Am. J. Hum. Genet. 63:1886–1897, 1998

A Discordant-Sibship Test for Disequilibrium and Linkage: No Need for Parental Data

Steve Horvath and Nan M. Laird

Department of Biostatistics, Harvard School of Public Health, Boston

Summary

The sibship disequilibrium test (SDT) is designed to detect both linkage in the presence of association and association in the presence of linkage (linkage disequilibrium). The test does not require parental data but requires discordant sibships with at least one affected and one unaffected sibling. The SDT has many desirable properties: it uses all the siblings in the sibship; it remains valid if there are misclassifications of the affectation status; it does not detect spurious associations due to population stratification; asymptotically it has a χ^2 distribution under the null hypothesis; and exact *P* values can be easily computed for a biallelic marker. We show how to extend the SDT to markers with multiple alleles and how to combine families with parents and data from discordant sibships. We discuss the power of the test by presenting sample-size calculations involving a complex disease model, and we present formulas for the asymptotic relative efficiency (which is approximately the ratio of sample sizes) between SDT and the transmission/disequilibrium test (TDT) for special family structures. For sib pairs, we compare the SDT to a test proposed both by Curtis and, independently, by Spielman and Ewens. We show that, for discordant sib pairs, the SDT has good power for testing linkage disequilibrium relative both to Curtis's tests and to the TDT using trios comprising an affected sib and its parents. With additional sibs, we show that the SDT can be more powerful than the TDT for testing linkage disequilibrium, especially for disease prevalence >.3.

Received March 16, 1998; accepted for publication September 16, 1998; electronically published December 7, 1998.

Address for correspondence and reprints: Prof. Nan M. Laird, Department of Biostatistics, Harvard School of Public Health, 667 Huntington Avenue, Boston, MA 02115. E-mail: laird@hsph.harvard.edu

0002-9297/98/6306-0034$02.00

Introduction

Family-based association tests between a marker and a disease locus have become popular mainly because of two reasons: (*a*) in the case of tight linkage between a marker and a disease locus, one can find an association even when it is difficult to detect linkage (Spielman and Ewens 1993; Risch and Merikangas 1996); and (*b*) these tests protect against detection of spurious associations that are due to population stratification. The transmission/disequilibrium test (TDT) (Spielman and Ewens 1993) is a prime example of such a test. It can be used as a linkage test; it can also be used as a test of linkage disequilibrium (Ott 1989), provided that either only one sib per family is used or the test is adjusted to take correlation between sibs into account (Cleves et al. 1997; Martin et al. 1997).

Many family-based association tests compare the alleles or genotypes transmitted versus those that were not transmitted to an affected child. These tests can be quite powerful, but they have one serious limitation: in general, they are applicable only if the marker data on both parents are available, although partial information can sometimes be used if only one parent is available (Curtis and Sham 1995; Curtis 1997). Since it is difficult to obtain parental data for late-onset diseases, there is a need for family-based association tests that do not require parental data.

Curtis (1997) has introduced a discordant-sibship test for association that compares the allele frequencies of sib pairs that are sampled from discordant sibships according to the following procedure: for each discordant sibship, randomly sample one affected sibling (case) and then choose (randomly, if necessary) an unaffected sibling (control) whose genotype is maximally different from that of the case. The sampling of maximally discordant sib pairs avoids the introduction of correlation terms arising from the use of multiple sibs; it may lead to a loss of some information, especially when there are several affected siblings.

Spielman and Ewens (1998) have introduced the sib-TDT (S-TDT), which tests for linkage by using discordant-sibship data. The S-TDT is similar in spirit to the Mantel-Haenszel test (see Laird et al. 1998 [in this issue])

and provides a general test for linkage. It will only be a valid test of linkage disequilibrium either for sib pairs or when one also assumes that there is no linkage, since it requires siblings to be independent, under the null hypothesis. For sib pairs, it is identical to the test suggested by Curtis (1997). Boehnke and Langefeld (1998) have considered tests using discordant sib pairs, focusing on the multiallelic case. They use Pearson-type χ^2 statistics for homogeneity and symmetry, evaluating the test significance by using premutation distributions.

In this study, we introduce a discordant-sibship test that uses the data of all the affected and unaffected siblings. It can be used both as a linkage test and as a linkage-disequilibrium test; it allows the computation of exact *P* values; and it can be quite powerful. We call it "SDT" (sibship disequilibrium test), because it will be particularly useful as a disequilibrium test and because of its structural similarity to the TDT. In subsequent sections, we compare the SDT to the S-TDT and Curtis's test in the case of sib pairs. Following Spielman and Ewens (1998), we also compare our test to the TDT. If the SDT compares favorably, then it can make sense to use study designs that do not ascertain parents. Comparing the SDT only to the S-TDT (or to Curtis's test) could lead to a situation in which the SDT is as powerful as the S-TDT (or Curtis's test) but in which all tests compare unfavorably with the TDT; here one should make every effort to obtain parental information. In most situations, the natural issue to address is how many more families (with discordant sibs) should be sampled if parents are not available.

The SDT Test

SDT for Discordant Sibships

We begin with the case of a biallelic marker with alleles denoted as "0" and "1." For each sibship, denote by "m^1_A (m^1_U)" the mean number of 1 alleles among the affected (unaffected) siblings; that is,

$$\begin{aligned} m^1_A &= (\text{total number of 1 alleles} \\ &\qquad \text{among the affecteds})/n_A \\ m^1_U &= (\text{total number of 1 alleles} \\ &\qquad \text{among the unaffecteds})/n_A \ , \end{aligned} \tag{1}$$

where n_A (n_U) denotes the number of affected (unaffected) members in the sibship.

For example, if a sibship has three affected sibs with genotypes (1,1), (1,0), and (1,1) and one unaffected sib with genotype (0,0), then $m^1_A = 5/3$ and $m^1_U = 0/1$. Let d^1 denote the difference $m^1_A - m^1_U$.

We define the SDT to be a (nonparametric) sign test

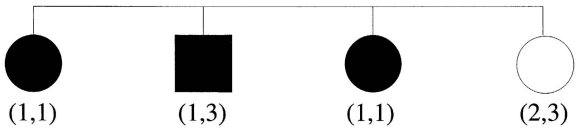

Figure 1 Sibship marker-genotype configuration

on these differences. If $d^1 = 0$ for a sibship, then we discard that sibship from the analysis. Let b be the number of sibships for which $d^1 > 0$ and let c be the number for which $d^1 < 0$. We define the SDT statistic as

$$T = (b - c)^2/(b + c) \ . \tag{2}$$

In Appendices A–C, we provide a proof that the SDT statistic asymptotically has a $\chi^2_{(1)}$ distribution under the null hypothesis of no linkage or no linkage disequilibrium (H_0:$\Delta(\theta - 1/2) = 0$). The exact version of the sign test allows us to compute exact *P* values; for example, a two-sided exact *P* value is given by

$$P_{\text{two-sided}} = 2\min\left[\sum_{i=0}^{b}\binom{b+c}{i}\left(\frac{1}{2}\right)^{b+c}, \sum_{i=b}^{b+c}\binom{b+c}{i}\left(\frac{1}{2}\right)^{b+c}\right] .$$

Extending the SDT to Multiple Alleles

For a marker with two alleles, the SDT is a (nonparametric) sign test on differences d^1. Similarly, we define the SDT for a marker with m alleles (denoted as "0,1, … ,$m - 1$") to be a multivariate sign test on quantities d^j, as follows: d^j ($j = 0, \dots, m - 1$) denotes the difference $m^j_A - m^j_U$, where $m^j_A(m^j_U)$ equals the average number of j alleles in the affected (unaffected) members of the sibship (see eq. [1]).

Figure 1 shows a sibship with three affected sibs and one unaffected sib. For the sibship in figure 1, we find that $m^1_A = 5/3$ and $m^1_U = 0/1$, $m^2_A = 0/3$, $m^2_U = 1/1$, $m^3_A = 1/3$, and $m^3_U = 1/1$ and, hence, $d^1 = 5/3$, $d^2 = -1/1$ and $d^3 = -2/3$. Note that $\Sigma_j m^j_A = \Sigma_j m^j_U = 2$ implies $d^0 = -\Sigma_{j=1}^m d^j$ and that we therefore can drop d^0 without losing any information.

There are several multivariate sign tests, but we define the multiallelic SDT by using the most popular one, known as the "component" sign test (Bickel 1965; Randles 1989): Let $\mathbf{S}' = (S^1, \dots, S^{m-1})'$, where $S^j = \Sigma_{i=1}^N \text{sgn}(d^j_i)$, d^j_i denotes the difference for the ith sibship, and $\text{sgn}(d) = 1(0, -1)$ as $d > (=, <) 0$. The test rejects H_0:$E(S) = 0$ for large values of $T = \mathbf{S}'\mathbf{W}^{-1}\mathbf{S}$, where the matrix $\mathbf{W}$ has elements $W_{jk} = \Sigma_{i=1}^N \text{sgn}(d^j_i)\text{sgn}(d^k_i)$ ($j,k = 1, \dots, m - 1$). One can verify that, in the case of a biallelic marker, T reduces to the biallelic SDT given in equa-

tion (2). In Appendix B, we show that, under the null hypothesis of no linkage or equilibrium, T asymptotically has a $\chi^2_{(m-1)}$ distribution.

A Class of Association Tests for Discordant Sib Pairs

A Class of Discordant-Sib-Pair Tests

In this section we focus on the case in which the data consist of N discordant sib pairs because other tests of association are available for this case and because analytic results are possible. We study only two marker alleles, 0 and 1. For each sib pair, we obtain a pair of numbers (i,j), where i denotes the number of 1 alleles in the affected sib and j denotes the number of 1 alleles in the unaffected sib ($i,j = 0,1,2$). Intuitively, we expect that, if the 1 allele is positively associated with the disease, then there will be more pairs (2,0) than pairs (0,2). Also, we expect to see more (2,1) or (1,0) pairs than (1,2) or (0,1) pairs. Denote by b_2 the number of pairs (2,0), and denote by c_2 the number of pairs (0,2). Further, denote by b_1 the number of pairs (2,1) or (1,0) and denote by c_1 the number of pairs (1,2) or (0,1). We will now introduce a class of test statistics T_x that can be used to test the null hypothesis of either no linkage or no linkage disequilibrium (H_0:$\Delta(\theta - 1/2) = 0$):

$$T_x = \frac{b_1 - c_1 + x(b_2 - c_2)}{\sqrt{b_1 + c_1 + x^2(b_2 + c_2)}} ,$$

where $x > 0$. Under the null hypothesis of no linkage or linkage disequilibrium, T_x has an asymptotic $N(0,1)$ distribution. T_2 corresponds to the test statistic introduced by Curtis (1997), which is equal to the S-TDT in the case of sib pairs, whereas T_1 corresponds to the SDT. Note that T_1 is based on simply the number of sib pairs with more 1 alleles in the affected sibling than in the unaffected sibling, whereas the T_2 test gives twice the weight to (2, 0) sib pairs than to (2, 1) or (1, 0) sib pairs. As noted by a referee (Duncan Thomas), T_2 is a score test of H_0 under an additive genetic model for risk. Similar expressions can be derived for score statistics when dominant or recessive models are assumed (e.g., see Schaid 1996). T_1 is not a score test for any genetic model, but it has many attractive properties, including simplicity and ease of extension to multiple sibs.

Comparison of the SDT to Both the S-TDT and the Curtis (1997) Test, in the Case of Sib Pairs

Assume that, for T_1 and T_2, one needs to collect on average of n_1 and n_2 sib pairs, respectively, to reject the null hypothesis, at significance level α with power β. Then, one way to compare the efficiencies is to form the quotient n_1/n_2. In general, it is difficult to find an analytic expression for this quotient; an approximation to it, known as the "asymptotic relative efficiency" (ARE[T_2:T_1]; Serfling 1981, pp. 317–319), may be defined, in the limiting cases, as either the linkage-disequilibrium parameter $\Delta \to 0$ or the recombination parameter $\theta \to 1/2$. In Appendix D, we also review the ARE, but here we focus on the results.

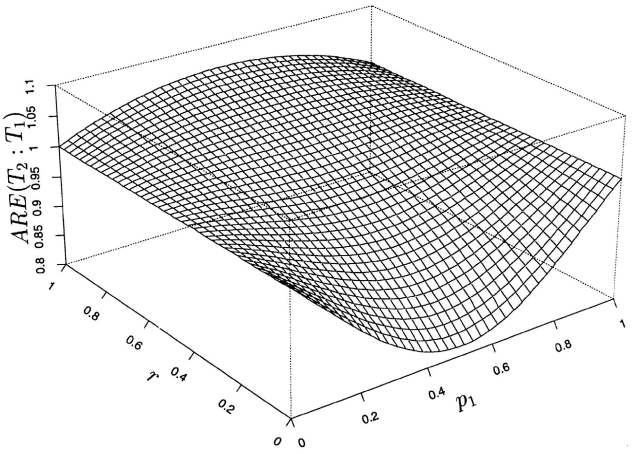

Figure 2 Graph of ARE(T_2:T_1), which is the expected number of discordant sib pairs required for T_1, divided by the expected number of discordant sib pairs required for T_2. p_1 is the marker-allele frequency, and r is defined by $r = (1 - K_O)/(1 - K_P)$.

We begin with the case in which linkage between gene and marker has already been established and one wants to test the null hypothesis of no linkage disequilibrium (H_0:$\Delta = 0$). Then, for $\theta \neq 1/2$, the ARE is given by

$$\mathrm{ARE}(T_2{:}T_1) = \frac{4 - 6p_1(1 - p_1)W}{[2 - p_1(1 - p_1)]^2} ,$$

where

$$W = [16(K_O - K_S)\theta^2(1 - \theta)^2 + 4(K_P - K_O) \times\theta(1 - \theta) + 1 - K_P]/ [4(K_P - 2K_S + K_O)\theta(1 - \theta) + 2 - (K_P + K_O)] .$$

Here p_1 is the allele frequency of the 1 allele, K_P is the disease prevalence when an arbitrary single-gene model is assumed, and K_O is the offspring relative risk when a single-gene model is assumed. The ARE has been defined such that T_2 is more powerful than T_1, if $\mathrm{ARE}(T_2{:}T_1) > 1$.

Since the ARE changes very little with θ, we set $\theta = 0$, so that $W = (1 + r)^{-1}$, where $r = (1 - K_O)/(1 - K_P)$. In figure 2 we have plotted the ARE as a function of r and p_1. For fixed p_1, the ARE is maximized when r reaches its maximum of 1, at $K_O = K_P$. In this case, ARE is maximized at 1.06 for $p_1 = 1/2$. Again, for fixed p_1, the ARE is minimized when r reaches its minimum of

0, as K_O approaches 1. In this case, ARE reaches its minimum of 0.820 when $p_1 = 1/2$. Large values of K_O imply that K_P is large (see Suarez et al. 1976), and the minimum corresponds to an unrealistic situation. For realistic diseases, the calculations suggest that the power of the SDT is very similar to that of Curtis's test (and the S-TDT), when Δ and θ are close to 0 and only sib pairs are used.

When one is testing for linkage (H_0:$\theta = 1/2$), the ARE for $\Delta \neq 0$ is given by

$$\begin{aligned}\mathrm{ARE}(T_2{:}T_1) = {} & [4 - 3p_1(1-p_1)] \\ & \times p_D(1-p_D)(K_O - K_P)\Big/ \\ & \Big\{\sqrt{K_S - K_O}(2p_1 - 1)\Delta + \\ & \sqrt{(K_O - K_P)p_D(1-p_D)[2 - p_1(1-p_1)]}\Big\}^2 ,\end{aligned}$$

where p_D is the disease-allele frequency and K_S is the sibling recurrence risk under the single-gene model. When $K_S = K_O$, then the ARE is maximized at $p_1 = 1/2$, with a maximum of 1.06, and it is minimized, with a minimum of 1, as p_1 approaches 1 or 0. In conclusion, when one is testing linkage, T_2 is always slightly more powerful than T_1, but, when one is testing equilibrium, neither test is uniformly more powerful. The two tests have similar power in most situations. The features that distinguish the SDT are its simplicity and the ease with which it can be generalized to sibships with several affected and unaffected siblings and still remain a valid test of association.

Comparison of T_x *to the TDT*

One might expect that the TDT is always more powerful than a test based only on sibs, but we will show here that this is not always the case: unaffected sibs can carry considerable information about both linkage and linkage disequilibrium, when the disease prevalence is high. We explore this issue by comparing the number of family trios (an affected sib and its parents) that need to be collected for the TDT versus the number of discordant sib pairs that need to be collected for the SDT to reject the null hypotheses specified below. In another section, we determine the number of sibship triplets required for a specific complex disease. The ARE, ARE(TDT,T_x), is defined such that it is approximately equal to the number of discordant sib pairs, divided by the number of family trios, that are required to reject H_0 at a given significance level and a given power for Δ close to 0.

The ARE can be compared to different benchmarks κ: the TDT is more cost efficient than T_x when ARE > κ. Denote the cost of ascertaining a family trio and a discordant sib pair as "A(trio)" and "A(pair)," respectively, and denote the cost of genotyping an individual as "G." Then one might reasonably define $\kappa = [A(\text{trio}) + 3G]/[A(\text{pair}) + 2G]$. For example, if the ascertainment costs are negligible and the cost is driven by genotyping, then $\kappa = 3/2$.

We begin with a case in which linkage has been established and one tests linkage equilibrium, H_0:$\Delta = 0$. When the methods described in Appendix D are used, the ARE between the TDT and T_2, for $\theta \neq 1/2$, is given by

$$\begin{aligned}\mathrm{ARE}(\mathrm{TDT}{:}T_2) = {} & 4(K_P - 2K_S + K_O)(1 - K_S) \\ & \times\theta(1-\theta) + (1 - K_S)(2 - K_O - K_P) . \qquad (3)\end{aligned}$$

Equation (3) can be approximated by setting $\theta = 0$, to obtain $\mathrm{ARE}(\mathrm{TDT}{:}T_2|\theta = 0) = (1 - K_S)(2 - K_P - K_O)$. How much can the sample sizes differ between the tests? For a fixed K_P, the ARE approaches $\mathrm{ARE}_{\max}(\mathrm{TDT},T_2) = 2(1 - K_P)^2$ as K_S and K_O approach K_P. This is >1 for $K_P < .3$ and >3/2 for $K_P < .13$; as $K_P \to 0$, it approaches a maximum of 2. The ARE approaches $\mathrm{ARE}_{\min} = 0$ as K_S (and hence K_P and K_O) approach 1. The last result can be explained as follows: when almost every sibling of an affected child is also affected, then discordant sib pairs carry a lot of information; thus, tests that are based on them can be much more powerful than the TDT, which is applied to affecteds only. However, high values of K_S are only possible when the disease prevalence K_P is high (Suarez et al. 1976), and in this situation it could be better in a TDT analysis to use (or include) unaffected sibs. Thus we have the interesting result that the asymptotic efficiency of the TDT can only be better than that of T_2 by a factor of 2, for small K_P and K_S and K_O close to K_P. On the other hand, for diseases with high prevalence, there is potential for considerable loss of efficiency when the discordant sibs are not used.

The ARE between the TDT and T_1 is given by

$$\mathrm{ARE}(\mathrm{TDT},T_1) = \mathrm{ARE}(\mathrm{TDT},T_2)\mathrm{ARE}(T_2,T_1) . \qquad (4)$$

For $p_1 = 1/2$ the ARE approaches $\mathrm{ARE}_{\max}(\mathrm{TDT},T_1) = 2.12(1 - K_P)^2$ as K_S and K_O approach K_P, and it approaches $\mathrm{ARE}_{\min} = 0$ as K_S approaches 1.

When one is testing for linkage (H_0:$\theta = 1/2$), the ARE between the TDT and T_2 is given by

$$\begin{aligned}\mathrm{ARE}(\mathrm{TDT}{:}T_2) = {} & \left[4\sqrt{p_D(1-p_D)}p_1(1-p_1)(1-K_S)^2\right]\Big/ \\ & \left[2\sqrt{p_D(1-p_D)}p_1(1-p_1) - \Delta(1-2p_1)\sqrt{K_O/K_P - 1}\right] .\end{aligned}$$

If $p_1 = 1/2$, this equation reduces to $\mathrm{ARE}(\mathrm{TDT}{:}T_2|p_1 = 1/2) = 2(1 - K_S)^2$. Thus, if $p_1 = 1/2$, the TDT is more powerful than T_2 if $K_S < 1 - .707 \times \sqrt{\kappa}$. The ARE be-

tween the TDT and T_1 when one is testing for linkage can again be computed with the aid of equation (4).

In summary, when K_P—and thus K_O and K_S—are large, there is considerable information in the unaffected sibs, and tests based on discordant sibs may be more powerful than those based on family trios. This is not surprising, since the TDT effectively compares the allele distribution among affecteds versus that in the population. If the disease is rare, this can be a very powerful test, but it will be low in power if the disease is common. Recall that K_O and K_S are defined in terms of a single-gene model; for multigene models, they should not be interpreted generally as recurrence risks but as recurrence risks resulting from considering solely the gene in question.

Some Sample-Size Calculations

It is difficult to get simple expressions for the ARE between either the TDT or the S-TDT and the SDT for multiple sibs. Here we present the average sample sizes required by SDT and TDT for special situations: Assume that we test the null hypothesis $H_0{:}\Delta = 0$ versus $H_A{:}\Delta = \Delta_{max}, \theta = 0$; the alternative is satisfied if the marker allele is a disease allele. Again, we assume a biallelic marker and disease model, but we restrict the penetrances of the disease locus to satisfy a genotypic relative-risk model that has been described by Risch and Merikangas 1996: $f_{01} = f_{00}\gamma$ and $f_{11} = f_{00}\gamma^2$, where f_{00}, f_{01}, and f_{11} denote P(affected), given 0, 1, or 2 disease alleles, and $\gamma \geq 1$ is the genotypic risk ratio associated with the disease allele. The sibling and offspring recurrence risks λ_S and λ_O are defined by $\lambda_S = K_S/K_P$ and $\lambda_O = K_O/K_P$, respectively. For the multiplicative penetrance model, we get $\lambda_S = (1 + .5w)^2$ and $\lambda_O = 1 + w$, where $w = p_D(1 - p_D)(\gamma - 1)^2/(p_D\gamma + 1 - p_D)^2$, $K_P = f_{00}(p_D\gamma + 1 - p_D)^2$, and p_D is the disease-allele frequency (Risch and Merikangas 1996). In Appendix E we describe how to compute sample sizes for the SDT.

In table 1 we list (*a*) the values of λ_S and K_P for the different disease-locus parameters that we consider here and (*b*) the expected sample size required for the TDT to reject H_0 with a two-sided test at $Z_\alpha = 2.80$ with power $\beta = .80$ ($Z_\beta = .84$) when H_A is true. The conservative α level corresponds to correcting (by the Bonferroni method) an α level of .05 for 10 comparisons. We also give the ratio of asymptotic sample sizes for the SDT, based on discordant-sib-pair families, relative to the TDT, based on family trios. These figures are in good agreement with the formulas for the ARE, even though they are for different values of Δ under H_A and are derived differently.

In table 2, the TDT is applied to nuclear families that consist of family trios, as in table 1. The SDT is now applied in two different situations: the first consists of sibship triplets with one affected sib and two unaffected sibs (AUU sibships), and the second consists of triplets with two affected sibs and one unaffected sib (AAU sibships). Note that in these cases the number genotyped per family is equal. Although additional affected sibs can be added to test for linkage by use of the TDT, incorporating them to test for linkage disequilibrium is more complex (Martin et al. 1997). We here use the TDT as a benchmark, since the sample sizes are easy to obtain by means of the formulas of Risch and Merikangas (1996).

For a disease-locus model with the specific penetrance functions described above, the sample sizes for the AAU pairs (table 2) show that the SDT compares favorably with the TDT, when K_P is large.

Discussion

The SDT can be viewed as a (nonparametric) sign test that compares, within each family, the average number of alleles in affected sibs versus those in unaffected sibs. For discordant sib pairs, several family-based tests of linkage that are also valid in testing for linkage disequilibrium can be constructed. The critical issue is how to extend these tests to multiple sibships; tests of linkage may not be valid tests of linkage disequilibrium, because of correlation between siblings, when $\theta < 1/2$ and $\Delta = 0$.

The SDT is also a valid test of linkage, but there are alternatives—for example, the S-TDT (Spielman and Ewens 1998) with arbitrary discordant sibships, conditional logistic regression (Self et al. 1991) or Curtis's test (1997). The SDT is particularly well suited as a linkage-disequilibrium test ($H_O{:}\Delta = 0$), because it avoids having to account for correlation between the siblings.

How does the SDT fare with misclassifications—that is, cases in which affected sibs are classified as unaffected sibs, and vice versa? With misclassifications, the absolute value of the expected values of $m_A^j - m_U^j$ is reduced. This leads to a bias toward the null hypothesis with a reduction in power, but the test remains valid.

It is straightforward to combine TDT and SDT when the data consist of a mixture of families with and without parental information. Let b_{TDT} (c_{TDT}) denote the number of times that a heterozygous parent transmits the 1 allele to an affected sib, and let b_{SDT} (c_{TDT}) denote the number of discordant sibships without parental information, where $m_A^1 - m_U^1 > (<) 0$. Define $b = b_{TDT} + b_{SDT}$ and $c = c_{TDT} + c_{SDT}$. One can show that the statistic $Z^2 = (b - c)^2/(b + c)$ has a $\chi^2_{(1)}$ distribution under the null hypothesis of no linkage. If the families with parental information consist of family trios, then Z^2 has also a $\chi^2_{(1)}$ distribution under the null hypothesis of no linkage disequilibrium. Our own experience suggests that the

Table 1

Sample-Size Ratios for Various Values of γ_S and K_P

				Sample-Size Ratios (SDT:TDT)							
				$f_{11} = .1$		$f_{11} = .3$		$f_{11} = .5$		$f_{11} = .7$	
γ	p_D	λ_S	No. of Trios[a]	K_P	R^b	K_P	R^b	K_P	R^b	K_P	R^b
4	.01	1.086	348	.007	1.97	.020	1.89	.033	1.82	.046	1.74
4	.1	1.537	48	.011	2.04	.032	1.90	.053	1.75	.074	1.60
4	.5	1.392	33	.039	1.97	.117	1.61	.195	1.24	.273	.94
4	.8	1.128	71	.072	1.82	.217	1.34	.361	.93	.506	.58
2	.01	1.010	1,974	.026	1.89	.077	1.66	.128	1.44	.179	1.23
2	.1	1.076	236	.030	1.92	.091	1.64	.151	1.39	.212	1.15
2	.5	1.114	116	.056	1.88	.169	1.43	.281	1.03	.394	.70
2	.8	1.050	217	.081	1.77	.243	1.24	.405	.79	.567	.43
1.5	.01	1.003	6,662	.045	1.81	.135	1.45	.225	1.13	.314	.85
1.5	.1	1.021	765	.049	1.84	.147	1.45	.245	1.10	.343	.80
1.5	.5	1.040	328	.069	1.83	.208	1.31	.347	.88	.486	.53
1.5	.8	1.021	574	.087	1.74	.261	1.16	.436	.70	.610	.35

[a] No. of families required for significance level $\alpha = .005$ and power $\beta = .8$, when both parents are available, $\theta = 0$, and $\Delta = \Delta_{max}$.

[b] Expected number of discordant sib pairs, divided by the expected number of family trios, for a two-sided test with significance level $\alpha = .005$ and power $\beta = .8$.

Table 2

Sample-Size Ratios for Family Trios

			Sample-Size Ratios (SDT:TDT)							
			$f_{11} = .1$		$f_{11} = .3$		$f_{11} = .5$		$f_{11} = .7$	
γ	P_D	No. of Trios[a]	AUU	AAU	AUU	AAU	AUU	AAU	AUU	AAU
4	.01	348	1.35	1.14	1.35	1.06	1.35	.99	1.35	.92
4	.1	48	1.46	1.58	1.42	1.38	1.38	1.17	1.33	1.00
4	.5	33	1.46	2.73	1.24	1.97	1.03	1.36	.85	.88
4	.8	71	1.25	3.28	.94	2.20	.66	1.35	.42	.73
2	.01	1,974	1.31	1.79	1.24	1.47	1.17	1.19	1.10	.96
2	.1	236	1.37	1.87	1.26	1.50	1.16	1.18	1.05	.91
2	.5	116	1.40	2.33	1.13	1.60	.88	1.03	.66	.61
2	.8	217	1.26	2.78	.92	1.71	.61	.96	.36	.46
1.5	.01	6,662	1.27	2.04	1.13	1.48	.98	1.04	.84	.71
1.5	.1	765	1.32	2.04	1.14	1.46	.96	1.02	.79	.68
1.5	.5	328	1.37	2.19	1.05	1.42	.76	.85	.51	.46
1.5	.8	574	1.26	2.49	.90	1.46	.58	.77	.31	.34

[a] Defined as table 1.

[b] Ratio is the expected number of discordant-sibship trios, divided by the expected number of family trios.

SDT is useful as an independent test even when sufficient parental data are available to allow one to perform the standard TDT by using family trios.

Acknowledgment

We would like to thank Dave Curtis, Warren Ewens, Kathy Lunetta, John Rogus, Duncan Thomas, Xiaolin Wang, Marsha Wilcox, and, especially, Xin Xu for their valuable help. Support for N.M.L. was provided by National Institutes of Health (NIH) grants GM29745 and MH 59536; support for S.H. was provided by NIH grant GM29745.

Appendix A

Conditional Probability of Marker Data

For proving that the SDT is a valid test of either H_0:$\Delta = 0$ or H_0:$\theta = 1/2$, the general strategy is to first derive a formula for the probability of the marker distribution in an arbitrary nuclear family, given the affectation status of the sibs; this is done in here, in Appendix A. Then, in Appendixes B and C, the formula is used to

show that, for an arbitrary discordant sibship, $P(d^1 > 0 | d^1 \neq 0) = 1/2$ if either $\theta = 1/2$ or $\Delta = 0$.

To obtain the required probability distribution, we extend the ordered notation introduced by Thomson (1995), to allow for families with arbitrary numbers of affected and unaffected sibs. The ordered notation gives a straightforward description of the passage of transmitted and nontransmitted marker alleles through a family. The basic feature is that the marker-locus genotype of the first (i.e., affected) sib is used to order the four parental marker alleles. The genotype of the father is described such that the marker allele transmitted to the first sib is listed before the nontransmitted allele, and the genotype of the mother is described similarly. Marker alleles in all subsequent sibs are described relative to that in the first sib. Figure A1 shows the ordered notation in the case in which the family has three affected sibs and one unaffected sib. In the following, we will label the alleles by the letter "I"; for example, the marker alleles of the father are I_1 and I_2, those of the mother are I_3 and I_4, and the marker genotype of the first sib is I_1,I_3. For the biallelic case, the letter "I" has been chosen to point out that I can also be considered as an indicator variable that indicates whether the 1 allele is present. The four possible genotypes of the mth additional affected sib in these cases are described relative to that of the first sib, by the four-dimensional vector $\mathbf{A}^m$. Similarly, we describe the genotypes of the nth unaffected sib by the four-dimensional vector $\mathbf{U}^n$. Again, as is shown in figure A1, $I_1 = 1$, $I_2 = 0$, $I_3 = 1$, $I_4 = 0$; and $\mathbf{A}^1 = (1,0,1,0)$, meaning that the first sib inherited I_1 and I_3, $\mathbf{A}^2 = (1,0,0,1)$, meaning that the sib inherited I_1 and I_4, etc. Note that, whereas Thomson uses "δ" to label siblings, we use "A" and "U" in all our formulas, to clearly denote affectation status.

We are now ready to give an explicit expression of the conditional probability of observing the family marker data, given the number of affected and unaffected sibs. We begin with an example: we denote by $\mathrm{MS}(\theta)\mathbf{A}_{(I_1,I_2,I_3,I_4)}(1,0,0,1)(1,0,1,0)[0,1,0,1]$ the probability of observing the marker data in figure A1, given that the sibs consist of three affected and one unaffected. Note that parentheses are used for affected and that square brackets are used for unaffected sibs. In general, we denote by $\mathrm{MS}(\theta)\mathbf{A}_I\mathrm{A}^2 \ldots \mathbf{A}^{n_\mathrm{A}}\mathrm{U}^1 \ldots \mathbf{U}^{n_\mathrm{U}}$ the probability of observing the marker data described by the four-dimensional vectors $I,\mathbf{A}^2 \ldots \mathbf{A}^{n_\mathrm{A}}\mathbf{U}^1 \ldots \mathbf{U}^{n_\mathrm{U}}$, given that there are n_A affected sibs and n_U unaffected sibs.

Before we can give an explicit expression of $\mathrm{MS}(\theta)\mathbf{A}_I\mathrm{A}^2 \ldots \mathbf{A}^{n_\mathrm{A}}\mathrm{U}^1 \ldots \mathbf{U}^{n_\mathrm{U}}$ in terms of the parameters of the disease and the marker locus, we have to introduce more notation. We assume a single-gene model, with two disease alleles indexed by D_0, D_1, and m marker alleles indexed by $j = 0,1, \ldots m-1$. The marker-allele frequencies are denoted "p_j" and the disease frequencies

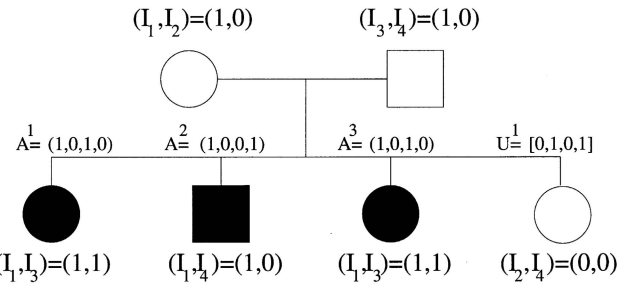

Figure A1 Nuclear-family marker-genotype configuration

are denoted by "p_{D_0}" and "p_{D_1}." The penetrance fractions are f_{00}, f_{01}, and f_{11}—and are expressed as P(affected)—for 0, 1, and 2 D_1 alleles, respectively. The linkage-disequilibrium parameter will be defined, for the biallelic-marker case, as $\Delta = P(\text{marker allele} = 1 \text{ and disease allele} = 1) - p_1 p_{D_1}$. We define the following functions of the disease-allele penetrances and the recombination fraction θ:

$$\begin{aligned}
f_{(1,0,1,0)(r,s,t,u)}(\theta) &= (1-\theta)^2 f_{rt} + \theta(1-\theta) f_{ru} \\
&\quad + \theta(1-\theta) f_{st} + \theta^2 f_{su} , \\
f_{(1,0,0,1)(r,s,t,u)}(\theta) &= (1-\theta)^2 f_{ru} + \theta(1-\theta) f_{rt} \\
&\quad + \theta(1-\theta) f_{su} + \theta^2 f_{st} , \\
f_{(0,1,1,0)(r,s,t,u)}(\theta) &= (1-\theta)^2 f_{st} + \theta(1-\theta) f_{su} \\
&\quad + \theta(1-\theta) f_{rt} + \theta^2 f_{ru} , \\
f_{(0,1,0,1)(r,s,t,u)}(\theta) &= (1-\theta)^2 f_{su} + \theta(1-\theta) f_{st} \\
&\quad + \theta(1-\theta) f_{ru} + \theta^2 f_{rt} . \qquad \text{(A1)}
\end{aligned}$$

Let $\mathbf{A}^r = (\mathrm{A}^r_1,\mathrm{A}^r_2,\mathrm{A}^r_3,\mathrm{A}^r_4)$ be the vector that codes the marker-genotype data for the rth affected sib. For convenience, we will often drop the subscripts (r,s,t,u) in equation (A1) and simply use "$f_{\mathrm{A}^m}(\theta)$"; for example, $f_{(1,0,1,0)}(\theta) = f_{(1,0,1,0)(r,s,t,u)}(\theta)$. Here is a final piece of notation, involving penetrances, that we need for equation (A3):

$$\phi_{rstu} = (f_{rt} + f_{ru} + f_{st} + f_{su})/4 . \qquad \text{(A2)}$$

Similarly, we define $\bar{\phi}_{rstu} = (\bar{f}_{rt} + \bar{f}_{ru} + \bar{f}_{st} + \bar{f}_{su})/4$, where $\bar{f} = 1 - f$. The probability $\mathrm{PAU}(n_\mathrm{A},n_\mathrm{U})$ of a sibship with n_A affected and n_U unaffected members is given by

$$\mathrm{PAU}(n_A,n_U) = \sum_{r,s,t,u=0}^{1} 4^{n_A+n_U-1} f_{rt}\phi_{rstu}^{n_A-1}\bar{\phi}_{rstu}^{n_U} p_{D_r}p_{D_s}p_{D_t}p_{D_u} \ . \quad \text{(A3)}$$

By generalizing the results of Thomson (1995), we can express $\mathrm{MS}(\theta)\mathbf{A}_I\mathrm{A}^2 \ldots \mathbf{A}^{n_A}\mathrm{U}^1 \ldots \mathbf{U}^{n_U}$ as

$$\sum_{r,s,t,u=0}^{1} \Big[f_{A^1}(\theta) \ldots f_{A^{n_A}}(\theta)\bar{f}_{U^1}(\theta) \ldots \bar{f}_{U^{n_U}}(\theta) \times k_{I_1r}k_{I_2s}k_{I_3t}k_{I_4u}p_{D_r}p_{D_s}p_{D_t}p_{D_u}\Big] \Big/ \mathrm{PAU}(n_A,n_U) \ , \quad \text{(A4)}$$

where k_{I_1r} equals the conditional probability that marker allele I_1 will be observed on a haplotype that contains disease allele D_r. If marker and disease locus are in linkage equilibrium, then $k_{I_1r} = p_{I_1}$.

Appendix B

Distribution of SDT, under H_0

We begin with the case of a biallelic marker. We outline how to prove that, under H_0:$\Delta(\theta - 1/2) = 0$, the biallelic SDT statistic (eq. [2]) has an asymptotic $\chi^2_{(1)}$ distribution. By definition, b (c) counts the number of sibships for which $d^1 = m_A^1 - m_U^1 > 0$ ($d^1 < 0$), given that $d^1 \neq 0$. Define $p_>$ ($p_<$) as the probability that $d^1 > 0$ ($d_1 < 0$), and define π as the conditional probability that $d^1 > 0$, given that $d_1 \neq 0$; that is,

$$\pi = p_>/(p_> + p_<) \ . \quad \text{(B1)}$$

We will show that, under H_0:$\Delta(\theta - 1/2) = 0$, $\pi = 1/2$ for any discordant sibship. Therefore, given $b + c$, b has the binomial distribution $b \sim \mathrm{binomial}(b + c,1/2)$, and one can use the central-limit theorem to show that the SDT statistic $T = (b - c)^2/(b + c)$ asymptotically has a $\chi^2_{(1)}$ distribution. To show that $\pi = 1/2$, we will prove that $p_> = p_<$ (see eq. [B1]). Note that $p_>$ can be expressed as

$$p_> = P(d^1 > 0 | n_A \text{ affected}, n_U \text{ unaffected}) \ ; \quad \text{(B2)}$$

an analogous equation exists for $p_<$. Thus, we need to show that, under H_0,

$$P(d^1 > 0 | n_A,n_U) = P(d^1 < 0 | n_A,n_U) \ . \quad \text{(B3)}$$

We express $P(d^1 > 0 | n_A)$ explicitly as

$$\sum_I \sum_{A^2,\ldots,A^{n_A}} \sum_{U^1,\ldots,U^{n_U}} \mathrm{MS}(\theta)\mathrm{A}_I\mathrm{A}^2 \ldots \mathrm{A}^{n_A}\mathrm{U}^1 \ldots \mathrm{U}^{n_U} I(d^1 > 0) \ ,$$

where summation over a four-dimensional vector denotes the sum over all four possible index values —$(I_1,0,I_3,0)$, $(I_1,0,0,I_4)$, $(0,I_2,I_3,0)$, and $(0,I_2,0,I_4)$ —and where $I(d^1 > 0)$ is an indicator variable that is 1 if $d^1 > 0$ and is 0 otherwise; an analogous equation exists for $p_<$.

To be able to proceed in the proof, we introduce the concept of a "mirror image" for marker-data probabilities: The "mirror image" $\mathrm{MS}(\theta)\mathbf{A}_{\bar{I}}\mathrm{A}^2 \ldots \mathbf{A}^{n_A}\mathrm{U}^1 \ldots \mathbf{U}^{n_U}$ of $\mathrm{MS}(\theta)\mathrm{A}_{(I_1,I_2,I_3,I_4)}\mathrm{A}^2 \ldots \mathrm{A}^{n_A}\mathrm{U}^1 \ldots \mathrm{U}^{n_U}$ is defined by setting

$$\mathrm{MS}(\theta)\mathbf{A}_{\bar{I}}\mathrm{A}^2 \ldots \mathbf{A}^{n_A}\mathrm{U}^1 \ldots \mathbf{U}^{n_U} = \mathrm{MS}(\theta)\mathrm{A}_{(I_2,I_1,I_4,I_3)}\mathrm{A}^2 \ldots \mathrm{A}^{n_A}\mathrm{U}^1 \ldots \mathrm{U}^{n_U} \ .$$

Note that the interchange of I_1 and I_2 simultaneously with I_3 and I_4 and then letting this permutation determine transmissions to all sibs is a special case of a permutation distribution described by Martin et al. (1997) for the parent-known case and testing of H_0:$\Delta = 0$. Making the permutation simultaneous in both parents avoids dealing with cases in which transmission status cannot be determined—that is, two (0,1) parents and two (0,1) sibs.

Note that the marker-data probability associated with figure A1 contributes to $p_>$ (since $d^1 = 5/3 - 0/1 > 0$), whereas the marker-data probability associated with figure B1 contributes to $p_<$ (since $d^1 = 1/3 - 2/1 < 0$). In Appendix C we show that this is not a coincidence but is a general property of mirror images of marker data: if $\mathrm{MS}(\theta)\mathbf{A}_I\mathrm{A}^2 \ldots \mathbf{A}^{n_A}\mathrm{U}^1 \ldots \mathbf{U}^{n_U}$ contributes to $p_>$ ($p_<$), then $\mathrm{MS}(\theta)\mathbf{A}_{\bar{I}}\mathrm{A}^2 \ldots \mathbf{A}^{n_A}\mathrm{U}^1 \ldots \mathbf{U}^{n_U}$ contributes to $p_<$ ($p_>$). Mirror images have the following two important properties proved in Appendix C:

1. If there is no association between marker and disease allele ($\Delta = 0$), then

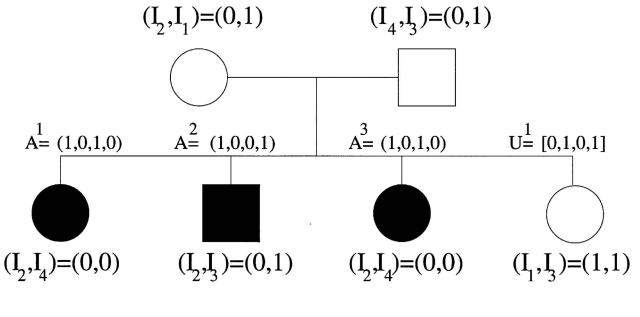

Figure B1 Mirror image of marker data in figure A1

$$\mathrm{MS}(\theta)\mathbf{A}_I\mathrm{A}^2 \dots \mathbf{A}^{n_\mathrm{A}}\mathrm{U}^1 \dots \mathbf{U}^{n_\mathrm{U}} = \mathrm{MS}(\theta)\mathbf{A}_{\bar{I}}\mathrm{A}^2 \dots \mathbf{A}^{n_\mathrm{A}}\mathrm{U}^1 \dots \mathbf{U}^{n_\mathrm{U}} \ .$$

2. If there is no linkage between the marker and disease locus ($\theta = 1/2$), then

$$\mathrm{MS}(\theta)\mathbf{A}_I\mathrm{A}^2 \dots \mathbf{A}^{n_\mathrm{A}}\mathrm{U}^1 \dots \mathbf{U}^{n_\mathrm{U}} = \mathrm{MS}(\theta)\mathbf{A}_{\bar{I}}\mathrm{A}^2 \dots \mathbf{A}^{n_\mathrm{A}}\mathrm{U}^1 \dots \mathbf{U}^{n_\mathrm{U}} \ .$$

We are now ready to complete the proof of equation (B3):

$$P(d^1 > 0 \mid n_\mathrm{A}, n_\mathrm{U}) = \sum_I \sum_{\mathrm{A}^2,\dots,\mathbf{A}^{n_\mathrm{A}}\,\mathrm{U}^1,\dots,\mathbf{U}^{n_\mathrm{U}}} \mathrm{MS}(\theta)\mathrm{A}_I\mathrm{A}^2 \dots \mathrm{A}^{n_\mathrm{A}}\mathrm{U}^1 \dots \mathrm{U}^{n_\mathrm{U}} \times \mathrm{PAU}(n_\mathrm{A}, n_\mathrm{U}) I(d^1 > 0) \ ,$$

and, forming the mirror image of each summand, we find that, if either $\Delta = 0$ or $\theta = 1/2$, then

$$P(d^1 > 0 \mid n_\mathrm{A}, n_\mathrm{U}) = \sum_{\bar{I}} \sum_{\mathrm{A}^2,\dots,\mathbf{A}^{n_\mathrm{A}}\,\mathrm{U}^1,\dots,\mathbf{U}^{n_\mathrm{U}}} \mathrm{MS}(\theta)\mathrm{A}_{\bar{I}}\mathrm{A}^2 \dots \mathrm{A}^{n_\mathrm{A}}\mathrm{U}^1 \dots \mathrm{U}^{n_\mathrm{U}} \times \mathrm{PAU}(n_\mathrm{A}, n_\mathrm{U}) I(d^1 > 0) = P(d^1 < 0 \mid n_\mathrm{A}, n_\mathrm{U}) \ .$$

For the multiallelic SDT, we assume a marker locus with m alleles and a disease locus with n alleles. We denote disease alleles by "$D_0, \dots, D_{n-1}$," and we denote their allele frequencies by "$p_{D_0}, \dots, p_{D_r}$"; we index disease alleles by $r = 0, \dots, n-1$. Analogous to the biallelic case, a measure, Δ_{jr}, of disequilibrium between marker allele j and disease allele D_r is defined by $\Delta_{jr} = k_{jr}p_{D_r} - p_j p_{D_r}$ $(j = 0, \dots, m-1, \ r = 0, \dots, n-1)$. The SDT is used to test the null hypothesis of no linkage ($\theta = 1/2$), or equilibrium $\Delta_{jr} = 0$—that is, $\mathrm{H}_0{:}\Delta_{jr}(\theta - 1/2) = 0$ $(j = 0, \dots, m-1, \ r = 0, \dots, n-1)$.

In a manner similar to what has been done in the biallelic case, one can show, under $\mathrm{H}_0{:}\Delta_{jr}(\theta - 1/2) = 0$ (for all $j = 0, \dots, m-1$, $r = 0, \dots, n-1$), it holds that $P(m^j_\mathrm{A} - m^j_\mathrm{U} > 0) = P(m^j_\mathrm{A} - m^j_\mathrm{U} < 0)$ $(j = 1, \dots, m)$. This implies that the median of d^j $(J = 1, \dots, m)$ is 0. This is the key assumption of the multivariate sign test, and one can use the multivariate central-limit theorem to show that the test statistic has an asymptotic $\chi^2_{(m-1)}$ distribution.

Appendix C

Properties of Mirror Images

PROPERTY 1. We will show that, if $\mathrm{MS}(\theta)\mathbf{A}_I\mathrm{A}^2 \dots \mathbf{A}^{n_\mathrm{A}}\mathrm{U}^1 \dots \mathbf{U}^{n_\mathrm{U}}$ contributes to $p_>$ $(p_<)$, then $\mathrm{MS}(\theta)\mathbf{A}_{\bar{I}}\mathrm{A}^2 \dots \mathbf{A}^{n_\mathrm{A}}\mathrm{U}^1 \dots \mathbf{U}^{n_\mathrm{U}}$ contributes to $p_<$ $(p_>)$. To do this, we need to express the values of m^1_A and m^1_U explicitly. Before we do this, let us consider the marker data in figure A1. Here, $m^1_\mathrm{A} = (1 + 1 + 1 + 0 + 1 + 1)/3$, which can be expressed as $m^1_\mathrm{A} = (I_1 + I_3 + I_1 + I_4 + I_1 + I_3)/3$ and, finally, as $m^1_\mathrm{A} = (\mathbf{A}^1 \cdot \mathbf{I} + \mathbf{A}^2 \cdot \mathbf{I} + \mathbf{A}^3 \cdot \mathbf{I})/3$, where $\mathbf{A}^1 = (1,0,1,0)$ and the dot $(\cdot)$ denotes the scalar product between four-dimensional vectors and $\mathbf{I} = (I_1, I_2, I_3, I_4) = (1,0,1,0)$. These simple formulas come about because, by labeling the alleles as "0" and "1," we are able to avoid indicator functions. In general, it holds that

$$m^1_\mathbf{A} = \frac{\sum_{m=1}^{n_\mathrm{A}} \mathbf{A}^m \cdot \mathbf{I}}{n_\mathrm{A}} \ ; \qquad \text{(C1)}$$

an analogous formula holds for m^1_U. Note that, since every parent transmits one allele to each sib, we have $\mathrm{A}^m_1 + \mathrm{A}^m_2 = \mathrm{A}^m_3 + \mathrm{A}^m_4 = 1 \quad (m = 1, \dots, n_\mathrm{A})$ and, hence, $\mathbf{A}^m \cdot \mathbf{I} = I_1 + I_2 + I_3 + I_4 - \mathbf{A}^m \cdot \tilde{\mathbf{I}}$ and $\mathbf{U}^n \cdot \mathbf{I} = I_1 + I_2 + I_3 + I_4 - \mathbf{U}^n \cdot \tilde{\mathbf{I}}$. These relations can be used to show that

$$\frac{\sum_{m=1}^{n_\mathrm{A}} \mathbf{A}^m \cdot \mathbf{I}}{n_\mathrm{A}} - \frac{\sum_{n=1}^{n_\mathrm{U}} \mathbf{U}^n \cdot \mathbf{I}}{n_\mathrm{U}} > 0$$

implies

$$\frac{\sum_{m=1}^{n_\mathrm{A}} \mathbf{A}^m \cdot \tilde{\mathbf{I}}}{n_\mathrm{A}} - \frac{\sum_{n=1}^{n_\mathrm{U}} \mathbf{U}^n \cdot \tilde{\mathbf{I}}}{n_\mathrm{U}} < 0 \ .$$

With the aid of equation (C1), one can see that this translates as follows: if $\mathrm{MS}(\theta)\mathbf{A}_I\mathrm{A}^2 \dots \mathbf{A}^{n_\mathrm{A}}\mathrm{U}^1 \dots \mathbf{U}^{n_\mathrm{U}}$ contributes to $p_>$ $(p_<)$, then $\mathrm{MS}(\theta)\mathbf{A}_{\bar{I}}\mathrm{A}^2 \dots \mathbf{A}^{n_\mathrm{A}}\mathrm{U}^1 \dots \mathbf{U}^{n_\mathrm{U}}$ contributes to $p_<$ $(p_>)$.

PROPERTY 2. We will show that, if there is no association between the marker allele (i.e., allele 1) and the disease allele (i.e., p_{D_1}) ($\Delta = 0$), then

$$\mathrm{MS}(\theta)\mathbf{A}_{(I_1,I_2,I_3,I_4)}\mathbf{A}^2 \dots \mathbf{A}^{n_\mathrm{A}}\mathbf{U}^1 \dots \mathbf{U}^{n_\mathrm{U}} = \mathrm{MS}(\theta)\mathbf{A}_{(I_2,I_1,I_4,I_3)}\mathbf{A}^2 \dots \mathbf{A}^{n_\mathrm{A}}\mathrm{U}^1 \dots \mathbf{U}^{n_\mathrm{U}} = \mathrm{MS}(\theta)\mathbf{A}_{\bar{I}}\mathbf{A}^2 \dots \mathbf{A}^{n_\mathrm{A}}\mathbf{U}^1 \dots \mathbf{U}^{n_\mathrm{U}} \ .$$

Recall that $k_{I_j r} = p_{I_j} + (-1)^{I_j}\Delta/p_{D_r}$. Therefore, $\Delta = 0$

implies that $k_{I_j r} = p_{I_j}$ ($j = 1 \ldots 4$). Using equation (A4), we conclude that

$$\begin{aligned}
&\mathrm{MS}(\theta)\mathbf{A}_I\mathbf{A}^2 \ldots \mathbf{A}^{n_\mathrm{A}}\mathbf{U}^1 \ldots \mathbf{U}^{n_\mathrm{U}} \\
&\quad = \sum_{r,s,t,u=0}^{1} [f_{\mathrm{A}^1}(\theta) \ldots f_{\mathrm{A}^{n_\mathrm{A}}}(\theta)\bar{f}_{\mathrm{U}^1}(\theta) \ldots \bar{f}_{\mathrm{U}^{n_\mathrm{U}}}(\theta) \\
&\qquad \times k_{I_1 r} k_{I_2 s} k_{I_3 t} k_{I_4 u} p_{D_r} p_{D_s} p_{D_t} p_{D_u}]/ \\
&\qquad \mathrm{PAU}(n_\mathrm{A}, n_\mathrm{U}) \\
&\quad = \sum_{r,s,t,u=0}^{1} [f_{\mathrm{A}^1}(\theta) \ldots f_{\mathrm{A}^{n_\mathrm{A}}}(\theta)\bar{f}_{\mathrm{U}^1}(\theta) \ldots \bar{f}_{\mathrm{U}^{n_\mathrm{U}}}(\theta) \\
&\qquad \times p_{I_1} p_{I_2} p_{I_3} p_{I_4} p_{D_r} p_{D_s} p_{D_t} p_{D_u}]/ \\
&\qquad \mathrm{PAU}(n_\mathrm{A}, n_\mathrm{U}) \\
&\quad = \sum_{r,s,t,u=0}^{1} [f_{\mathrm{A}^1}(\theta) \ldots f_{\mathrm{A}^{n_\mathrm{A}}}(\theta)\bar{f}_{\mathrm{U}^1}(\theta) \ldots \bar{f}_{\mathrm{U}^{n_\mathrm{U}}}(\theta) \\
&\qquad \times p_{I_2} p_{I_1} p_{I_4} p_{I_3} p_{D_r} p_{D_s} p_{D_t} p_{D_u}]/ \\
&\qquad \mathrm{PAU}(n_\mathrm{A}, n_\mathrm{U}) \\
&\quad = \mathrm{MS}(\theta)\mathrm{A}_{(I_2,I_1,I_4,I_3)}\mathrm{A}^2 \ldots \mathbf{A}^{n_\mathrm{A}}\mathrm{U}^1 \ldots \mathbf{U}^{n_\mathrm{U}} \\
&\quad = \mathrm{MS}(\theta)\mathbf{A}_{\bar{I}}\mathbf{A}^2 \ldots \mathbf{A}^{n_\mathrm{A}}\mathbf{U}^1 \ldots \mathbf{U}^{n_\mathrm{U}} \ .
\end{aligned}$$

PROPERTY 3. Here we show that, if there is no linkage between the marker and the disease locus ($\theta = 1/2$), then

$$\begin{aligned}
&\mathrm{MS}(1/2)\mathrm{A}_{(I_1,I_2,I_3,I_4)}\mathrm{A}^2 \ldots \mathrm{A}^{n_\mathrm{A}}\mathrm{U}^1 \ldots \mathrm{U}^{n_\mathrm{U}} \\
&\quad = \mathrm{MS}(1/2)\mathrm{A}_{(I_2,I_1,I_4,I_3)}\mathrm{A}^2 \ldots \mathrm{A}^{n_\mathrm{A}}\mathrm{U}^1 \ldots \mathrm{U}^{n_\mathrm{U}} \\
&\quad = \mathrm{MS}(1/2)\mathrm{A}_{\bar{I}}\mathrm{A}^2 \ldots \mathrm{A}^{n_\mathrm{A}}\mathrm{U}^1 \ldots \mathrm{U}^{n_\mathrm{U}} \ .
\end{aligned}$$

Note that $f_{(\mathrm{A}_1,\mathrm{A}_2,\mathrm{A}_3,\mathrm{A}_4)}(1/2) = \phi_{rstu}$ (see eq. [A2]). It is easily verified that

$$\phi_{rstu} = \phi_{srut} \ . \tag{C2}$$

We define $C = 4^{n_a+n_\mathrm{U}}\mathrm{PAU}(n_\mathrm{A}, n_\mathrm{U})$. Then we conclude that

$$\begin{aligned}
&\mathrm{MS}(1/2)\mathbf{A}_I\mathrm{A}^2 \ldots \mathbf{A}^{n_\mathrm{A}}\mathbf{U}^1 \ldots \mathbf{U}^{n_\mathrm{U}} \\
&\quad = \frac{1}{C}\sum_{r,s,t,u=0}^{1} f_{\mathrm{A}^1}(1/2) \ldots f_{\mathrm{A}^{n^\mathrm{A}}}(1/2)\bar{f}_{\mathrm{U}^1}(1/2) \ldots \bar{f}_{\mathrm{U}^{n_\mathrm{U}}}(1/2) \\
&\qquad k_{I_1 r} k_{I_2 s} k_{I_3 t} k_{I_4 u} p_{D_r} p_{D_s} p_{D_t} p_{D_u} \\
&\quad = \frac{1}{C}\sum_{r,s,t,u=0}^{1} \phi_{rstu}^{n_\mathrm{A}}\bar{\phi}_{rstu}^{n_\mathrm{U}} k_{I_1 r} k_{I_2 s} k_{I_3 t} k_{I_4 u} p_{D_r} p_{D_s} p_{D_t} p_{D_u} \\
&\quad = \frac{1}{C}\sum_{s,rt,u=0}^{1} \phi_{rstu}^{n_\mathrm{A}}\bar{\phi}_{rstu}^{n_\mathrm{U}} k_{I_2 r} k_{I_1 s} k_{I_4 t} k_{I_3 u} p_{D_r} p_{D_s} p_{D_t} p_{D_u} \ ,
\end{aligned}$$

and, using symmetry relations (C2), we get

$$= \frac{1}{C}\sum_{s,r,t,u=0}^{1} \phi_{srut}^{n_\mathrm{A}}\bar{\phi}_{srut}^{n_\mathrm{U}} k_{I_2 s} k_{I_1 r} k_{I_4 u} k_{I_3 t} p_{D_r} p_{D_s} p_{D_t} p_{D_u} \ ;$$

and, relabeling the indices—r as s and s as r—we get

$$\begin{aligned}
&= \frac{1}{C}\sum_{r,s,t,u=0}^{1} \phi_{rstu}^{n_\mathrm{A}}\bar{\phi}_{rstu}^{n_\mathrm{U}} k_{I_2 r} k_{I_1 s} k_{I_3 t} k_{I_4 u} p_{D_r} p_{D_s} p_{D_t} p_{D_u} \\
&= MS(1/2)\mathbf{A}_{(I_2,I_1,I_4,I_3)}\mathbf{A}^2 \ldots \mathbf{A}^{n_\mathrm{A}}\mathbf{U}^1 \ldots \mathbf{U}^{n_\mathrm{U}} \\
&= MS(1/2)_{\bar{I}}\mathbf{A}^2 \ldots \mathbf{A}^{n_\mathrm{A}}\mathbf{U}^1 \ldots \mathbf{U}^{n_\mathrm{U}} \ .
\end{aligned}$$

Appendix D

Computation of the ARE

We will briefly review some results regarding the ARE of test statistics; then we will compare T_1 and T_2 in the biallelic setting. We assume the null hypothesis $\mathrm{H}_0{:}\Delta = 0$. If, on average, we need n_1 (n_2) sib pairs in order to reject H_0 at a given level with a given power, then $\mathrm{ARE}(T_2{:}T_1)$ is approximately equal to n_1/n_2. Strictly speaking, ARE (also known as "Pitman efficiency") equals the ratio of sample sizes that give the same asymptotic power function, under sequences of local alternatives. In our case, a sequence of local alternatives is defined by $\mathrm{H}_\mathrm{A}{:}\Delta_N = \delta/\sqrt{N}$. If, under the sequence of alternatives, T_x converges, in distribution, to a $N(\delta\mu_x,1)$ distribution ($x = 1,2$), then the ARE is defined to be $\mathrm{ARE}(T_2{:}T_1) = (\mu_2/\mu_1)^2$. Now we rewrite the test statistics in a form that allows us to compute μ_1 and μ_2: $T_{x,N} = \sqrt{N}[W_{x,N} - \mu(0)]/\tilde{\sigma}$, where $W_{x,N} = [b_1 - c_1 + x(b_2 - c_2)]/N$, $\mu(0) = 0$ and $\tilde{\sigma}^2 = [b_1 + c_1 + x^2(b_2 + c_2)/N]$. Next, we cite a result that shows how to compute the μ_1 and μ_2 (see Serfling 1981, pp. 317–319): Consider testing $\mathrm{H}_0{:}\Delta = 0$ versus $\mathrm{H}_\mathrm{A}{:}\Delta > 0$. Suppose that W_N is a statistic such that $\sqrt{N}[W_N - \mu(\Delta)]$ converges, in probability, to $N[0,\sigma^2(\Delta)]$ on some interval about $\Delta = 0$, where $\sigma^2(\Delta)$ and $\mu(\Delta)$ are continuously differentiable. Set $T_N = \sqrt{N}[W_N - \mu(0)]/\tilde{\sigma}$, where $\tilde{\sigma}^2$ converges in probability to $\sigma^2(0)$, under H_0. Then, under a sequence of local alternatives, $\Delta_N = \delta/N$, for $\delta > 0$, T_N converges, in distribution, to $N(\delta\mu_T,1)$, where $\mu_T = [\partial\mu(0)/\partial\Delta][1/\sigma(0)]$. Let us define the following probabilities: $P_1 = P((1,0),(2,1))$, $Q_1 = P((0,1),(1,2))$, $P_2 = P((2,0))$, and $Q_2 = P((0,2))$. These probabilities are functions of the marker and disease-allele parameters described in section 3. Using the central-limit theorem, one can show that, for large sample sizes,

$$\sqrt{N}\left[\frac{1}{N}\begin{pmatrix} b_1 \\ c_1 \\ b_2 \\ c_2 \end{pmatrix} - \begin{pmatrix} P_1 \\ Q_1 \\ P_2 \\ Q_2 \end{pmatrix}\right]$$

$$\sim N(0, \begin{bmatrix} P_1(1-P_1) & -P_1Q_1 & -P_1P_2 & -P_1Q_2 \\ -Q_1P_1 & Q_1(1-Q_1) & -Q_1P_2 & -Q_1Q_2 \\ -P_2P_1 & -P_2Q_1 & P_2(1-P_2) & -P_2Q_2 \\ -Q_2P_1 & -Q_2Q_1 & -Q_2P_2 & Q_2(1-Q_2) \end{bmatrix} .$$

One can use the Δ method to show that, under H_0, $E(W_{x,N}) = 0$ and $V(W_{x,N}) = 2P_1 + 2x^2P_2$. Now we compute the ARE between T_x and the TDT. Let b (c) denote the number of times that a heterozygous parent transmits the 1 allele (0 allele) to an affected sib. Then the square root of the TDT statistic is given by $TDT_N = \sqrt{N}[W_N - \mu(0)]/\tilde{\sigma}$, where $W_N = (b - c)/N$, $\mu(0) = 0$, $\tilde{\sigma}^2 = (b + c)/N$, and N denotes the number of family trios. Under sequences of local alternatives, TDT_N converges, in distribution, to $N(\delta\mu_{TDT},1)$ where μ_{TDT} can be determined as above. We used the software package MAPLE to perform the calculations, since expressions for P_1, Q_1, P_2, and Q_2 are not readily available.

Many of the results given in the present article can be expressed in terms of three quantities—K_P, K_S, and K_O–which are functions of the disease-allele frequencies and penetrances. K_P is the disease frequency $K_P = p_D^2 f_{11} + 2p_D(1 - p_D)f_{10} + (1 - p_D)^2 f_{00}$. Here we assume $n = 2$, $P_{D_1} = P_D$, and $P_{D_0} = 1 - P_D$. If the disease is caused by a single disease gene, then K_S (K_O) is the incidence of affected sibs (affected offspring) of probands. Explicitly, we have $K_S = K_P + (V_A/2 + V_D/4)/K_P$ and $K_O = K_P + (V_A/2)/K_P$, where $V_A = 2p_D(1 - p_D)[p_D(f_{11} - f_{01}) + (1 - p_D)(f_{01} - f_{00})]^2$ and $V_D = p_D^2(1 - p_D)^2(f_{11} - 2f_{01} + f_{00})^2$. The relationship between K_P, K_S, and K_O has been explored by Suarez et al. (1976). For multifactorial diseases, K_S and K_O should not be interpreted as sibling and offspring recurrence risk but simply as expressions that comprise disease-allele frequencies and penetrance functions.

Appendix E

Sample-Size Calculations

Here we outline how we computed the sample sizes in table 2. Let us focus on the case of AUU sibships (table 2), since the calculations for AAU sibships are analogous. We want to compute the average number of sibships that need to be genotyped to reject H_0:$\Delta(\theta - 1/2) = 0$ when the true model is H_A:$\Delta = \Delta_{max}$, $\theta = 0$. The SDT (T) can be considered as a one-sample binomial test that compares π to its value under H_0 ($\pi_0 = 1/2$). The value of π can be computed with the aid of equations (B1)–(B3). Since the computations can become quite lengthy, we used the software package Maple to help us with symbolic manipulations. One can show (Rosner 1995) that the sample size needed to conduct a one-sided test with significance level α and power $1 - \beta$ is given by

$$n_{\text{eff}} = \frac{\pi_0(1 - \pi_0)\left[Z_{1-\alpha} + Z_{1-\beta}\sqrt{\frac{\pi_1(1-\pi_1)}{\pi_0(1-\pi_0)}}\right]^2}{(\pi_1 - \pi_0)^2} .$$

Note that n_{eff} is the required number of sibships with $m_A^1 \neq m_U^1$. However, we are interested in the average number n of sibships that need to be genotyped. Since $n_{\text{eff}} = n(p_> + p_<)$ (see Appendix B), we conclude that

$$n = \frac{\pi_0(1 - \pi_0)\left[Z_{1-\alpha} + Z_{1-\beta}\sqrt{\frac{\pi_1(1-\pi_1)}{\pi_0(1-\pi_0)}}\right]^2}{(p_> + p_<)(\pi_1 - \pi_0)^2} .$$

Electronic-Database Information

A C++ program that performs the (multiallelic) SDT on any number of markers is available, free of charge, via the anonymous address ftp://sph70-57.harvard.edu/XDT/

A program that does sample-size calculations for sibships of the form AU, AAU, and AUU is available, free of charge, from Xiaolin Wang (xaiolin@hsph.harvard.edu). The program requires the software package MAPLE.

References

Bickel PJ (1965) On some asymptotic competitors to Hotellings T^2. Ann Math Stat 36:160–173

Boehnke M, Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62:950–961

Cleves MA, Olson JM, Jacobs KB (1997) Exact transmission-disequilibrium tests for candidate gene testing and genomic screening with multiallelic markers. Genet Epidemiol 14: 337–347

Curtis D (1997) Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319–333

Curtis D, Sham PC (1995) A note on the application of the transmission disequilibrium test when a parent is missing. Am J Hum Genet 56:811–812

Laird NM, Blacker D, Wilcox M (1998) The sib transmission/disequilibrium test is a Mantel-Haenszel test. Am J Hum Genet 63:1915 (in this issue)

Mantel H, Haenszel W (1959) Statistical aspects of the analyses of data from retrospective studies. J Natl Cancer Inst 22:719–748

Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am J Hum Genet 61: 439–448

Ott J (1989) Statistical properties of the haplotype relative risk. Genet Epidemiol 6:127–130

Randles RH (1989) A distribution-free multivariate sign test based on interdirections. J Am Stat Assoc 84:1045–1050

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

Rosner B (1995) Fundamentals of biostatistics, 4th ed. Duxbury Press, Belmont, CA

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, Longton G, Kopecky KJ, Liang K-Y (1991) On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics 47:53–62

Serfling RJ (1981) Approximation theorems of mathematical statistics. John Wiley & Sons, New York, pp 317–319

Spielman RS, Ewens WJ (1993) Transmission/disequilibrium test (TDT) for linkage and linkage disequilibrium between disease and marker. Am J Hum Genet Suppl 53:863

——— (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458

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

Thomson G (1995) Analysis of complex human genetic traits: an ordered-notation method and new tests for mode of inheritance. Am J Hum Genet 57:474–486