

Linkage Detection Adaptive to Linkage Disequilibrium: The Disequilibrium Maximum-Likelihood–Binomial Test for Affected-Sibship Data

Jian Huang and Yanming Jiang

Department of Statistics and Actuarial Science, University of Iowa, Iowa City

Summary

It has been demonstrated in the literature that the transmission/disequilibrium test (TDT) has higher power than the affected-sib-pair (ASP) mean test when linkage disequilibrium (LD) is strong but that the mean test has higher power when LD is weak. Thus, for ASP data, it seems clear that the TDT should be used when LD is strong but that the mean test or other linkage tests should be used when LD is weak or absent. However, in practice, it may be difficult to follow such a guideline, because the extent of LD is often unknown. Even with a highly dense genetic-marker map, in which some markers should be located near the disease-predisposing mutation, strong LD is not inevitable. Besides the genetic distance, LD is also affected by many factors, such as the allelic heterogeneity at the disease locus, the initial LD, the allelic frequencies at both disease locus and marker locus, and the age of the mutation. Therefore, it is of interest to develop methods that are adaptive to the extent of LD. In this report, we propose a disequilibrium maximum-binomial-likelihood (DMLB) test that incorporates LD in the maximum-binomial-likelihood (MLB) test. Examination of the corresponding score statistics shows that this method adaptively combines two sources of information: (a) the identity-by-descent (IBD) sharing score, which is informative for linkage regardless of the existence of LD, and (b) the contrast between allele-specific IBD sharing score, which is informative for linkage only in the presence of LD. For ASP data, the proposed test has higher power than either the TDT or the mean test when the extent of LD ranges from moderate to strong. Only when LD is very weak or absent is the DMLB slightly less powerful than the mean test; in such cases, the TDT has essentially no power to detect linkage. Therefore, the DMLB test is an interesting approach to linkage detection when the extent of LD is unknown.

Received April 28, 1999; accepted for publication September 27, 1999; electronically published November 23, 1999.

Address for correspondence and reprints: Dr. Jian Huang, Department of Statistics and Actuarial Science, 241 Schaeffer Hall, University of Iowa, Iowa City, IA 52242. E-mail: jian-huang@uiowa.edu

© 1999 by The American Society of Human Genetics. All rights reserved. 0002-9297/1999/6506-0031\$02.00

Introduction

Recently, Abel et al. considered the maximum-likelihood-binomial (MLB) method for linkage analysis using affected-sibship data (Abel et al. 1998; Abel and Müller-Myhsok 1998). This method, based on the binomial distribution of parental alleles among affected offspring (Badner et al. 1984; Majumder and Pal 1987), can be applied to multiplex sibships in a natural way. For the models used in simulation studies by Abel et al., this test has correct type 1–error rate and has statistical power similar to that of the mean test and maximum-likelihood–score (MLS) test (Blackwelder and Elston 1985; Risch 1990).

The MLB and other linkage-analysis methods are based on the identity-by-descent (IBD)–sharing configuration, in which the allelic state is not taken into account. According to the formulation of the linkage likelihood by Whittemore (1996), given the affection status of the individuals and the structure of a pedigree and under the assumption of linkage equilibrium, the likelihood depends only on the IBD-sharing configuration. Thus, under linkage equilibrium, the IBD-sharing configuration is a sufficient statistic for the genetic parameters and contains all the information for linkage. However, this is no longer the case in the presence of linkage disequilibrium (LD), as has been illustrated by Clerget-Darpoux (1982) in a LOD-score analysis; she showed that the expected value of the LOD score is higher when LD is accounted for than when it is not.

The transmission/disequilibrium test (TDT) is designed to detect linkage in the presence of LD (Spielman et al. 1993). It is based on the idea that, if there is linkage, parents will preferentially transmit marker alleles in LD with disease to their affected offspring (Rubinstein et al. 1981; Falk and Rubinstein 1987; Ott 1989). The TDT can use families with a single affected offspring. For families with two or more affected offspring, the TDT can be thought of as using the contrast between *allele-specific* IBD-sharing scores—it examines both excessive sharing of alleles in positive LD with the disease and reduced sharing of alleles in negative LD with the disease.

If a disease-causing mutation originated in one or a few founders in a population, the alleles closely linked

to the mutation may be transmitted together with the mutation, for many generations. With the rapid progress of molecular technology, it is increasingly feasible to use a dense set of genetic markers, such as single-nucleotide polymorphisms, saturating the genome in a genomewide screen (Wang et al. 1998). Some of the markers should be located close to the disease-causing mutation and, hence, could be in LD with the mutation. It is desirable to exploit such potential LD in gene-mapping studies (Risch and Merikangas 1996).

Several studies have used affected-sib-pair (ASP) data to compare the mean test's power to detect linkage and the TDT's power to detect linkage (Risch and Merikangas 1996; Camp 1997; Xiong and Guo 1998; Tu and Whittemore 1999). In brief, the conclusion is that the TDT has higher power when LD is complete or near complete but that the mean test has higher power when LD is weak. Thus it seems clear that the TDT should be used when LD is strong but that the mean test or some other linkage test should be used when LD is weak or absent. However, it may be difficult to follow such a guideline in practice, because the extent of LD is usually unknown. Indeed, even with a highly dense genetic-marker map, in which some markers should be located near the disease-predisposing mutation, strong LD is not inevitable. Besides the genetic distance, LD is also a function of many additional factors, such as the initial LD, the allelic frequencies at both disease locus and marker locus, and the age of the mutation. (Devlin and Risch 1995; Guo 1997). How large an effect these factors have on LD is difficult to assess.

Allelic heterogeneity within a disease gene also has a strong impact on LD (Terwilliger and Weiss 1998). Many disease-predisposing genes have multiple alleles predisposing to disease, such as genes predisposing Alzheimer disease (Tysoe et al. 1998), breast cancer (Szabo and King 1997), and cystic fibrosis (Welsh and Smith 1995). This appears to be a common phenomenon in many diseases. Terwilliger and Weiss (1998) have listed a selection of 64 disease-predisposing loci with multiple alleles that have been published in *The American Journal of Human Genetics's* volumes 60–62 (1997–98). Previous work has shown that allelic heterogeneity decreases the degree of LD and can drastically reduce the power of the TDT (Terwilliger and Weiss 1998; Slager et al. 1999).

Therefore, it is of interest to develop methods that are adaptive to the extent of LD. These methods should make full use of LD if it is indeed present and should have power comparable to that of standard linkage methods, such as the mean test, when LD is weak or absent. To achieve this, they must efficiently combine the following two sources of information: (1) the standard IBD-sharing configuration, which is informative for linkage regardless of the existence of LD, and (2) the

allele-specific IBD-sharing configuration for marker alleles in positive or negative LD with the disease, which is informative for linkage only in the presence of LD.

In the present report, we propose a disequilibrium maximum-likelihood–binomial (DMLB) test for linkage. This method extends the MLB test, by incorporating LD, and appears to offer an approach to efficiently combine linkage information from the IBD-sharing and the allele-specific IBD-sharing scores. As does the original MLB method, the DMLB method uses families with multiple affected sibships, in a natural way. Because the DMLB method uses IBD-sharing information, its power to detect linkage is comparable to that of the mean test, for families with ASPs or multiplex-sibship data, when LD is weak or absent.

In the following, we first formulate the DMLB test and present its asymptotic distributions under the null hypothesis of no linkage. We then show, by examining its score statistics, that the DMLB test can be considered as an adaptive combination of the mean test and the TDT. We also demonstrate that, for a wide range of LD, the DMLB test has higher statistical power than do the TDT and the mean test, for ASP data.

The DMLB Test for Linkage

Consider a nuclear family with m affected children. Let m_1 be the number of affected children who have received marker allele B_1 from a heterozygous B_1B_2 parent. If the marker is not linked to the disease, then m_1 has a binomial distribution with parameters m and $.5$ (Majumder and Pal 1987; Abel et al. 1998). Thus, a test for linkage can be constructed by assessment of the departure, from $.5$, of the probability that allele B_1 of the heterozygous parent will be transmitted (Abel et al. 1998; Abel and Müller-Myshok 1998). For this heterozygous parent and the affected sibs, with an unknown phase and assuming linkage equilibrium, Abel et al. proposed the following likelihood as the basis for testing of linkage:

$$.5\alpha^{m_1}(1 - \alpha)^{m-m_1} + .5(1 - \alpha)^{m_1}\alpha^{m-m_1}, \quad (1)$$

where $\alpha = .5$ if there is no linkage and $\alpha > .5$ if there is linkage. Heuristically, α may be interpreted as the probability that an affected child has received the marker allele transmitted with the disease allele. The overall likelihood is simply the product of all the likelihoods over the heterozygous parents and sibships. Likelihood (1) can also be motivated by use of a recessive model with a phase-unknown double-backcross mating type. However, no such model specification is required for the likelihood ratio (LR) test to be valid, since it is correct under the null hypothesis of no linkage, and hence the test has the correct type 1–error rate for a given critical value.

A reviewer has pointed out that the binomial likelihood (1) is actually a composite likelihood and not a true likelihood, because, under the alternative hypothesis of linkage, the transmission of maternal alleles may not be independent of the transmission of paternal alleles. This comment also applies to likelihood (2), below.

Likelihood (1) is a mixture of two binomial distributions that are mirror images of each other. The mixing proportion .5 is based on the assumption of linkage equilibrium. For any fixed m and m_1 , this assumption results in a likelihood that is *least* informative for α , because there is maximum uncertainty about which component of the mixture distribution generates the data. However, in the presence of LD, the mixing proportion is no longer .5. This can be incorporated into the likelihood by introduction of a parameter, λ , representing the mixing proportion. Specifically, for the j th heterozygous parent and k th sibship, let m_{jk} be the number of affected children, and let m_{jk1} be number of affected children who have received the marker allele B_1 from the j th heterozygous parent. Then the form of the likelihood that generalizes (1) and that incorporates LD is

$$L_{jk}(\alpha, \lambda) = \lambda \alpha^{m_{jk1}} (1 - \alpha)^{m_{jk} - m_{jk1}} + (1 - \lambda)(1 - \alpha)^{m_{jk1}} \alpha^{m_{jk} - m_{jk1}} . \tag{2}$$

The overall likelihood is

$$L(\alpha, \lambda) = \prod_{jk} L_{jk}(\alpha, \lambda) . \tag{3}$$

The range of λ depends on our knowledge of the extent of LD. There are two situations: (i) if there is prior knowledge that allele B_1 is in positive LD with disease, such as in the case when previous population studies have indicated that the allele is associated with the disease of interest, then the range of λ is $.5 \leq \lambda \leq 1$; (ii) if there is no prior knowledge about which marker allele is in LD with disease, such as when one is conducting a genomewide screen, then the range of λ is $0 \leq \lambda \leq 1$. Accordingly, the hypotheses and the LR-test statistics are formulated as follows.

In the first case, the hypotheses are

$$H_0 : \alpha = .5, .5 \leq \lambda \leq 1 , \\ H_A : \alpha > .5, .5 \leq \lambda \leq 1 . \tag{4}$$

Let $(\hat{\alpha}_1, \hat{\lambda}_1)$ be the maximum-likelihood estimator of (α, λ) under the restriction that $.5 \leq \alpha \leq 1$ and $.5 \leq \lambda \leq 1$. The LR-test statistic is

$$\Lambda_1 = \frac{L(\hat{\alpha}_1, \hat{\lambda}_1)}{L(.5)} , \tag{5}$$

where $L(.5) = .5^M$, with $M = \sum_{jk} m_{jk}$.

In the second case, the hypotheses are

$$H_0 : \alpha = .5, 0 \leq \lambda \leq 1 , \\ H_A : \alpha > .5, 0 \leq \lambda \leq 1 . \tag{6}$$

Let $(\hat{\alpha}_2, \hat{\lambda}_2)$ be the maximum-likelihood estimator of (α, λ) under the restriction that $.5 \leq \alpha \leq 1$ and $0 \leq \lambda \leq 1$. The LR-test statistic is

$$\Lambda_2 = \frac{L(\hat{\alpha}_2, \hat{\lambda}_2)}{L(.5)} , \tag{7}$$

where $L(.5)$ is the same as above.

We note that the range of α is restricted to $\alpha \geq .5$. This is because the likelihood is symmetrical in the sense that $L(\alpha, \lambda) = L(1 - \alpha, 1 - \lambda)$, as has been pointed out by a reviewer. Consider the four quadrants of a unit square: $A = \{(\alpha, \lambda) : \alpha \in [.5, 1], \lambda \in [0, .5]\}$, $B = \{(\alpha, \lambda) : \alpha \in [.5, 1], \lambda \in [.5, 1]\}$, $C = \{(\alpha, \lambda) : \alpha \in [0, .5], \lambda \in [0, .5]\}$, and $D = \{(\alpha, \lambda) : \alpha \in [0, .5], \lambda \in [.5, 1]\}$. Because of symmetry, the likelihood is the same for the regions A and D and for the regions B and C . Thus, the likelihood need be evaluated only for the regions A and B , which are for $\alpha \geq .5$.

Under the null hypothesis of no linkage, the value of parameter λ is irrelevant, because the likelihood is a constant as long as $\alpha = .5$. Thus, the regularity conditions required for the χ^2 distributional results for an LR test are not satisfied, and the LR statistics do not have an asymptotic χ^2 distribution.

If the data set consists of families with a single affected child, the statistic $2\log\Lambda_1$ is asymptotically distributed as a 50:50 mixture of χ_0^2 and χ_1^2 distributions, where χ_0^2 denotes the degenerate distribution that puts probability 1.0 at 0 and where $2\log\Lambda_2$ is asymptotically distributed as a χ_1^2 distribution. For such data, it is shown, in the next section, that the LR tests Λ_1 and Λ_2 are equivalent to the one-sided and two-sided versions of the TDT, respectively. If most of the families in the data set have two or more affected children, then $2\log\Lambda_1$ is asymptotically distributed as $.25\chi_0^2 + .5\chi_1^2 + .25\chi_2^2$ —that is, a $.25:.5:.25$ mixture of χ_0^2 , χ_1^2 , and χ_2^2 —and $2\log\Lambda_2$ is asymptotically distributed as a 50:50 mixture of χ_1^2 and χ_2^2 . Formal statements for the distributions of Λ_1 and Λ_2 , as well as the proof, are given in Appendix A.

Note that, when parameter λ is introduced, the df increase not by 1, as in many standard problems, but only by approximately a quarter, for testing hypotheses (4), or approximately half, for hypotheses (6).

DMLB as an Adaptive Combination of the TDT and the Mean Test

In this section, we examine the relation between the TDT, the mean test, and the proposed DMLB test by considering the latter's score statistics. We first show that, for data consisting of simplex families, the DMLB is asymptotically equivalent to the TDT. For multiplex families, the DMLB test can be viewed as an efficient combination of the TDT and the mean test. It adaptively combines two aspects of the linkage information from the data: the IBD-sharing scores used in the mean test and the allele-specific IBD-sharing scores used in the TDT.

Families with a Single Affected Child

First, consider data consisting of families with a single affected child. The mean test cannot be applied to such data; however, both the TDT and the DMLB test can be used.

Let n_1 be the total number of heterozygous parents, and let n_{11} be total number of marker alleles B_1 transmitted from the heterozygous B_1B_2 parents. Likelihood (3) becomes

$$L(\alpha, \lambda) = [\lambda\alpha + (1 - \lambda)(1 - \alpha)]^{n_1} \times [(1 - \lambda)\alpha + \lambda(1 - \alpha)]^{n_1 - n_{11}} .$$

In this case, α and λ cannot be estimated simultaneously, because these two parameters are confounded. However, it is still possible to test the hypothesis $\alpha = .5$, provided that $\lambda \neq .5$. Let $\beta_2 = 2[\lambda\alpha + (1 - \lambda)(1 - \alpha) - .5] = 4(.5 - \lambda)(.5 - \alpha)$. Then $(1 - \lambda)\alpha + \lambda(1 - \alpha) = .5(1 - \beta_2)$. Likelihood (3) can be rewritten, up to a multiplicative constant, as

$$L(\beta_2) = (1 + \beta_2)^{n_{11}}(1 - \beta_2)^{n_1 - n_{11}} , \tag{8}$$

because, for any $0 \leq \lambda \leq 1$ and $\lambda \neq .5$, $\alpha = .5$ if and only if $\beta_2 = 0$. Therefore, testing $\alpha = .5$ is equivalent to testing $\beta_2 = 0$ —provided that $\lambda \neq .5$; that is, provided that LD exists.

The TDT can be derived as a score test corresponding to the DMLB test. The score statistic of likelihood (9) evaluated at $\beta_2 = 0$ is $2n_{11} - n_1$, and the Fisher information at $\beta_2 = 0$ is n_1 . Thus the score-test statistic is

$$\frac{(2n_{11} - n_1)^2}{n_1} = \frac{(n_{11} - n_{10})^2}{n_1} ,$$

where $n_{10} = n_1 - n_{11}$ is the number of B_2 alleles transmitted from the heterozygous parents. This is exactly the TDT statistic given by Spielman et al. (1993). According

to standard theory, the DMLB test, which is the LR test based on $L(\beta_2)$ in likelihood (9), is asymptotically equivalent to the TDT (Cox and Hinkley 1974); however, for samples that include families with two or more affected children, this equivalence no longer holds, as shown below for affected-sib-pair and mixed-sibship data.

Families with Two Affected Children

Suppose that there are n_2 heterozygous B_1B_2 parents in the data set. Let n_{20} be the number of heterozygous parents who transmitted B_2 to both children; let n_{21} be the number of heterozygous parents who transmitted B_1 to one child and B_2 to another child; and let n_{22} be the number of heterozygous parents who transmitted B_1 to both children. The likelihood of the DMLB test can be written as

$$L(\alpha, \lambda) = [\lambda(1 - \alpha)^2 + (1 - \lambda)\alpha^2]^{n_{20}}[\alpha(1 - \alpha)]^{n_{21}} \times [\lambda\alpha^2 + (1 - \lambda)(1 - \alpha)^2]^{n_{22}} .$$

It is difficult to work with this likelihood parametrized in terms of (α, λ) , because the null hypothesis corresponds to a set of points $\{(\alpha, \lambda) : \alpha = .5, 0 \leq \lambda \leq 1\}$. Thus, we reparameterize this likelihood, using $\beta_1 = 4(.5 - \alpha)^2, \beta_2 = 4(.5 - \alpha)(.5 - \lambda)$. Let $\beta = (\beta_1, \beta_2)$. The likelihood becomes

$$L(\beta) = (1 + \beta_1 - 2\beta_2)^{n_{20}}(1 - \beta_1)^{n_{21}}(1 + \beta_1 + 2\beta_2)^{n_{22}} . \tag{9}$$

By its definition and the range of α and λ , β must lie within the region $\mathcal{B} = \{(\beta_1, \beta_2) : 0 \leq \beta_1 \leq 1, 0 \leq \beta_2 \leq 1, \beta_2^2 \leq \beta_1\}$, for hypotheses (4), or must lie within the region $\mathcal{B} = \{(\beta_1, \beta_2) : 0 \leq \beta_1 \leq 1, -1 \leq \beta_2 \leq 1, \beta_2^2 \leq \beta_1\}$, for hypotheses (6). In both cases, the null hypothesis corresponds to a single point $(\beta_1, \beta_2) = (0, 0)$.

Without consideration of the restriction that β belongs to \mathcal{B} , the score-test statistic corresponding to likelihood (11) is

$$\frac{(n_{20} + n_{22} - n_{21})^2}{n_2} + \frac{2(n_{22} - n_{20})^2}{n_2} ,$$

in which the first term is the mean test statistic and the second term is the TDT statistic. Thus, without consideration of the restriction \mathcal{B} , the score test for likelihood (11) is simply the sum of the mean test and the TDT. Although the mean test statistic and the TDT are independent (Spielman et al. 1993), this simple addition of two independent test statistics results in a test with 2 df, and its power may be lower than that of either of the two original tests with 1 df.

Therefore, it is important to take into account the restriction \mathcal{B} . The resulting appropriate score statistic is

$$\frac{(n_{20} + n_{22} - n_{21})^2}{n_2} + \frac{2(n_{22} - n_{20})^2}{n_2} - \inf_{\beta \in \mathcal{B}} \left\{ \left(\frac{n_{20} + n_{22} - n_{21}}{\sqrt{n_2}} - \beta_1 \right)^2 + \left[\frac{\sqrt{2}(n_{22} - n_{20})}{\sqrt{n_2}} - \beta_2 \right]^2 \right\}.$$

This score statistic is asymptotically equivalent to the LR statistic. This form of restricted score statistic is motivated by equation (A1) in Appendix A, which is used in proving the asymptotic distribution of the LR statistics.

The positive homogeneous cone approximating \mathcal{B} at $\beta = 0$ can be used to replace \mathcal{B} in the expression above (Chernoff 1954). This simplifies computation and does not change the asymptotic distribution. For \mathcal{B} associated with hypotheses (4), the approximating cone is the first quadrant, $C = \{(\beta_1, \beta_2) : \beta_1 \geq 0, \beta_2 \geq 0\}$. For \mathcal{B} associated with hypotheses (6), the approximating cone is the half space $C = \{(\beta_1, \beta_2) : \beta_1 \geq 0\}$; that is, equivalently, the score-test statistic can be defined as

$$S = \frac{(n_{20} + n_{22} - n_{21})^2}{n_2} + \frac{2(n_{22} - n_{20})^2}{n_2} - \inf_{\beta \in C} \left\{ \left(\frac{n_{20} + n_{22} - n_{21}}{\sqrt{n_2}} - \beta_1 \right)^2 + \left(\frac{\sqrt{2}(n_{22} - n_{20})}{\sqrt{n_2}} - \beta_2 \right)^2 \right\}.$$

Because C has a simple form, S can be calculated explicitly, as follows. For testing hypotheses (4), the score statistic is

$$S_1 = \begin{cases} 0 & \text{if } n_{20} + n_{22} \leq n_{21} \\ & \text{and } n_{22} \leq n_{20}, \\ \frac{(n_{20} + n_{22} - n_{21})^2}{n_2} & \text{if } n_{20} + n_{22} > n_{21} \\ & \text{and } n_{22} \leq n_{20}, \\ \frac{2(n_{22} - n_{20})^2}{n_2} & \text{if } n_{20} + n_{22} \leq n_{21} \\ & \text{and } n_{22} > n_{20}, \\ \frac{(n_{20} + n_{22} - n_{21})^2}{n_2} & \text{if } n_{20} + n_{22} > n_{21} \\ + \frac{2(n_{22} - n_{20})^2}{n_2} & \text{and } n_{22} > n_{20}. \end{cases} \quad (10)$$

For testing hypotheses (6), the score statistic is

$$S_2 = \begin{cases} \frac{2(n_{22} - n_{20})^2}{n_2} & \text{if } n_{20} + n_{22} \leq n_{21}, \\ \frac{(n_{20} + n_{22} - n_{21})^2}{n_2} & \text{if } n_{20} + n_{22} > n_{21}. \end{cases} \quad (11)$$

By means of the argument given in the last paragraph of Appendix A, it can be directly verified that S_1 is asymptotically distributed as $.25\chi_0^2 + .5\chi_1^2 + .25\chi_2^2$ and that S_2 is asymptotically distributed as $.5\chi_1^2 + .5\chi_2^2$, for the respective hypotheses, (4) and (6), which agree with the asymptotic distributions of the LR tests.

Note that $n_{20} + n_{22} - n_{21}$ is a score based on IBD sharing that ignores the allelic state and that $n_{22} - n_{20}$ is the contrast of allele-specific IBD-sharing scores. It can be seen from equations (10) and (11) that the score statistics of the DMLB test *adaptively* combine these two aspects of linkage information. Furthermore, if there is linkage, we expect that $n_{20} + n_{22} > n_{21}$, and, if there is linkage and LD between marker allele B_1 and the disease, we expect that $n_{22} > n_{20}$. Thus the value of S_1 or S_2 under linkage and LD will likely be the sum of the mean test and the TDT statistics, but with $df < 2$. This gives an intuitive reason why the DMLB should have good power in comparison with the TDT or the mean test, for a broad range of LD.

Mixed-Sibship Data

In practice, a sample usually consists of families with sibships of variable sizes. The foregoing interpretation of the DMLB test from a score-test point of view continues to be applicable. To illustrate this, we consider a sample that has n_1, n_2 , and n_3 heterozygous parents with

one, two, and three affected sibs, respectively. Let (n_{10}, n_{11}) and (n_{20}, n_{21}, n_{22}) be defined as above. Use the following notation:

- n_{30} = number of parents who transmit B_2 to all three children ,
- n_{31} = number of parents who transmit B_1 to one child and B_2 to two children ,
- n_{32} = number of parents who transmit B_1 to two children and B_2 to one child ,
- n_{33} = number of parents who transmit B_1 to all three children .

Then, after some calculation using the reparametrization β , the likelihood of the data is $L(\beta) = L_1(\beta)L_2(\beta)L_3(\beta)$, where

$$\begin{aligned}
 L_1(\beta) &= (1 - \beta_2)^{n_{10}}(1 + \beta_2)^{n_{11}} , \\
 L_2(\beta) &= (1 + \beta_1 - 2\beta_2)^{n_{20}}(1 - \beta_1)^{n_{21}} \\
 &\quad \times (1 + \beta_1 + 2\beta_2)^{n_{22}} , \\
 L_3(\beta) &= [1 + 3\beta_1 - \beta_2(3 + \beta_1)]^{n_{30}} \\
 &\quad \times [1 - \beta_1 - \beta_2(1 - \beta_1)]^{n_{31}} \\
 &\quad \times [1 - \beta_1 + \beta_2(1 - \beta_1)]^{n_{32}} \\
 &\quad \times [1 + 3\beta_1 + \beta_2(3 + \beta_1)]^{n_{33}} .
 \end{aligned}$$

Let $s_1 = n_{20} + n_{22} + 3n_{30} + 3n_{33}$, $s_2 = n_{21} + n_{31} + n_{32}$, $t_1 = n_{11} + 2n_{22} + 3n_{33} + n_{32}$, $t_2 = n_{10} + 2n_{20} + 3n_{30} + n_{31}$. For the testing of hypotheses (4), the score-test statistic is

$$S_1 = \begin{cases} 0 & \text{if } s_1 \leq s_2 \text{ and } t_1 \leq t_2 , \\ \frac{(s_1 - s_2)^2}{n_2 + 3n_3} & \text{if } s_1 > s_2 \text{ and } t_1 \leq t_2 , \\ \frac{(t_1 - t_2)^2}{n_1 + 2n_2 + 3n_3} & \text{if } s_1 \leq s_2 \text{ and } t_1 > t_2 , \\ \frac{(s_1 - s_2)^2}{n_2 + 3n_3} + \frac{(t_1 - t_2)^2}{n_1 + 2n_2 + 3n_3} & \text{if } s_1 > s_2 \text{ and } t_1 > t_2 . \end{cases}$$

For the testing of hypotheses (6), the score statistic is

$$S_2 = \begin{cases} \frac{(t_1 - t_2)^2}{n_1 + 2n_2 + 3n_3} & \text{if } s_1 \leq s_2 , \\ \frac{(s_1 - s_2)^2}{n_2 + 3n_3} + \frac{(t_1 - t_2)^2}{n_1 + 2n_2 + 3n_3} & \text{if } s_1 > s_2 . \end{cases}$$

Again, S_1 and S_2 adaptively combine the test statistic $(s_1 - s_2)^2/(n_2 + 3n_3)$ based on IBD sharing and the TDT statistic $(t_1 - t_2)^2/(n_1 + 2n_2 + 3n_3)$. Again, either on the basis of the results of the asymptotic distributions of the LR-test statistics Λ_1 and Λ_2 or by direct verification, S_1 is asymptotically distributed as $.25\chi_0^2 + .5\chi_1^2 + .25\chi_2^2$, and S_2 is asymptotically distributed as $.5\chi_1^2 + .5\chi_2^2$.

Power Comparison of the TDT, the Mean Test, and DMLB, for ASP Data

For ASP data, the LR tests Λ_1 and Λ_2 , defined in formulas (5) and (7), are equivalent to the score tests S_1 and S_2 given by formulas (10) and (11). So the power of the DMLB test can be derived by means of the expressions for S_1 and S_2 . To obtain the noncentrality parameters required for power calculations, it is necessary to consider the transmission probabilities under linkage and LD. These probabilities are calculated as follows.

For a diallelic marker with two alleles, B_1 and B_2 , there are three informative parental mating types at the marker locus: (i) $B_1B_2 \times B_1B_1$, (ii) $B_1B_2 \times B_2B_2$, and (iii) $B_1B_2 \times B_1B_2$. For the mating type $B_1B_2 \times B_1B_1$, there are three possible types of ASPs with respect to the marker genotypes: (B_1B_1, B_1B_1) , (B_1B_1, B_1B_2) , and (B_1B_2, B_1B_2) . Let \mathcal{H} denote the event that at least one parent is heterozygous at the marker locus; that is, $\mathcal{H} = \{B_1B_2 \times B_1B_1, \text{ or } B_1B_2 \times B_2B_2, \text{ or } B_1B_2 \times B_1B_2\}$. We define the conditional probabilities as

$$\begin{aligned}
 p_{11} &= P(B_1B_1, B_1B_1, B_1B_2 \times B_1B_1 | \text{ASP}, \mathcal{H}) , \\
 p_{12} &= P(B_1B_1, B_1B_2, B_1B_2 \times B_1B_1 | \text{ASP}, \mathcal{H}) , \\
 p_{13} &= P(B_1B_2, B_1B_2, B_1B_2 \times B_1B_1 | \text{ASP}, \mathcal{H}) .
 \end{aligned}$$

These are the conditional probabilities of children's marker genotypes and parental marker genotypes, given that both children are affected and the event \mathcal{H} . We condition on the fact that both children are affected and the event \mathcal{H} , because only those families satisfying these two conditions are actually used in the test.

Similarly, for the mating type $B_1B_2 \times B_2B_2$, there are three possible types of ASPs: (B_1B_2, B_1B_2) , (B_1B_2, B_2B_2) , and (B_2B_2, B_2B_2) . We define the conditional probabilities as

$$\begin{aligned}
 p_{21} &= P(B_1B_2, B_1B_2, B_1B_2 \times B_2B_2 | ASP, \mathcal{H}) , \\
 p_{22} &= P(B_1B_2, B_2B_2, B_1B_2 \times B_2B_2 | ASP, \mathcal{H}) , \\
 p_{23} &= P(B_2B_2, B_2B_2, B_1B_2 \times B_2B_2 | ASP, \mathcal{H}) .
 \end{aligned}$$

For mating type $B_1B_2 \times B_1B_2$, there are six possible types of ASPs: (B_1B_1, B_1B_1) , (B_1B_1, B_1B_2) , (B_1B_1, B_2B_2) , (B_1B_2, B_1B_2) , (B_1B_2, B_2B_2) , and (B_2B_2, B_2B_2) . We define the conditional probabilities as

$$\begin{aligned}
 p_{31} &= P(B_1B_1, B_1B_1, B_1B_2 \times B_1B_2 | ASP, \mathcal{H}) , \\
 p_{32} &= P(B_1B_1, B_1B_2, B_1B_2 \times B_1B_2 | ASP, \mathcal{H}) , \\
 p_{33} &= P(B_1B_1, B_2B_2, B_1B_2 \times B_1B_2 | ASP, \mathcal{H}) , \\
 p_{34} &= P(B_1B_2, B_1B_2, B_1B_2 \times B_1B_2 | ASP, \mathcal{H}) , \\
 p_{35} &= P(B_1B_2, B_2B_2, B_1B_2 \times B_1B_2 | ASP, \mathcal{H}) , \\
 p_{36} &= P(B_2B_2, B_2B_2, B_1B_2 \times B_1B_2 | ASP, \mathcal{H}) .
 \end{aligned}$$

The calculation of these conditional probabilities is given in Appendix C.

Let p_2 , p_1 , and p_0 be the probabilities that a heterozygous parent transmits 2, 1, and 0 B_1 alleles to both children, conditional on both children being affected and there being at least one heterozygous parent in the family. When we consider all the possible mating types, we have

$$\begin{aligned}
 p_2 &= p_{11} + p_{21} + p_{31} + .5p_{32} + .25p_{34} , \\
 p_1 &= p_{12} + p_{22} + .5p_{32} + p_{33} + .5p_{34} + .5p_{35} , \\
 p_0 &= p_{13} + p_{23} + .25p_{34} + .5p_{35} + p_{36} .
 \end{aligned}$$

These probabilities are functions of the recombination fraction (θ), LD coefficient, penetrances, and gene frequencies at both disease and marker loci.

Let the sample size n be the number of heterozygous parents from families with both an ASP and at least one heterozygous parent. These families are those who are actually used in the computation of the test statistics. When the expressions for the score statistics S_1 and S_2 defined by expressions (10) and (11) are used, the non-centrality parameters are $\eta_1 = 2\sqrt{n}(.5 - p_1)$ and $\eta_2 = \sqrt{2n}(p_2 - p_0)$.

Let $Y_1 \sim N[\eta_1, 4p_1(1 - p_1)]$ and $Y_2 \sim N[\eta_2, 2[p_0 + p_2 - (p_0 - p_2)^2]]$ be two independent normal random variables. Let c_1 be the critical value of the DMLB test for a given type 1-error rate. For the one-sided hypotheses (4), the power of the DMLB can be approximated by

$$\begin{aligned}
 &P(Y_1 > \sqrt{c_1})P(Y_2 \leq 0) + P(Y_2 > \sqrt{c_1})P(Y_1 \leq 0) \\
 &\quad + P(Y_1^2 + Y_2^2 > c_1, Y_1 > 0, Y_2 > 0) .
 \end{aligned}$$

For the two-sided hypotheses (6), the power of the DMLB test can be approximated by

$$\begin{aligned}
 &P(|Y_2| > \sqrt{c_1})P(Y_1 \leq 0) \\
 &\quad + P(Y_1^2 + Y_2^2 > c_1, Y_1 > 0) .
 \end{aligned}$$

These two expressions can be derived directly from expressions (10) and (11) and by invoking the central limit theorem.

The power of the TDT can be approximated by $P(|Y_2| > \sqrt{c_2})$, and the power of the mean test can be approximated by $P(Y_1 > \sqrt{c_3})$.

When there is no linkage, $p_1 = .5$ and $p_0 = p_2 = .25$, regardless of whether LD is present. When there is no LD, $p_0 = p_2$, regardless of whether linkage is present. Thus, η_1 can be considered as a measure of linkage, and η_2 can be considered as a ‘‘product measure’’ of linkage and LD; here we call η_2 a ‘‘product measure’’ because it is equal to 0 if either linkage or LD does not exist. On the basis of the foregoing expressions for the power functions, we see that both η_1 and η_2 contribute to the power of the DMLB; however, only η_1 contributes to the power of the mean test, and only η_2 contributes to the power of the TDT.

We now give examples of the sample sizes required to achieve 80% power when the DMLB, the TDT, and the mean test are used, when LD ranges from 0 to the maximum extent. For brevity, we consider only the two-sided hypotheses (6). The type 1-error rate is set at .0001. The critical value for the DMLB test, $c_1 = 17.38$, is computed on the basis of its asymptotic distribution $.5\chi_1^2 + .5\chi_2^2$. The critical value for the TDT is $c_2 = 15.14$, corresponding to the χ_1^2 distribution, and the critical value for the mean test is $c_3 = 13.83$, corresponding to the $.5\chi_0^2 + .5\chi_1^2$ distribution.

The sample sizes required to achieve 80% power for the TDT and the mean test can be calculated by means of the following equation: $n = [(.84\sigma + c)/\mu]^2$, where, for the TDT, $c = \sqrt{c_2} = 3.89$, $\mu = \sqrt{2}(p_2 - p_0)$, and $\sigma^2 = 2[p_2 + p_0 - (p_2 - p_0)^2]$ and, for the mean test, $c = \sqrt{c_3} = 3.72$, $\mu = 2(.5 - p_1)$, and $\sigma^2 = 4p_1(1 - p_1)$. There is no simple formula for sample-size calculation for the DMLB test. However, because power is an increasing function of sample size, it is easy to find the n that gives 80% power numerically.

We consider four genetic models: recessive, dominant, additive, and multiplicative. In each model, the θ between marker and disease locus is set to be 0. Let f_0, f_1 , and f_2 be the penetrances of disease genotypes dd, Dd , and DD , respectively, where D is the disease-causing

allele. The relative genotypic risks (GRRs) are defined to be $r_1 = f_1/f_0$ and $r_2 = f_2/f_0$. We consider the following GRR values in the power calculation: for the recessive model, $r_1 = 1$, $r_2 = 4$; for the dominant model, $r_1 = r_2 = 4$; for the additive model, $r_1 = 4$ and $r_2 = 7$; and, for the multiplicative model, $r_1 = 4$ and $r_2 = 16$. Three disease-allele frequencies—.10, .20, and .50—and two frequencies—.20 and .50—of marker allele B_1 are considered. Therefore, in each model, for the given θ of 0 and the GRRs, there are six combinations of disease- and marker-allele frequencies. For each of these six combinations, LD in the range of 0%–100% of the maximum possible extent, with a step size of 5%, is used in the calculation. Figure 1 shows the sample size versus the percentage of the maximum possible LD under the recessive model, for each of the six configurations. Similarly, figures 2–4 show the sample size versus the percentage of the maximum possible LD, for the dominant, additive, and multiplicative models, respectively. In each plot, the unbroken line is for the DMLB test, the dashed line for the TDT, and the dotted line for the mean test.

These plots show that, for all the models considered, the DMLB test requires smaller sample sizes to achieve 80% power than does the TDT, for the whole range of LD. For both the DMLB test and the TDT, as the percentage of maximum LD changes from 0% to 100%, the range of the sample sizes needed to achieve 80% power becomes larger. When LD is maximum or nearly so, the power of the DMLB test and that of the TDT are similar; however, when LD is less than maximum, the DMLB test has greater power than does the TDT. The difference ranges from appreciable to substantial if the percentage of the maximum possible LD is <75%. When LD is very weak or absent, the mean test has higher power than both the TDT and the DMLB; in this case, however, for the models considered here, the sample sizes needed to achieve 80% power are large and beyond the range of most studies.

We have also considered models with different parameter values, such as GRRs of 2.0 and 1.5 and families with more than two affected children. The results (not shown) are similar to those shown in figures 1–4. Thus the conclusions above appear to hold quite generally.

Discussion

The DMLB test is a test for linkage. The only purpose for the incorporation of LD into the likelihood is to increase statistical power in the detection of linkage. By the equivalence between the LR and score tests examined above (in the section entitled “DMLB as an Adaptive Combination of the TDT and the Mean Test”), incorporation of LD into the likelihood amounts to adaptive combination of information from IBD-sharing and allele-specific IBD-sharing scores.

An important property of the TDT is that it is not affected by population admixture. More precisely, the TDT has correct type 1 error when data are ascertained in an admixture population. This is also true for the DMLB test, because it uses children’s marker-genotype data, given the parental marker genotypes. This can also be seen from the form of the likelihood; for example, consider the case in which the population consists of two subpopulations and in which the coefficients of LD of the two subpopulations are different. This does not affect the type 1 error of the LR tests given by expressions (5) and (7), because, under the null hypothesis $\alpha = .5$, the parameter λ disappears from the likelihood. Thus the value of λ has no effect on the distribution of the test statistic when there is no linkage. The fact that the DMLB is not affected by population admixture can also be seen from the corresponding score statistics described above (in the section entitled “DMLB as an Adaptive Combination of the TDT and the Mean Test”).

An advantage of the DMLB test is that it can explicitly incorporate locus heterogeneity, in a way similar to that of the heterogeneity LOD score (Smith 1961; Ott 1991). This is useful, because locus heterogeneity is common in complex diseases. Incorporation of locus heterogeneity into the test should increase the power, if it indeed exists. It is of interest to study the properties and performance of the heterogeneity DMLB test in comparison with the test when locus heterogeneity is not taken into account. This is a problem that we intend to consider in the future. Note that it is not possible to incorporate locus heterogeneity into the TDT or into the score tests that have been given above (in the section entitled “DMLB as an Adaptive Combination of the TDT and the Mean Test”).

The DMLB is described for a marker with two alleles. One way to generalize it to a marker with more than two alleles is to focus on one allele at a time and to treat the other alleles as the second allele, as is done in the report by Spielman et al. (1993), and then to use the maximum of the resultant DMLB statistics (either in the form of LR statistics [given above, in the section entitled “The DMLB Test for Linkage”] or in the form of score statistics [given above, in the section entitled “DMLB as an Adaptive Combination of the TDT and the Mean Test”]), as is done in the report by Schaid (1996). The distribution of this maximum statistic may not be tractable analytically; however, an empirical P value can be obtained by simulations, by means of an approach described by Schaid (1996). Several extensions of the TDT, for multiple alleles, have been considered (Sham and Curtis 1995; Schaid 1996; Spielman and Ewens 1996). It would be worthwhile to consider analogous extensions of the DMLB test, for a marker with multiple alleles.

In the formulation of the DMLB test, it is assumed

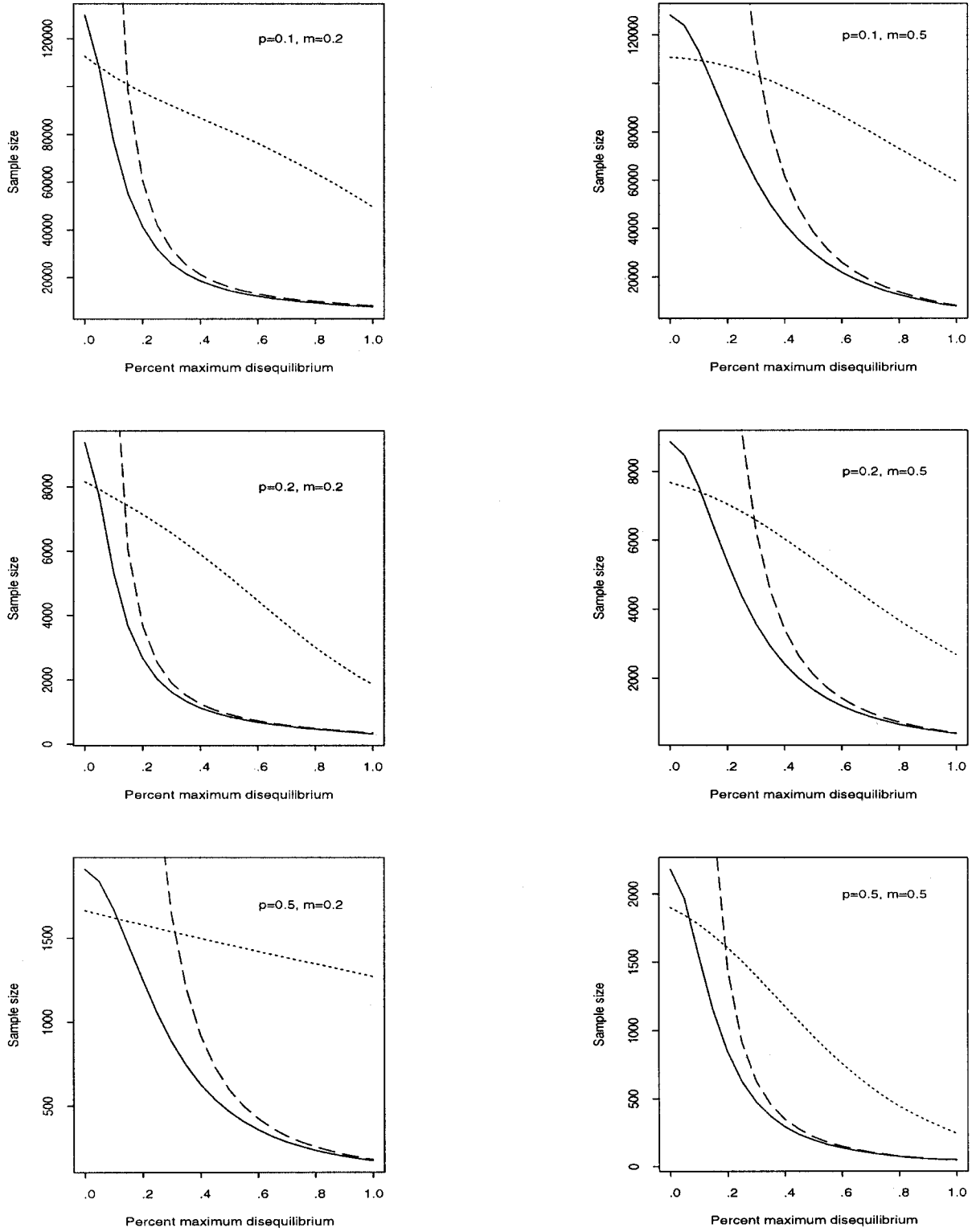


Figure 1 Recessive model: sample sizes required to achieved 80% power, with a type 1 error of .0001. The unbroken line is for the DMLB test, the dashed line is for the TDT, and the dotted line is for the mean test. The plots are for the six combinations of disease-allele frequencies .10, .20, and .50 and marker-allele (B_1) frequencies .20 and .50. In each case, $\theta = 0$ and the GRRs are $r_1 = 1$ and $r_2 = 4$.

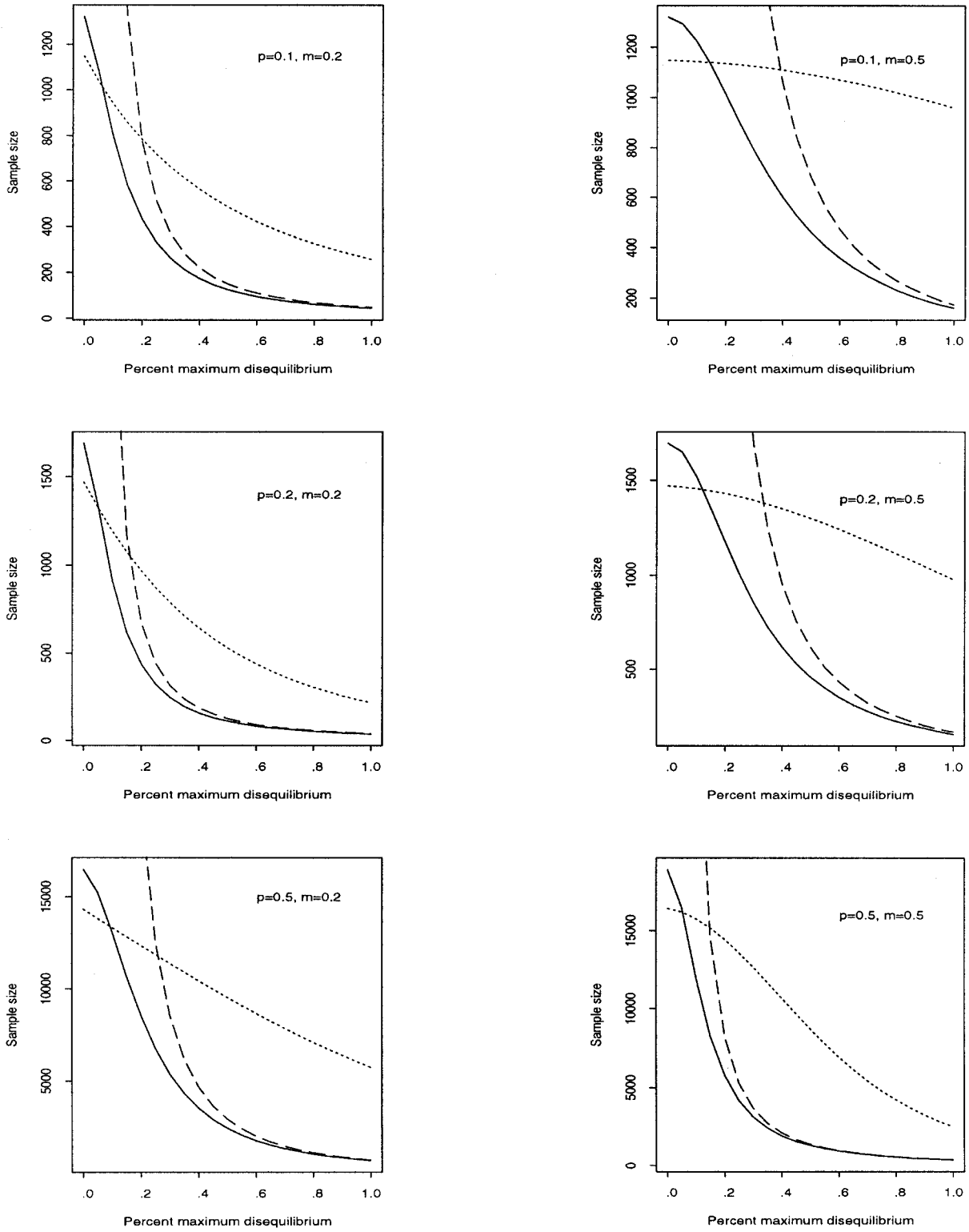


Figure 2 Dominant model: sample sizes required to achieve 80% power, with a type 1 error of .0001. The unbroken line is for the DMLB test, the dashed line is for the TDT, and the dotted line is for the mean test. The plots are for the six combinations of disease-allele frequencies .10, .20, and .50 and marker-allele (B_1) frequencies .20 and .50. In each case, $\theta = 0$ and the GRRs are $r_1 = 4$ and $r_2 = 4$.

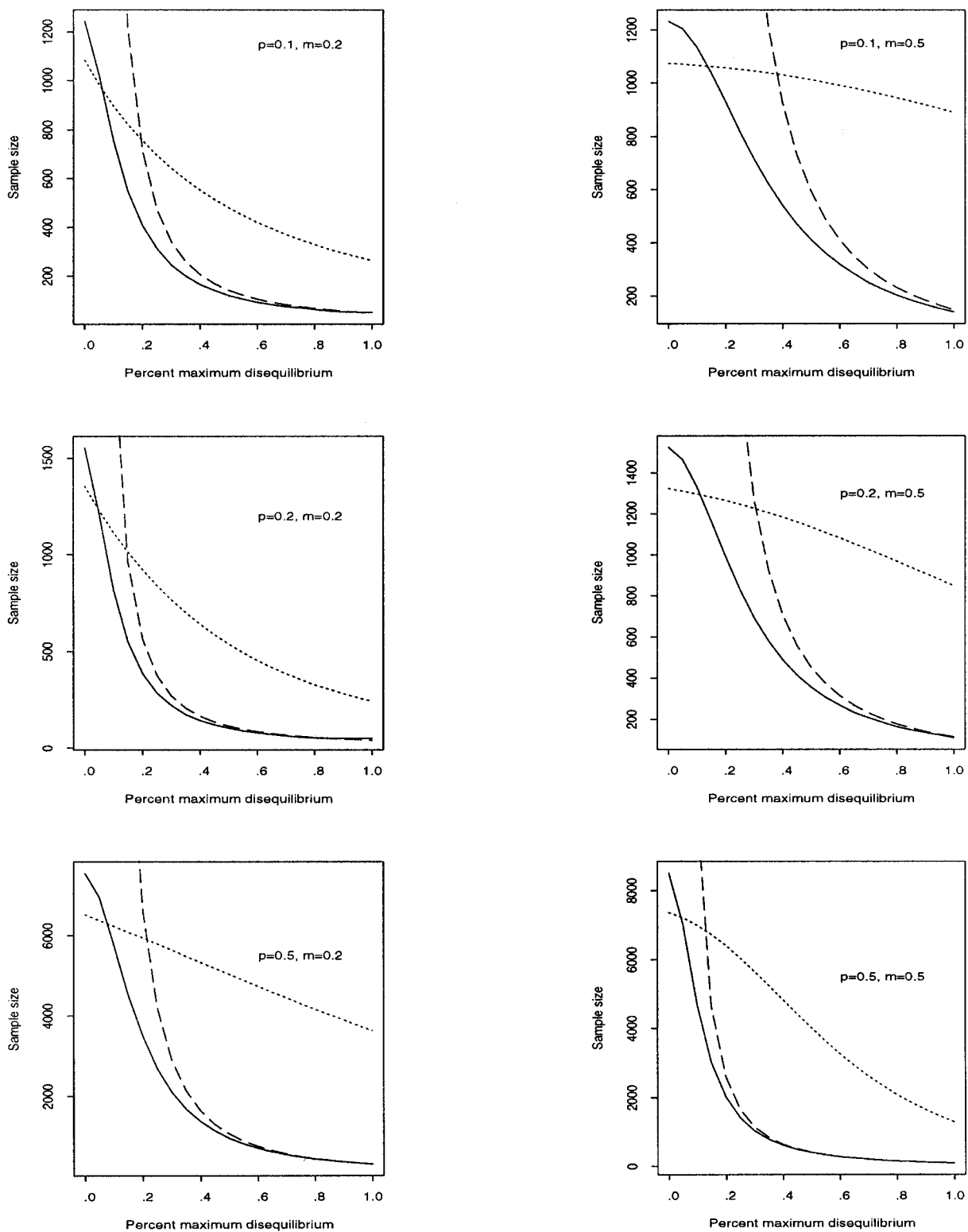


Figure 3 Additive model: sample sizes required to achieved 80% power, with a type 1 error of .0001. The unbroken line is for the DMLB test, the dashed line is for the TDT, and the dotted line is for the mean test. The plots are for the six combinations of disease-allele frequencies .10, .20, and .50 and marker-allele (B_1) frequencies .20 and .50. In each case, $\theta = 0$ and the GRRs are $r_1 = 4$ and $r_2 = 7$.

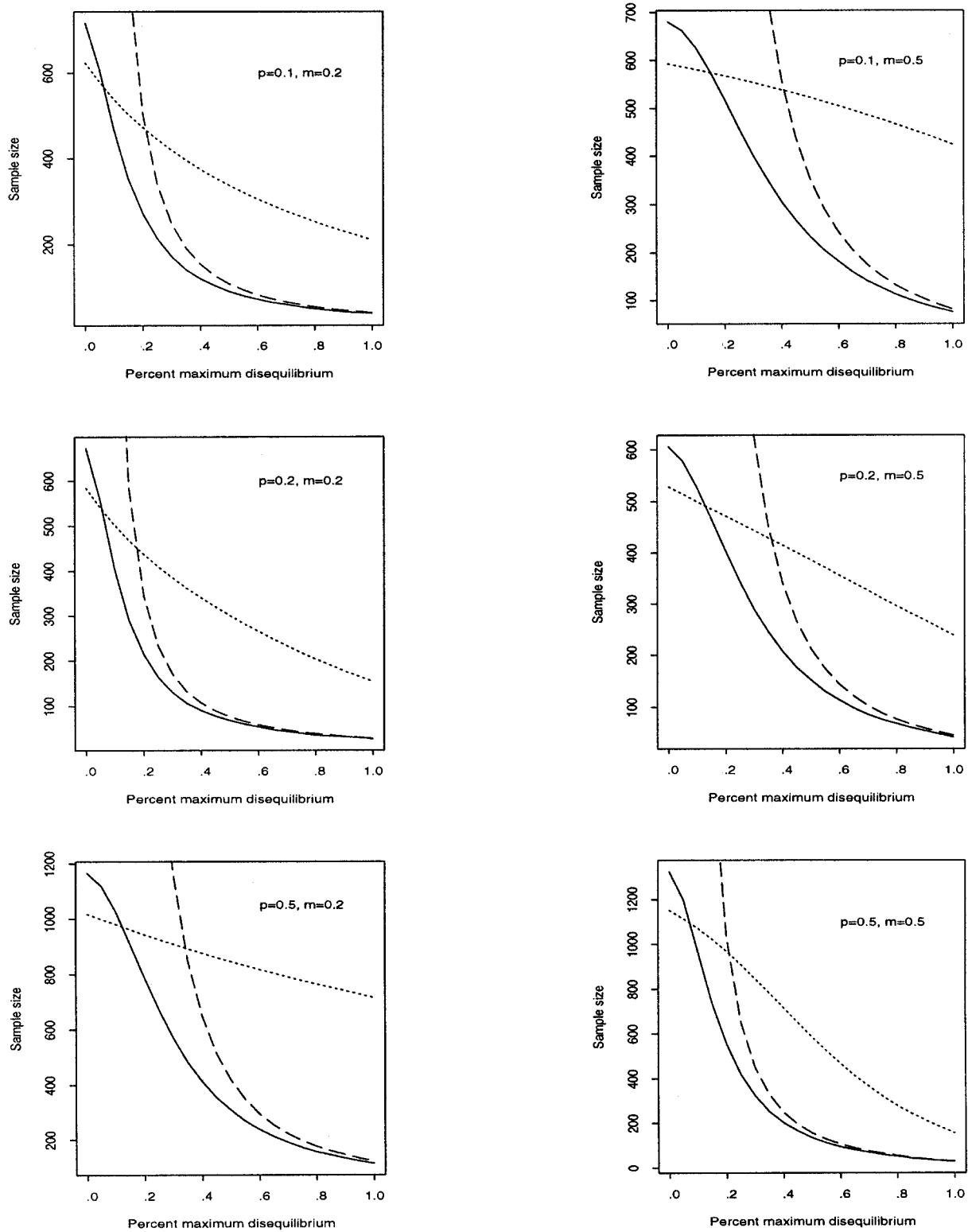


Figure 4 Multiplicative model: sample sizes required to achieved 80% power, with a type 1 error of .0001. The unbroken line is for the DMLB test, the dashed line is for the TDT, and the dotted line is for the mean test. The plots are for the six combinations of disease-allele frequencies .10, .20, and .50 and marker-allele (B_1) frequencies .20 and .50. In each case, $\theta = 0$ and the GRRs are $r_1 = 4$ and $r_2 = 16$.

that the parental marker genotypes can be completely known. In practice, this may not be possible, especially for diseases with delayed age at onset, such as Alzheimer disease. Part or all of the parental genotypes may be missing. Several methods have been proposed to use affected and unaffected children when parental genotypes are not available (Boehnke and Langefeld 1998; Horvath and Laird 1998; Spielman and Ewens 1998; Knapp 1999). It would be interesting to extend the DMLB test to the situation when parental genotypes are not available.

In summary, the DMLB test adaptively combines link-

age information from the IBD-sharing and allele-specific IBD-sharing scores and, in statistical power, compares favorably with the mean test and with the TDT, for a broad range of LD. Therefore, the DMLB test appears to be an interesting approach to linkage detection when the extent of LD is unknown.

Acknowledgments

We thank Veronica Vieland and Susan Slager for many helpful discussions and comments. This work is supported by National Institute of Mental Health grant K01-01541.

Appendix A

Asymptotic Distribution of the LR-Test Statistics

We first give a formal statement of the asymptotic distribution of the LR-test statistics.

PROPOSITION 6.1. Let n_j be the number of heterozygous parents with j affected children, $1 \leq j \leq J$, with J being a finite integer. Let $n = \sum_{j=1}^J n_j$. Suppose that the limits $\tau_j = \lim (n_j/n), 1 \leq j \leq J$, exist.

1. If $\tau_j = 0$ for all $j \geq 2$, then under the null hypothesis, as $n \rightarrow \infty$, $2 \log \Lambda_1 \rightarrow_d .5\chi_0^2 + .5\chi_1^2$ and $2 \log \Lambda_2 \rightarrow_d \chi_1^2$. Here \rightarrow_d denotes convergence in distribution.

2. If there is at least one $\tau_j > 0$ for some $j \geq 2$, then, under the null hypothesis, as $n \rightarrow \infty$, $2 \log \Lambda_1 \rightarrow_d .25\chi_0^2 + .5\chi_1^2 + .25\chi_2^2$, and $2 \log \Lambda_2 \rightarrow_d .5\chi_1^2 + .5\chi_2^2$.

The condition on τ_j in part 1 of Proposition 6.1 means that the data essentially consist of families with a single affected offspring; the condition on τ_j in part 2 of Proposition 6.1 means that the data consist of a nonnegligible number of families with either affected sib pairs or multiple affected sibships.

Part 1 of Proposition 6.1 follows from standard results. For part 2, we only prove the result for Λ_1 . The result for Λ_2 can be shown similarly. Let n_k be the number of heterozygous parents with k affected children, where $1 \leq k \leq K$. Here K is a finite integer. Let $n = \sum_{k=1}^K n_k$. Suppose that $\tau_k = \lim (n_k/n)$ exists. Then the limiting version of the log likelihood (divided by n) is $\ell(\alpha, \lambda; x) = \sum_{k=1}^K \tau_k \ell_k(\alpha, \lambda; x)$, where $\ell_k = \lim_{n \rightarrow \infty} \log L_k$ is the log-likelihood function for a heterozygous parent with k affected children and where

$$L_k(\alpha, \lambda; x) = \lambda \alpha^x (1 - \alpha)^{k-x} + (1 - \lambda) \alpha^{k-x} (1 - \alpha)^x, \quad x = 0, 1, 2, \dots, k; \quad k \geq 1 .$$

The technical difficulty in the establishment of the asymptotic distribution of the LR test is that the null hypothesis does not correspond to a single point in the parameter space but, rather, to a set of points on the boundary $\{(\alpha, \lambda) : \alpha = .5, .5 \leq \lambda \leq 1\}$. As in the report by Chernoff and Lander (1995), we reparametrize the model, to avoid this difficulty. For the current problem, a convenient reparametrization is $\beta_1 = 4(.5 - \alpha)^2$ and $\beta_2 = 4(.5 - \alpha)(.5 - \lambda)$. Let $\beta = (\beta_1, \beta_2)$. Then the null hypothesis corresponds to a single point $(0, 0)$. The parameter space becomes

$$\mathcal{B} = \{(\beta_1, \beta_2) : 0 \leq \beta_1 \leq 1, 0 \leq \beta_2 \leq 1, \beta_2^2 \leq \beta_1\} .$$

In terms of the parameter β , the likelihood can be written as

$$L_k(\beta; x) = (1 - \beta_1)^x [(1 + \beta_2 \beta_1^{-1/2})(1 - \beta_1^{1/2})^{k-2x} + (1 - \beta_2 \beta_1^{-1/2})(1 + \beta_1^{1/2})^{k-2x}] .$$

Note that the new null-hypothesis point $(0,0)$ is on the boundary of this parameter space. In Appendix B, it is shown that the Fisher information matrix at 0, denoted by $\mathbf{I}(0)$, is a diagonal matrix with both diagonal elements strictly positive. Thus we can write $\mathbf{I}(0) = \text{diag}(a_1^2, a_2^2)$, where $a_1 > 0$ and $a_2 > 0$. After transformation $\beta^* = I^{-1/2}(0)\beta$, the new parameter space becomes

$$\mathcal{B}^* = \{(\beta_1, \beta_2) : 0 \leq \beta_1 \leq a_1^{-1}, 0 \leq \beta_2 \leq a_2^{-1}, \beta_2^2 \leq (a_1/a_2^2)\beta_1\} .$$

The tangent cone of \mathcal{B}^* at $(0, 0)$ is the first quadrant $C_0 = \{(\beta_1, \beta_2) : \beta_1 \geq 0, \beta_2 \geq 0\}$. On the basis of either the result reported by Chernoff (1954) or that reported by Self and Liang (1987), $2\log\Lambda_1$ converges, in distribution, to

$$X \equiv \| Z \|^2 - \inf_{x \in C_0} \| Z - x \|^2 \stackrel{d}{=} .25\chi_0^2 + .5\chi_1^2 + .25\chi_2^2 , \tag{A1}$$

where $Z \sim N(0, I_2)$. The second equality in expression (A1) follows because (i) the probability of Z falling in the first quadrant is .25, and, given that Z falls in the first quadrant, $X \sim \chi_2^2$; (ii) the probability of Z falling in the second or fourth quadrant is .5, and, given that Z falls in the second or fourth quadrant, $X \sim \chi_1^2$; and (iii) the probability of Z falling in the third quadrant is .25, and, given that Z falls in the third quadrant, $X = 0$, with $P = 1.0$. In addition, we also use the fact that the length of a normal random vector with mean 0 is independent of its direction.

Appendix B

Positive Definiteness of the Information Matrix

We now show that $I(0)$ is a diagonal matrix and is strictly positive definite. To this end, it suffices to show that the Fisher information matrix $I_k(0)$ corresponding to L_k at $\beta = 0$ is diagonal and strictly positive definite for $k \geq 2$, since $I(0) = \sum_{k=1}^K \tau_k I_k(0)$, where $K \geq 2$. When the binomial expansion is used,

$$L_k(\beta; x) = \begin{cases} 2^{-k}(1 - \beta_1)^x \left[\sum_{i=0, i \text{ even}}^{k-2x} \binom{k-2x}{i} \beta_1^{i/2} - \beta_2 \sum_{i=1, i \text{ odd}}^{k-2x} \binom{k-2x}{i} \beta_1^{(i-1)/2} \right] & \text{if } x \leq [(k-1)/2] \\ 2^{-k}(1 - \beta_1)^{k-x} \left[\sum_{i=0, i \text{ even}}^{2x-k} \binom{2x-k}{i} \beta_1^{i/2} + \beta_2 \sum_{i=1, i \text{ odd}}^{2x-k} \binom{2x-k}{i} \beta_1^{(i-1)/2} \right] & \text{if } x \geq k - [(k-1)/2] \\ 2^{-k}(1 - \beta_1)^{k/2} & \text{if } x = k/2, k \text{ even} \end{cases} .$$

This expression is needed in the computation of the Fisher information matrix. Let $\ell_k = \log L_k$. After some tedious calculation,

$$\left. \frac{\partial^2 \ell_k(\beta; x)}{\partial \beta_1^2} \right|_{\beta=0} = \begin{cases} -x + 2 \binom{k-2x}{4} - \left(\frac{k-2x}{2} \right)^2 & \text{if } x \leq [(k-1)/2] \\ -(k-x) + 2 \binom{2x-k}{4} - \left(\frac{2x-k}{2} \right)^2 & \text{if } x \geq k - [(k-1)/2] \\ -k/2 & \text{if } k \text{ is even and } x = k/2 \end{cases} ,$$

$$\frac{\partial^2 \ell_k(\beta; x)}{\partial \beta_1 \partial \beta_2} \Big|_{\beta=0} = \begin{cases} -\binom{k-2x}{3} + (k-2x)\binom{k-2x}{2} & \text{if } x \leq [(k-1)/2] \\ \binom{2x-k}{3} - (2x-k)\binom{2x-k}{2} & \text{if } x \geq k - [(k-1)/2] \\ 0 & \text{if } k \text{ is even and } x = k/2 \end{cases}, \tag{B1}$$

and

$$\frac{\partial^2 \ell_k(\beta; x)}{\partial \beta_2^2} \Big|_{\beta=0} = -(k-2x)^2 .$$

Let

$$I_k(0) = \begin{bmatrix} a_{11}(k) & a_{12}(k) \\ a_{12}(k) & a_{22}(k) \end{bmatrix}$$

be the Fisher information matrix. The entries of $I_k(0)$ can be obtained by taking expectations of the second derivatives. First,

$$a_{11}(k) = -E \left[\frac{\partial^2 \ell_k(\beta; x)}{\partial \beta_1^2} \Big|_{\beta=0} \right] = \begin{cases} 2^{-k+1} \sum_{i=0}^{[(k-1)/2]} \binom{k}{i} \left[i - 2 \binom{k-2i}{4} + \binom{k-2i}{2} \right] & \text{if } k \text{ is odd} \\ 2^{-k} \left\{ 2 \sum_{i=0}^{[(k-1)/2]} \binom{k}{i} \left[i - 2 \binom{k-2i}{4} + \binom{k-2i}{2} \right]^2 + (k/2) \binom{k}{k/2} \right\} & \text{if } k \text{ is even} \end{cases} .$$

By the symmetry of expression(B1), it can be shown that

$$a_{12}(k) = E \left[\frac{\partial^2 \ell_k(\beta; x)}{\partial \beta_1 \partial \beta_2} \Big|_{\beta=0} \right] = 0 .$$

For entry a_{22} ,

$$a_{22}(k) = -E \left[\frac{\partial^2 \ell_k(\beta; x)}{\partial \beta_2^2} \Big|_{\beta=0} \right] = 2^{-k} \sum_{i=0}^k \binom{k}{i} (k-2i)^2 = k ,$$

where the following two identities are used:

$$\sum_{i=0}^k \binom{k}{i} i = k2^{k-1}$$

and

$$\sum_{i=0}^k \binom{k}{i} i^2 = k(k-1)2^{k-2} + k2^{k-1} .$$

So the Fisher information matrix is

$$I_k(0) = \begin{bmatrix} a_{11}(k) & 0 \\ 0 & k \end{bmatrix}.$$

Finally, to show that $I_k(0)$ is strictly positive definite when $k \geq 2$, it suffices to show that $a_{11}(k) > 0$ for every $k \geq 2$. For $k \geq 2$ and $0 \leq i \leq [(k - 1)/2]$,

$$i - 2\binom{k-2i}{4} + \binom{k-2i}{2}^2 = i + \frac{1}{12}(k - 2i)(k - 2i - 1)[-(k - 2i - 2)(k - 2i - 3) + 3(k - 2i)(k - 2i - 1)] \geq 0.$$

When $i = 0$, this term becomes

$$-2\binom{k}{4} + \binom{k}{2}^2 = \frac{1}{12}(k)(k - 1)[-(k - 2)(k - 3) + 3(k)(k - 1)] > 0, \text{ if } k \geq 2.$$

Thus $a_{11}(k) > 0$ when $k \geq 2$. This shows that $I_k(0)$ is strictly positive definite.

Appendix C

Computation of the Conditional Probabilities for the Section entitled "Power Comparison, between the TDT, the Mean Test, and the DMLB, for ASP Data"

Table C1

Conditional Probabilities of Marker Genotype of Affected Child, Given Parental Mating Type

MATING TYPE	CONDITIONAL PROBABILITY OF MARKER GENOTYPE OF AFFECTED CHILD ^a		MATING-TYPE PROBABILITY
	B_1B_1	B_1B_2	
$DB_1/DB_2 \times DB_1/DB_1$	$\frac{1}{2}f_2$	$\frac{1}{2}f_2$	$2P^3(DB_1)P(DB_2)$
$DB_1/dB_2 \times DB_1/DB_1$	$\frac{1}{2}(f_2(1 - \theta) + f_1\theta)$	$\frac{1}{2}(f_2\theta + f_1(1 - \theta))$	$2P^3(DB_1)P(dB_2)$
$dB_1/DB_2 \times DB_1/DB_1$	$\frac{1}{2}(f_1(1 - \theta) + f_2\theta)$	$\frac{1}{2}(f_1\theta + f_2(1 - \theta))$	$2P(DB_1)^2P(dB_1)P(DB_2)$
$dB_1/dB_2 \times DB_1/DB_1$	$\frac{1}{2}f_1$	$\frac{1}{2}f_1$	$2P(DB_1)^2P(dB_1)P(dB_2)$
$DB_1/DB_2 \times DB_1/dB_1$	$\frac{1}{4}(f_2 + f_1)$	$\frac{1}{4}(f_2 + f_1)$	$4P^2(DB_1)P(dB_1)P(DB_2)$
$DB_1/dB_2 \times DB_1/dB_1$	$\frac{1}{4}(f_2(1 - \theta) + f_1 + f_0\theta)$	$\frac{1}{4}(f_2\theta + f_1 + f_0(1 - \theta))$	$4P^2(DB_1)P(dB_1)P(dB_2)$
$dB_1/DB_2 \times DB_1/dB_1$	$\frac{1}{4}(f_0(1 - \theta) + f_1 + f_2\theta)$	$\frac{1}{4}(f_2(1 - \theta) + f_1 + f_0\theta)$	$4P(DB_1)P^2(dB_1)P(DB_2)$
$dB_1/dB_2 \times DB_1/dB_1$	$\frac{1}{4}(f_1 + f_0)$	$\frac{1}{4}(f_1 + f_0)$	$4P(DB_1)P^2(dB_1)P(dB_2)$
$DB_1/DB_2 \times dB_1/dB_1$	$\frac{1}{2}f_1$	$\frac{1}{2}f_1$	$2P(DB_1)P(DB_2)P^2(dB_1)$
$DB_1/dB_2 \times dB_1/dB_1$	$\frac{1}{2}(f_0\theta + f_1(1 - \theta))$	$\frac{1}{2}(f_0(1 - \theta) + f_1\theta)$	$2P(DB_1)P^2(dB_1)P(dB_2)$
$dB_1/DB_2 \times dB_1/dB_1$	$\frac{1}{2}(f_0(1 - \theta) + f_1\theta)$	$\frac{1}{2}(f_0\theta + f_1(1 - \theta))$	$2P^3(dB_1)P(DB_2)$
$dB_1/dB_2 \times dB_1/dB_1$	$\frac{1}{2}f_0$	$\frac{1}{2}f_0$	$2P^3(dB_1)P(dB_2)$

^a $f_0, f_1,$ and f_2 are the penetrances of the disease genotypes $dd, Dd,$ and $DD,$ respectively. θ is between the disease and marker locus.

In the calculation below, we assume that (a) mating is random in the population, (b) Hardy-Weinberg equilibrium holds in the parental generation, (c) the population is homogeneous with respect to all the genetic parameters, (d) the disease locus has two alleles, d and D , where D is the disease-causing allele or increases the risk of disease, and (e) the affection status in children is independent conditional on their genotypes at the disease locus.

For any column vectors \mathbf{u} and \mathbf{v} of k dimensions, define the following symbolic operations: $\mathbf{u}^m = (u_1^m, \dots, u_k^m)^t$, and $\mathbf{u}^o\mathbf{v} = (u_1v_1, \dots, u_kv_k)^t$. We will reserve \mathbf{u}^t to denote the transpose of \mathbf{u} .

Table C2

Conditional Probabilities of Marker Genotype of Affected Child, Given Parental Mating Type

CONDITIONAL PROBABILITIES OF MARKER GENOTYPE OF AFFECTED CHILD				
MATING TYPE	B_1B_1	B_1B_2	B_2B_2	MATING-TYPE PROBABILITY
$DB_1/DB_2 \times DB_1/DB_2$	$\frac{1}{4}f_2$	$\frac{1}{2}f_2$	$\frac{1}{4}f_2$	$P^2(DB_1)P^2(DB_2)$
$DB_1/dB_2 \times DB_1/DB_2$	$\frac{1}{4}[f_2(1-\theta) + f_1\theta]$	$\frac{1}{4}(f_2 + f_1)$	$\frac{1}{4}[f_2\theta + f_1(1-\theta)]$	$2P^2(DB_1)P(DB_2)P(dB_2)$
$dB_1/DB_2 \times DB_1/DB_2$	$\frac{1}{4}[f_1(1-\theta) + f_2\theta]$	$\frac{1}{4}(f_1 + f_2)$	$\frac{1}{4}[f_1\theta + f_2(1-\theta)]$	$2P(DB_1)P(dB_1)P^2(DB_2)$
$dB_1/dB_2 \times DB_1/DB_2$	$\frac{1}{4}f_1$	$\frac{2}{4}f_1$	$\frac{1}{4}f_1$	$2P(DB_1)P(DB_2)P(dB_1)P(dB_2)$
$DB_1/dB_2 \times DB_1/dB_2$	$\frac{1}{4}[f_2(1-\theta)^2 + f_1\theta(1-\theta) + f_1\theta(1-\theta) + f_0\theta^2]$	$\frac{2}{4}[f_2\theta(1-\theta) + f_1(1-\theta)^2 + f_1\theta^2 + f_0\theta(1-\theta)^2]$	$\frac{1}{4}[f_2\theta^2 + f_1\theta(1-\theta) + f_1\theta(1-\theta) + f_0(1-\theta)^2]$	$P^2(DB_1)P^2(dB_2)$
$dB_1/DB_2 \times DB_1/dB_2$	$\frac{1}{4}[f_1(1-\theta)^2 + f_0\theta(1-\theta) + f_2\theta(1-\theta) + f_1\theta^2]$	$\frac{1}{4}\{4f_1\theta(1-\theta) + f_0(1-\theta)^2 + f_0\theta^2 + f_2[\theta^2 + (1-\theta)^2]\}$	$\frac{1}{4}[f_1\theta^2 + f_0\theta(1-\theta) + f_2\theta(1-\theta) + f_1(1-\theta)^2]$	$2P(DB_1)P(DB_2)P(dB_1)P(dB_2)$
$dB_1/DB_2 \times dB_1/dB_2$	$\frac{1}{4}[f_0(1-\theta)^2 + f_1\theta(1-\theta) + f_1\theta(1-\theta) + f_2\theta^2]$	$\frac{2}{4}[f_0\theta(1-\theta) + f_1(1-\theta)^2 + f_1\theta^2 + f_2\theta(1-\theta)]$	$\frac{1}{4}[f_0\theta^2 + 2f_1\theta(1-\theta) + f_2(1-\theta)^2]$	$P^2(dB_1)P^2(DB_2)$
$dB_1/dB_2 \times DB_1/dB_2$	$\frac{1}{4}(f_1(1-\theta) + f_0\theta)$	$\frac{1}{4}(f_1 + f_0)$	$\frac{1}{4}[f_1\theta + f_0(1-\theta)]$	$2P(DB_1)P(dB_1)P^2(dB_2)$
$dB_1/dB_2 \times dB_1/DB_2$	$\frac{1}{4}(f_0(1-\theta) + f_1\theta)$	$\frac{1}{4}(f_0 + f_1)$	$\frac{1}{4}[f_0\theta + f_1(1-\theta)]$	$2P(dB_2)P^2(dB_1)P(dB_2)$
$dB_1/dB_2 \times dB_1/dB_2$	$\frac{1}{4}f_0$	$\frac{2}{4}f_0$	$\frac{1}{4}f_0$	$P^2(dB_1)P^2(dB_2)$

NOTE.—See footnote to table 1.

The calculation below is applicable to general multiplex-sibship data. First consider a nuclear family with parental mating type $B_1B_2 \times B_1B_1$ and s affected children, where $s \geq 1$. Let f_0 , f_1 , and f_2 be the conditional probabilities of being affected, given the disease genotypes dd , dD , and DD , respectively. Let \mathbf{g}_{11} and \mathbf{g}_{12} denote the column vectors whose elements are listed, in table C1, under the “ B_1B_1 ” and “ B_1B_2 ” column heads, respectively.

Let ψ_1 denote the column vector of mating-type probabilities given, in table C1, under the “Mating-Type Probability” column head. Let k be the number of affected children with marker B_1B_1 . We have

$$\begin{aligned} &P(k \ B_1B_1 \ \text{sibs}, (s - k) \ B_1B_2 \ \text{sibs}, B_1B_2 \times B_1B_1 | s \ \text{affected sibs}, \mathcal{H}) \\ &= P(k \ B_1B_1 \ \text{sibs}, (s - k) \ B_1B_2 \ \text{sibs} | s \ \text{affected sibs}, B_1B_2 \times B_1B_1) P(B_1B_2 \times B_1B_1 | s \ \text{affected sibs}, \mathcal{H}) . \end{aligned} \quad (C1)$$

The conditional probability of offsprings’ marker data, given that they are affected and that their parental marker genotypes are known, can be computed by means of the following equations:

$$\begin{aligned} &P(k \ B_1B_1 \ \text{sibs}, (s - k) \ B_1B_2 \ \text{sibs} | s \ \text{affected sibs}, B_1B_2 \times B_1B_1) \\ &= \frac{P(k \ B_1B_1 \ \text{sibs}, (s - k) \ B_1B_2 \ \text{sibs}, s \ \text{affected sibs}, B_1B_2 \times B_1B_1)}{P(s \ \text{affected sibs}, B_1B_2 \times B_1B_1)} , \end{aligned} \quad (C2)$$

where

$$P(k \ B_1B_1 \ \text{sibs}, (s - k) \ B_1B_2 \ \text{sibs}, s \ \text{affected sibs}, B_1B_2 \times B_1B_1) = \binom{s}{k} [(\mathbf{g}_{11}^k)^\circ (\mathbf{g}_{12}^{s-k})]^t \psi_1 ,$$

and

$$P(s \ \text{affected sibs}, B_1B_2 \times B_1B_1) = [(\mathbf{g}_{11} + \mathbf{g}_{12})^s]^t \psi_1 . \quad (C3)$$

For a family with parental marker genotype $B_1B_2 \times B_1B_2$ and s affected sibs, suppose that there are k_1 sibs with marker B_1B_1 , k_2 sibs with marker B_1B_2 , and $k_3 = s - k_1 - k_2$ sibs with marker B_2B_2 . Let $\mathbf{g}_{2.11}$, $\mathbf{g}_{2.12}$, and $\mathbf{g}_{2.22}$ be the column vectors whose elements are listed, in table C2, under the “ B_1B_1 ,” “ B_1B_2 ,” and “ B_2B_2 ” column heads, respectively. Let ψ_2 be the vector of mating-type probabilities listed, in table C2, under the “Mating Type Probabilities.” column head. Then,

$$P(k_1 \ B_1B_1, k_2 \ B_1B_2, k_3 \ B_2B_2 | s \ \text{affected sibs}, B_1B_2 \times B_1B_2) = \frac{s!}{k_1! k_2! k_3!} \frac{[\mathbf{g}_{2.11}^{k_1} \mathbf{g}_{2.12}^{k_2} \mathbf{g}_{2.22}^{k_3}]^t \psi_2}{[(\mathbf{g}_{2.11} + \mathbf{g}_{2.12} + \mathbf{g}_{2.22})^s]^t \psi_2} .$$

The conditional probability of a particular mating type, given that there are s affected sibs and that at least one parent is heterozygous, can be calculated on the basis of the equations given above; for example,

$$P(B_1B_2 \times B_1B_1 | s \ \text{affected sibs}, \mathcal{H}) = \frac{P(s \ \text{affected sibs}, B_1B_2 \times B_1B_1)}{P(s \ \text{affected sibs}, \mathcal{H})} , \quad (C4)$$

where the numerator can be calculated by means of expression (C3), and where the denominator is

$$\begin{aligned} P(s \ \text{affected sibs}, \mathcal{H}) &= P(s \ \text{affected sibs}, B_1B_2 \times B_1B_1) + P(s \ \text{affected sibs}, B_1B_2 \times B_2B_2) \\ &\quad + P(s \ \text{affected sibs}, B_1B_2 \times B_1B_2) . \end{aligned}$$

Now the conditional probabilities defined in the section entitled “Power Comparison, between the TDT, Mean Test, and DMLB, for ASP Data” can be computed easily, by means of equations (C1), (C2), and (C4) together with tables C1 and C2. Table 1 is for the mating type of one heterozygous parent, and table C2 is for the mating type in which both parents are heterozygous.

References

- Abel L, Alcais A, Mallet A (1998) Comparison of four sib-pair linkage methods for analyzing sibships with more than two affecteds: interest of the binomial-maximum-likelihood approach. *Genet Epidemiol* 15:371–390
- Abel L, Müller-Myhsok B (1998) Robustness and power of the maximum-likelihood-binomial and maximum-likelihood-score methods, in multipoint linkage analysis of affected sibship data. *Am J Hum Genet* 63:638–647
- Badner JA, Chakravatri A, Wagener DK (1984) A test of non-random segregation. *Genet Epidemiol* 1:329–340
- Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet Epidemiol* 2:85–97
- Boehnke M, Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *Am J Hum Genet* 62:950–961
- Camp NJ (1997) Genomewide transmission/disequilibrium testing: consideration of the genotype relative risks at disease loci. *Am J Hum Genet* 61:1424–1430
- Chernoff H (1954) On the distribution of the likelihood ratio. *Ann Math Stat* 25:573–578
- Chernoff H, Lander E (1995) Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *J Stat Plann Inference* 43:19–40
- Clerget-Darpoux F (1982) Bias of the estimated recombination fraction and lod score due to an association between a disease gene and a marker gene. *Ann Hum Genet* 46:363–372
- Cox DR, Hinkley DV (1974) *Theoretical statistics*. Halsted Press, New York
- Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for finescale mapping. *Genomics* 29:311–322
- Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculation. *Ann Hum Genet* 51:227–233
- Guo SW (1997) Linkage disequilibrium measures for fine-scale mapping: a comparison. *Hum Hered* 47:301–314
- Horvath S, Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am J Hum Genet* 63:1886–1897
- Knapp M (1999) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. *Am J Hum Genet* 64:861–870
- Majumder PP, Pal N (1987) Non-random segregation: uniformly most powerful test and related considerations. *Genet Epidemiol* 4:277–287
- Ott J (1989) Statistical properties of the haplotype relative risk. *Genet Epidemiol* 6:127–130
- Ott J (1991) *Analysis of human genetic linkage*, rev ed. John Hopkins University Press, Baltimore
- Risch N (1990) 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
- Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273:1516–1517
- Rubinstein P, Walker M, Carpenter C, Carrier C, Krassner J, Falk C, Ginsberg F (1981) Genetics of HLA disease associations: the use of the haplotype relative risk (HRR) and the “haplotype-delta” (Dh) estimates in juvenile diabetes from three racial groups. *Hum Immunol* 3:384
- Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* 13:423–449
- Self SG, Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J Am Stat Assoc* 82:605–610
- Sham PC, Curtis D (1995) An extended transmission/equilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet* 59:323–336
- Slager SL, Huang J, Vieland VJ. The effect of allelic heterogeneity on the power of the transmission disequilibrium test. *Genet Epidemiol* (in press)
- Smith CAB (1961) Homogeneity test for linkage data. *Proc Sec Int Congr Hum Genet* 1:212–213
- Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage equilibrium and association. *Am J Hum Genet* 59:983–989
- (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* 62:450–458
- Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506–516
- Szabo CI, King M-C (1997) Population genetics of BRCA1 and BRCA2. *Am J Hum Genet* 60:1013–1020
- Terwilliger JD, Weiss KM (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr Opin Biotechnol* 9:578–594
- Tu I-P, Whittemore AS (1999) Power of association and linkage tests when disease alleles are unobserved. *Am J Hum Genet* 64:641–649
- Tysoe C, Whittaker J, Xuereb J, Cairns NJ, Cruts M, van Broeckhoven C, Wilcock G, et al (1998) A presenilin-1 truncating mutation is present in two cases with autopsy-confirmed early-onset Alzheimer disease. *Am J Hum Genet* 62:70–76
- 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
- Welsh MJ, Smith AE (1995) Cystic fibrosis. *Sci Am* 273(6):52–59
- Whittemore AS (1996) Genome scanning for linkage: an overview. *Am J Hum Genet* 59:704–716
- Xiong M, Guo SW (1998) The power of linkage detection by the transmission-disequilibrium tests. *Hum Hered* 48:295–312