disease allele is segregating when there are sporadic cases than when there are no sporadic cases. Since test statistics based on affected sibs are more powerful when only one disease allele is present in each family, the required sample sizes under assortative mating for these statistics are smaller when there is sporadic disease than when there is no sporadic disease.

Epistasis and Recessive Expression

The required sample sizes for the test statistics based on (1,1) pairs, (1,1,1) trios, and (0,1,1) trios in the case of recessive expression or epistasis (tables 6 and 7) are less than the corresponding sample sizes for dominant expression (table 3). Thus, for these test statistics, multiple disease loci acting independently with dominant expression are a greater impediment to detecting linkage than are loci with recessive expression or loci acting epistatically.

Penetrance and the Selection of Test Statistics

For high penetrances, high degrees of assortative mating, and no sporadic disease (table 5), the T_{001} statistic for (0,0,1) trios is the most powerful. But when the penetrance is about .6 or less, the T_{001} statistic performs poorly. In comparison, the T_{011} statistic for (0,1,1) trios is powerful at low and high penetrances. At penetrances greater than or equal to .5, the T_{011} statistic is more powerful than the T_{11} statistic for (1,1) pairs; however, for penetrances less than or equal to .3, T_{11} is superior to T_{011}.

The optimal coefficients b_j^* (eq. [10]) for the test statistic for (0,1,1) trios vary for different penetrance functions. The b_j coefficients for T_{011} are specifically chosen to give optimal, or near optimal, power for penetrances greater than or equal to .5, for disease loci with dominant expression. The lower power of the T_{011} statistic relative to the T_{11} statistic at low penetrances is therefore not surprising. However, since (0,1,1) trios contain a (1,1) pair, for a specified linkage hypothesis (i.e., specified penetrance function, etc.), it is always possible to construct a test statistic for (0,1,1) trios that has power greater than or equal to the power of any statistic for (1,1) pairs. But when the penetrance is low, the unaffected member of the trio contains very little information, and optimizing the statistic for (0,1,1) trios under this condition gives only slight improvement in power over the statistic for (1,1) pairs. Thus, when the penetrance is unknown,

it is a better strategy to use the b_j that we have chosen, which are optimal for penetrances greater than or equal to .5, for analyzing data from (0,1,1) trios. The (1,1) pairs contained in these trios can then, of course, also be evaluated using the T_{11} statistic.

In contrast to the case of (0,1,1) and (0,0,1) trios, the optimal coefficients b_j^* for the test statistics for (1,1) and (0,1) pairs and for (1,1,1) trios do not vary with the value of the penetrance, when linkage hypotheses with disease loci with dominant expression are considered. For the models represented in tables 3–5, the sample size requirements of the T_{11}, T_{01}, and T_{111} statistics are within 1% of the requirements of the corresponding Neyman-Pearson most powerful tests.

Schaid and Nick (1990, 1991) showed that, for dominant expression, the power of T_{11} was very close to optimal for one disease locus, with or without sporadic cases, and random mating. Our results show that for T_{11}, along with T_{01} and T_{111}, this is also true for multiple disease loci with dominant expression, with or without sporadic cases, and assortative mating.

The power of the test statistics T_{11}, T_{01}, T_{111}, T_{011}, and T_{001} are farther from their optimal levels when recessive and epistatic penetrance functions are considered. However, the sample-size requirements are lower in these cases than they are for dominant penetrance functions. Thus, it is a better strategy, when the penetrance function is unknown, to use statistics optimized for the dominant case.

Coincidentally, although the b_j coefficients for T_{111} were chosen on the basis of comparison with the Neyman-Pearson most powerful test, with this choice of b_j, $T_{111}+2$ is identical to the Green and Woodrow (1977) statistic for three affected sibs.

Another test statistic for (0,1,1) trios has been considered elsewhere. Suarez et al. (1982) examined the power of a statistic for (0,1,1) trios that was derived from a general scoring method that Alter and Quevedo (1979) used for families with multiple affected members. In our notation (see table 1A), this test statistic is

$$Z = \frac{\hat{p}_1 + \hat{p}_2}{\hat{p}_1 + \hat{p}_2 + \hat{p}_5 + \hat{p}_6 + \hat{p}_7}.$$

Suarez et al. (1982) found that the power of this test statistic was less than that of a statistic for (1,1) pairs. However, the relative inefficiency of the Z statistic is due to its specific configuration rather than to an inherent lack of power in (0,1,1) trios.

Mate-Selection and Disease Phenotypes

The distinction between the mate-selection phenotype and the disease phenotype is important, since, in general, the penetrance functions and concordance rates for the two phenotypes will be different. Our methodology can easily be adjusted to compute the required sample size when the mate-selection phenotype and the disease phenotype have different penetrance functions and concordance rates. In this case, the concordance rate, penetrance function, and prevalence used for computing the genotype distribution under assortative mating would represent the mate-selection phenotype. The penetrance function representing the disease phenotype would then be used for the computation of the probability distributions of the test statistics for linkage (eq. [9]).

Conclusion

Genetic heterogeneity and assortative mating, features of many important common disorders, may make it difficult to collect, for any single disorder, a sample of affected sib pairs that is adequate to detect linkage of disease-susceptibility loci to marker loci. In this setting, collection of data from sib trios with one or two affected individuals may provide the basis for successful detection of linkage. Even under conditions of random mating, the additional effort to collect data for sib trios may be worthwhile.

Acknowledgments

This work was supported by a grant from the North Carolina Alcoholism Research Agency. We thank Ruth Heim, Charles Chase, Daphne Morrell, Lawrence Kupper, and Grazyna Janic for reading an earlier version of this manuscript and for their many helpful suggestions.

References

Alter M, Quevedo J (1979) Genetic segregation of multiple sclerosis and histocompatibility (HLA) haplotypes. J Neurol 222:67–74

Amos CI, Elston RC (1989) Robust methods for the detection of genetic linkage for quantitative data from pedigrees. Genet Epidemiol 6:349–360

Amos CI, Elston RC, Wilson AF, Bailey-Wilson JE (1989) A more powerful robust sib-pair test of linkage for quantitative traits. Genet Epidemiol 6:435–449

Blackwelder WC, Elston RC (1982) Power and robustness of sib-pair linkage tests and extension to larger sibships. Commun Stat Theor Methods 11:449–484

——— (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2:85–97

Cavalli-Sforza LL, King M-C (1986) Detecting linkage for genetically heterogeneous diseases and detecting heterogeneity with linkage data. Am J Hum Genet 38:599–616

Crow JF, Felsenstein J (1968) The effect of assortative mating on the genetic composition of a population. Eugenics Q 15:85–97

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542

Goldin LR, Gershon ES (1988) Power of the affected-sib-pair method for heterogeneous disorders. Genet Epidemiol 5:35–42

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

Jacob T, Bremer DA (1986) Assortative mating among men and women alcoholics. J Stud Alcohol 47:219–222

Karlin S (1968) Equilibrium behavior of population genetic models with non-random mating. J Appl Prob 5:231–313

Karlin S, Scudo FM (1969) Assortative mating based on phenotype. II. Two autosomal alleles without dominance. Genetics 63:499–510

Knapp M (1991) A powerful test of sib-pair linkage for disease susceptibility. Genet Epidemiol 8:141

Lange K (1976) Stable gene equilibria for mixtures of random and assortative mating. Math Biosci 29:49–57

Merikangas KR (1982*a*) Assortative mating among in-patients with primary affective disorder. Psychol Med 12: 753–764

——— (1982*b*) Assortative mating for psychiatric disorders and psychological traits. Arch Gen Psychiatry 39:1173–1180

O'Donald P (1960) Assortative mating in a population in which two alleles are segregating. Heredity 15:389–396

Palmour RM, Bonnet F, Marcouillier M, Smith AJK, Ervin FR, Dongier M, Fujiwara TM, et al (1991) Genetic epidemiology of alcoholism in a Canadian outpatient population. Am J Hum Genet 49 [Suppl]: A478

Penrose LS (1935) The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Ann Eugenics 6:133–138

Ploughman LM, Boehnke M (1989) Estimating the power of a proposed linkage study for a complex genetic trait. Am J Hum Genet 44:543–551

Risch N (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241

Schaid DJ, Nick TG (1990) Sib-pair linkage tests for disease susceptibility loci: common tests vs. the asymptotically most powerful test. Genet Epidemiol 7:359–370

——— (1991) Reply to "A powerful test of sib-pair linkage for disease susceptibility." Genet Epidemiol 8:142–143

Scudo FM, Karlin S (1969) Assortative mating based on phenotype. I. Two autosomal alleles with dominance. Genetics 63:479–498

Stark AE (1977) A model of assortative mating with partial dominance. Heredity 39:91–95

Suarez B, O'Rourke D, Van Eerdewegh P (1982) Power of the affected-sib-pair method to detect disease susceptibility loci of small effect: an application to multiple sclerosis. Am J Med Genet 12:309–326

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

Suarez B, Van Eerdewegh P (1984) A comparison of three affected-sib-pair scoring methods to detect HLA-linked disease susceptibility genes. Am J Med Genet 18:135–146

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

Wright S (1921) Systems of mating. III. Assortative mating based on somatic resemblance. Genetics 6:144–161