A Sibship Test for Linkage in the Presence of Association: The Sib Transmission/Disequilibrium Test

Richard S. Spielman[1] and Warren J. Ewens[2]

[1]Department of Genetics, University of Pennsylvania School of Medicine, and [2]Department of Biology, University of Pennsylvania, Philadelphia

Summary

Linkage analysis with genetic markers has been successful in the localization of genes for many monogenic human diseases. In studies of complex diseases, however, tests that rely on linkage disequilibrium (the simultaneous presence of linkage and association) are often more powerful than those that rely on linkage alone. This advantage is illustrated by the transmission/disequilibrium test (TDT). The TDT requires data (marker genotypes) for affected individuals and their parents; for some diseases, however, data from parents may be difficult or impossible to obtain. In this article, we describe a method, called the "sib TDT" (or "S-TDT"), that overcomes this problem by use of marker data from unaffected sibs instead of from parents, thus allowing application of the principle of the TDT to sibships without parental data. In a single collection of families, there might be some that can be analyzed only by the TDT and others that are suitable for analysis by the S-TDT. We show how all the data may be used jointly in one overall TDT-type procedure that tests for linkage in the presence of association. These extensions of the TDT will be valuable for the study of diseases of late onset, such as non–insulin-dependent diabetes, cardiovascular diseases, and other diseases associated with aging.

Introduction

The transmission/disequilibrium test (TDT) was introduced as a method for identification of markers closely linked to genes that contribute to disease susceptibility (Spielman et al. 1993); if sufficiently closely linked, these markers are likely to be in linkage disequilibrium with the disease-susceptibility genes. (We use a narrow definition of the term "linkage disequilibrium," meaning the presence of both linkage and association, although past usage of this term has not always included linkage.) Provided that association is present, the TDT often has more power than conventional linkage tests, but, since it uses within-family comparisons only, it is not affected by aspects of population structure that can lead to associations in the absence of linkage (Ewens and Spielman 1995). The TDT is particularly suited to the testing of markers that may be at or very closely linked to genes that influence the risk for a complex disease (Lander and Kruglyak 1995; Risch and Merikangas 1995).

The TDT uses data from families in which marker genotypes are known for the father, the mother, and the affected offspring, but only parents who are heterozygous for marker alleles are considered. Since the TDT tests for unequal transmission of alleles from the parents to affected offspring, it cannot be performed if genotypic data for the parents are not available.

When diseases with onset in adulthood or in old age are studied, it may be impossible to obtain genotypes for markers in the parents of the affected offspring. This difficulty has limited the applicability of the TDT. In this article, we describe an alternative test that is analogous to the original TDT. Instead of using marker data from affected offspring and their parents, this method compares the marker genotypes in affected and unaffected offspring. For this reason, we call this new method the "sib TDT," or "S-TDT." The S-TDT does not reconstruct parental genotypes and does not depend on estimates of allele frequencies.

In some studies there will be families for which parental genotypes are available, other families for which genotypes of unaffected sibs are available but those of the parents are not, and still others for which both kinds of data are available. We will show how data from these three types of families can be combined into one overall test.

Methods

The data consist of marker genotypes for sibships that meet two requirements: (1) there must be at least one

Received August 1, 1997; accepted for publication November 25, 1997; electronically published February 13, 1998.

Address for correspondence and reprints: Dr. Warren J. Ewens, Department of Biology, University of Pennsylvania, Philadelphia, PA 19104-6018. E-mail: wewens@sas.upenn.edu

0002-9297/98/6202-0030$02.00

affected and one unaffected member in the sibship; and (2) the members of the sibship must not all have the same genotype (a sibship with one affected and one unaffected member, with different marker genotypes—that is, the "minimal" sibship [see Discussion]—meets these requirements). The initial units of observation are the marker genotypes of the offspring for each family, as in table 1.

In essence, the S-TDT determines whether the marker allele frequencies among affected offspring differ significantly from the frequencies among their unaffected sibs. Disease association without linkage will not result in such differences; unless linkage is also present (including in the case in which the marker itself is responsible), the frequencies will be the same in the affected and the unaffected sibs, apart from random-sampling effects. Thus, the null hypothesis is that the disease and the marker are not linked.

Table 1

Genotypes for a Locus with Alleles M_1, M_2, and M_3, in Three Fabricated Sibships

	No. of Sibs with Genotype				M_1 Alleles in "Affected" Sibs, by Chance[a]	
Sibship and Sib Status	M_1M_1	M_1M_2	M_2M_2	M_1M_3	Mean	Variance
1 (7 sibs):						
Affected	2	1	...	...		
Unaffected	...	2	...	2	3.8571	.4082
2 (5 sibs):						
Affected	...	1	...	...		
Unaffected	...	1	2	1	.6000	.2400
3 (4 sibs):						
Affected	1	...	...	...		
Unaffected	...	1	2	...	.7500	.6875

[a] The mean and variance of the number of M_1 alleles among "affected" sibs were calculated by use of the terms summed in equations (1) and (2) (see text).

The Permutation Procedure

To compare marker allele frequencies in affected and unaffected offspring, it would be natural to accumulate allele totals over families, with results like those in table 2. For data of this type, the customary method of testing for departures from the null hypothesis would be by χ^2 analysis. However, a χ^2 used with these aggregated data is not valid, because of nonindependence of the observations on sibs from the same family. To provide a valid test, we adopted a within-family Monte Carlo permutation procedure, performed by a computer program. Consider a family with a affected and u unaffected sibs, each with a known marker genotype. To determine what differences between affected and unaffected sibs would be produced by chance, we permute the observed genotypes within sibships as follows. Ignoring actual affected status, we choose a of the $a + u$ sibs *at random* and assign them to the "affected" category; the remainder are assigned to the "unaffected" category (the allocation is of genotypes, not alleles, since permutation of individual alleles could result in the creation of genotypes not seen in the sibship.) Because the permutation is carried out within families, potential problems resulting from population structure are eliminated, as is true of the original TDT. For one "replicate," the randomization is performed for within-family data, and then the resulting numbers of alleles in affected and unaffected sibs are totaled over the families. The procedure gives a simulation result in the same form as that used in table 2. This procedure is repeated a large number of times (replicates), and, after each replicate, a table analogous to table 2 is constructed. These randomly generated replicates provide the "null" distribution for a test of linkage.

To simplify the discussion, we first consider the case either in which there are only two marker alleles, M_1 and M_2, or in which one marker (e.g., M_1) is of particular interest and all other marker alleles are grouped as M_2. The case of three or more markers is discussed later (see Discussion).

The number of M_1 alleles among the individuals randomly chosen as "affected" is used to test for linkage. The *P* value is assigned by noting the proportion of replicates in which this number is equal to, or more extreme than, the observed value in the actual data. The level of statistical significance is given by this "empirical" *P* value.

The z Score Procedure

With a sufficiently large number of replicates, the permutation procedure will provide precise *P* values. The properties of the procedure can be illustrated more easily by use of a statistical analytical method that, with a large sample, is essentially identical to the computer simulation for permutation testing.

In the sibship described above, with a affected and u unaffected sibs, the total number of sibs is $t = a + u$. Suppose that in this sibship the number of sibs who are of genotype M_1M_1 is r and the number of sibs who are of genotype M_1M_2 is s, where M_2 represents all alleles other than M_1. In the permutation procedure, the

Table 2

Total Number of Alleles in Affected and Unaffected Members of Sibships in Table 1

	No. of Alleles			
Sib Status	M_1	M_2	M_3	Total
Affected	8	2	0	10
Unaffected	7	12	3	22

number of M_1 alleles among "affected" sibs has mean $(2r + s)a/t$ and variance $au[4r(t - r - s) + s(t - s)]/[t^2(t - 1)]$, under the null hypothesis of no linkage between disease and marker. These formulas are derived from the hypergeometric distribution; the derivation is given in Appendix A.

The overall mean A and variance V of the number of M_1 alleles among "affected" sibs, when totaled over all families in the data, are given by simple summation:

$$A = \sum (2r + s)a/t \quad (1)$$

and

$$V = \sum au[4r(t - r - s) + s(t - s)]/[t^2(t - 1)] \ , \quad (2)$$

where, in both cases, summation is over all families in the sample.

For example, in the three sibships of table 1, summation of the means and variances for allele M_1 gave the totals $A = 5.2071$ and $V = 1.3357$ (see Appendix A). By using the usual method, we calculated a z score for allele M_1, from A, V, and the total observed number (Y) of M_1 alleles among affected sibs in the actual data set: $z = (Y - A)/\sqrt{V}$. From the z score, an approximate P value can be calculated, by use of a normal distribution approximation. It is customary to make a continuity correction, and the P value is then calculated from $z' = (|Y - A| - ½)/\sqrt{V}$. For allele M_1 in the data given in table 1, this expression yielded $z' = 1.9839$.

The calculations described above can be performed easily if the data are entered into a spreadsheet program (a program in Excel is available from R.S.S.). Thus, the z score approach, which is a highly accurate approximation, is a simple and attractive alternative to the permutation procedure and eliminates the need for computer simulation. As we show below, the z score method also makes it easy to combine results from the TDT and the S-TDT.

Combining the TDT and the S-TDT into an Overall Test

We now show how the TDT and the S-TDT procedures can be combined. Of course, for each family, the genotype of at least one affected offspring must be available. The data for other family members define three groups of families: (*i*) genotypes are available for both parents but not for unaffected sibs; (*ii*) genotypes are available for unaffected sibs but not for both parents; and (*iii*) genotypes are available for both parents and for unaffected sibs.

Group (*iii*) families meet the requirements for both the TDT and the S-TDT. We expect that, in sufficiently large samples, the TDT is at least as powerful as the S-TDT, for cases in which either test could be used (see Appendix B). We therefore combine families in group (*iii*) with those in group (*i*) and ignore the unaffected sibs in the families in group (*iii*), thus treating the combined group exactly as in the original TDT. Henceforth, group (*i*) refers to this combined group. As above, we first consider the case of only two marker alleles, M_1 and M_2, and describe extension to the case of more than two marker alleles in the Discussion section.

The appropriate test for families in group (*i*) is the TDT, which is usually performed with a (McNemar test) χ^2 statistic. However, for our present purposes, it is more convenient (and equivalent) to use as the test statistic the number (X) of transmissions of allele M_1 from the n M_1M_2 parents in this group to their affected children. When marker and disease are unlinked, X has a binomial distribution with mean $n/2$ and variance $n/4$.

The S-TDT procedure is appropriate for families in group (*ii*). As described in the discussion following equations (1) and (2), the test statistic is the number (Y) of M_1 alleles among affected sibs. When marker and disease are unlinked, Y has a distribution with mean A and variance V, given in equations (1) and (2).

When data are available from both group (*i*) and group (*ii*), the natural test statistic is W, the sum of X and Y. Under the null hypothesis that disease and marker are unlinked, W has mean A_{comb} and variance V_{comb}, given by

$$A_{\text{comb}} = n/2 + A \quad (3)$$

and

$$V_{\text{comb}} = n/4 + V \ , \quad (4)$$

respectively. The test of significance is now performed with a z statistic, calculated from the formula

$$z = (W - A_{\text{comb}})/\sqrt{V_{\text{comb}}} \ . \quad (5)$$

The null hypothesis that disease and marker are unlinked is rejected if z departs significantly from zero, as judged by standard z tables (as with the S-TDT procedure, a correction for continuity should be made).

Results

We applied the S-TDT and the combined procedure to several sets of data that were analyzed previously by use of the TDT.

Data from Genetic Analysis Workshop 9 (GAW9)

First, we analyzed the data of GAW9 (Hodge 1995). In GAW9, computer-simulated genotype data were generated for 200 nuclear families, for a large number of

Table 3

Total Number of M7 Alleles of Marker D5G23 in Affected and Unaffected Members of Sibships Informative for the S-TDT, among 200 GAW9 Families

	No. of Alleles		
Sib Status	M7	Other	Total
Affected	161	161	322
Unaffected	230	486	716
Total	391	647	1,038

Note.—For the number of M7 alleles in the observed number of affected sibs, mean $A = 124.3$, and variance $V = 30.99$. See text for details.

multiallelic marker loci, and the families were ascertained on the basis of (simulated) disease. Strong associations (no recombination) were introduced between the disease and allele M8 of marker D1G31 and allele M7 of marker D5G23 (the two marker loci are unlinked). We and other participants in GAW9 analyzed the data, using the conventional TDT; for comparison with the present analysis of the S-TDT, our results (McGinnis et al. 1995), which are representative, are quoted here. The TDT χ^2 for allele M8 of D1G31 was 31.25, and that for allele M7 of D5G23 was 58.07. Both of these results are clearly highly significant, even allowing for testing of multiple alleles. Thus the TDT accurately identified the two markers that were associated and linked with disease.

To assess the S-TDT procedure, we used the same simulated sibships used by GAW9; however, we ignored the genotypes of the parents but included the genotypes of the unaffected sibs. We performed the S-TDT for three marker loci: the two linked sites (D1G31 and D5G23, discussed above) and one marker (D5G21) not associated with disease. We give detailed results for only one of the linked markers, D5G23.

The total numbers of alleles observed in the affected and unaffected sibs in the 200 simulated sibships are shown in table 3. Genotypes were permuted within sibships, as described in Methods. The excess of allele M7 in affected sibs was greater than that seen in any of the 50,000 random permutations generated for the S-TDT; this gives an empirical one-sided P value of <.00002. The corresponding z' score was 6.50 ($P \approx 4 \times 10^{-11}$), obtained by comparison of the observed number of M7 alleles among affected sibs with the expected number of M7 alleles, calculated as 124.3 by use of equation (1). The results for D1G31 were similar; allele M8 was identified correctly by permutation. Furthermore, the marker that was not in linkage disequilibrium (D5G21) did not yield significant results; for the six alleles at this locus, the smallest P value was .17. (Since our goal in this article is to illustrate the method, we have not corrected for testing of multiple alleles. The testing of multiple marker alleles is discussed below.)

The remarkable statistical power demonstrated above is due partly to the fact that the GAW9 data were simulated for the testing of very large genetic effects detected at the two disease loci and partly to the large sample size. Family studies of the size of the GAW9 study, with 200 sibships and many unaffected sibs, often will not be available, and for a smaller number of sibships there is a corresponding loss of power. However, we found that, for the GAW9 data, highly significant evidence for linkage was always obtained by use of various sets of 50 families randomly chosen from the 200 in the data set (results not shown).

The Insulin Gene 5′ VNTR

We also tested the permutation procedure with data for a real disease. For this purpose, we used data on insulin-dependent diabetes mellitus (IDDM) and on the variation at the VNTR marker adjacent to the insulin gene; these data were analyzed elsewhere by use of the conventional TDT (Spielman et al. 1993). The VNTR shows linkage with IDDM, and, in many case-control studies, the smallest of the three classes (class 1) of alleles detectable by Southern blotting has been shown to be associated with IDDM (Bell et al. 1984; Cox and Spielman 1989; Julier et al. 1991). Following the common practice used in previous analyses, we grouped class 2 and class 3 together as "other." In the conventional TDT analysis, the class 1 allele was transmitted 78 times, and other alleles were transmitted 46 times ($\chi^2 = 8.26$) (Spielman et al. 1993); the normal approximation to the binomial gives $z' = 2.78$, with $P = .0027$ (one-sided test, because only an excess of the class 1 allele was of interest).

For the S-TDT, we used the same families analyzed previously by the TDT; there were data for 76 sibships with at least one affected and one unaffected member, and, of these, only 45 had marker-genotype differences among the sibs. The S-TDT procedure was performed

Table 4

Total Number of Class 1 and Other Alleles of the Insulin Gene 5′ VNTR in Affected and Unaffected Members of IDDM Sibships Informative for the S-TDT

	No. of Alleles		
Sib Status	Class 1	Other	Total
Affected	151	43	194
Unaffected	122	62	184
Total	273	105	378

Note.—For the number of class 1 alleles in the observed number of affected sibs, mean $A = 140.67$, and variance $V = 12.45$. See text for details.

for these 45 sibships. The numbers of alleles observed in affected and unaffected sibs are given in table 4.

For these data, the one-sided empirical P value from the permutation procedure was $647/250,000 = .0026$. Using the parallel z score approach, we calculated, by use of equations (1) and (2), the mean A and variance V of the number of class 1 alleles among "affected" sibs as 140.67 and 12.45, respectively. The z' score, obtained when these values were used with the observed total 151, was 2.79 ($P = .0026$, one-sided test); the agreement with the P value from the permutation procedure illustrates the accuracy of the z score method.

In this example, the z' score for the S-TDT appears to agree almost exactly with that obtained from the conventional TDT ($z' = 2.78$, $P = .0027$). However, such close agreement should not be expected in general, since the two tests do not use the same data.

Combining the S-TDT and the TDT

Among the families typed for the VNTR, there were many that were not suitable for the S-TDT, either because all offspring were affected or because all had the same marker genotype. In eight of these families, however, one or both parents were heterozygous for the marker, so that the original TDT was applicable. We illustrate the calculations for the combined TDT by adding the data from these families to those from the 45 families considered in the previous section.

There were 21 transmissions from heterozygous parents to affected offspring: 14 transmissions were of the class 1 allele. From the properties of binomial distribution, the mean and variance for the TDT component are $21/2 = 10.5$ and $21/4 = 5.25$, respectively. Combining these values with the mean and variance given above for the S-TDT component, we calculated $A_{\text{comb}} = 151.17$ (equation [3]) and $V_{\text{comb}} = 17.70$ (equation [4]). The observed total number of class 1 alleles among affected offspring (TDT and S-TDT results combined) is $X + Y = 14 + 151 = 165 = W$. Use of equation (5), with a continuity correction, leads to $z' = 3.17$ and $P = .0008$ (one-sided test). In this case, the z score was increased, and the result illustrates the potential increase in power achieved by the combination of data from the TDT and the S-TDT. However, this conclusion is not necessarily a general one, and, in practice, z scores resulting from the combined TDT may be larger or smaller than those obtained from either component alone.

Discussion

The S-TDT extends the concept of the original TDT to sibships for which the parents' genotypes are unknown. We now compare the properties of the TDT with those of the S-TDT and of the combined procedure described above. For simplicity, we continue to consider only the case of two marker alleles, M_1 and M_2.

Validity of the S-TDT as a Test of Association

The S-TDT has been proposed above as a test of *linkage* between marker and disease; this also was the primary intended use for the TDT. However, the TDT is known to be valid also as a test of *association,* provided the data are entirely from simplex families (one affected offspring; one or both parents may be heterozygous for the marker). We now find the analogous requirement for the S-TDT to be valid as a test of association.

To do this, we consider the smallest sibships that can give data for the S-TDT. This "minimal" configuration for the S-TDT consists of exactly one affected and one unaffected sib, with different marker genotypes. In the S-TDT procedure, if there is no association between disease and marker, the two possible genotypic assignments for the affected and the unaffected sib are equally likely in these sibships. Since this property is what is simulated by the S-TDT permutation procedure, the S-TDT is valid as a test of association for families of this type. Thus, for the S-TDT to be valid as a test of association, a sufficient requirement is that the sibships all be of the minimal configuration.

For sibships that do not have the minimal configuration, the S-TDT is not valid as a test of association; for these configurations, it cannot be assumed, even in the absence of association, that sibs with the same genotype are no more likely than sibs of different genotypes to be concordant for disease status. Thus, the only sibships that provide a valid test of association are those with the minimal configuration.

Requirements for Identity of the TDT and S-TDT

The minimal configuration for the TDT is a special case of the simplex family; in addition to having exactly one affected offspring, the family must have only one parent heterozygous for the marker. Consider a sample in which all families have one heterozygous (M_1M_2) parent, one homozygous (e.g., M_2M_2) parent, one affected offspring, and one unaffected offspring. For the TDT we ignore the genotypic data from the unaffected offspring, whereas for the S-TDT we do not have the genotypic data from the parents. Thus, when viewed from the perspective of the TDT, these families are of the minimal TDT configuration, and, when viewed from the perspective of the S-TDT, they are of the minimal S-TDT configuration.

Thus, the data from families of this type can be analyzed by use of either the TDT or the S-TDT. The testing procedure is the same in both cases: the test statistic is the number of occurrences of M_1 in the affected sibs,

Table 5

Power of the S-TDT and the TDT, for Sibships with One M_1M_2 Parent, One M_2M_2 Parent, One Affected Sib, and 2, 3, or 4 Unaffected Sibs

	Power of the S-TDT, for Families with									
	3 Sibs		4 Sibs			5 Sibs				Power of the TDT
T[a]	(2,1)	(1,2)	(3,1)	(2,2)	(1,3)	(4,1)	(3,2)	(2,3)	(1,4)	
.500	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
.525	.233	.247	.194	.259	.230	.185	.248	.258	.213	.259
.550	.573	.617	.502	.639	.568	.443	.612	.636	.522	.639
.575	.867	.900	.799	.915	.862	.721	.896	.913	.819	.915
.600	.980	.989	.952	.992	.979	.917	.988	.991	.963	.992

NOTE.—The number of M_1M_2 and M_2M_2 sibs in each sibship is given in parentheses. The type I error rate is 5%. Sample size is $F = 400$ families.

[a] The probability that an M_1M_2 parent transmitted the M_1 allele to an affected offspring.

and, in both cases, this number has a binomial distribution with parameters of ½ and F (the number of families in the data), when disease and marker are unlinked. Thus, for data from only families of this type, the TDT and the S-TDT have identical properties, even though they are, in general, different procedures. In particular, for data of this type, both procedures are valid as tests of *association,* and they also have the same properties, and when used as tests of *linkage* with these data, the TDT and the S-TDT have identical properties, including the same power and significance levels. (We discuss another case of identity in the next section, in connection with table 5.)

A second property common to the TDT and the S-TDT concerns data from pedigrees that contain several sibships—for example, sets of cousins in two sibships or a set of sibs and their aunts and/or uncles. In these pedigrees, the TDT is valid as a test of linkage but not as a test of association, when based on data from all heterozygous parents and their offspring (Spielman and Ewens 1996). The corresponding results for the S-TDT are data, from separate sibships, that also can be combined to give a test that is valid for linkage but not for association.

Power: Comparison with the TDT

We noted above that some families may meet the requirements for both the TDT and the S-TDT. For these families, the procedure with more statistical power should be used. We now investigate the relative power of the TDT and the S-TDT.

Families in which both parents are homozygous for the marker may not be used in the TDT or in the S-TDT. For families in which both parents are heterozygous, we expect the TDT to be more powerful than the S-TDT, since, in the TDT, each affected offspring contributes information on two transmissions. Thus, the most interesting comparison of power is for families in which one parent is heterozygous and one is homozygous. We therefore consider only this case. For the purpose of exposition, we assume a uniform sample: in every family, the mating type is $M_1 M_2 \times M_2 M_2$, s number of sibs are M_1M_2, and $t - s$ are M_2M_2.

In the first example, we assumed the presence of one affected sib in every family. Table 5 gives the power for the TDT and the S-TDT, for various marker genotype configurations for the sibs (construction of table 5 is described in Appendix B.)

Two main conclusions can be drawn from this table and from the calculations in Appendix B. First, for the family structure considered, the S-TDT is generally less powerful than the TDT, although the difference in power is often small. Second, Appendix B shows that, when there are equal numbers of M_1M_2 and M_2M_2 sibs in each family (table 5 shows the case of two of each genotype), the TDT and the S-TDT procedures are identical and thus have the same power. The more dissimilar these two numbers are, the less powerful the S-TDT is, relative to the TDT.

For one case in table 5 (the "[2,2]" case of the S-TDT), the results of the TDT and the S-TDT were identical; this fact led us to ask whether this identity occurs in the S-TDT whenever there are equal numbers of M_1M_2 and M_2M_2 sibs. We considered the case in which, for every family, the mating type is $M_1M_2 \times M_2M_2$ (as used above) but in which there are *two* affected sibs. Power calculations for this case of equal numbers of M_1M_2 and M_2M_2 sibs, derived by a method similar to that described above, are given in table 6 (we assumed $F = 200$, to give the same number of affected offspring as in table 5). In this case we found that the TDT and the S-TDT do not have equal power; the TDT has marginally more power than the S-TDT, for all the cases considered. This may be confirmed generally by use of calculations analogous to those given in Appendix B.

These calculations concern only a small proportion of all possible family configurations and are intended to cover only situations in which a simple and direct com-

Table 6

Power of the S-TDT and the TDT, for Sibships with One M_1M_2 Parent, One M_2M_2 Parent, and Two Affected Sibs

	Power of the S-TDT, for Families with			
T^a	4 Sibs	6 Sibs	8 Sibs	Power of the TDT
.500	.050	.050	.050	.050
.525	.204	.226	.236	.259
.550	.497	.559	.583	.638
.575	.795	.854	.875	.915
.600	.952	.976	.982	.992

NOTE.—In each family, there are equal numbers of M_1M_2 and M_2M_2 sibs. The type I error rate is 5%. Sample size is $F = 200$ families.

[a] The probability that an M_1M_2 parent transmitted the M_1 allele to an affected sib.

parison can be made between the powers of the TDT and the S-TDT procedures. For all cases reported in tables 5 and 6, the power of the S-TDT was less than or equal to that of the TDT. It appears that, for sufficiently large sample sizes, the TDT is always more powerful than the S-TDT. For this reason, we recommend use of the TDT for those families that can be analyzed by either method.

Bias with Missing Parents

Curtis and Sham (1995) have shown that a bias can arise in the TDT if the genotype for one parent is missing, even though it is clear which marker allele the available (heterozygous) parent transmitted to an affected child; in these cases, the data should not be used for the TDT. For families of this type there might be marker information on unaffected sibs. If so, these families can be included in group (ii), since there is no corresponding bias in the S-TDT.

Multiple Alleles and Multiple Loci

For the theory described above, we focused on one marker allele, M_1. In several applications this is natural: for the insulin gene VNTR example discussed above, there was prior interest in the class 1 allele. If there are more than two alleles at the marker locus but if no allele is of special interest, a more complex procedure is necessary.

Table 7

Approximate Significance Points of $z_{\max}$, the z Score Largest in Absolute Value

	$z_{\max}$, (and Bonferroni Approximation), for Type I Error of			
k	5%	1%	.1%	.01%
2	1.96 (1.96)	2.58 (2.58)	3.29 (3.29)	3.89 (3.89)
3	2.34 (2.39)	2.91 (2.93)	3.56 (3.58)	4.15 (4.15)
4	2.47 (2.50)	.02 (3.02)	3.65 (3.65)	4.25 (4.25)
5	2.55 (2.58)	3.08 (3.09)	3.71 (3.72)	4.26 (4.26)
10	2.80 (2.81)	3.29 (3.29)	3.89 (3.89)	4.42 (4.42)

Before discussing multiple alleles in the S-TDT and the combined test, we summarize the corresponding procedures already developed for the TDT. Two broad approaches have been proposed. In the first (Schaid 1996), all alleles other than M_1 are grouped as "non-M_1," and a conventional two-allele TDT is then performed. If there are k marker alleles, this procedure is repeated for each of the other $k - 1$ alleles, so that a total of k TDT χ^2 statistics are calculated. The test statistic is maxTDT, the largest of the k TDT values calculated. In a previous study (Ewens and Spielman, 1997) we gave a table of significance points for maxTDT.

For the present discussion it is more convenient (and equivalent) to use a "$z_{\max}$" statistic instead of maxTDT. We calculated the numbers of $M_1, \ldots, M_k$ alleles transmitted from heterozygous parents to their affected offspring and determined the corresponding z scores by comparison with the number of times each is observed with its permutation expected value. In contrast with the case in which one allele (M_1) was of interest a priori, no allele is of special interest in this case. Thus, we adopted a two-sided procedure and chose the quantity $z_{\max}$, the z score that is largest in absolute value, as the test statistic. Approximate significance points for this statistic were found by simulation and are given in table 7; significance points found from the Bonferroni correction for multiple testing are given in parentheses (table 7 shows that, for practical purposes, the Bonferroni values are very accurate and provide a slightly conservative test).

The TDT $z_{\max}$ approach for the testing of multiple alleles generalizes naturally for the S-TDT and for the combined test. For the S-TDT, a z score was calculated for each of the k alleles, as described in Methods, and the largest absolute z score was chosen as the $z_{\max}$ score. Similarly, for the combined test, k different z scores were computed from equation (5), and the largest absolute z score was chosen as the $z_{\max}$ score. The significance of the $z_{\max}$ was determined by reference to table 7.

A second approach is a simultaneous, or joint, test of all k marker alleles. For the TDT, this was done with a χ^2 statistic with $k - 1$ df (Schaid 1996; Spielman and Ewens 1996) or with a logistic regression model (Duffy 1995; Harley et al. 1995; Sham and Curtis 1995).

For the S-TDT and the combined test, there are two drawbacks to the approach of testing multiple alleles jointly. First, a significant effect for one (or several) marker alleles might be obscured by the presence of many other alleles with little or no association (Spielman and Ewens 1996); this drawback also applies to the TDT. Second, it is not apparent how to combine the TDT and the S-TDT procedures when alleles are tested jointly. For these reasons we recommend the $z_{\max}$ approach. Ultimately, however, the best approach will also depend on the extent to which one marker allele is associated with the disease more strongly than any of the other marker alleles.

The above discussion concerns the adjustment for multiple testing when more than two *alleles* occur at the marker locus. A further correction for multiple testing is needed if multiple marker *loci* are tested (Lander and Kruglyak 1995). As in the case of the TDT (Ewens and Spielman, 1997), a Bonferroni correction should be used for this situation.

Conclusion

Recent studies (Risch and Merikangas 1996) have emphasized the importance of linkage-disequilibrium analysis as a means of localization of genes for complex diseases. The TDT has proved to be a powerful test for this purpose. However, the TDT relies on data from parental genotypes. In practice, one or both parents may be unavailable for study, and this is especially likely for diseases of adult or late onset. In this category are numerous complex diseases in which a genetic contribution is suspected: diabetes; cardiovascular diseases, including hypertension and stroke; adult psychiatric diseases; and other diseases specifically associated with aging, especially Alzheimer disease and Parkinson disease. In diseases such as these, limitations on the TDT imposed by lack of data from parents are circumvented by the procedures for the S-TDT and the combined TDT, presented in this article.

Acknowledgments

The idea for a TDT-like test for sibships without data from parents was suggested to R.S.S. by Glenys Thomson. We thank P. Concannon, N. Kaplan, H. Kazazian, R. Nussbaum, E. Thompson, J. Murray, and B. Weir for their comments on the manuscript. Ralph McGinnis edited the manuscript and provided GAW9 and VNTR data; Philip Wexler performed initial tabulations; and Kevin Howard programmed and completed the simulations. Our research is supported by National Institutes of Health grants GM21135 (to W.J.E.) and DK46618 and DK47481 (to R.S.S.).

Appendix A

Let x be the number of M_1M_1 sibs and y the number of M_1M_2 sibs who are classified as "affected," under the random permutation within a sibship. Then, given the totals r, s, a, u, and t, as defined in the text, x may be regarded as an entry in a 2 × 2 contingency table with marginal totals a, u, r, and $t-r$ and therefore has a hypergeometric distribution with mean ra/t and variance $r(t-r)au/[t^2\ (t-1)]$. Similarly, y has a hypergeometric distribution with mean sa/t and variance $s(t-s)au/[t^2(t-1)]$. Furthermore, the covariance of x and y is $-rsau/[t^2(t-1)]$. The number of M_1 alleles among the "affected" sibs, under random permutation, is $2x+y$, and by standard statistical formulas this number has mean 2(mean of x) + (meanof y), leading to the value $(2r+s)a/t$, given in equation (1). The variance of $2x+y$ is $4\mathrm{Var}(x)+\mathrm{Var}(y)+4\mathrm{Cov}(x,y)$, and, by use of the above formulas, this leads to the variance formula given in equation (2). These formulas give the means and variances for sibships 1, 2, and 3 in table 1.

Appendix B

For a simple example of the power comparison of the TDT and the S-TDT, it is convenient to assume that all families in the data are of the same type. Specifically, we assume that each family has t sibs, one of whom is affected, that each family has one heterozygous M_1M_2 parent and one homozygous M_2M_2 parent, and that in each family s sibs are M_1M_2 and $t-s$ sibs are M_2M_2. For these families the power comparison of the TDT and the S-TDT can be made through a single parameter, T, the probability that the heterozygous parent in each family transmitted the M_1 allele to the affected sib. We assume that the TDT and the S-TDT are both performed as (one-sided) tests for an excess ($T > ½$) of transmissions of the M_1 allele to affected offspring.

We denote the number of families by F. The TDT statistic is X, the number of transmissions of M_1 from the F M_1M_2 parents to the F affected offspring. When marker and disease are unlinked, X has a binomial distribution with mean $F/2$ and variance $F/4$. By use of a normal distribution approximation and a 5% type I error, the hypothesis of no linkage is rejected when

$$X \geqslant F/2 + 1.645\sqrt{(F/4)}\ . \qquad \text{(B1)}$$

Suppose now that disease and marker are linked and that the probability that an M_1M_2 parent transmitted the M_1 allele to an affected child is T, which is no longer necessarily ½. In this case, X has a binomial (F, T) distribution, and the power of the test is the probability that the inequality (B1) occurs when X has this distribution. A z transformation shows that this is the probability that a standard normal random variable exceeds the value

$$\left(\tfrac{1}{2}-T\right)\sqrt{F}/\sqrt{T(1-T)} + 1.645/\sqrt{4T(1-T)}\ . \qquad \text{(B2)}$$

The S-TDT procedure takes the values s and $t-s$ as given. In this procedure, the test statistic is the number Y of M_1 alleles, among the affected sibs, that are necessarily transmitted from the heterozygous parents. When disease and marker are unlinked, it follows from equations (1) and (2) and from the genotypic composition assumed above that the mean of $Y = sF/t$ and the variance of $Y = s(t-s)F/t^2$. We therefore reject the hypothesis of no linkage when

$$Y \geqslant sF/t + 1.645\sqrt{s(t-s)F/t^2}\ . \qquad \text{(B3)}$$

Suppose now that disease and marker are linked and that the probability that an M_1M_2 parent transmitted the M_1 allele to an affected child is T (as above). To calculate the power of this test, we must make an assumption about the probability that an M_1M_2 parent transmitted the M_1 allele to an unaffected sib. We assume here that this probability is ½, independently for each unaffected sib and independent of the allele transmitted to the affected sib. (The exact value depends on penetrance and thus generally is unknown. For genes that make a small contribution to the disease, the value ½ will provide a sufficiently close approximation.) Under this assumption, the mean and variance of Y are

$$sFT/[Ts + (1-T)(t-s)] \qquad \text{(B4)}$$

and

$$s(t-s)FT(1-T)/[Ts + (1-T)(t\text{-}s)]^2\ , \qquad \text{(B5)}$$

respectively. The power of the test is the probability that the inequality (B3) occurs, given that Y has a normal distribution, with mean and variance as in equations (B4) and (B5). This is the probability that a standard normal random variable exceeds the value

$$\sqrt{s(t-s)}\left(\tfrac{1}{2} - T\right)\sqrt{F}/\left[\tfrac{1}{2}t\sqrt{T(1-T)}\right]$$
$$+\,1.645\,[Ts + (1-T)(t-s)]/t\sqrt{T(1-T)}\ . \qquad \text{(B6)}$$

When F, the number of families in the sample, is large, the leading term in equations (B2) and (B6) is the first term. This term is negative, and, since it is always true that $\sqrt{s(t-s)} \leqslant ½t$ (because a geometric mean is always less than or equal to the corresponding arithmetic mean), the leading term in equation (B2) is more negative than the leading term in equation (B6). This implies that the probability from equation (B2) exceeds that from equation (B6), for a sufficiently large sample size. Thus the power of the TDT exceeds that of the S-TDT. For example, if $T = .6$, $F = 400$, $t = 5$, and $s = 3$, then the probabilities from equations (B2) and (B6) are .992 and .988, respectively; these values are found in the last row of table 5.

Note that when $s = t - s$ (equal numbers of M_1M_2 and M_2M_2 sibs in the family), expressions (B2) and (B6) are equal, so that for this case the TDT and the S-TDT have equal power. This is demonstrated by the "2,2" column in table 5.

References

Bell GI, Horita S, Karam JH (1984) A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. Diabetes 33:176–183

Cox NJ, Spielman RS (1989) The insulin gene and susceptibility to IDDM. Genet Epidemiol 6:65–69

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

Duffy D (1995) Screening a 2 cM genetic map for allelic association: a simulated oligogenic trait. Genet Epidemiol 12:595–600

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57:455–464

——— Disease associations and the transmission/disequilibrium test (1997) In: Dracopoli NC (ed) Current protocols in human genetics. Suppl 15. Wiley, New York, pp. 1.12.1–1.12.13

Harley JB, Moser KL, Neas BR (1995) Logistic transmission modeling of simulated data. Genet Epidemiol 12:607–612

Hodge SE (1995) An oligogenic disease displaying weak marker associations: a summary of contributions to problem 1 of GAW9. Genet Epidemiol 12:545–554

Julier C, Hyer RN, Davies J, Merlin F, Soularue P, Briant L, Cathelineau G, et al (1991) Insulin-IGF2 region on chromosome 11p encodes a gene implicated in HLA-DR4-dependent diabetes susceptibility. Nature 354:155–159

Lander ES, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

McGinnis RE, Ewens WJ, Spielman RS (1995) The TDT reveals linkage and linkage disequilibrium in a rare disease. Genet Epidemiol 12:637–640

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

Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13:423–450

Sham PC, Curtis DR (1995) An extended transmission/disequilibrium test (TDT) for multi-allelic marker loci. Ann Hum Genet 59:323–336

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983–989

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