Am. J. Hum. Genet. 61:439–448, 1997

Tests for Linkage and Association in Nuclear Families

E. R. Martin,[1,2] N. L. Kaplan,[1] and B. S. Weir[2]

[1]Biostatistics Branch, NIEHS, Research Triangle Park, NC; and [2]Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh

Summary

The transmission/disequilibrium test (TDT) originally was introduced to test for linkage between a genetic marker and a disease-susceptibility locus, in the presence of association. Recently, the TDT has been used to test for association in the presence of linkage. The motivation for this is that linkage analysis typically identifies large candidate regions, and further refinement is necessary before a search for the disease gene is begun, on the molecular level. Evidence of association and linkage may indicate which markers in the region are closest to a disease locus. As a test of linkage, transmissions from heterozygous parents to all of their affected children can be included in the TDT; however, the TDT is a valid χ^2 test of association only if transmissions to unrelated affected children are used in the analysis. If the sample contains independent nuclear families with multiple affected children, then one procedure that has been used to test for association is to select randomly a single affected child from each sibship and to apply the TDT to those data. As an alternative, we propose two statistics that use data from all of the affected children. The statistics give valid χ^2 tests of the null hypothesis of no association or no linkage and generally are more powerful than the TDT with a single, randomly chosen, affected child from each family.

Introduction

Traditional methods of linkage analysis can be useful for identification of candidate regions for disease-susceptibility loci, but these regions are often quite large, and, therefore, further refinement of location is desirable before an attempt to map the gene physically is made. One approach for narrowing of the region of interest is to test for associations between the disease locus and markers in the candidate region (e.g., see Copeman et al. [1995]; Wang et al. [1996]). Following Spielman and Ewens (1996), we use the term "linkage disequilibrium" if there is association in the presence of linkage, to differentiate this case from association that occurs without linkage, as a result of other factors such as population stratification. In the absence of forces such as selection, mutation, and drift, in a random mating population, linkage disequilibrium generally is found only between tightly linked loci; therefore, evidence of linkage disequilibrium would suggest that the marker is physically close to a disease locus.

The transmission/disequilibrium test (TDT) was introduced by Spielman et al. (1993) as a test for linkage in the presence of association, but it also can be used as a test of association in the presence of linkage—that is, linkage disequilibrium—if the data consist of nuclear families with a single affected child (Spielman and Ewens 1996). As a test of linkage, the TDT is a valid χ^2 test, meaning that it has the correct significance level, even if families with multiple affected children are in the sample. However, it is well known that if there is linkage and no association, then affected siblings cannot be treated independently, and, consequently, the TDT is not a valid χ^2 test of linkage disequilibrium, for nuclear families with more than one affected child. A simple solution is to select randomly one affected child from each family and to apply the results for simplex families (e.g., see Wang et al. [1996]). This strategy, however, is less than optimal, since much of the data may be discarded, and a procedure that uses all of the affected individuals should be more powerful.

In this article, we propose two statistics that use data from all affected siblings from independent nuclear families and that are approximately χ^2 distributed when there is no linkage disequilibrium. We show how to construct the statistics for sib-pair data and how to combine sib-pair and simplex families, in a single analysis. We present evidence that the statistics give valid χ^2 tests for linkage disequilibrium and examine the power of the tests, for several alternative models. In addition, we compare the power of these tests to that of the alternative of sampling one affected child from each family and using the TDT.

Methods

In a recent editorial, Spielman and Ewens (1996) presented the statistic T_{mhet}, to generalize the TDT to multi-

Received February 11, 1997; accepted for publication May 12, 1997.

Address for correspondence and reprints: Dr. Eden R. Martin, Biostatistics Branch, NIEHS, P.O. Box 12233, Research Triangle Park, NC 27709. E-mail: eden@addax.niehs.nih.gov

0002-9297/97/6102-0024$02.00

allelic markers. The statistic uses data from nuclear families with at least one affected child. Parents and their affected children are genotyped at a marker locus with m alleles, and only parents who are heterozygous at the marker locus are included in the analysis. Let n_{ij} be the number of parents that transmit allele i and do not transmit allele j, to an affected child. Then,

$$T_{\text{mhet}} = \frac{m-1}{m} \sum_{i=1}^{m} \frac{(n_{i.} - n_{.i})^2}{n_{i.} + n_{.i}}, \qquad (1)$$

where $n_{i.} = \Sigma_{j=1}^{m} n_{ij}$ and $n_{.i} = \Sigma_{j=1}^{m} n_{ji}$. Since all parents are heterozygous at the marker locus, $n_{ii} = 0$ for all i.

If only nuclear families with a single affected child are sampled, then T_{mhet} has approximately a χ^2 distribution with $(m - 1)$ df, under the composite null hypothesis of no linkage or no association (i.e., no linkage disequilibrium) (Spielman and Ewens 1996; Kaplan et al. 1997). This result is not true, however, if there are families with multiple affected children in the sample. If one is testing only the simple hypothesis of no linkage, then T_{mhet}, calculated by use of all of the transmissions from each family in the sample, is still a valid χ^2 test. In fact, the families even can come from an extended pedigree. The complication caused by use of families with multiple affected children arises when the simple hypothesis of no association is tested, since the transmissions from a parent to each of his or her affected children are correlated if there is linkage, even if there is no association. Because of these correlations within families, T_{mhet} does not lead to a valid χ^2 test of the simple null hypothesis of no association or of the composite null hypothesis of no association or no linkage.

The statistics that we present focus on the set of transmissions from a parent to his or her affected children, rather than focusing on the individual transmissions to each child, as is the case with T_{mhet}. The motivation for this is that, conditional on the parental genotypes, the set of transmissions from one parent to his or her affected offspring is independent of the set of transmissions from the other parent, if there is no linkage disequilibrium. Kaplan et al. (1997) showed this for families with a single affected child, and their argument can be generalized easily if there are several affected children. In this way, we are able to maintain the independence property and thereby to obtain valid χ^2 tests.

Sib Pairs

We begin by examining the case in which the families in the sample each have two affected children. We first consider a marker locus with two alleles, M_1 and M_2, and suppose that there are h parents in the sample who are heterozygous at the marker locus. For the kth heterozygous parent, we define a random variable X_k as

$$X_k = \begin{cases} 2, & \text{if } M_1 \text{ is transmitted to both affected children;} \\ 1, & \text{if } M_1 \text{ is transmitted to one affected child and } M_2 \text{ to the other;} \\ 0, & \text{if } M_2 \text{ is transmitted to both affected children.} \end{cases}$$

In the appendix, we derive the distribution of X_k and show that, when there is no linkage disequilibrium, $\Pr(X_k = 2) = \Pr(X_k = 0)$ for all k. In addition, arguments similar to those in Kaplan et al. (1997) can be used to show that the X_{ij}'s are independent under the null hypothesis of no linkage disequilibrium. It follows that, under the null hypothesis, the X_{ij}'s are independent and identically distributed, with $E(X_k) = 1$ and with $\text{Var}(X_k) = E(X_k - 1)^2$.

Let s_{11} be the number of heterozygous parents that transmit M_1 to both affected children; let s_{22} be the number of heterozygous parents that transmit M_2 to both affected children; and let s_{12} be the number of heterozygous parents that transmit M_1 to one affected child and M_2 to the other. The sample mean can be written as $\bar{X} = (\Sigma_{k=1}^{h} X_k)/h = (2s_{11} + s_{12})/h$. An unbiased estimator of the variance of X_k is $\hat{V} = [\Sigma_{k=1}^{h} (X_k - 1)^2]/h = (s_{11} + s_{22})/h$. It follows from the central limit theorem that

$$Z^2 = \left(\frac{\sqrt{h}(\bar{X} - 1)}{\sqrt{\hat{V}}} \right)^2 = \frac{(s_{11} - s_{22})^2}{s_{11} + s_{22}}$$

is approximately a central χ^2 random variable with 1 df, when there is no linkage disequilibrium.

With sib-pair data, when there are only two alleles, we can write equation (1) as $T_{\text{mhet}} = 2(s_{11} - s_{22})^2/(s_{11} + s_{12} + s_{22})$. This allows us to write Z^2 as T_{sp}: $T_{\text{sp}} = (h/2h^*)T_{\text{mhet}}$, where $h = s_{11} + s_{12} + s_{22}$ is the number of heterozygous parents and where $h^* = s_{11} + s_{22}$ is the number of heterozygous parents who transmit the same allele to both affected offspring. Alternatively, we can define $T^*_{\text{mhet}} = 2(s_{11} - s_{22})^2/(s_{11} + s_{22})$, which is T_{mhet} calculated by use of only heterozygous parents who transmit the same allele to both affected siblings. This leads to a second form for Z^2, T_{su}: $T_{\text{su}} = T^*_{\text{mhet}}/2$. The two statistics T_{sp} and T_{su} are identical when there are two marker alleles, but important differences emerge when the marker has more than two alleles.

For a marker with $m > 2$ alleles, the development is more involved. For simplicity, we assume here that $m = 3$. The development for a larger m is analogous. To simplify the notation, we let N(0, 1) denote the normal distribution, with mean 0 and variance 1, and MVN (**0, I**) denote the multivariate normal distribution, with mean vector **0** and variance-covariance matrix **I**, the identity matrix.

To understand how the statistics for testing association are constructed, it is helpful to first explain why

Table 1

Transmitted and Nontransmitted Alleles for a Three-Allele Marker Locus

	Nontransmitted Allele		
Transmitted Allele	M_1	M_2	M_3
M_1	...	n_{12}	n_{13}
M_2	n_{21}	...	n_{23}
M_3	n_{31}	n_{32}	...

Note.—n_{ij} is the number of parents that transmit allele M_i but do not transmit allele M_j to an affected child. Transmissions from homozygous parents are not included.

T_{mhet} leads to an approximately valid χ^2 test of linkage. The data from heterozygous parents can be arranged in a 3×3 contingency table (table 1). We can write T_{mhet} as $T_{\text{mhet}} = \mathbf{Y'Y}$, where $\mathbf{Y}$ is a 3×1 vector with ith component

$$Y_i = \sqrt{\frac{2}{3}} \frac{n_{i.} - n_{.i}}{\sqrt{n_{i.} + n_{.i}}}, \qquad i = 1, 2, \text{ or } 3 .$$

The key point is to recognize that we can write Y_i as a linear combination of random variables that are independent and approximately N (0, 1), when there is no linkage. If we let

$$Z_{ij} = \frac{n_{ij} - n_{ji}}{\sqrt{n_{ij} + n_{ji}}}, \qquad 1 \leq i < j \leq 3 , \tag{2}$$

then we can rewrite Y_i as

$$Y_1 = \sqrt{\frac{2}{3}} \left(\sqrt{\frac{n_{12} + n_{21}}{n_{1.} + n_{.1}}} Z_{12} + \sqrt{\frac{n_{13} + n_{31}}{n_{1.} + n_{.1}}} Z_{13} \right) ;$$

$$Y_2 = \sqrt{\frac{2}{3}} \left(-\sqrt{\frac{n_{12} + n_{21}}{n_{2.} + n_{.2}}} Z_{12} + \sqrt{\frac{n_{23} + n_{32}}{n_{2.} + n_{.2}}} Z_{23} \right) ;$$

and

$$Y_3 = \sqrt{\frac{2}{3}} \left(-\sqrt{\frac{n_{13} + n_{31}}{n_{3.} + n_{.3}}} Z_{13} - \sqrt{\frac{n_{23} + n_{32}}{n_{3.} + n_{.3}}} Z_{23} \right) .$$

Note that $T_{\text{mhet}} = \mathbf{Y'Y} = \mathbf{Z'K'KZ}$, where

$$\mathbf{Z} = \begin{bmatrix} Z_{12} \\ Z_{13} \\ Z_{23} \end{bmatrix}$$

and where $\mathbf{K}$ is the 3×3 matrix

$$\mathbf{K} = \sqrt{\frac{2}{3}} \begin{bmatrix} \sqrt{\frac{n_{12} + n_{21}}{n_{1.} + n_{.1}}} & \sqrt{\frac{n_{13} + n_{31}}{n_{1.} + n_{.1}}} & 0 \\ -\sqrt{\frac{n_{12} + n_{21}}{n_{2.} + n_{.2}}} & 0 & \sqrt{\frac{n_{23} + n_{32}}{n_{2.} + n_{.2}}} \\ 0 & -\sqrt{\frac{n_{13} + n_{31}}{n_{3.} + n_{.3}}} & -\sqrt{\frac{n_{23} + n_{32}}{n_{3.} + n_{.3}}} \end{bmatrix} \tag{3}$$

It is important to notice the form of $\mathbf{K}$. In particular, for each row, the sum of squared elements is 1.

When there is no linkage, $\mathbf{Z}$ is approximately MVN ($\mathbf{0}$, $\mathbf{I}$), and it follows from theorem 4.6 in the work by Graybill (1961) that if $\mathbf{K'K}$ is idempotent, then the distribution of T_{mhet} is approximately χ^2 with df equal to the rank of $\mathbf{K'K}$. Only in the special case in which $n_{12} + n_{21} = n_{13} + n_{31} = n_{23} + n_{32}$ is it true that $\mathbf{K'K}$ is idempotent, with a rank of 2; however, the simulations from the study by Kaplan et al. (1997) provide evidence that the χ^2 approximation can be used even if the values for $n_{ij} + n_{ji}$ are not all equal. This suggests that if we take any vector that is approximately MVN ($\mathbf{0}$, $\mathbf{I}$) and multiply it by a matrix that has the same form as $\mathbf{K}$, then the product of the resulting vector and its transpose will be approximately χ^2 with 2 df. It is this observation that guides us in constructing the χ^2 tests for association.

As previously noted, the transmissions from a parent to each of his or her affected children are correlated when there is linkage. As a result, the Z_{ij}'s in equation (2) generally are not N(0, 1), under the null hypothesis of no association. However, we can construct statistics that are approximately N(0, 1) in a manner similar to that used for the two-allele case. Assume that there are h_{ij} parents with genotype M_iM_j ($1 \leq i < j \leq 3$) and that each parent has two affected children. For the kth parent with genotype M_iM_j, define

$$X_{ijk} = \begin{cases} 2, & \text{if } M_i \text{ is transmitted to both affected children;} \\ 1, & \text{if } M_i \text{ is transmitted to one affected child and } M_j \text{ to the other;} \\ 0, & \text{if } M_j \text{ is transmitted to both affected children.} \end{cases} \tag{4}$$

Similar to the two-allele case, if there is no linkage disequilibrium, then, conditional on the parent having heterozygous genotype M_iM_j, $\Pr(X_{ijk} = 2) = \Pr(X_{ijk} = 0)$ and $E(X_{ijk}) = 1$. If $\bar{X}_{ij} = \sum_{k=1}^{h_{ij}} X_{ijk}/h_{ij}$ and if $\hat{V}_{ij} = \sum_{k=1}^{h_{ij}} (X_{ijk} - 1)^2/h_{ij}$ is an unbiased estimator of V_{ij}, the variance of X_{ijk}, then

$$Z_{ij}^{u} = \frac{\sqrt{h_{ij}}(\bar{X}_{ij} - 1)}{\sqrt{\hat{V}_{ij}}}$$

is approximately N(0, 1). Furthermore, Z_{12}^{u}, Z_{13}^{u}, and Z_{23}^{u} are independent; therefore, $\mathbf{Z}^{u}$, the 3×1 vector of these components, is approximately MVN ($\mathbf{0}$, $\mathbf{I}$) when there is no linkage disequilibrium. Let $s_{ii,jj}$ be the number of M_iM_j parents that give allele M_i (and not allele M_j) to both children, and, similarly, let $s_{jj,ii}$ be the number that give allele M_j (and not allele M_i) to both children. It is convenient to define the sums $s_{ii.} = \Sigma_{j \neq i}\, s_{ii,jj}$ and $s_{.ii} = \Sigma_{j \neq i}\, s_{jj,ii}$. We can write

$$Z_{ij}^{u} = \frac{s_{ii,jj} - s_{jj,ii}}{\sqrt{s_{ii,jj} + s_{jj,ii}}}.$$

We define the matrix

$$\mathbf{K}^{u} = \sqrt{\frac{2}{3}} \begin{bmatrix} \sqrt{\frac{s_{11,22} + s_{22,11}}{s_{11.} + s_{.11}}} & \sqrt{\frac{s_{11,33} + s_{33,11}}{s_{11.} + s_{.11}}} & 0 \\ -\sqrt{\frac{s_{11,22} + s_{22,11}}{s_{22.} + s_{.22}}} & 0 & \sqrt{\frac{s_{22,33} + s_{33,22}}{s_{22.} + s_{.22}}} \\ 0 & -\sqrt{\frac{s_{11,33} + s_{33,11}}{s_{33.} + s_{.33}}} & -\sqrt{\frac{s_{22,33} + s_{33,22}}{s_{33.} + s_{.33}}} \end{bmatrix}.$$

The matrix $\mathbf{K}^{u}$ has the same form as equation (3), and so it follows from our earlier discussion that, when there is no linkage disequilibrium, $\mathbf{Z}^{u\prime}\mathbf{K}^{u\prime}\mathbf{K}^{u}\mathbf{Z}^{u}$ is approximately χ^2 with 2 df. Furthermore, we can write $T_{su} = \mathbf{Z}^{u\prime}\mathbf{K}^{u\prime}\mathbf{K}^{u}\mathbf{Z}^{u} = T_{mhet}^{*}/2$.

We can derive an alternative statistic by noting that if there is no linkage disequilibrium, then arguments similar to those in the appendix can be used to show that the distribution of X_{ijk} does not depend on i and j, and so $V_{12} = V_{13} = V_{23}$. Hence, we can use a pooled estimate of the variance,

$$\hat{V}^{p} = \frac{\sum_{i=1}^{2}\sum_{j=i}^{3}\sum_{k=1}^{h_{ij}}(X_{ijk} - 1)^2}{\sum_{i=1}^{2}\sum_{j=i}^{3} h_{ij}} = \frac{h^*}{h},$$

where the total number of heterozygous parents is $h = h_{12} + h_{13} + h_{23}$ and where the number of heterozygous parents that give the same allele to both affected children is h^*. We then write

$$Z_{ij}^{p} = \frac{\sqrt{h_{ij}}(\bar{X}_{ij} - 1)}{\sqrt{\hat{V}^{p}}} = \frac{\sqrt{h}(s_{ii,jj} - s_{jj,ii})}{\sqrt{h^* h_{ij}}}.$$

We define the matrix

$$\mathbf{K}^{p} = \sqrt{\frac{2}{3}} \begin{bmatrix} \sqrt{\frac{h_{12}}{h_{1.}}} & \sqrt{\frac{h_{13}}{h_{1.}}} & 0 \\ -\sqrt{\frac{h_{12}}{h_{2.}}} & 0 & \sqrt{\frac{h_{23}}{h_{2.}}} \\ 0 & -\sqrt{\frac{h_{13}}{h_{3.}}} & -\sqrt{\frac{h_{23}}{h_{3.}}} \end{bmatrix}.$$

Again, $\mathbf{K}^{p}$ has the same form as equation (3); therefore, when there is no linkage disequilibrium, $\mathbf{Z}^{p\prime}\mathbf{K}^{p\prime}\mathbf{K}^{p}\mathbf{Z}^{p}$ is approximately χ^2 with 2 df. In addition, $T_{sp} = \mathbf{Z}^{p\prime}\mathbf{K}^{p\prime}\mathbf{K}^{p}\mathbf{Z}^{p} = (h/2h^*)\,T_{mhet}$. It is important to note that T_{su} and T_{sp} generalize immediately to markers with more than three alleles. All that is required is that the general form of T_{mhet} in equation (1) be used.

The form of T_{sp} suggests a simple derivation of the distribution of the statistic, under the null hypothesis of no linkage. We know from previous results that, for a marker with m alleles, T_{mhet} is approximately χ^2 with $(m - 1)$ df (Spielman and Ewens 1996; Kaplan et al. 1997). In addition, if there is no linkage, then h^*/h converges in probability to $\frac{1}{2}$; therefore, T_{sp} also must be approximately χ^2 with $(m - 1)$ df, when there is no linkage. Unfortunately, no such simple argument exists for the case of linkage but no association.

In families in which the parents and both affected offspring have the same heterozygous genotype, one cannot determine whether a heterozygous parent transmits the same allele or different alleles to each of his or her affected children, and so the calculation of T_{sp} is problematic. For example, if parents and offspring all have marker genotype M_1M_2, then it is impossible to tell if one parent transmitted an M_1 to both children and the other parent transmitted an M_2 to both children or if both parents transmitted an M_1 to one child and an M_2 to the other. Such a family causes no difficulties in the calculation of T_{mhet}, but we do not know whether to add two observations or no observations to h^*. Under the null hypothesis of no linkage, either both parents transmit the same allele or both transmit different alleles to the affected siblings, with equal probability; therefore, a simple solution is to add the expected contribution from the family, under the null hypothesis, which means the addition of one observation to h^*, for each such family. With highly polymorphic markers, the number of these families is likely to be small relative to the number of unambiguous families, and so the error incurred by substitution of the expected value will be negligible. Alternatively, if the number of ambiguous families is not small, which may be the case with a biallelic marker locus, then the contribution of these

families can be approximated with the expected value of the contribution.

Combination of Data from Affected Sib Pairs and Singletons

The T_{sp} statistic can be generalized for the use of data from all independent nuclear families, regardless of the number of affected siblings, although the resulting statistic is not a simple function of T_{mhet}. Again, to simplify the arguments, the development here is for a marker locus with three alleles. Suppose that there are h_{Xij} parents with marker genotype M_iM_j and with two affected children. Let X_{ijk} be defined as in equation (4). In addition, let there be h_{Yij} parents with marker genotype M_iM_j and with a single affected child. We define, for the kth such parent, the random variable

$$Y_{ijk} = \begin{cases} 1, & \text{if } M_i \text{ and not } M_j \text{ is transmitted to the affected child;} \\ 0, & \text{if } M_j \text{ and not } M_i \text{ is transmitted to the affected child.} \end{cases}$$

We define

$$\bar{W}_{ij} = \frac{\sum_{k=1}^{h_{Xij}} X_{ijk} + \sum_{k=1}^{h_{Yij}} Y_{ijk}}{h_{Xij} + h_{Yij}}$$

and

$$Z_{ij} = \frac{\sqrt{h_{Xij} + h_{Yij}}\left(\bar{W}_{ij} - \dfrac{h_{Xij} + \frac{h_{Yij}}{2}}{h_{Xij} + h_{Yij}}\right)}{\sqrt{\dfrac{h_{Xij}\hat{V}_X^P + h_{Yij}\hat{V}_Y^P}{h_{Xij} + h_{Yij}}}},$$

which is approximately N(0, 1) if there is no linkage disequilibrium. We use estimates of the variances, $\hat{V}_X^P = h_X^*/h_X$ and $\hat{V}_Y^P = 1/4$, pooled over heterozygous parental genotypes, where h_X is the total number of heterozygous parents who have two affected children and where h_X^* is the total number of heterozygous parents who transmit the same allele to both children. We can rewrite

$$Z_{ij} = \frac{(2s_{Xii,jj} + s_{Yi,j} - 2s_{Xjj,ii} - s_{Yj,i})}{\sqrt{\dfrac{h_X^*}{h_X}4h_{Xij} + h_{Yij}}},$$

where $s_{Xii,jj}$ is, as before, the number of parents with genotype M_iM_j who have two affected children and who give M_i to both children, and where $s_{Yi,j}$ is the number of parents with genotype M_iM_j who have a single affected child and who transmit M_i to that child.

We define the matrix **K** by

$$\mathbf{K} = \sqrt{\frac{2}{3}}\begin{pmatrix} \sqrt{\dfrac{2h_{X12} + h_{Y12}}{2h_{X1.} + h_{Y1.}}} & \sqrt{\dfrac{2h_{X13} + h_{Y13}}{2h_{X1.} + h_{Y1.}}} & 0 \\ -\sqrt{\dfrac{2h_{X12} + h_{Y12}}{2h_{X2.} + h_{Y2.}}} & 0 & \sqrt{\dfrac{2h_{X23} + h_{Y23}}{2h_{X2.} + h_{Y2.}}} \\ 0 & -\sqrt{\dfrac{2h_{X13} + h_{Y13}}{2h_{X3.} + h_{Y3.}}} & -\sqrt{\dfrac{2h_{X23} + h_{Y23}}{2h_{X3.} + h_{Y3.}}} \end{pmatrix}.$$

Then, as before, the χ^2 statistic for the test for association is

$$T_{sp} = \mathbf{Z'K'KZ} = \left(\frac{2}{3}\right)\sum_{i=1}^{3}\frac{\left(\sum_{j\neq i}\sqrt{\dfrac{2h_{Xij} + h_{Yij}}{\frac{h_X^*}{h_X}4h_{Xij} + h_{Yij}}}(2s_{Xii,jj} + s_{Yi,j} - 2s_{Xjj,ii} - S_{Yj,i})\right)^2}{2h_{Xi.} + h_{Yi.}}.$$

Note that, when there are only singletons, the test statistic reduces to T_{mhet} and that, when there are only sib pairs, it reduces to $h_X T_{mhet}/h_X^*$. For general m, the statistic is modified by replacement of the $2/3$ with $(m - 1)/m$ and by extension of the sums from 1 to m.

Similar arguments can be used to combine families with larger affected sibships, so that all nuclear families can be used in a single analysis. We do not present these derivations in this article.

Monte Carlo Test

An alternative to the use of χ^2 critical values is the use of Monte Carlo randomization techniques, to determine an empirical P value. These methods particularly may be useful when samples are small, since Monte Carlo tests always attain significance levels close to the nominal level. We use the same procedure outlined by Kaplan et al. (1997). The only additional point is that the labels "transmitted" and "nontransmitted" must be permuted for each set of sibs as a whole rather than for each sib independently. The justification for this is that if a pair of affected sibs has a heterozygous parent with genotype M_iM_j, then, when there is no linkage disequilibrium, the probability that the parent transmits M_i to both children is equal to the probability that the parent transmits M_j to both children and the probability that the parent transmits M_i to the first child and M_j to the second is equal to the probability that the parent transmits M_j to the first child and M_i to the second.

The procedure is equivalent to the Monte Carlo–Markov Chain (MCMC) method of Cleves et al. (1997).

They recommend use of the statistic T^0_{mb}, which is simply $mT_{\text{mhet}}/(m-1)$. For sib pairs, they exclude the parents who transmit different marker alleles to each child, since these parents are uninformative. The Monte Carlo test using the statistic T_{su} is equivalent to their MCMC test. We find very little difference in the power of the Monte Carlo test that uses T_{sp} and the one that uses T_{su}, and so either can be used in the Monte Carlo test.

Simulations

Simulations were used to estimate the significance levels and the power of the tests discussed. In all cases, the nominal significance level used was .05. We used a five-allele marker locus with unimodal (one allele with frequency .7 and the rest with frequency .075), bimodal (two alleles with frequency .35 and the rest with frequency .1), or uniform (all alleles with frequency .2) allele frequencies among chromosomes carrying the normal allele. Even though our analytic development is for a three-allele marker locus, we chose to use a five-allele marker for our simulations because this is, perhaps, more realistic for the microsatellite markers currently used in genetic mapping. A biallelic disease locus, with the disease allele D_1 having frequency .05, was used. Three models of inheritance were examined—recessive ($f_{11} = .4$ and $f_{12} = f_{21} = f_{22} = 0$), additive ($f_{11} = .0181$, $f_{12} = f_{21} = 0.0091$, and $f_{22} = .0001$), and multiplicative ($f_{11} = .012$, $f_{12} = f_{21} = .003$, and $f_{22} = .00075$), where f_{ij} is the penetrance (i.e., the probability that an individual with genotype D_iD_j is affected with the disease). These models were chosen so that the disease prevalence was about .001.

In all our simulations, we let the marker be linked completely to the disease locus, since we were interested in the behavior of the tests when there is linkage. Significance levels were estimated by simulation of data with no association. In particular, marker allele frequencies for chromosomes with the disease allele were set equal to those for the normal chromosomes. Power estimates were calculated with haplotype frequencies that were chosen in order to give varying degrees of association measured by an index I^*. The definition of I^* and a discussion of how haplotype frequencies were selected are given in the study by Kaplan et al. (1997). All simulated data consists of independent nuclear families with two affected children.

Results

We first demonstrated that both T_{sp} and T_{su} are valid χ^2 tests of association, when the data are composed of sib pairs. The significance levels in table 2 were estimated for the tests using five-allele markers with unimodal, bimodal, or uniform distributions of allele frequencies. Estimates are shown for sample sizes of 25, 50, and 100 families with two affected children. Note that the actual number of heterozygous parents is a random variable depending on the model parameters. There is complete linkage between disease and marker loci but no association between alleles at the two loci. To estimate each significance level, we calculated the proportion of 10,000 simulated data sets that had statistics larger than the χ^2 critical value.

For small samples, both T_{sp} and T_{su} gave tests that are conservative, and, in almost every example, the test using T_{su} was more conservative than the test using T_{sp}. This result is not surprising, since T_{sp} uses the pooled variance estimate, which is more accurate than independent estimation of variances, when samples are small. In either case, when tables are sparse, a Monte Carlo test can be used to correct for the conservativeness. In table 3, we show, for a sample size of 25 families, that the suggested Monte Carlo procedure gave valid tests that were less conservative than the χ^2 tests. Only the results for the unimodal marker allele are shown. The bimodal and uniform markers gave similar results. To estimate each P value, 99 pseudosamples were drawn. Again, 10,000 P values were generated for each significance-level estimate.

We examined the relative power of the T_{su} and T_{sp} χ^2 tests and found little difference in the powers of the two tests. These results are not included, for the sake of brevity. With a sample size of 100 families with affected sib pairs, there was no difference between the powers of T_{su} and of T_{sp}. For smaller sample sizes, T_{sp} was somewhat more powerful than T_{su}, as would be expected, on the basis of the significance-level results. With the Monte Carlo test, there was no difference in the powers of T_{su} and T_{sp}. The Monte Carlo test can be slightly more powerful than the χ^2 test, when the sample size is small. In light of these results, the following simulations consider only T_{sp} and use the χ^2 critical values to conduct the test.

The TDT is a valid χ^2 test of association if a single affected child is chosen randomly from each family; however, this test can be much less powerful than the test using T_{sp}. This was demonstrated, by simulation, for the unimodal marker. Using 100 families with two affected children, we applied T_{sp}. We also randomly sampled an affected child from each sib pair and calculated T_{mhet}. Both procedures provided valid χ^2 tests of linkage disequilibrium, and their estimated powers are compared in figure 1. For the multiplicative and additive models, T_{sp} can have a great deal more power than T_{mhet} calculated from a random sample of children. For the recessive model, the powers of the two tests were almost identical. This last result is not surprising, since, with a recessive disease, complete linkage between the marker and disease loci, and an infrequent disease allele, almost all sib pairs will receive the same marker allele from

Table 2

Estimates of Significance Levels for the T_{sp} and T_{su} χ^2 Tests Based on 10,000 Simulated Data Sets, for a Five-Allele Marker

	Significance Level[a]					
	$F = 100$		$F = 50$		$F = 25$	
Marker and Model	T_{sp}	T_{su}	T_{sp}	T_{su}	T_{sp}	T_{su}
Unimodal:						
Recessive	.0495	.0493	.0493	.0494	.0426	.0428
Multiplicative	.0503	.0491	.0487	.0453	.0454	.0344
Additive	.0510	.0495	.0485	.0477	.0411	.0339
Bimodal:						
Recessive	.0493	.0490	.0483	.0486	.0466	.0452
Multiplicative	.0490	.0481	.0462	.0452	.0438	.0388
Additive	.0494	.0494	.0485	.0470	.0472	.0434
Uniform:						
Recessive	.0490	.0492	.0462	.0461	.0454	.0457
Multiplicative	.0501	.0501	.0456	.0449	.0441	.0409
Additive	.0482	.0476	.0473	.0470	.0446	.0428

[a] For each data set, F families with two affected children were sampled. The nominal significance level is .05.

their parent. When each parent transmits the same allele to both affected children, T_{sp} is equal to $T_{mhet}/2$. This is exactly the result obtained when a child was randomly sampled from each pair. So, under the recessive model with complete linkage, the two statistics should be approximately equal.

The estimates of power in figure 1, do not always vary monotonically with I^*. For example, in the additive model, the marker with $I^* = .0474$ has an estimated power of .9791, whereas the marker with $I^* = .0651$ has an estimated power of .9563. Kaplan et al. (1997) showed that the value of I^* can be used to predict the power of the χ^2 test using T_{mhet}, for families with a single affected child. This is not the case if the families have affected sib pairs. However, the plots in figure 1 suggest that I^* could be used to rank markers, in terms of power, with little error.

Table 3

Estimates of Significance Levels for the T_{sp} and T_{su} Monte Carlo Tests Based on 10,000 Simulated Data Sets, for a Five-Allele Marker

	Significance Level[a]	
Marker and Model	T_{sp}	T_{su}
Unimodal:		
Recessive	.0506	.0506
Multiplicative	.0498	.0503
Additive	.0478	.0482

[a] For each data set, 25 families with two affected children were sampled, and P values were calculated by the drawing of 99 pseudosamples. The nominal significance level is .05.

Discussion

The TDT originally was proposed by Spielman et al. (1993) as a test of linkage in the presence of association. The test uses parents who are heterozygous at the marker locus and compares the frequencies of marker alleles that are transmitted to affected children with the frequencies of marker alleles that are not transmitted. If there is no linkage, then these frequencies should be equal. It is important to note that, when testing for linkage, transmissions to all the affected children can be included in the test even if the children are related. For a marker with m alleles, T_{mhet} is approximately χ^2 with $(m - 1)$ df when there is no linkage (Spielman and Ewens 1996; Kaplan et al. 1997).

If the sample consists of only nuclear families with a single affected child, then the TDT also can be used to test for association in the presence of linkage. However, the TDT is not a valid χ^2 test of association if families in the sample have several affected sibs, since the transmissions to affected sibs are not independent when there is linkage. One way to deal with this problem is to sample randomly one affected child from each family and to use the TDT. This method discards much of the information in the sample and, therefore, is not as powerful as the method proposed in this paper. The method that we propose considers the set of transmissions to affected sibs

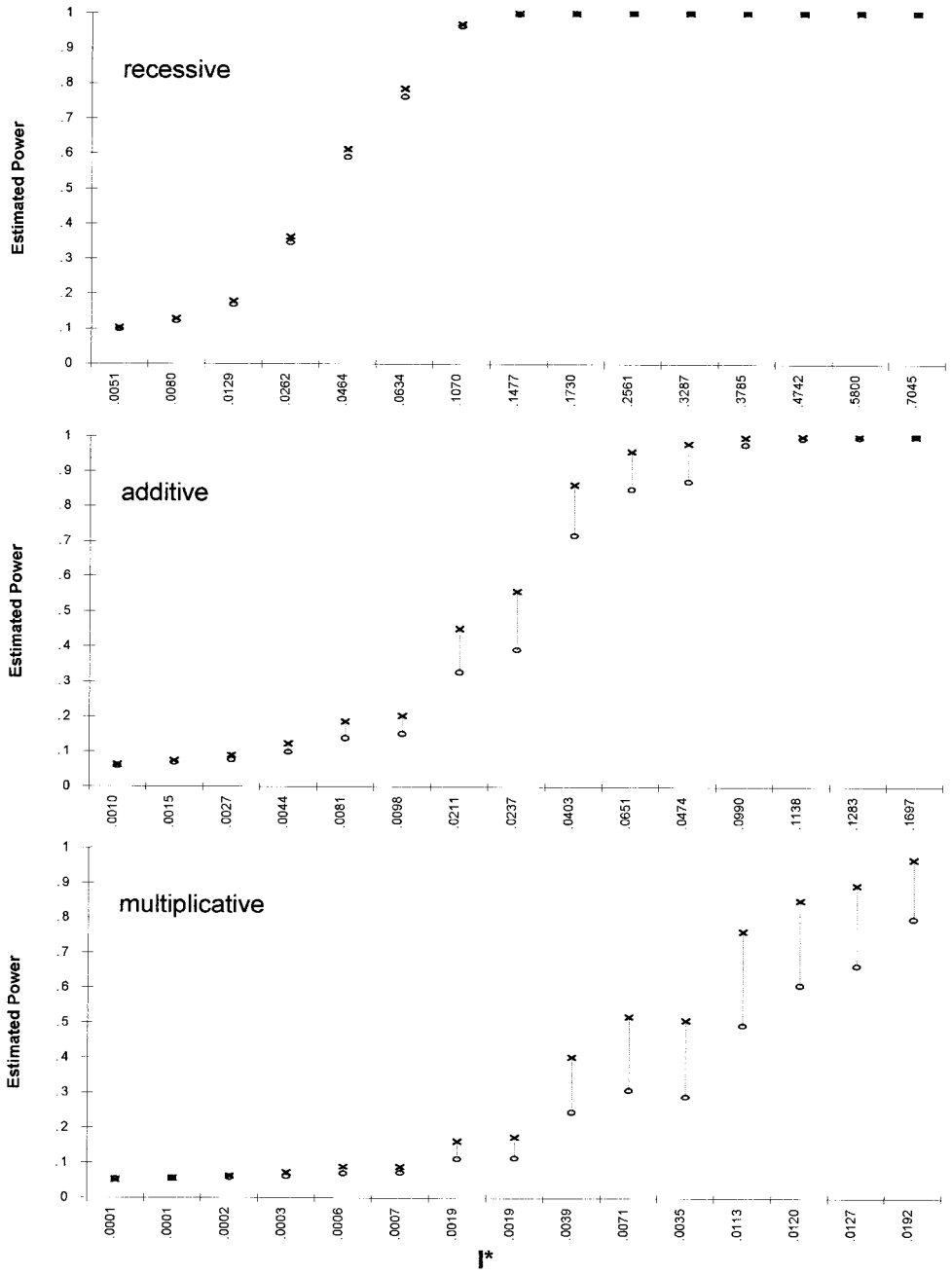

Figure 1 Estimated powers of the T_{sp} χ^2 test (indicated by "×") and the T_{mhet} χ^2 test performed by random sampling of a child from each family (indicated by an oval [○]). We simulated data for 100 families with two affected children each, assuming complete linkage between a disease locus and a five-allele unimodal marker locus and varying degrees of association, measured by I^*. Estimates are shown for three models of inheritance. For each power estimate, 10,000 data sets were simulated, and .05 was the nominal significance level used in both tests.

in a family rather than the transmissions to each child separately. In this way, the independence property is retained, and statistics that lead to valid χ^2 tests of association in the presence of linkage can be constructed.

The two statistics that we define differ only in how the variance V_{ij} is estimated. For T_{sp}, the data are pooled to obtain a single estimate of V_{ij}, whereas, for T_{su}, the data are not pooled and each V_{ij} is estimated individually. Both statistics lead to valid χ^2 tests of linkage disequilibrium, and simulation studies indicate that their powers are similar, with T_{sp} being slightly more powerful when the sample size is small. This is not unexpected, since the pooled estimate of the variance should be more accurate than the individual estimates.

An alternative to the use of χ^2 critical values is the use of a Monte Carlo procedure, to estimate P values. This procedure is the same as the one described by Kaplan et al. (1997), provided that the labels "transmitted" and "nontransmitted" are permuted for the set of affected sibs rather than for each sib independently. Since the Monte Carlo procedure always leads to a test with a significance level close to the nominal value, we recommend its use if the sample size is small, to guard against an overly conservative test.

The tests for linkage disequilibrium presented in this paper are for independent nuclear families. However, in practice, the data may consist of extended pedigrees with several affected individuals. In principle, it is possible to generalize the arguments discussed here, in order to construct a single test that accommodates all affected relatives. For instance, one can estimate the mean and the variance between transmissions to affected cousin pairs, just as was done for affected sib pairs, and the include these data in the statistic, with appropriate weights. If the data consist of several extended pedigrees, then this may be a worthwhile exercise that presumably would lead to a more powerful test of linkage disequilibrium.

Acknowledgments

This work was supported, in part, by NIH grants P01 GM45344, T32 GM08443, and R01 NS23360.

Appendix

Consider a marker locus with m alleles, $M_1, M_2, \ldots, M_m$, having frequencies $q_1, q_2, \ldots, q_m$. Suppose that there is a disease locus with d alleles, $D_1, D_2, \ldots, D_d$, having frequencies $p_1, p_2, \ldots, p_d$. We denote the conditional probability that a gamete carries marker allele M_i, given that it carries allele D_r as $\Pr(M_i|D_r)$. Let θ be the recombination fraction between the disease and the marker loci. We denote the penetrances, the probability that an individual is affected, given that he or she has genotype D_rD_s, as f_{rs}. Finally, we introduce the following notation: $f^*_{rtu} = (f_{rt} + f_{ru})/4$ and $B^{rs}_{ij} = \Pr(M_i|D_r)\Pr(M_j|D_s) + \Pr(M_i|D_s)\Pr(M_j|D_r)$.

Let the probability that a parent transmits M_i to both children and does not transmit M_j, given that he or she transmits D_r to both children and does not transmit D_s, be $T^{(rs;rs)}_{ij;ij}$. The probability $T^{(rs;sr)}_{ij;ij}$, $T^{(rs;rs)}_{ij;ji}$, and $T^{(rs;sr)}_{ij;ji}$ are defined analogously. When random mating is assumed, these probabilities can be written as

$$T^{(rs;rs)}_{ij;ij} = T^{(rs;sr)}_{ij;ji} = \Pr(M_i|D_r)\Pr(M_j|D_s)(1-\theta)^2 + \Pr(M_i|D_s)\Pr(M_j|D_r)\theta^2$$

and

$$T^{(rs;sr)}_{ij;ij} = T^{(rs;rs)}_{ij;ji} = B^{rs}_{ij}(1-\theta)\theta .$$

Using these probabilities, we can calculate the probability that a parent with genotype M_iM_j transmits M_i to both affected children and does not transmit M_j, given that he or she has two affected children:

$$P_{ij;ij} = P_{ji;ji} = \frac{\sum_{r,s,t,u} p_r p_s p_t p_u f^*_{rtu}(f^*_{rtu}T^{rs;rs}_{ij;ij} + f^*_{stu}T^{rs;sr}_{ij;ij})}{\sum_{r,s,t,u} p_r p_s p_t p_u f^*_{rtu}(f^*_{rtu} + f^*_{stu})B^{rs}_{ij}};$$

and we can calculate the probability that a parent with genotype M_iM_j transmits M_i and not M_j to one child and M_j and not M_i to the other, given that he or she has two affected children:

$$P_{ij;ji} = p_{ji;ij} = \frac{\sum_{r,s,t,u} p_r p_s p_t p_u f^*_{rtu}(f^*_{rtu}T^{rs;rs}_{ij;ji} + f^*_{stu}T^{rs;sr}_{ij;ji})}{\sum_{r,s,t,u} p_r p_s p_t p_u f^*_{rtu}(f^*_{rtu} + f^*_{stu})B^{rs}_{ij}}.$$

When there is no linkage, $\theta = 1/2$, and these probabilities reduce to

$$P_{ij;ij} = P_{ij;ji} = P_{ji;ji} = P_{ji;ij} = 1/4 . \quad \text{(A1)}$$

When there is no association, $\Pr(M_i|D_r) = \Pr(M_i|D_s) = q_i$, and the above probabilities become

$$P_{ij;ij} = P_{ji;ji} = \frac{1}{4} + (1-2\theta)^2 \times \frac{\sum_{r,s,t,u} p_r p_s p_t p_u f^*_{rtu}(f^*_{rtu} - f^*_{stu})}{4\sum_{r,s,t,u} p_r p_s p_t p_u f^*_{rtu}(f^*_{rtu} + f^*_{stu})} \quad \text{(A2)}$$

and

$$P_{ij;ji} = P_{ji;ij} = \frac{1}{4} - (1-2\theta)^2 \times \frac{\sum_{r,s,t,u} p_r p_s p_t p_u f^*_{rtu}(f^*_{rtu} - f^*_{stu})}{4\sum_{r,s,t,u} p_r p_s p_t p_u f^*_{rtu}(f^*_{rtu} + f^*_{stu})} . \quad \text{(A3)}$$

For a marker locus with two alleles, $\Pr(X_k = 2) = P_{12;12}$, $\Pr(X_k = 1) = P_{12;21} + P_{21;12}$, and $\Pr(X_k = 0) = P_{21;21}$. It follows from equations (A1), (A2), and (A3) that, when there is no linkage disequilibrium, $\Pr(X_k = 2) = \Pr(X_k = 0)$. In addition, the probabilities given in equations (A2) and (A3) depend only on the θ, the pene-

trances, and the disease allele frequencies but are independent of the allele frequencies at the marker locus. Consequently, when there is no linkage disequilibrium, the same probabilities hold, even for a marker locus with multiple alleles.

References

Cleves MA, Olson JM, Jacobs KB. Exact transmission-disequilibrium tests for candidate gene testing and genomic screening with multiallelic markers. Genet Epidemiol (in press)

Copeman JB, Cucca F, Hearne CM, Cornall RJ, Reed PW, Ronningen KS, Undlien DE, et al (1995) Linkage disequilibrium mapping of a type 1 diabetes susceptibility gene (IDDM7) to chromosome 2q31-q33. Nat Genet 9:80–85

Graybill FA (1961) An introduction to linear statistical models. McGraw-Hill, New York

Kaplan NL, Martin ER, Weir BS (1997) Power studies for the transmission/disequilibrium tests with multiple alleles. Am J Hum Genet 60:691–702

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

Wang S, Detera-Wadleigh SD, Coon H, Sun C-e, Goldin LR, Duffy DL, Byerley WF, et al (1996) Evidence of linkage disequilibrium between schizophrenia and the *SCA1* CAG repeat on chromosome 6p23. Am J Hum Genet 59:731–736