Am. J. Hum. Genet. 64:641–649, 1999

Power of Association and Linkage Tests When the Disease Alleles Are Unobserved

I-Ping Tu and Alice S. Whittemore

Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA

Summary

Genomewide association studies have been advocated as a promising alternative to genomewide linkage scans for detection of small-effect genes in complex diseases. Comparisons of power and sample size between the two strategies have shown considerable advantages for the association studies. These comparisons assume that the set of markers includes the exact disease-related polymorphism. A concern, however, is that the power of an association study decreases when this is not the case, because of discrepant allele frequencies and less-than-maximum disequilibrium between the disease-related polymorphism and its nearest marker. Here, we quantify this concern by comparing the sample sizes needed by the two strategies when the markers exclude the disease-related polymorphism. For affected sib pairs and their parents, we found that incomplete disequilibrium and differing allele frequencies can have substantial negative impact on the power of association studies, resulting, in some circumstances, in little gain and even in loss of power, compared with linkage analysis. We provide some guidelines for choosing between strategies, for the detection of genes for complex diseases.

Introduction

For some diseases with complex genetic etiologies, conflicting results have emerged from case-control studies of association between specific markers and disease, compared with linkage analyses based on allele sharing within families. Whereas the case-control studies have shown strong associations, the linkage tests have proved negative (Terwillinger and Ott 1992; Spielman et al. 1993). To explain this phenomenon, Risch and Merikangas (1996) suggested that allele-sharing linkage tests have poor power, compared with tests for association, and that a genomewide search for associations involving several diallelic markers within each gene (spaced ~1 kb apart) is more sensitive than multipoint scanning for linkage. To support this argument, Risch and Merikangas (1996) compared the number of families needed for an affected–sib pair linkage test to that needed for an association test based on the transmission/disequilibrium-test (TDT) statistic (Spielman et al. 1993). They showed substantial reductions in sample size for the association test, compared with the allele-sharing test, provided (1) that a multiplicative model specifies the penetrances of three genotypes at a given disease locus and (2) that the set of diallelic markers includes this locus.

With regard to the first of these assumptions, Camp (1997) found similar sample-size reductions for additive, dominant, and recessive penetrance models. With regard to the second assumption, Muller-Myhsok and Abel (1997) noted that, if less-than-maximum linkage disequilibrium exists between the disease locus and the closest diallelic marker and if the allele frequencies at the disease and marker loci differ substantially, the sample sizes necessary for the association test would be increased substantially. These authors recently elaborated on their comments with sample-size estimates for the TDT (Abel and Muller-Myhsok 1998). These estimates were limited to families containing a single affected offspring, with penetrances determined by a multiplicative model, and they were not compared with sample sizes needed for allele-sharing tests.

Here, we further quantify the comments of Muller-Myhsok and Abel (1997) by comparing the sample sizes needed for association and allele-sharing tests, under several different penetrance models and several different sets of assumptions about the extent of linkage disequilibrium and the discrepancy of allele frequencies between disease and marker loci. We found that less-than-maximum disequilibrium and differing frequencies between disease and marker alleles can have substantial negative impact on the power of the association test, resulting in relatively little gain and sometimes even in loss of power, compared with the allele-sharing test.

Received September 10, 1998; accepted for publication November 25, 1998; electronically published February 4, 1999.

Address for correspondence and reprints: Dr. Alice S. Whittemore, Department of Health Research and Policy, Stanford University School of Medicine, Redwood Building, Room T204, Stanford, CA 94305-5405. E-mail: asw@osiris.stanford.edu

0002-9297/99/6402-0036$02.00

Methods

We assume that N pairs of affected sibs and their parents have been typed at a set of closely spaced diallelic markers, such as single-nucleotide polymorphisms, on an autosomal chromosome. We also assume that each of the two alleles at any marker has a frequency of $\geqslant 10\%$ in the parental population. We wish to test the null hypothesis that the chromosome contains no disease locus, using tests with good power against the alternative hypothesis that it contains a single disease locus. The disease locus may be distinct from all the markers; however, we assume that the markers are so closely spaced that the probability of recombination between the disease locus and its nearest marker is 0. We assume that, at the disease locus, one allele or group of alleles, *A,* is associated with increased risk of the disease, and we use *a* to denote the collection of wild-type alleles at the locus.

We consider test statistics of the form $T = \max_t S_t$, where S_t is a function of the family genotypes at marker locus t and where $\max_t$ denotes the maximum over the marker loci on the chromosome. We determine a threshold value c such that the null probability that T exceeds c is ~.05. For the reasons given in Appendix A, we set $c = 4.12$ for the allele-sharing test and $c = 5.33$ for the association test. We then calculate the number N of families needed for an 80% probability that T exceeds c when the chromosome contains a disease locus. To do so, we approximate this probability by the quantity

$$1 - \Phi\left(\frac{c - \sqrt{N}\mu}{\sigma}\right) \ . \qquad (1)$$

Here, Φ is the standard Gaussian cumulative distribution function, and $\sqrt{N}\mu$ and σ^2 are the mean and variance, respectively, of the statistic S_t, at the marker locus t closest to the disease locus. (Appendix A includes a justification of approximation [1].) Equating expression (1) to the desired power, 80%, and solving for N indicates that the number of families needed is approximately

$$N = \left(\frac{c + .84\sigma}{\mu}\right)^2 \ . \qquad (2)$$

Thus, to determine the sample size N for each of the two tests, we must specify the quantities μ and σ^2.

Linkage

For linkage, we use the means statistic $S_t = (X_{2,t} - X_{0,t})/\sqrt{N/2}$ for the statistic S_t. Here, $X_{i,t}$ denotes the number of sib pairs sharing i alleles identical by descent at locus t, for $i = 0, 1, 2$ (Suarez et al. 1978). Like Risch and Merikangas (1996) and Camp (1997), we assume that the sibs' identity by descent at any marker can be determined unambiguously. This is plausible because the markers are assumed to be so closely spaced that the four parental haplotypes are distinguishable in any chromosomal region.

To determine μ and σ^2 in equation (2), we note that, when the chromosome contains a disease gene, the mean and variance of S_τ, at locus τ of the nearest marker, equal their values at the disease locus, because the probability of recombination between the disease locus and its nearest marker essentially is 0. The mean and variance at the trait locus are $\sqrt{N}\mu$ and σ^2, respectively, with

$$\mu = \sqrt{2}(\pi_2 - \pi_0) \ , \qquad (3)$$

and

$$\sigma^2 = 2\pi_2(1 - \pi_2) + 2\pi_0(1 - \pi_0) + 4\pi_2\pi_0 \ . \qquad (4)$$

Here, π_i denotes the probability that an affected sib pair shares i alleles identical by descent at the disease locus, for $i = 0, 1, 2$. Appendix B includes equations for π_i in terms of the frequencies and penetrances of the two alleles at the disease locus.

Association

For the association test, we use the TDT statistic at marker locus t for S_t. This statistic is calculated by counting certain types of parental meioses. Only meioses from parents who are heterozygous at locus t are included in the count. Specifically, we arbitrarily label the two marker alleles "*B*" and "*b*." For the kth family, x_{kt} denotes the number of meioses in which a heterozygous parent transmits allele B to an offspring, minus the number of meioses in which a heterozygous parent transmits allele b. For example, $x_{kt} = 4$ for a family with two heterozygous parents and two homozygous *BB* offspring. In contrast, $x_{kt} = 2$ for a family with one heterozygous parent and two homozygous *BB* offspring, $x_{kt} = -2$ for a family with two heterozygous parents, one heterozygous offspring, and one homozygous *bb* offspring, and $x_{kt} = 0$ if both parents are homozygous or if all family members are heterozygous. The TDT statistic at locus t is $S_t = (|\sum_k x_{kt}|)/\sqrt{4N\text{hêt}}$, where hêt is the proportion of heterozygotes among the $2N$ parents, an estimate of the prevalence het of *Bb* heterozygotes in the parental population.

To determine the mean $\sqrt{N}\mu$ and variance σ^2 of the statistic S_τ at the marker locus $t = \tau$ closest to the disease-causing polymorphism, we use h_i to denote the probability that a family contains i heterozygous parents, given that both offspring are affected, for $i = 0, 1, 2$. Then, het $= h_2 + \frac{1}{2}h_1$. Also, let τ_{ij} denote the probability that

two affected offspring having j heterozygous parents receive i copies of allele B from these parents, for $i = 0, 1, ..., 4$ and $j = 1, 2$. Table 1 gives the distribution of x_{kt} in terms of these probabilities. Straightforward calculation by use of this distribution shows that, for a large N, the asymptotic distribution of S_τ is the nonnegative part of a Gaussian distribution with mean $\sqrt{N}\mu$ and variance σ^2. Here,

$$\mu = \text{het}^{-1/2}[h_2(2\tau_{42} - 2\tau_{02} + \tau_{32} - \tau_{12}) + h_1(\tau_{21} - \tau_{01})] , \quad (5)$$

and

$$\sigma^2 = \text{het}^{-1}[h_2(4\tau_{42} + 4\tau_{02} + \tau_{32} + \tau_{12}) + h_1(\tau_{21} + \tau_{01})] - \text{het}^{-1}[h_2(2\tau_{42} - 2\tau_{02} + \tau_{32} - \tau_{12}) + h_1(\tau_{21} - \tau_{01})]^2 . \quad (6)$$

Appendix C includes equations for h_i and τ_{ij} in terms of the penetrances at the disease locus, the allele frequencies at the two loci, and the extent of disequilibrium between them.

Results

Figures 1–4 show the sample sizes needed for the linkage and association tests, for penetrances governed by additive, multiplicative, dominant, and recessive models, respectively. Sample sizes are shown for various frequencies of disease allele A and allele B of the marker nearest the disease locus and under various assumptions about the extent of disequilibrium between the disease and marker alleles.

The penetrance models all specify the penetrances by $f_i = F(\alpha + \beta c_i)$, for $i = 0, 1, 2$, with $c_0 = 0$. Here, F is the exponential function for the multiplicative model and the identity function for the other models. For the multiplicative and additive models, $c_i = i$; for the recessive model, $c_1 = 0$ and $c_2 = 1$, whereas $c_1 = c_2 = 1$ for the dominant model. We chose the constants α and β so that the population attributable risk (PAR) would be 33% for all models (the PAR is the proportion of all disease occurrences that would be eliminated if everyone had the normal genotype at the disease locus).

Table 1

Distribution of the Differences between the Numbers of *B* and *b* Alleles Transmitted, by Heterozygous Parents, to Two Affected Sibs

Difference	Probability
4	$h_2\tau_{42}$
2	$h_2\tau_{32} + h_1\tau_{21}$
0	$h_0 + h_1\tau_{11} + h_2\tau_{22}$
−2	$h_2\tau_{12} + h_1\tau_{01}$
−4	$h_2\tau_{02}$

NOTE.—B and b denote the two alleles at the marker closest to the disease locus.

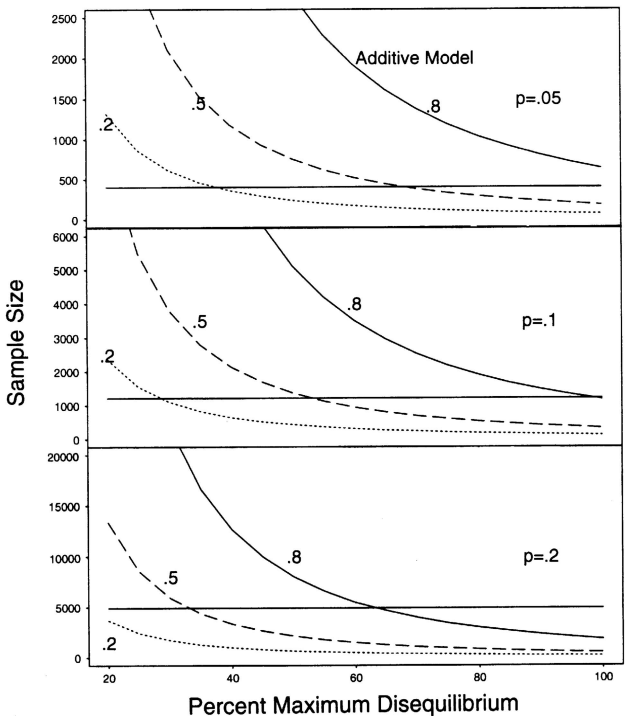

Figure 1 Sample sizes needed, for 80% power, by the affected-sib-pair linkage test (*straight line*) and the association test (*solid, dashed, and dotted curved lines*). An additive genetic model is assumed. The top, middle, and bottom panels give sample sizes needed for disease-allele frequencies of 5%, 10%, and 20%, respectively; in each case, the penetrances are determined so that the disease gene accounts for one-third of all disease occurrences. In each panel, the sample sizes for the association test are plotted against the percentage of maximum disequilibrium between the disease allele and the marker allele in positive disequilibrium with it, when the disease allele has a frequency of 80% (*solid curved line*), 50% (*dashed curved line*), and 20% (*dotted curved line*).

For figures 1–4, the sample sizes for both tests were obtained by use of equation (2), with μ and σ given by equation (B4) for the linkage test and by equations (5) and (6) for the association test. We assumed Hardy-Weinberg proportions for parental genotypes at both the disease locus and its nearest marker. Under this assumption, μ and σ, as well as N for the linkage test, depend on the frequencies and penetrances of the two alleles at the disease locus. For the association test, μ and σ also depend on the frequencies of alleles at the marker locus and the extent of disequilibrium between the two loci. In figures 1–4, this is expressed as a percentage of the maximum possible disequilibrium. Risch

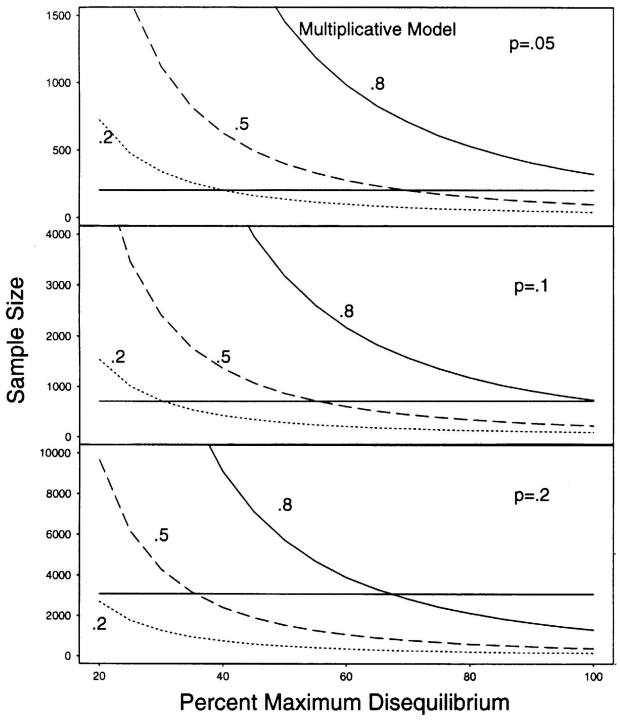

Figure 2 Sample sizes based on a multiplicative model. See legend to figure 1 for details.

and Merikangas (1996) and Camp (1997) calculated sample sizes by assuming that alleles *A* and *B* are identical, which implies equal frequencies for the two alleles and maximum disequilibrium between them. Under these circumstances, the sample sizes shown are roughly comparable with those calculated by those authors, despite some approximations and errors in their variance equations (discussed in Appendix C).

Figures 1–4 show that less-than-maximum disequilibrium and unequal frequencies between alleles *A* and *B* can substantially increase the sample sizes needed for the association test. As expected from the work of Risch and Merikangas (1996), the situation most favorable for the association test occurs when the disease allele is common (figs. 1–4, *bottom panels*) and when the marker allele in positive disequilibrium with it has roughly the same frequency (*dotted curved line*). However, when the marker allele is more common than the disease allele (*dashed and solid curved lines*), the two must be in fairly tight disequilibrium for the association test to beat the linkage test. Moreover, the balance of power tips increasingly toward the linkage test as the frequency of the disease allele drops (figs. 1–4, *top and middle panels*). Even in the case of maximum disequilibrium between the two loci, discrepancies in the frequencies of trait and marker alleles reduce power, with the magnitude of the reduction increasing as the amount of discrepancy increases.

To gain perspective on the power reduction for the association test that is due to less-than-maximum disequilibrium and to discrepancies in allele frequencies, it is helpful to express the disequilibrium coefficient in terms of the percentage of all *B*-bearing chromosomes that also carry the disease allele *A*. In the optimal situation for the association test—that is, when alleles *A* and *B* coincide—100% of all *B*-bearing chromosomes carry allele *A*. Table 2 presents results when the frequency of *A* is less than or equal to that of *B*. Table 2 shows that, for rare *A* alleles (overall frequency 1%) and relatively common *B* alleles, as little as 1% of the *B*-bearing chromosomes carry allele *A*. In contrast, for equally common *A* alleles (overall frequency 20%) and maximum disequilibrium, 100% of the *B*-bearing chromosomes carry *A*. However, table 2 also shows that, even when *A* and *B* are both common and have the same frequency, less-than-maximum disequilibrium can reduce the prevalence of allele *A* among the *B*-bearing chromosomes. Since diallelic markers typically are chosen to have a frequency of 10%–50% in the population (Wang et al. 1998), the presence of "false positives" (chromosomes carrying alleles *B* and *a*) and "false negatives" (chromosomes carrying alleles *b* and *A*) can have

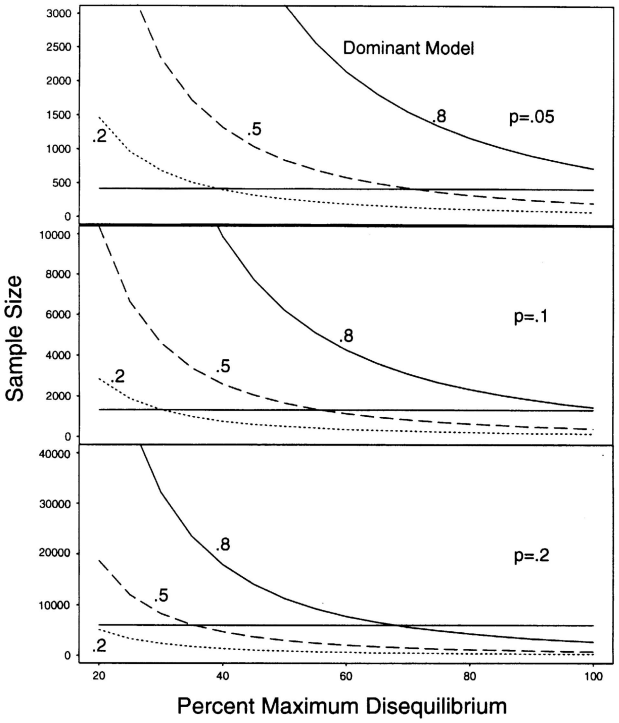

Figure 3 Sample sizes based on a dominant model. See legend to figure 1 for details.

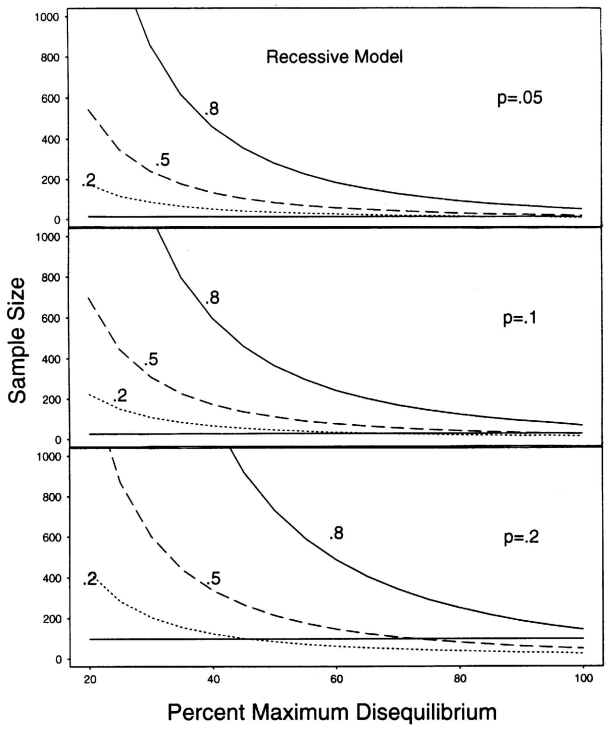

Figure 4 Sample sizes based on a recessive model. See legend to figure 1 for details.

a negative impact on the effectiveness of the association test. The situation is closely related to that of exposure misclassification in studies of disease cases and disease-free controls.

Discussion

When planning a gene-identification study using affected sib pairs and their parents, which analytic strategy should be chosen—linkage or association? The results in figures 1–4 show that the answer depends on the anticipated frequency of the disease-related allele in the population under study and its relation to the allele frequencies of the markers used. The answer also depends strongly on the extent of linkage disequilibrium between the disease locus and its nearest marker. As a general rule, even slight departures from maximum linkage disequilibrium will tilt the balance in favor of the linkage test, when the frequencies of disease and marker alleles are highly discrepant—for example, if the disease allele is rare (<5% frequency in the population under study) and the positively associated allele at the marker locus is common. If the marker allele has a frequency as high as 80%, the linkage test is almost always more powerful.

Recently, Abel and Muller-Myhsok (1998) calculated the number of single-affected-offspring families needed, by the TDT, for 80% power under a multiplicative model and under various assumptions about the disparity of allele frequencies and the extent of disequilibrium at marker and disease loci. The sample-size increases required as allele-frequency disparity increases and as disequilibrium decreases agree well with those found here for the multiplicative model.

In their response to the remarks of Muller-Myhsok and Abel (1997), Risch and Merikangas (1997) noted strong disequilibrium and comparable allele frequencies among individuals in U.S. white populations, at least in the regions surrounding some disease loci (e.g., the apolipoprotein E region involved in Alzheimer disease and the VNTR region on chromosome 11p). In such situations the association test should do well, relative to the linkage test; however, there is need to determine whether these examples hold more generally. Recent data on haplotypes at 88 loci within a 9.7-kb region of the lipoprotein lipase gene in three populations (African Americans, non-Hispanic U.S. whites, and Finns) suggest that the strength of disequilibrium varies considerably from one pair of loci to another (fig. 5 in the report by Clark et al. [1998]).

If the objective is to evaluate a particular marker polymorphism of a candidate gene, it would be helpful to estimate the marker-allele frequencies in the parental population, before the study is started. These estimates then could be combined with a range of plausible frequencies for the (unobserved) actual disease-related polymorphism and with a range of plausible estimates for the disequilibrium coefficient between the two alleles, to estimate the sample sizes needed. Both the allele frequencies of the marker and disease polymorphisms and the strength of disequilibrium can vary by ethnicity and country of origin (Ingles et al. 1997). The results presented here show that these factors can have a substantial effect on the power of the association test. Therefore,

Table 2

Percentage of All *B*-Bearing Chromosomes That Carry a Disease Allele

Disease-Allele Frequency (%)	Marker-Allele Frequency (%)	Disequilibrium Coefficient, by Percentage of Maximum		
		33	67	100
1	20	2.3	3.7	5.0
	50	1.3	1.7	2.0
	80	1.1	1.2	1.3
10	20	23.3	36.7	50.0
	50	13.3	16.7	20.0
	80	10.8	11.7	12.5
20	20	46.7	73.3	100.0
	50	26.7	33.3	40.0
	80	21.7	23.3	25.0

such variation may account for the observed lack of consistency in findings of association, across different national and ethnic groups.

Further work is needed so that the sample sizes calculated here can be generalized to include (1) missing parental genotypes and (2) more than two alleles at the marker loci. Missing parental genotypes for some of the affected sib pairs will increase the sample sizes needed by both linkage and association tests. The relative power loss for one test, compared with that for the other, is an area in need of research (also see Risch and Teng, 1998). Similarly, the effects of multiple marker alleles on both tests need to be examined.

Acknowledgment

This research was supported by National Institutes of Health grant R35 CA47448.

Appendix A

Threshold and Power

Threshold

For linkage, we obtained a threshold value c, using the Gaussian approximation proposed by Feingold et al. (1993), with the intermarker distances equal to 0. According to this approximation, the threshold value c satisfies the equation

$$.05 = 1 - \Phi(c) + \beta L \phi(c) \ . \qquad \text{(A1)}$$

Here, ϕ is the standard Gaussian density, Φ is the standard Gaussian cumulative distribution function, the rate β is .04/cm, and the total chromosome length L is 3,500 cm. Solving equation (A1) for c gives $c = 4.12$.

For association, we assume statistical independence of the TDT statistics S_t at different marker loci t. We show in Appendix C that, under the null hypothesis, the asymptotic marginal distribution of each S_t is the nonnegative part of a standard Gaussian distribution. Thus, we use the arguments of Risch and Merikangas (1996) to choose the threshold value $c = 5.33$.

Power

To determine the probability that the statistic T exceeds c—that is, that S_t exceeds c at some marker locus t on the chromosome containing the disease locus—we approximate each of the two processes S_t by using a Gaussian process (see Feingold et al. [1993]). We also write

$$\begin{aligned} P(T \geq c) &= P(\max S_t \geq c) \\ &= P(S_\tau \geq c) + P(S_\tau < c, \max > c) \ . \end{aligned} \qquad \text{(A2)}$$

Here, τ is the locus of the marker closest to the disease locus. Since the noncentrality of S_t is maximized at $t = \tau$ and decays approximately exponentially as t moves away from τ, it is plausible that the first term on the right side of equation (A2) contributes most to the probability that T exceeds c. Accordingly, the approximation is

$$P(T \geq c) \approx P(S_\tau \geq c) \approx 1 - \Phi\left(\frac{c - \sqrt{N}\mu}{\sigma}\right) ,$$

where $\sqrt{N}\mu$ and σ^2 are the mean and variance, respectively, of S_τ.

Appendix B

Allele-Sharing Probabilities for Affected Sibs

To determine the affected sibs' allele-sharing probability π_i at the trait locus, for $i = 0, 1, 2$, we use the usual variance-component decomposition (Crow and Kimura 1970):

$$\begin{aligned} \pi_0 &= \tfrac{1}{4}(1 - \lambda) \ , \\ \pi_1 &= \tfrac{1}{2}(1 - \xi) \ , \\ \pi_2 &= \tfrac{1}{4}(1 + \lambda + 2\xi) \ . \end{aligned} \qquad \text{(B1)}$$

In these equations,

$$\begin{aligned} \lambda &= \frac{\tfrac{1}{2}V_A + \tfrac{1}{4}V_D}{K^2 + \tfrac{1}{2}V_A + \tfrac{1}{4}V_D} \ , \\ \xi &= \frac{\tfrac{1}{4}V_D}{K^2 + \tfrac{1}{2}V_A + \tfrac{1}{4}V_D} \ , \end{aligned} \qquad \text{(B2)}$$

where K is the population prevalence of the disease and V_A and V_D are the additive and dominance variance components, respectively. The latter are defined by

$$\begin{aligned} V_A &= 2pq[p(f_2 - f_1) + q(f_1 - f_0)]^2 \ , \\ V_D &= p^2q^2(f_2 - 2f_1 + f_0)^2 \ , \end{aligned} \qquad \text{(B3)}$$

where f_i is the penetrance for individuals with i copies of the disease-susceptibility allele A, for $i = 0, 1, 2$, and

where $p = 1 - q$ is the frequency of allele A. Substitution of equation (B1) into equations (3) and (4) gives

$$\mu = \frac{\lambda + \xi}{\sqrt{2}} \ ,$$

$$\sigma^2 = 1 + \xi - \tfrac{1}{2}(\lambda + \xi)^2 \ . \qquad \text{(B4)}$$

Substitution of equation (B3) into equation (B2) and of equation (B2) into equation (B4) gives the quantities μ and σ^2 in terms of the frequencies and penetrances of the two alleles at the disease locus.

Appendix C

Parental Heterozygosity and Transition Probabilities

Here, we describe the parental heterozygosity probability h_i and the transmission probability τ_{ij} at the marker closest to the disease-causing polymorphism. Specifically, we express these probabilities in terms of the frequencies and penetrances of the disease-susceptibility alleles, the marker-allele frequencies, and the extent of disequilibrium between alleles at the two loci. Our equations differ from those of Risch and Merikangas (1996) and Camp (1997) in two respects. First, unlike these authors, we did not assume conditional independence of parental genotypes at the disease locus, given the sibs' affected status. Although such independence holds for the multiplicative model assumed by Risch and Merikangas (1996), it gives only an approximate variance for other models, such as the additive, recessive, and dominant models considered by Camp (1997). Second, we did not assume, as did these previous authors, that the contributions of the two sibs to the statistic x_{kt} in table 1 are independent; this assumption holds under the null hypothesis but not under the alternative hypothesis.

Parental Heterozygosity Probabilities

H_j denotes the event that a family contains j parents who are heterozygous at the diallelic marker nearest the disease-causing polymorphism, for $j = 0, 1, 2$, and ASP denotes the event that the two offspring are affected. Also, G_A denotes the number of copies of the disease allele A carried by a parent, and $v_{Ai} = P(G_A = i)$, for $i = 0, 1, 2$. We use a similar notation—that is, G_B and v_{Bi}—for the number of copies of allele B at the marker. By Bayes rule,

$$h_2 \equiv P(\mathrm{H}_2|\mathrm{ASP}) = \frac{v_{B1}^2 c_{11}}{P(\mathrm{ASP})} \ , \qquad \text{(C1)}$$

where c_{ij} is the probability that both offspring are affected, given that their parents have B genotypes i and j, for $i, j = 0, 1, 2$. Similarly,

$$h_1 \equiv P(\mathrm{H}_1|\mathrm{ASP}) = \frac{2v_{B1}(v_{B0}c_{10} + v_{B2}c_{12})}{P(\mathrm{ASP})} \ , \qquad \text{(C2)}$$

and

$$h_0 \equiv P(\mathrm{H}_0|\mathrm{ASP}) = 1 - h_1 - h_2 \ .$$

We assume that parental mating is random with respect to the parents' genotypes at both the disease and marker loci. We also assume that the two offsprings' phenotypes are conditionally independent, given their genotypes at the disease locus. Then, the probability c_{ij} can be written as

$$c_{ij} = P(\mathrm{ASP}|G_B = i, G_B = j) = \sum_{k=0}^{2}\sum_{l=0}^{2} \alpha_{lk}^2 r_{ki} r_{lj} \ ,$$

where

$$\alpha_{lk} = \alpha_{kl}$$
$$= P(\text{one offspring is affected}|G_A = k, G_A = l)$$

and $r_{ij} = P(G_A = i|G_B = j)$. Values for α_{lk} are given as follows: $\alpha_{02} = f_1$ and $\alpha_{11} = \frac{1}{4}(f_0 + 2f_1 + f_2)$; otherwise, $\alpha_{lk} = \frac{1}{2}(f_l + f_k)$. Values for r_{ij} are given as follows:

$$r_{i2} = \binom{2}{i} p_B^i q_B^{2-i} \ , \ i = 0, 1, 2 \ ;$$

$r_{21} = p_B p_b$, $r_{11} = p_B q_b + p_b q_B$, and $r_{01} = q_B q_b$; and

$$r_{i0} = \binom{2}{i} p_b^i q_b^{2-i} \ , \ i = 0, 1, 2 \ .$$

Here $p_B = 1 - q_B = P(A|B) = p + \delta/P$ and $p_b = 1 - q_b = P(A|b) = p - \delta/Q$. Also, $p = 1 - q$ and $P = 1 - Q$ are the frequencies of alleles A and B, respectively, and $\delta = P(AB) - pP$ is the disequilibrium coefficient between them. The probability that both offspring are affected is

$$P(\mathrm{ASP}) = \sum_{i=0}^{2}\sum_{j=0}^{2} \alpha_{ij}^2 v_{Ai} v_{Aj} \ .$$

Transmission Probabilities

To describe the parent-offspring transmission probability τ_{ij}, we use T_i to denote the event that the heterozygous parents in a family transmit i copies of allele B

to the two offspring, for $i = 0, 1, ..., 4$. With this notation, we may write

$$\tau_{ij} = P(\mathrm{T}_i|\mathrm{ASP}, \mathrm{H}_j) = \frac{P(\mathrm{H}_j, \mathrm{T}_i)P(\mathrm{ASP}|\mathrm{H}_j, \mathrm{T}_i)}{P(\mathrm{ASP})h_j} .$$

Here,

$$P(\mathrm{H}_1, \mathrm{T}_i) = \binom{2}{i}(\tfrac{1}{2})^2 2v_{B1}(1 - v_{B1}),\ i = 0, 1, 2 ,$$

$$P(\mathrm{H}_2, \mathrm{T}_i) = \binom{4}{i}(\tfrac{1}{2})^4 v_{B1}^2,\ i = 0, 1, ..., 4 . \tag{C3}$$

To complete the description of τ_{ij}, we must compute $P(\mathrm{ASP}|\mathrm{H}_1, \mathrm{T}_i)$, for $i = 0, 1, 2$, and $P(\mathrm{ASP}|\mathrm{H}_2, \mathrm{T}_i)$, for $i = 0, 1, ..., 4$. To do so, we use $\mathrm{P} \overset{A}{\rightarrow} k$ to denote the event that a parent transmits a total of k copies of allele A to the two offspring, with a similar definition for $\mathrm{P} \overset{B}{\rightarrow} k$. With this notation,

$$P(\mathrm{ASP}|\mathrm{H}_j, \mathrm{T}_i) = \sum_{k=0}^{2}\sum_{l=0}^{2} \Delta_{kl}\xi_{klij} ,$$

where $\Delta_{kl} = \Delta_{lk} = P(\mathrm{ASP}|\mathrm{M} \overset{A}{\rightarrow} k, \mathrm{F} \overset{A}{\rightarrow} l)$ and M and F denote the mother and father, respectively. Also, $\xi_{klij} = P(\mathrm{M} \overset{A}{\rightarrow} k, \mathrm{F} \overset{A}{\rightarrow} l|\mathrm{H}_j, \mathrm{T}_i)$. Values for Δ_{kl} are given as follows: $\Delta_{22} = f_2^2$, $\Delta_{21} = f_2 f_1$, and $\Delta_{20} = f_1^2$; and $\Delta_{11} = \frac{1}{2}(f_2 f_0 + f_1^2)$, $\Delta_{10} = f_1 f_0$, and $\Delta_{00} = f_0^2$. Also, $\xi_{kli1} = \frac{1}{2}(\rho_{ki}\psi_l + \rho_{li}\psi_k)$, for $i = 0, 1, 2$, where $\rho_{rs} = P(\mathrm{P} \overset{A}{\rightarrow} r|G_B = 1, \mathrm{P} \overset{B}{\rightarrow} s)$ and $\psi_r = P(\mathrm{P} \overset{A}{\rightarrow} r|G_B \neq 1)$. Values for ρ_{rs} are shown in the following table:

	s		
r	0	1	2
0	q_b	$q_B q_b$	q_B
1	0	$p_B q_b + p_b q_B$	0
2	p_b	$p_B p_b$	p_B

In addition,

$$\psi_0 = w_0(q_b^2 + \tfrac{1}{2}p_b q_b) + w_2(q_B^2 + \tfrac{1}{2}p_B q_B) ,$$

$$\psi_1 = w_0 p_b q_b + w_2 p_B q_B ,$$

$$\psi_2 = w_0(p_b^2 + \tfrac{1}{2}p_b q_b) + w_2(p_B^2 + \tfrac{1}{2}p_B q_B) ,$$

where $w_i = v_{Bi}/(v_{B0} + v_{B2})$, for $i = 0, 2$. Finally, values for ξ_{kli2}, for $i = 0, 1, ..., 4$, are shown below:

i	ξ_{kli2}
0	$\rho_{k0}\rho_{l0}$
1	$(\rho_{k0}\rho_{l1} + \rho_{k1}\rho_{l0})/2$
2	$(4\rho_{k1}\rho_{l1} + \rho_{k0}\rho_{l2} + \rho_{k2}\rho_{l0})/6$
3	$(\rho_{k2}\rho_{l1} + \rho_{k1}\rho_{l2})/2$
4	$\rho_{k2}\rho_{l2}$

We assume Hardy-Weinberg proportions for the parental marker genotypes. Under this assumption, $v_{A2} = p^2$, $v_{A1} = 2pq$, and $v_{A0} = q^2$, where $p = 1 - q$ is the frequency of allele A. Similarly, $v_{B2} = P^2$, $v_{B1} = 2PQ$, and $v_{B0} = Q^2$, where $P = 1 - Q$ is the frequency of allele B. Therefore, the heterozygosity probability h_j and the transition probability τ_{ij} depend only on the frequencies p and P of alleles A and B, respectively; their disequilibrium coefficient δ for the parental population; and the penetrances f_0, f_1, and f_2 at the disease locus.

We conclude by noting that, under the null hypothesis of no disease gene on the chromosome, the transmission probability τ_{ij} at any marker locus t reduces to $\tau_{ij} = P(\mathrm{H}_j, \mathrm{T}_i)/h_j$. By substituting equations (C1–C3) into this equation and using the results in equations (5) and (6), we obtain for S_t a null asymptotic mean of 0 and an asymptotic variance of 1. By invoking the central limit theorem, we see that, for a large N, the marginal distribution of each statistic S_t is asymptotically the nonnegative part of a standard Gaussian.

References

Abel L, Muller-Myhsok B (1998) Maximum likelihood expression of the transmission/disequilibrium test and power considerations. Am J Hum Genet 63:664–667

Camp NJ (1997) Genomewide transmission/disequilibrium testing: consideration of the genotypic relative risks at disease loci. Am J Hum Genet 61:1424–1430

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, et al (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63: 595–612

Crow JF, Kimura M (1970) An introduction to population genetics theory. Burgess, Minneapolis

Feingold E, Brown PO, Siegmund D (1993) Gaussian models for genetic linkage analysis using complete high resolution maps of identity by descent. Am J Hum Genet 53:234–251

Ingles SA, Haile RW, Henderson BE, Kolonel LN, Nakaichi G, Shi C-Y, Yu MC, et al (1997) Strength of linkage disequilibrium between two vitamin D receptor markers in five ethnic groups: implications for association studies. Cancer Epidemiol Biomarkers Prev 6:93–98

Muller-Myhsok B, Abel L (1997) Genetic analysis of complex disease. Science 275:1328–1329

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

——— (1997) Response to "Genetic analysis of complex diseases" letters. Science 275:1329–1330

Risch N, Teng J (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling. Genome Res 8:1273–1288

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitius (IDDM). Am J Hum Genet 52:506–516

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

Terwilliger JD, Ott J (1992) A haplotype-based "haplotype relative risk" approach to detecting allelic associations. Hum Hered 42:337–346

Wang DG, Fan J-B, Siao C-J, Berno A, Young P, Sapolsky R, Ghandour G, et al (1998) Large scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1081