Am. J. Hum. Genet. 64:861–870, 1999

The Transmission/Disequilibrium Test and Parental-Genotype Reconstruction: The Reconstruction-Combined Transmission/ Disequilibrium Test

Michael Knapp

Institute for Medical Statistics, University of Bonn, Bonn

Summary

Spielman and Ewens recently proposed a method for testing a marker for linkage with a disease, which combines data from families with and without information on parental genotypes. For some families without parental-genotype information, it may be possible to reconstruct missing parental genotypes from the genotypes of their offspring. The treatment of such a reconstructed family as if parental genotypes have been typed, however, can introduce bias. In the present study, a new method is presented that employs parental-genotype reconstruction and corrects for the biases resulting from reconstruction. The results of an application of this method to a real data set and of a simulation study suggest that this approach may increase the power to detect linkage.

Introduction

The transmission/disequilibrium test (TDT), introduced by Spielman et al. (1993), is a simple and powerful method to detect linkage between a marker and a disease-susceptibility locus, in the presence of linkage disequilibrium between the two loci. It considers parents heterozygous for a certain allele (*A*) at the marker locus and counts the number of times that such parents transmitted *A* to their affected offspring. Therefore, the TDT requires that the affected offspring and also their parents have been typed at the marker locus. The availability of parental marker genotypes can pose a problem, especially when the disease of interest has a late age of onset. For this reason, Spielman and Ewens recently proposed a method called the "sib TDT" (S-TDT), which does not require parental marker genotypes, but instead uses marker genotypes of unaffected siblings. Their approach compares the observed number of *A* alleles in affected children with the number expected with no linkage, conditioned on the observed distribution of marker genotypes in the whole sibship. They also show how to combine data from families in which parental genotypes are available with data from families in which genotypes of unaffected sibs are available but genotypes of parents are not (Spielman and Ewens 1998). In the present article, the procedure will be called the "combined TDT" (C-TDT).

For some families without parental-genotype information, it may be possible to reconstruct parental genotypes from the genotypes of their offspring. Spielman and Ewens (1996, but see also 1999) have suggested that one can treat these reconstructed families as if parental genotypes have been typed. Curtis (1997), however, shows that such a procedure can introduce bias. To give a simple example, consider a family with one affected and one unaffected child, and assume that the genotypes of both parents are actually *AB*. Parental genotypes can be reconstructed only if one child has genotype *AA* and the other has genotype *BB*. If there is no linkage between the marker and the disease, one-half of the time the affected sib will have genotype *AA*, whereas in the other half of such families, the affected sib will have *BB*. Therefore, the null expectation of the number of alleles *A* in the affected sib is 1, but the null variance is also 1, which is two times larger than the variance of *A* in an affected offspring of a double-heterozygous *AB* mating with typed parents. Because of this increased variance, treating such a reconstructed family as if parental genotypes have been typed will inflate the type I error rate of the TDT.

Curtis (1997) claims that correcting this bias would require the knowledge of population marker–allele frequencies. Such reliance on population frequencies would not be opportune, however, because a key benefit of the TDT is lost. On the other hand, deducing the parental genotypes, when possible, is a quite natural and attrac-

Received August 28, 1998; accepted for publication January 7, 1999; electronically published February 18, 1999.

Address for correspondence and reprints: Dr. Michael Knapp, Institute for Medical Statistics, University of Bonn, Sigmund-Freud-Strasse 25, D-53105 Bonn, Germany. E-mail: knapp@imsdd.meb.uni-bonn.de

0002-9297/99/6403-0024$02.00

Table 1

First and Second Moment of *T* in Families with Reconstructed Parental Genotypes, When Both Parental Genotypes Are Missing

Parental Mating Type	Condition (R)	$E_{H_0}(T \mid R)$	$E_{H_0}(T^2 \mid R)$
$AB \times AB$	$N^{AA} > 0$ and $N^{BB} > 0$	n_a	$n_a\left(n_a + \frac{1}{2}\right) + n_a\left[\frac{(5-2n_a)3^{n_c-2}-2^{n_c-1}}{4^{n_c}-2\cdot 3^{n_c}+2^{n_c}}\right]$
$AB \times AC$	($N^{AA} > 0$ and $N^{BC} > 0$) or ($N^{AA} > 0$ and $N^{AB} > 0$ and $N^{AC} > 0$)	$n_a\left(1 + \frac{3^{n_c-1}-2^{n_c}+1}{4^{n_c}-3^{n_c}-2^{n_c+1}+3}\right)$	$n_a\left[\frac{(n_a+1/2)4^{n_c}-(4n_a+2)3^{n_c-2}-(9n_a+1)2^{n_c-1}+6n_a}{4^{n_c}-3^{n_c}-2^{n_c+1}+3}\right]$
$AB \times BC$	($N^{BB} > 0$ and $N^{AC} > 0$) or ($N^{BB} > 0$ and $N^{AB} > 0$ and $N^{BC} > 0$)	$\frac{n_a}{2}\left(1 - \frac{3^{n_c-1}-2^{n_c}+1}{4^{n_c}-3^{n_c}-2^{n_c+1}+3}\right)$	$n_a\left[\frac{(n_a+1)4^{n_c-1}-(4n_a+2)3^{n_c-2}-(n_a+1)2^{n_c-2}+n_a}{4^{n_c}-3^{n_c}-2^{n_c+1}+3}\right]$
$AB \times CD$	($N^{AC} > 0$ and $N^{BD} > 0$) or ($N^{AD} > 0$ and $N^{BC} > 0$)	$\frac{n_a}{2}$	$\frac{n_a^2}{4} + \frac{n_a}{4}\left[\frac{4^{n_c-1}-(n_a+1)2^{n_c-1}+n_a}{(2^{n_c-1}-1)^2}\right]$

tive approach for a geneticist. Parental-genotype reconstruction generally will use information from affected and (when available) unaffected sibs. Thus, discussion of the appropriate analysis, when reconstruction is carried out, is best done in the context of the C-TDT. In the present article, this procedure is called the "reconstruction combined TDT" (RC-TDT).

In the present article, how to reconstruct parental genotypes and, at the same time, how to allow for the biases involved in this reconstruction are described. The RC-TDT is then applied to a data set used by Spielman and Ewens (1998), and the results are compared with those of the TDT and the S-TDT. Finally, a simulation study that uses genetic models described elsewhere (Boehnke and Langefeld 1998) is used to evaluate the power and the true size of the RC-TDT and S-TDT.

Methods

Notation

It will be assumed that there is a specific allele (A) at the marker locus that is of particular interest. Because there are at most four different alleles segregating in a single family, and because families without allele A are uninformative for the present purpose, it is sufficient to consider a marker locus with four alleles A, B, C, and D. Thus, B, C, and D may denote different alleles, across families.

The sample consists of m nuclear families (parents and children). For $1 \leq i \leq m$, n_{ai} denotes the number of affected children, n_{ui} denotes the number of unaffected children, and $n_{ci} = n_{ai} + n_{ui}$ denotes the size of the sibship for family i. In each family, all children have been typed at the marker locus, but the marker genotypes for at least one parent in some families may be unavailable. Let $N_{ai}^g(N_{ui}^g)$ be random variables, denoting the number of affected (or unaffected) children with genotype g in family i. Small letters (i.e., n_{ai}^g and n_{ui}^g) are used to denote the observed values of N_{ai}^g and N_{ui}^g. Further, let $N_i^g = N_{ai}^g + N_{ui}^g$ and $n_i^g = n_{ai}^g + n_{ui}^g$ denote the random variable and the observed number of children with genotype g in family i, respectively. T_i denotes the number of A alleles in affected children (i.e., $T_i = 2N_{ai}^{AA} + N_{ai}^{AB} + N_{ai}^{AC} + N_{ai}^{AD}$).

Parental-Genotype Reconstruction

In some families, it will be possible to reconstruct missing parental marker genotypes on the basis of the observed genotypes in the offspring. As the example in the introduction illustrates, the null expectation and variance of the number of alleles A, transmitted by heterozygous parents to their affected children in such reconstructed families, do not necessarily equal the corresponding expressions for completely typed families. However, it is possible to calculate the null expectation and variance of T_i, conditional on the event that missing parental genotypes could be reconstructed. Two situations have to be distinguished: (1) when both parental genotypes are missing and (2) when one parental genotype is missing but the other parental genotype has been typed.

When both parental genotypes are missing.—A nec-

Table 2

First and Second Moment of T in Families with Reconstructed Parental Genotypes, When Only One Parental Genotype Is Missing

Parental Mating Type	Condition (R)	$E_{H_0}(T \mid R)$	$E_{H_0}(T^2 \mid R)$
$AA \times AB$	$N^{AA} > 0$ and $N^{AB} > 0$	$\frac{3}{2} n_a$	$\frac{9}{4} n_a^2 + \frac{n_a}{4} \cdot \frac{2^{n_c-1} - n_a}{2^{n_c-1} - 1}$
$AB \times AB$	$N^{AA} > 0$ and $N^{BB} > 0$	n_a	$n_a\left(n_a + \frac{1}{2}\right) + n_a\left[\frac{(5-2n_a)3^{n_c-2} - 2^{n_c-1}}{4^{n_c} - 2 \cdot 3^{n_c} + 2^{n_c}}\right]$
$AB \times AC$	($N^{AA} > 0$ and $N^{BC} > 0$) or ($N^{AA} > 0$ and $N^{AC} > 0$)	$n_a\left(1 + \frac{3^{n_c-1} - 2^{n_c-1}}{4^{n_c} - 3^{n_c} - 2^{n_c} + 1}\right)$	$n_a\left[\frac{(n_a + 1/2)4^{n_c} - (4n_a + 2)3^{n_c-2} - (9n_a + 1)2^{n_c-2} + n_a}{4^{n_c} - 3^{n_c} - 2^{n_c} + 1}\right]$
$AB \times BC$	($N^{BB} > 0$ and $N^{AC} > 0$) or ($N^{BB} > 0$ and $N^{BC} > 0$)	$\frac{n_a}{2}\left(1 - \frac{3^{n_c-1} - 1}{4^{n_c} - 3^{n_c} - 2^{n_c} + 1}\right)$	$n_a\left[\frac{(n_a + 1)4^{n_c-1} - (4n_a + 2)3^{n_c-2} - (n_a + 1)2^{n_c-2} + n_a}{4^{n_c} - 3^{n_c} - 2^{n_c} + 1}\right]$
$AB \times CD$	($N^{AC} > 0$ or $N^{BC} > 0$) and ($N^{AD} > 0$ or $N^{BD} > 0$)	$\frac{n_a}{2}$	$\frac{n_a}{4}(n_a + 1)$
$BB \times AB$	$N^{AB} > 0$ and $N^{BB} > 0$	$\frac{n_a}{2}$	$\frac{n_a^2}{4} + \frac{n_a}{4} \cdot \frac{2^{n_c-1} - n_a}{2^{n_c-1} - 1}$
$BB \times AC$	$N^{AB} > 0$ and $N^{BC} > 0$	$\frac{n_a}{2}$	$\frac{n_a^2}{4} + \frac{n_a}{4} \cdot \frac{2^{n_c-1} - n_a}{2^{n_c-1} - 1}$
$BC \times AB$	($N^{AB} > 0$ or $N^{AC} > 0$) and ($N^{BB} > 0$ or $N^{BC} > 0$)	$\frac{n_a}{2}$	$\frac{n_a^2}{4} + \frac{n_a}{4} \cdot \frac{2^{n_c-1} - n_a}{2^{n_c-1} - 1}$
$BC \times AD$	($N^{AB} > 0$ or $N^{AC} > 0$) and ($N^{BD} > 0$ or $N^{CD} > 0$)	$\frac{n_a}{2}$	$\frac{n_a^2}{4} + \frac{n_a}{4} \cdot \frac{2^{n_c-1} - n_a}{2^{n_c-1} - 1}$

essary condition of a missing parental genotype that can be reconstructed is that it is heterozygous. Thus, when both parental genotypes are missing, in order for them to be reconstructed, both genotypes must be heterozygous. Further, in the present context, only those families in which at least one parent is heterozygous for allele A are of interest. Four different parental mating types have to be distinguished: (1) both parents are heterozygous for allele A, with the same genotype (e.g., $AB \times AB$); (2) both parents are heterozygous for allele A, but with different genotypes (e.g., $AB \times AC$); (3) both parents are heterozygous for some allele other than A, but one parent is heterozygous for A (e.g., $AB \times BC$); and (4) one parent is heterozygous for allele A, and there are four different parental alleles (e.g., $AB \times CD$). These parental mating types are listed in table 1. The second column of table 1 presents a necessary and sufficient condition, for the observed marker genotypes in the offspring, to allow reconstruction of the parental mating type. For example, if in family i there is at least one child with genotype AA (i.e., $N_i^{AA} > 0$), and at least one child with genotype BB (i.e., $N_i^{BB} > 0$), then both parents in this family are AB. (Note that the family index i has been dropped in table 1.) These mating types with corresponding conditions were listed previously by Curtis (1997); however, that study presents a different condi-

Table 3

Test-Statistic Values and Sample Sizes of TDT, S-TDT, and RC-TDT, for GAW9 Data

	TDT		S-TDT[a]		RC-TDT[b]					
					Both		Paternal		Maternal	
ALLELE	z'	n[c]	z'	n[d]	z'	n[e]	z'	n[e]	z'	n[e]
D1G31M8	5.48	67	4.81	50	5.31	50(30)	5.23	51(39)	5.19	54(39)
D5G23M7	7.55	152	6.50	107	6.91	107(49)	7.09	111(75)	6.82	109(62)

[a] Both parental genotypes are missing in all families.

[b] Both = both parental genotypes are missing in all families; Paternal = only the paternal genotype is missing in all families; Maternal = only the maternal genotype is missing in all families.

[c] No. of families with at least one parent heterozygous for the allele of interest.

[d] No. of families suitable for S-TDT.

[e] No. of families suitable for RC-TDT (no. of families in which parental genotypes could be reconstructed is given in parentheses).

tion for mating type $AB \times CD$. He requires that, for reconstruction of this mating type, at least three different genotypes be observed in the offspring, whereas the condition presented in table 1 is less stringent. Indeed, if $N_i^{AC} > 0$ and $N_i^{BD} > 0$ (or $N_i^{AD} > 0$ and $N_i^{BC} > 0$), then the possibility remains that the mating type is $AD \times CB$ (or $AC \times BD$), instead of $AB \times CD$. For the purposes of the present paper, however, it only is necessary to decide whether a parent is homozygous for allele *A*, heterozygous for allele *A*, or has two alleles different from *A*.Thus, the condition presented in table 1 is sufficient. For an exact reconstruction of parental genotypes, the condition given by Curtis (1997) is appropriate.

Columns 3 and 4 of table 1 contain expressions for the conditional expectation of T_i and T_i^2, provided that parental-genotype reconstruction has been possible. To illustrate the method used to obtain these expressions, Appendix A presents the details of the derivation for the first parental mating type (i.e., $AB \times AB$). Since

$$\mathrm{Var}_{H_0}(T_i \mid R) = E_{H_0}(T_i^2 \mid R) - [E_{H_0}(T_i \mid R)]^2 \ ,$$

table 1 can be used to obtain the conditional null variance of T_i. For example, this variance becomes

$$\frac{n_{ai}}{2} + n_{ai}\left[\frac{(5 - 2n_{ai})3^{n_{ci}-2} - 2^{n_{ci}-1}}{4^{n_{ci}} - 2 \cdot 3^{n_{ci}} + 2^{n_{ci}}}\right] \qquad (1)$$

for an $AB \times AB$ mating. This formula can be compared with the variance formula $n_{ai}/2$, which applies when parental genotypes are available directly. Note that the value of equation (1) is $< n_{ai}/2$, for either $n_{ai} \geq 3$ or for $n_{ai} = 2$ and $n_{ui} \leq 1$, whereas it is $> n_{ai}/2$ for either $n_{ai} = 1$ or for $n_{ai} = 2$ and $n_{ui} \geq 2$. This illustrates that, when reconstructed $AB \times AB$ families are treated as if parental genotypes have been typed, sometimes the actual type I error rate will exceed the nominal significance level and sometimes it will be less. The direction of the effect depends on the observed number of affected and unaffected children (i.e., n_{ai} and n_{ui}) in the family.

When one parental genotype is missing.—The S-TDT does not distinguish between families for which both parental genotypes are missing and families with only one missing genotype. To reconstruct parental genotypes, however, such partial information can be taken into account. Table 2, which is organized analogously to table 1, lists all nine parental mating types with at least one parent heterozygous for allele *A*, which may be reconstructable from the genotypes observed in the children. For each of these nine mating types, the first genotype denotes the genotype of the typed parent, whereas the second genotype has to be reconstructed. The second column of table 2 lists the condition on genotypes of the children that makes it possible to reconstruct the missing parental genotype. The condition given for mating type $BC \times AB$ in table 2 is not sufficient for an exact reconstruction of the missing parental genotype (for example, if $N_i^{AB} > 0$ and $N_i^{BC} > 0$, then the missing parental genotype can be either AB or AC), but it is sufficient to make the determination that the missing parent is heterozygous for allele *A* (see also the fourth parental mating type in table 1). The technique used to derive the conditional expectations of T_i and T_i^2 is very similar to that used to obtain the results given in table 1. It should be noted that, in some rare cases (i.e., when $n_{ai} = 1$ and the parental mating type is $AA \times AB$, $BB \times AB$, $BB \times AC$, $BC \times AB$, or $BC \times AD$), these conditional expectations are identical to the corresponding expressions for typed parents (i.e., these reconstructed families can be treated as if both parents have been typed).

C-TDT with Reconstructed Parental Genotypes (RC-TDT)

Suppose there are m nuclear families, with at least one affected child (i.e., $n_{ai} \geq 1$ for all $1 \leq i \leq m$). Each family belongs to one of the following five categories:

1. Both parents have been typed, and at least one parent is heterozygous for allele *A*.

2. Only a single parent has been typed, but the ge-

Table 4

Simulated True Type I Error Rates of the S-TDT and of RC-TDT for Complete and Partially Missing Parental Genotypes, for Sibships with at least One Affected Sib

Sibship Size (No. of Families) and Test[a]	Error Frequency at Nominal α = [b]		
	.05	.01	.001
2 (300):			
S-TDT	.041	.007	.0003
RC-TDT, paternal	.043	.009	.0004
4 (150):			
S-TDT	.047	.011	.0014
RC-TDT:			
Both	.050	.012	.0014
Paternal	.051	.012	.0016
6 (100):			
S-TDT	.055	.011	.0007
RC-TDT:			
Both	.057	.015	.0010
Paternal	.060	.015	.0012

[a] Both = both parental gentoypes are missing in all families; Paternal = only the paternal genotype is missing in all families.
[b] Determined on the basis of the dominant model with $f_{DD} = .2$.

notype of the missing parent can be reconstructed, and at least one parent is heterozygous for allele *A*.

3. Both parental genotypes are missing but can be reconstructed, and at least one parent is heterozygous for allele *A*.

4. At least one parental genotype is missing and cannot be reconstructed, but the condition for the S-TDT is satisfied (i.e., there is at least one affected and at least one unaffected child in this family, not all of the children possess the same genotype, and allele *A* occurs in the genotype of at least one child.)

5. All families not belonging to categories 1–4.

Families in category 5 are discarded from the analysis. For the remaining families, let e_i and v_i denote the appropriate null expectation and variance of T_i, such that:

(i) for families in category 1, $e_i = \frac{n_{ai}}{2}$ (or $e_i = n_{ai}$) and $v_i = \frac{n_{ai}}{4}$ (or $v_i = \frac{n_{ai}}{2}$), when only a single parent (or both parents) is (are) heterozygous for allele *A*;

(ii) for families in category 2, e_i and v_i are calculated from table 2;

(iii) for families in category 3, e_i and v_i are calculated from table 1; and

(iv) for families in category 4, e_i and v_i are calculated from equations (1) and (2), given by Spielman and Ewens (1998).

For these circumstances, the test statistic of the RC-TDT is given by the equation $\sum (T_i - e_i)/\sqrt{\sum v_i}$, in which the summation is over all families in categories 1–4. The distribution of this statistic is approximately the standard normal distribution under the null hypothesis of no linkage.

Simulation Study

To verify that the RC-TDT has an appropriate size, and to compare the power of the RC-TDT with the power of the S-TDT, a simulation study was conducted. The design of these simulations closely followed the approach used by Boehnke and Langefeld (1998). In brief, the disease locus possessed two alleles, *D* and *d*, with frequencies p and q, respectively, and penetrances $1 \geq f_{DD} \geq f_{Dd} \geq f_{dd} \geq 0$, not all equal. The penetrance f_{DD} for the predisposing genotype was $f_{DD} = .2$, .5, or .8. Dominant ($f_{DD} = f_{Dd}$), additive ($f_{Dd} = (f_{DD} + f_{dd})/2$), and recessive ($f_{Dd} = f_{dd}$) models were simulated; for each model, a disease prevalence K_P of 5%, and an attributable fraction of 50%, were assumed. The disease allele frequency p that resulted for each of the disease models was given by Boehnke and Langefeld (1998), in their table 2.

For the marker locus, only the first marker described

Table 5

Simulated Power of the S-TDT and RC-TDT, for Sibships with at Least One Affected Sib (α = .001)

	300 Families with Two Sibs		150 Families with Four Sibs			100 Families with Six Sibs		
				RC-TDT[a]			RC-TDT[a]	
Model	S-TDT	Paternal	S-TDT	Both	Paternal	S-TDT	Both	Paternal
D1	.62	.68	.58	.62	.64	.51	.57	.58
D2	.64	.72	.85	.88	.88	.86	.89	.89
D3	.64	.76	.97	.98	.98	.98	.98	.98
A1	.62	.68	.51	.56	.56	.39	.44	.44
A2	.60	.69	.65	.71	.72	.63	.70	.71
A3	.62	.69	.80	.84	.85	.81	.85	.84
R1	.56	.57	.51	.55	.55	.38	.42	.45
R2	.59	.59	.67	.69	.68	.63	.65	.67
R3	.57	.62	.80	.82	.82	.80	.80	.80

[a] Paternal = only the paternal genotype is missing in all families; Both = both parental genotypes are missing in all families.

by Boehnke and Langefeld (1998) was considered. This marker consisted of six codominant alleles, with population frequencies of .4, .2, .1, .1, .1, and .1, and was completely linked to the disease locus. Also according to the procedure of Boehnke and Langefeld (1998), the haplotype frequencies were set to yield a frequency difference of C, for the first marker allele, between randomly selected affected and unaffected individuals. It was also assumed that all remaining marker allele frequencies are reduced proportionately in affected individuals. With these conditions, the population frequency h_{kD} of the haplotype, with marker allele k and disease allele D, then becomes

$$h_{1D} = \frac{K_P[a_1 + C(1 - K_P)] - a_1(pf_{Dd} + qf_{dd})}{p(f_{DD} - f_{Dd}) + q(f_{Dd} - f_{dd})} ,$$

and

$$h_{kD} = a_k\left(\frac{p - h_{1D}}{1 - a_1}\right), \quad \text{for} \quad k \geq 2 ,$$

with a_k denoting the population frequency of marker allele k. The value $C = .15$ was used to compare the power of the RC-TDT with the S-TDT, and the value $C = .0$ was used to verify that the RC-TDT gave appropriate significance levels.

Each simulated sample consisted of families with an identical number of sibs (n_c) in each family (with n_c = 2, 4, or 6), which were ascertained on the basis of the presence of an affected child. The number of families per sample was $600/n_c$ (i.e., each sample consisted of a total of 600 children). To assess the power of the tests, 500 replicate samples were generated, under 27 different simulation conditions (i.e., for each combination of mode of inheritance [dominant, additive, and recessive]; for penetrance, f_{DD} [.2, .5, and .8]; and for sib number, n_c [2, 4, and 6]). For each replicate sample, the statistics obtained with the S-TDT and with the RC-TDT were calculated. For the RC-TDT, it was assumed that either no parental marker information was available, or that only maternal marker information was available. To evaluate the true size of the tests for a small nominal significance level, such as $\alpha = .001$, >500 replicates were required. For this purpose, it was decided to generate R = 10,000 replicate samples for each n_c, but only for the dominant model, with $f_{DD} = .2$.

Results

Analysis of Data from Genetic Analysis Workshop 9 (GAW9)

Spielman and Ewens (1998) analyzed data from the GAW9, using the TDT and the S-TDT procedures. In the current paper, these data are analyzed with the RC-TDT procedure, described previously. GAW9 data consisted of 200 nuclear families, each of which has at least one affected child. Twenty-five families contained only a single affected child and no unaffected sib. In the remaining 175 families, at least one unaffected sib was present. Only 16 families had more than one affected child. Complete marker information was available, for each individual in every family, at 360 marker loci along 6 chromosomes. The disease model used to generate these data assumed the presence of four disease alleles with additive effects, each located on a separate chromosome. Two of the disease alleles coincided with alleles at distributed marker loci: allele M8 of marker D1G31, and allele M7 of marker D5G23. Hodge (1995) describes GAW9 data in detail.

The first column of table 3 contains the results that we obtained by using the conventional TDT (in terms of z' scores with continuity correction) and the numbers (n) of families suitable for TDT analysis. The second column of table 3 contains the results that we obtained by using the S-TDT when genotypes of the parents were ignored in all families. The third column of table 3 contains z' scores obtained with the RC-TDT when all parental genotypes were ignored. Since there were no families with more than one affected child and without an unaffected sib, the families suitable for this RC-TDT analysis were exactly the same as those analyzed with the S-TDT. For both markers, the RC-TDT gave larger z' scores than the S-TDT. This is particularly true for allele M8, of marker D1G31, for which missing parental genotypes could be reconstructed in 60% of the 50 families suitable for S-TDT analysis, at which the difference in z' scores obtained with the TDT and the RC-TDT was quite small. In this instance, the availability of unaffected children could nearly compensate for the missing parental-genotype information.

To examine the performance of the RC-TDT in a situation in which partial information on parental genotypes is available, in all families, either only the maternal (column 4 of table 3) or only the paternal (column 5 of table 3) genotype was assumed to be known. In this situation, there was a small number of additional families that could be analyzed with the RC-TDT but that could not be included for S-TDT analysis. Typically, these were families with a parental mating type $AB \times CD$, in which the typed AB parent transmitted the same allele to all children, whereas the untyped CD parent transmitted each allele to at least one child. Such a family allowed the reconstruction of the missing parental genotype and therefore was suitable for RC-TDT analysis but was not suitable for the S-TDT. Thus, the total number of families suitable for RC-TDT, in instances when partial information on parental genotypes is available, was larger than the number of families suitable for

the S-TDT (or for the RC-TDT, without both parental genotypes). Also, in some of the families suitable for the S-TDT, parental-genotype reconstruction was not possible when both parental genotypes were missing, but was possible when only a single parental genotype was unknown. The numbers in parentheses in columns 4 and 5 of table 3 are the total number of families for which the missing parental genotype could be reconstructed. Surprisingly, however, the increase in parental marker genotype information generally was not reflected in an accompanying increase of the z' score obtained with the RC-TDT. An increased z' score was present only when solely maternal genotypes were available for marker D5G23. It is noteworthy that, in all instances, the z' scores in columns 4 and 5 are still larger than those obtained with the S-TDT.

Simulated Size and Power

Table 4 presents estimates of the true type I error rate, at nominal significance levels of $\alpha = .05$, .01, and .001. Since these estimates were obtained from $R = 10{,}000$ replicates, their standard deviations are .0022, .001, and .003. No results are given for $n_c = 2$ without parental genotypes. In this case, the S-TDT and RC-TDT are identical, as can be seen when the the formulas given in table 1 are compared with the formulas given by Spielman and Ewens (1998). The data in table 4 indicate that the true size of the RC-TDT is slightly larger than the true type I error rate of the S-TDT. For $n_c = 2$, all tests tend to be conservative, whereas for larger sibship sizes, the true type I error rate exceeds the nominal α. A possible explanation for this observation is that the sample size (i.e., number of families) was smaller for larger sibship sizes. The simulations support the validity of approximating the null distribution of z' scores with a standard normal distribution for S-TDT and for RC-TDT.

Power estimates at significance level of $\alpha = .001$ are presented in table 5, for nine disease models. These disease models are denoted by "D," "A," and "R" for the mode of inheritance (i.e., dominant, additive, and recessive) and "1," "2," and "3" for the value of f_{DD} (i.e., .2, .5, and .8). In instances for which there is no parental-genotype information available, application of the RC-TDT instead of the S-TDT results in a modest but consistent gain of power. Generally, an additional but quite small power increase is obtained in instances for which only a single parental genotype is missing.

Discussion

An attractive feature of the S-TDT proposed by Spielman and Ewens (1998) is that it allows a joint analysis of families in which parental genotypes are available and of families in which no parental genotypes but genotypes of unaffected sibs are available. If an increasing degree of marker polymorphism and an increasing size of the sibship are present, then there is an increasing probability that missing parental genotypes can be uniquely determined from the genotypes of the children. For example, parental-genotype reconstruction is possible for each of the three families presented in table 1 of Spielman and Ewens (1998). As noted by Curtis (1997), it is erroneous to treat these reconstructed families as if parental genotypes have been typed. The present paper shows that such an approach may inflate the type I error rate or may decrease the power to detect linkage, with the direction of the effect depending on the number of affected and unaffected children in the family. Both kinds of bias, however, can be avoided, by use of the appropriate null expectation and variance, supplied in tables 1 and 2 of the present paper. The evidence provided by the application of the RC-TDT procedure to data of GAW9, as well as the results obtained from simulations, support the hypothesis that parental-genotype reconstruction improves the power of family-based association analysis.

Acknowledgment

The author is deeply indebted to Warren Ewens, for many helpful discussions and for his generosity of support, which have greatly contributed to this work.

Appendix

Calculations of $\mathbf{E}_{H_0}(\boldsymbol{T}_i \mid \boldsymbol{N}_i^{AA} \cdot \boldsymbol{N}_i^{BB} > \mathbf{0})$ and $\mathbf{Var}_{H_0}(\boldsymbol{T}_i \mid \boldsymbol{N}_i^{AA} \cdot \boldsymbol{N}_i^{BB} > \mathbf{0})$

In this appendix, the details of the calculation of $E_{H_0}(T_i \mid N_i^{AA} \cdot N_i^{BB} > 0)$ and of $Var_{H_0}(T_i \mid N_i^{AA} \cdot N_i^{BB} > 0)$ are presented.

Calculation of $P_{H_0}(N_i^{AA} > 0$ *and* $N_i^{BB} > 0)$

$$P_{H_0}(N_i^{AA} > 0 \text{ and } N_i^{BB} > 0) = 1 - P_{H_0}(N_i^{AA} = 0 \text{ or } N_i^{BB} = 0)$$
$$= 1 - P_{H_0}(N_i^{AA} = 0) - P_{H_0}(N_i^{BB} = 0) + P_{H_0}(N_i^{AA} = 0 \text{ and } N_i^{BB} = 0)$$
$$= 1 - 2\left(\frac{3}{4}\right)^{n_{ci}} + \left(\frac{1}{2}\right)^{n_{ci}} . \quad \text{(A1)}$$

Calculation of $P_{H_0}(T_i = c \cap N_i^{AA} > 0 \cap N_i^{BB} > 0)$ *for* $0 \leq c \leq 2n_{ai}$

We have

$$P_{H_0}(T_i = c \cap N_i^{AA} > 0 \cap N_i^{BB} > 0) = P_{H_0}(T_i = c) - P_{H_0}[T_i = c \cap (N_i^{AA} = 0 \cup N_i^{BB} = 0)]$$
$$= \binom{2n_{ai}}{c}\left(\frac{1}{2}\right)^{2n_{ai}} - P_{H_0}[T_i = c \cap (N_i^{AA} = 0 \cup N_i^{BB} = 0)] \quad \text{(A2)}$$

To calculate the second term in equation (A2), three cases have to be distinguished. Each of these cases relies on the fact that $N_i^{AA} = 0$ or $N_i^{BB} = 0$, together with $T_i = c$, fixes the values for N_{ai}^{AA}, N_{ai}^{AB}, and N_{ai}^{BB}.

Case 1, $c = n_{ai}$*:*

$$P_{H_0}[T_i = n_{ai} \cap (N_i^{AA} = 0 \cup N_i^{BB} = 0)] = P_{H_0}[N_{ai}^{AA} = 0 \cap N_{ai}^{AB} = n_{ai} \cap (N_{ui}^{AA} = 0 \cup N_{ui}^{BB} = 0)]$$
$$= \left(\frac{1}{2}\right)^{n_{ai}}\left[2\left(\frac{3}{4}\right)^{n_{ui}} - \left(\frac{1}{2}\right)^{n_{ui}}\right] . \quad \text{(A3)}$$

Case 2, $0 \leq c < n_{ai}$*:*

$$P_{H_0}[T_i = c \cap (N_i^{AA} = 0 \cup N_i^{BB} = 0)] = P_{H_0}(N_{ai}^{AA} = 0 \cap N_{ai}^{AB} = c \cap N_{ai}^{BB} = n_{ai} - c \cap N_{ui}^{AA} = 0)$$
$$= \binom{n_{ai}}{c}\left(\frac{1}{2}\right)^{c}\left(\frac{1}{4}\right)^{n_{ai}-c}\left(\frac{3}{4}\right)^{n_{ui}} . \quad \text{(A4)}$$

Case 3, $n_{ai} < c \leq 2n_{ai}$*:*

$$P_{H_0}[T_i = c \cap (N_i^{AA} = 0 \cup N_i^{BB} = 0)] = P_{H_0}(N_{ai}^{AA} = c - n_{ai} \cap N_{ai}^{AB} = 2n_{ai} - c \cap N_{ai}^{BB} = 0 \cap N_{ui}^{BB} = 0)$$
$$= \binom{n_{ai}}{c - n_{ai}}\left(\frac{1}{4}\right)^{c-n_{ai}}\left(\frac{1}{2}\right)^{2n_{ai}-c}\left(\frac{3}{4}\right)^{n_{ui}} . \quad \text{(A5)}$$

Calculation of $P_{H_0}(T_i = c | N_i^{AA} > 0$ *and* $N_i^{BB} > 0)$

When equations (A1)–(A5) are combined,

$$P_{H_0}(T_i = c \mid N_i^{AA} > 0 \text{ and } N_i^{BB} > 0) = \frac{\binom{2n_{ai}}{c}\left(1/2\right)^{2n_{ai}} - \binom{n_{ai}}{c}\left(1/2\right)^{2n_{ai}-c}\left(3/4\right)^{n_{ui}}}{1 - \left(1/2\right)^{n_{ci}}\left[2\left(3/2\right)^{n_{ci}} - 1\right]} \text{ for } 0 \leqslant c < n_{ai} ,$$

$$P_{H_0}(T_i = c \mid N_i^{AA} > 0 \text{ and } N_i^{BB} > 0) = \frac{\binom{2n_{ai}}{n_{ai}}\left(1/2\right)^{2n_{ai}} - \left(1/2\right)^{n_{ci}}\left[2\left(3/2\right)^{n_{ui}} - 1\right]}{1 - \left(1/2\right)^{n_{ci}}\left[2\left(3/2\right)^{n_{ci}} - 1\right]} \text{ for } c = n_{ai} , \tag{A6}$$

$$P_{H_0}(T_i = c \mid N_i^{AA} > 0 \text{ and } N_i^{BB} > 0) = \frac{\binom{2n_{ai}}{c}\left(1/2\right)^{2n_{ai}} - \binom{n_{ai}}{c-n_{ai}}\left(1/2\right)^{c}\left(3/4\right)^{n_{ui}}}{1 - \left(1/2\right)^{n_{ci}}\left[2\left(3/2\right)^{n_{ci}} - 1\right]} \text{ for } n_{ai} < c \leqslant 2n_{ai} .$$

Calculation of $E_{H_0}(T_i|N_i^{AA} > 0$ *and* $N_i^{BB} > 0)$

Since

$$\sum_{c=0}^{2n_{ai}} \binom{2n_{ai}}{c}\left(\frac{1}{2}\right)^{2n_{ai}} c = n_{ai} ,$$

$$\sum_{c=0}^{n_{ai}-1} \binom{n_{ai}}{c}\left(\frac{1}{2}\right)^{2n_{ai}-c}\left(\frac{3}{4}\right)^{n_{ui}} c = \left(\frac{3}{4}\right)^{n_{ui}}\left(\frac{1}{2}\right)^{n_{ai}} \sum_{c=1}^{n_{ai}} \binom{n_{ai}}{c}\left(\frac{1}{2}\right)^{c} (n_{ai} - c) ,$$

$$\sum_{c=n_{ai}+1}^{2n_{ai}} \binom{n_{ai}}{c-n_{ai}}\left(\frac{1}{2}\right)^{c}\left(\frac{3}{4}\right)^{n_{ui}} c = \left(\frac{3}{4}\right)^{n_{ui}}\left(\frac{1}{2}\right)^{n_{ai}} \sum_{c=1}^{n_{ai}} \binom{n_{ai}}{c}\left(\frac{1}{2}\right)^{c} (n_{ai} + c) ,$$

$$\sum_{c=1}^{n_{ai}} \binom{n_{ai}}{c}\left(\frac{1}{2}\right)^{c} = \left(\frac{3}{2}\right)^{n_{ai}} - 1 ,$$

it follows from equation (A6) that

$$E_{H_0}(T_i \mid N_i^{AA} > 0 \text{ and } N_i^{BB} > 0) = n_{ai} .$$

Calculation of $\mathrm{Var}_{H_0}(T_i|N_i^{AA} > 0$ *and* $N_i^{BB} > 0)$

Since

$$\sum_{c=0}^{2n_{ai}} \binom{2n_{ai}}{c} \left(\frac{1}{2}\right)^{2n_{ai}} c^2 = n_{ai}\left(n_{ai} + \frac{1}{2}\right),$$

$$\sum_{c=0}^{n_{ai}-1} \binom{n_{ai}}{c} \left(\frac{1}{2}\right)^{2n_{ai}-c} \left(\frac{3}{4}\right)^{n_{ui}} c^2 = \left(\frac{3}{4}\right)^{n_{ui}} \left(\frac{1}{2}\right)^{n_{ai}} \sum_{c=1}^{n_{ai}} \binom{n_{ai}}{c} \left(\frac{1}{2}\right)^{c} (n_{ai} - c)^2,$$

$$\sum_{c=n_{ai}+1}^{2n_{ai}} \binom{n_{ai}}{c-n_{ai}} \left(\frac{1}{2}\right)^{c} \left(\frac{3}{4}\right)^{n_{ui}} c^2 = \left(\frac{3}{4}\right)^{n_{ui}} \left(\frac{1}{2}\right)^{n_{ai}} \sum_{c=1}^{n_{ai}} \binom{n_{ai}}{c} \left(\frac{1}{2}\right)^{c} (n_{ai} + c)^2,$$

$$\sum_{c=1}^{n_{ai}} \binom{n_{ai}}{c} \left(\frac{1}{2}\right)^{c} c^2 = \left(\frac{3}{2}\right)^{n_{ai}-2} n_{ai}\left(\frac{n_{ai}}{4} + \frac{1}{2}\right),$$

it follows that

$$\begin{aligned} &\mathrm{E}_{H_0}(T_i^2 \mid N_i^{AA} > 0 \text{ and } N_i^{BB} > 0) \\ &= \frac{n_{ai}\left(n_{ai} + 1/2\right) - n_{ai}(5n_{ai} + 1)\left(4/9\right)\left(3/4\right)^{n_{ci}} + n_{ai}^2\left(1/2\right)^{n_{ci}}}{1 - \left(1/2\right)^{n_{ci}}\left[2\left(3/2\right)^{n_{ci}} - 1\right]} \\ &= n_{ai}\left(n_{ai} + \frac{1}{2}\right) + \frac{(5 - 2n_{ai})n_{ai}\left(3/4\right)^{n_{ci}}\left(1/9\right) - n_{ai}\left(1/2\right)^{n_{ci}+1}}{1 - \left(1/2\right)^{n_{ci}}\left[2\left(3/2\right)^{n_{ci}} - 1\right]}. \end{aligned}$$

Thus,

$$\begin{aligned} &\mathrm{Var}_{H_0}(T_i \mid N_i^{AA} N_i^{BB} > 0) \\ &= \frac{n_{ai}}{2} + n_{ai}\frac{(5 - 2n_{ai})\left(3/2\right)^{n_{ci}}\left(1/9\right) - 1/2}{2^{n_{ci}} - 2\left(3/2\right)^{n_{ci}} + 1}. \end{aligned}$$

References

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

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

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

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

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

———(1999) TDT Clarification. Am J Hum Genet 64: 667–668

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