

*Research Article*

# **The Transmission Disequilibrium/Heterogeneity Test with Parental-Genotype Reconstruction for Refined Genetic Mapping of Complex Diseases**

**Jing Han and Yongzhao Shao**

*Division of Biostatistics, NYU School of Medicine, New York University, 650 First Avenue, 5th Floor, New York, NY 10016, USA*

Correspondence should be addressed to Yongzhao Shao, [yongzhao.shao@nyumc.org](mailto:yongzhao.shao@nyumc.org)

Received 2 March 2012; Accepted 1 May 2012

Academic Editor: Xiaohua Douglas Zhang

Copyright © 2012 J. Han and Y. Shao. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In linkage analysis for mapping genetic diseases, the transmission/disequilibrium test (TDT) uses the linkage disequilibrium (LD) between some marker and trait loci for precise genetic mapping while avoiding confounding due to population stratification. The sib-TDT (S-TDT) and combined-TDT (C-TDT) proposed by Spielman and Ewens can combine data from families with and without parental marker genotypes (PMGs). For some families with missing PMG, the reconstruction-combined TDT (RC-TDT) proposed by Knapp may be used to reconstruct missing parental genotypes from the genotypes of their offspring to increase power and to correct for potential bias. In this paper, we propose a further extension of the RC-TDT, called the reconstruction-combined transmission disequilibrium/heterogeneity (RC-TDH) test, to take into account the identical-by-descent (IBD) sharing information in addition to the LD information. It can effectively utilize families with missing or incomplete parental genetic marker information. An application of this proposed method to Genetic Analysis Workshop 14 (GAW14) data sets and extensive simulation studies suggest that this approach may further increase statistical power which is particularly valuable when LD is unknown and/or when some or all PMGs are not available.

## **1. Introduction**

Genetic linkage analysis is an important step in localizing and identifying genes in the chromosomes that underlie many human diseases and other traits of interest. A brief overview of commonly used statistical methods for linkage analysis including recently developed model-free and model-based methods for mapping qualitative- and quantitative-trait loci, can be found in Shao [1]. For more extensive discussions on linkage analysis, readers can consult Ott [2].

Mapping genes that underlie complex diseases is of great current interest. The essence of linkage analysis is to identify statistical association between the inheritance of a complex genetic disease phenotype and inheritance of specific pieces of genetic material (called marker alleles). Many complex diseases including cancers have an inheritable component. For marker alleles that are associated with inheritance of complex diseases, it is common that the transmission probabilities of a marker allele of interest vary across heterozygous parents, due to locus heterogeneity, etiological heterogeneity, and many other complexities and/or combinations of them [3, 4]. Under such transmission heterogeneity, the transmission likelihood generally has the form of mixture models with many parameters [4, 5]. It can be shown that the efficient score test of such mixture likelihood includes two parts, one part related to transmission disequilibriums reflected by existence of linkage disequilibrium (LD) and the other related to transmission heterogeneity in the form of excessive dispersion in sharing of genetic markers as might be inferred from identical by descent (IBD) patterns (e.g., allele-sharing patterns among affected sib-pairs).

The transmission/disequilibrium test (TDT) developed by Spielman et al. [6] uses the LD information between some marker and disease loci for precise genetic mapping while avoiding confounding due to population stratification. It has been extended in multiple directions to meet the need for mapping complex traits, for example [7, 8]. In particular, missing parental genetic marker genotypes are very common for studying diseases with late onset. The sib-TDT (S-TDT) and combined-TDT (C-TDT) proposed by Spielman and Ewens [9] can deal with families without parental marker genotypes (PMGs) and can combine with data from families having PMG available. For some families with missing PMG, the reconstruction-combined TDT (RC-TDT) proposed by Knapp [10, 11] may be used to reconstruct missing PMG from the genotypes of their offspring to increase power of the C-TDT with a correction for potential bias in using reconstructed PMG [12].

An attractive feature of the RC-TDT is that it utilizes the missing PMG that can be uniquely determined from the genotypes of the children and corrects potential biases resulting from using reconstructed PMG by employing appropriate null expectation and variance, supplied in Tables 1 and 2 of Knapp [10]. Similar to the TDT and C-TDT, the RC-TDT is powerful only when there is strong LD. Usually LD is unknown, and it is difficult to measure, thus it is generally desirable to combine LD information with information on allele sharing obtained based on IBD patterns [5, 13].

For fine mapping of complex genetic disorders, Shao [4] derived a general mixture likelihood for allele transmission under various transmission disequilibrium and/or heterogeneity and further proposed a *transmission disequilibrium/heterogeneity* (TDH) test to efficiently combine the transmission disequilibrium and heterogeneity information to maximize the power for detecting linkage using genetic data from nuclear families. The TDH test was shown to be an efficient score test of the general mixture likelihood derived in Shao [4] which is a summation of two parts, a *transmission/disequilibrium test* (TDT) part which utilizes the LD information and a *transmission heterogeneity test* (THT) part that utilizes IBD-sharing information. To see that the THT utilizes IBD-sharing information, it should be pointed out that general mixture likelihood contains the mixture binomial likelihood discussed in Huang and Jiang [13] and Lo et al. [5], and the test statistic of the classical mean test for affected sib-pairs (ASPs) is a special case of the THT statistic with  $n_{ai} = 2$  in Shao [4]. The classical mean test for affected sib-pairs is the most well-known IBD sharing-based linkage test [14]. The THT is applicable to general sibship and thus can be regarded as an extension of the classical mean test for affected sib-pairs.

**Table 1:** Moments of  $T_i$  under  $H_0$ .

PMG	$ET_i$	$\text{Var}(T_i)$	$\text{Var}[(T_i - ET_i)^2]$
AB/AA	$3n_{ai}/2$	$n_{ai}/4$	$n_{ai}(n_{ai} - 1)/8$
AB/AB	$n_{ai}$	$n_{ai}/2$	$n_{ai}(2n_{ai} - 1)/4$
AB/BB	$n_{ai}/2$	$n_{ai}/4$	$n_{ai}(n_{ai} - 1)/8$

**Table 2:** Distribution of  $T_i$  when one PMG is missing but reconstructible.

PMG	Condition (R)	Range of $T_i$	$P_{H_0}(T_i = c   R)$
AB × AB	$N^{AA} > 0$ and $N^{BB} > 0$	$0 \leq c < n_a$	$\frac{\binom{2n_a}{c} (1/2)^{2n_a} - \binom{n_a}{c} (1/2)^{2n_a-c} (3/4)^{n_a}}{1 - (1/2)^{n_c} [2(3/2)^{n_c} - 1]}$
		$c = n_a$	$\frac{\binom{2n_a}{n_a} (1/2)^{2n_a} - (1/2)^{n_c} [2(3/2)^{n_c} - 1]}{1 - (1/2)^{n_c} [2(3/2)^{n_c} - 1]}$
		$n_a < c \leq 2n_a$	$\frac{\binom{2n_a}{c} (1/2)^{2n_a} - \binom{n_a}{c-n_a} (1/2)^c (3/4)^{n_a}}{1 - (1/2)^{n_c} [2(3/2)^{n_c} - 1]}$
AA × AB	$N^{AA} > 0$ and $N^{AB} > 0$	$c = n_a$	$\frac{\binom{n_a}{c-n_a} (1/2)^{n_a} - (1/2)^{n_c}}{1 - 2(1/2)^{n_c}}$
		$n_a < c < 2n_a$	$\frac{\binom{n_a}{c-n_a} (1/2)^{n_a}}{1 - 2(1/2)^{n_c}}$
		$c = 2n_a$	$\frac{\binom{n_a}{c-n_a} (1/2)^{n_a} - (1/2)^{n_c}}{1 - 2(1/2)^{n_c}}$
BB × AB	$N^{AB} > 0$ and $N^{BB} > 0$	$c = 0$	$\frac{\binom{n_a}{c} (1/2)^{n_a} - (1/2)^{n_c}}{1 - 2(1/2)^{n_c}}$
		$0 < c < n_a$	$\frac{\binom{n_a}{c} (1/2)^{n_a}}{1 - 2(1/2)^{n_c}}$
		$c = n_a$	$\frac{\binom{n_a}{c} (1/2)^{n_a} - (1/2)^{n_c}}{1 - 2(1/2)^{n_c}}$

In practice, parental marker genotypes are often incomplete for many genetic studies particularly for late onset diseases. Only using families with complete parental marker genotype information would lead to throwing away a large portion of the useful data and can also lead to biases. It is thus crucially important to make the TDH test applicable to families with missing or incomplete parental marker genotype information. In this paper, we develop a transmission disequilibrium/heterogeneity test with parental-genotype reconstruction, which utilizes both the LD information and the IBD-sharing information and can combine families with or without PMG information.

The transmission disequilibrium/heterogeneity test with parental-genotype reconstruction (RC-TDH) will be introduced in the next section. In Section 3, the RC-TDH test is applied to a data set from GAW14, and the results are compared with those of the RC-TDT. Finally, simulation studies that use common genetic models [5, 15] are carried out to compare the power and the true size of the RC-TDT and RC-TDH test. The numerical results suggest that RC-TDH test may greatly increase the statistical power which is particularly valuable whenever LD levels are unknown and/or whenever there is missing PMG information as in studying of a disease with late age of onset.

It should be pointed out that the main comparison made in this paper will be between RC-TDT and RC-TDH. We will not formally compare them with the classical IBD-based linkage tests such as those implemented in Genehunter and other softwares. The main rationale is as follows. We are mainly interested in fine mapping of genetic variants that underlie complex diseases, where the classical linkage tests are known to have low power because they do not utilize LD information effectively. With the rapid advancement of biotechnology, it is now feasible and affordable to use dense genetic markers, for example, the single nucleotide polymorphisms (SNPs), for genomewide linkage scan. With a large number of dense genetic markers (e.g., SNPs) some of the markers can be expected to fall into the LD block of the causal genetic variants; thus LD would generally exist to some degree for many markers. Thus the TDT and TDH tests would have power advantage over classical linkage tests which only effectively utilize the IBD information.

## 2. Method

### 2.1. Notation

It will be assumed that there are two alleles  $A$  and  $B$  at the marker locus, and allele  $A$  is of particular interest. Let  $n_{ai}$  denote the number of affected children, let  $n_{ui}$  denote the number of unaffected children, and let  $n_{ci} = n_{ai} + n_{ui}$  denote the size of the sibship for family  $i$ . In each family, all children have been typed at the marker locus, but the PMG may or may not be available. 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.,  $T_i = 2N_{ai}^{AA} + N_{ai}^{AB}$ ). The notation introduced here is consistent with Knapp [10, 11] and Han [16].

### 2.2. The TDH Test with Complete PMG

For completeness, we first consider the case when PMG are observed along with children's marker genotypes. Let  $x_i$  be the number of alleles  $A$  transmitted by the  $i$ th marker heterozygous parent to the affected children. When the exact number  $x_i$  of marker alleles  $A$

transmitted to affected children cannot be determined as might happen in families with two heterozygous parents, then  $T_i$  can be used to replace  $x_i$ . Using  $T_i$  in families with ambiguous transmissions, the TDT statistic can be written as  $T_D = T_d^2$  where

$$T_d = \sum_i \frac{T_i - ET_i}{\sqrt{\sum_i \text{Var}(T_i)}}. \quad (2.1)$$

The transmission heterogeneity test (THT) statistic is denoted as  $T_H = T_h^2$  where

$$T_h = \frac{\max\left\{\sum_i [(T_i - ET_i)^2 - \text{Var}(T_i)], 0\right\}}{\sqrt{\sum_i \text{Var}[(T_i - ET_i)^2]}}, \quad (2.2)$$

where the moments of  $T_i$  under  $H_0$  given the parental marker genotypes (PMGs) are summarized in Table 1.

The transmission disequilibrium/heterogeneity (TDH) test is based on the following test statistic [4]:

$$T_{DH} = T_D + T_H. \quad (2.3)$$

In terms of statistical optimality, it can be shown that the TDH test is the efficient score test from the mixture likelihood function under transmission disequilibrium and heterogeneity [4]. In theory, the efficient score test is known to be locally most powerful.

### 2.3. The Reconstruction-Combined TDH (RC-TDH) Test

When at least one parent with missing PMG, Knapp [10] proposed a reconstruction-combined TDT (RC-TDT) to reconstruct PMG from the genotypes of their offspring and correct for the biases resulting from using reconstructed PMG. To improve the power to detect linkage, we propose the reconstruction-combined TDH test (RC-TDH) using the following test statistic:

$$\frac{[\sum (T_i - e_i)]^2}{\sum v_i} + \frac{\left[\max\left\{\sum \left((T_i - e_i)^2 - E_{H_0}[(T_i - e_i)^2 | R]\right), 0\right\}\right]^2}{\sum \text{Var}_{H_0}[(T_i - e_i)^2 | R]}, \quad (2.4)$$

where  $T_i$  denotes the number of marker alleles  $A$  in affected children, and  $e_i = E_{H_0}(T_i | R)$ ,  $v_i = \text{Var}_{H_0}(T_i | R)$  denote the appropriate null expectation and variance of  $T_i$ , respectively, as can be found in Tables 1 and 2 of Knapp [10]. In the RC-TDH statistic, the first term is the RC-TDT statistic of Knapp [10] and the second term is the RC-THT statistic with the restriction. To get the appropriate null expectation  $\text{Var}_{H_0}[(T_i - E_{H_0}(T_i | R))^2 | R]$ , we need to derive the conditional distribution of  $T_i$  given the constraint for reconstruction  $R$ .

When one parental genotype is missing and reconstructible, the conditional probabilities of  $T_i$  are listed in Table 2. Note that the family index  $i$  has been dropped in the formula in Table 2. In the first column, the first parental genotype is typed and the second

one is reconstructed. The second column presents a necessary and sufficient condition, for the observed marker genotypes in the offspring, to allow reconstruction of the parental genotypes. The details of the derivation are provided in Han [16].

When both parental genotypes are missing, the reconstruction condition and the conditional probabilities of  $T_i$  are the same as that of one parental genotype is missing and the known parental genotype is  $AB$ .

When 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), the distribution of  $T_i$  can be calculated using the affected and unaffected children genotypes by the hypergeometric distribution. The details are provided in the Appendix section.

As in C-TDT and RC-TDT, families not belonging to the previous categories will be ignored.

### 3. Application to Genetic Analysis Workshop 14 Data

The proposed RC-TDH test was applied to a Genetic Analysis Workshop 14 (GAW14) dataset to compare the power with that of RC-TDT. The GAW14 simulated data were generated by Dr. David Greenberg. A behavioral disorder has been simulated in multiple replicates of four different populations/groups. There are 100 families in the Aipotu, Karnagar, and Danacaa data sets. There are 100 replicates for each data set. The results of power comparison of RC-TDH with RC-TDT to analyze the linkage between the trait b disease allele and the marker B01T0561 are presented in Table 3. This trait has incomplete penetrance with  $f_{DD} = 30\%$ . Application of the RC-TDH is illustrated in Table 3 with 50% and 100% missing parental genotypes. The power is based on type I error at 0.05 level.

## 4. Simulation

### 4.1. Simulation Set-Up

Simulation studies are conducted to compare the powers of the proposed RC-TDH test with the RC-TDT. To attain the correct type I error rates, we directly simulated the critical values under the null hypothesis of no linkage, in which  $\theta$  (recombination frequency) = 0.5. In the simulations for the null distribution, 1,000,000 replicates of samples of nuclear families are generated and the empirical critical values are obtained. Based on 500 independent replicates and the empirical critical values, we estimate the power of the tests using the relative frequencies of the simulated test statistics which exceed the empirical critical values.

To generate the family-based data, as in earlier work [5], we consider two biallelic loci: one disease locus (with disease allele  $D$  and normal allele  $d$ ) and one marker locus (with allele  $A$  and  $B$ ). The frequency for disease allele  $D$  is  $p_D$  and for marker allele  $A$  is  $p_A$ . The linkage disequilibrium is the deviation of the frequency of  $DA$  haplotype from its equilibrium value (expected by chance). Define the  $LD$  parameter as

$$\Delta = \frac{p_{DA} - p_D \cdot p_A}{\min(p_D \cdot p_B, p_d \cdot p_A)}. \quad (4.1)$$

In our simulations, we assume  $A$  is the allele in  $LD$  with  $D$ . Thus, the range of the  $LD$  parameter  $\Delta$  is in  $[0, 1]$ , in which 0 indicates linkage equilibrium. There are three penetrance parameters,  $f_{DD}$ ,  $f_{Dd}$ , and  $f_{dd}$ , corresponding to three possible disease genotypes.

**Table 3:** Power comparison of the RC-TDH test with RC-TDT using GAW14 data.

Population	100% PMG Missing		50% PMG Missing	
	RC-TDT	RC-TDH	RC-TDT	RC-TDH
Aipotu	0.27	0.58	0.57	1.00
Karnagar	0.14	0.33	0.46	1.00
Danacaa	0.37	0.86	0.74	1.00

In the study of 100% PMG missing, we ignore all the parental marker genotypes. In the study of 50% PMG missing, we use 50% families with parental marker genotypes and 50% families without parental marker genotypes.

Simulation study 1 closely followed the approach used by Boehnke and Langefeld [15]. For each model, a disease prevalence  $K_p$  of 5% was assumed. The disease allele frequency  $p$  that resulted from each of the disease models can be calculated by  $K_p = p^2 f_{DD} + 2p(1-p)f_{Dd} + (1-p)^2 f_{dd}$ . Summary of the parameters used in this simulation study is in Table 4.

Summary of the parameters used in simulation study 2 is in Table 5. Four commonly used disease models are used here: dominant ( $f_{Dd} = f_{DD}$ ), additive ( $f_{Dd} = (f_{DD} + f_{dd})/2$ ), multiplicative ( $f_{Dd} = \sqrt{f_{DD} \cdot f_{dd}}$ ), and recessive ( $f_{Dd} = f_{dd}$ ) models.

#### 4.2. Simulation Results

Table 6 presents estimates of the critical values for RC-TDH at significance levels of .05, .01, and .001. Table 7 presents the estimates of the true type I error rate, at nominal significance levels of .05, .01, and .001. The simulations support the validity of approximating the null distribution with a standard normal distribution for RC-TDT.

The results of simulation study 1 are shown in Table 8. The disease models are denoted by "D," "A," and "R" for the mode of inheritance (i.e., dominant, additive, and recessive); "1" and "2" for the value of  $f_{DD}$  (i.e., 1.0 and 0.5). The presented results come from the simulations with 4 sibs in each family, which have the same trend as those with 2 or 6 sibs in each family. In instances for which there is no parental genotype information available, application of the RC-TDH instead of the RC-TDT results in a consistent gain of power, especially when linkage disequilibrium is weak.

We conducted simulation study 2 to compare the power of the proposed RC-TDH test with that of RC-TDT according to linkage disequilibrium in different scenarios based on Table 5, such as tight linkage versus weak linkage, full penetrance versus incomplete penetrance. Each simulated sample consists of families with an identical number of sibs ( $n_c$ ) in each family (with  $n_c = 3$ ), which are ascertained on the basis of the presence of an affected child. Each sample consists of a total of 600 children. Half of the 200 families have complete PGM, and half of the families without PGM. To assess the power of the tests, 500 replicate samples are generated, under different simulation scenarios. For each replicate sample, the statistics obtained with the proposed RC-TDH and with the RC-TDT were calculated.

To compare power of the RC-TDH with that of the RC-TDT at different  $LD$  levels, we set the range of  $LD$  between 0 and 1, recombination fraction at 0.01, the frequency of allele  $D$  at 0.1, the frequency of allele  $A$  at 0.5, penetrance for genotype  $DD$  at full penetrance 1, penetrance for genotype  $dd$  at 0.01, and then the penetrance for genotype  $Dd$  can be determined by the modes of inheritance. The results in Table 9 and Figure 1 show that the power increases with  $LD$ , and the proposed RC-TDH is more powerful than RC-TDT, especially when  $LD$  is weak as in scenario 1 of Table 4.

**Table 4:** Parameters used in simulation study 1.

Scenario	Mode	$p_D$	$p_A$	$f_{DD}$	$f_{dd}$	$f_{Da}$
1	Dominant	0.013	0.4	1.0	0.025	1.000
2		0.016	0.4	0.8	0.025	0.800
3		0.027	0.4	0.5	0.025	0.500
4		0.074	0.4	0.2	0.025	0.200
5	Additive	0.026	0.4	1.0	0.025	0.513
6		0.032	0.4	0.8	0.025	0.413
7		0.053	0.4	0.5	0.025	0.263
8		0.143	0.4	0.2	0.025	0.113
9	Recessive	0.160	0.4	1.0	0.025	0.025
10		0.180	0.4	0.8	0.025	0.025
11		0.229	0.4	0.5	0.025	0.025
12		0.378	0.4	0.2	0.025	0.025

**Table 5:** Parameters used in simulation study 2.

Scenario	$\theta$	$p_A$	$p_D$	$f_{dd}$	$f_{DD}$	$f_{Da}$			
						Dom	Rec	Add	Mul
S1	.01	.50	.10	.01	1.0	1.0	.01	.505	.100
S2	.10	.50	.10	.01	1.0	1.0	.01	.505	.100
S3	.01	.10	.10	.01	1.0	1.0	.01	.505	.100
S4	.10	.10	.10	.01	1.0	1.0	.01	.505	.100
S5	.01	.50	.10	.01	0.5	0.5	.01	.255	.071
S6	.10	.50	.10	.01	0.5	0.5	.01	.255	.071
S7	.01	.10	.10	.01	0.5	0.5	.01	.255	.071
S8	.10	.10	.10	.01	0.5	0.5	.01	.255	.071

Penetrance is the conditional probability of observing a phenotype given a specified disease genotype. In scenario 1, we set  $f_{DD}$  (the penetrance for a subject whose marker genotype is  $DD$ ) at 1, which is an idealistic penetrance. To compare the power of the proposed RC-TDH with that of its competitor under different penetrance,  $f_{DD}$  is varied from full penetrance to incomplete penetrance 0.5, which is more realistic. The results in Table 9 and Figure 2 show that the proposed RC-TDH has better power than RC-TDT with half penetrance for genotype  $DD$  individuals as in scenario 5 of Table 5.

In summary, our simulation results show that the proposed RC-TDH is generally more powerful than RC-TDT for a broad range of  $LD$ , the tightness of the linkage, and across disease models.

## 5. Discussion

For mapping complex diseases, it is common that the transmission probabilities of a marker allele of interest vary across heterozygous parents, due to locus heterogeneity, etiological heterogeneity, and many other complexities and/or combinations of them [3, 4]. Under such transmission heterogeneity, the transmission likelihood generally has the form of mixture models with many parameters, and the efficient score test has two parts in the form of a TDH test [4]. This paper studies a TDH test which allows the inclusion of reconstructed



**Table 6:** Simulated critical values for RC-TDH.

Sibship size	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
3	5.23	8.59	14.04
4	5.40	9.08	15.43
6	5.52	9.82	16.91

Note: determined on the basis of the dominant model with  $f_{DD} = 0.2$  (Scenario 4 in Table 4).

**Table 7:** Simulated true type I error rates of the RC-TDT and of RC-TDH.

Sibship size	$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.001$	
	RC-TDT	RC-TDH	RC-TDT	RC-TDH	RC-TDT	RC-TDH
3	0.0490	0.0502	0.0094	0.0100	0.0008	0.0010
4	0.0485	0.0499	0.0097	0.0099	0.0010	0.0010
6	0.0503	0.0497	0.0101	0.0100	0.0008	0.0010

Determined on the basis of the dominant model with  $f_{DD} = 0.2$  (scenario 4 in Table 4).

**Table 8:** Powers of RC-TDT and RC-TDH in simulation study 1.

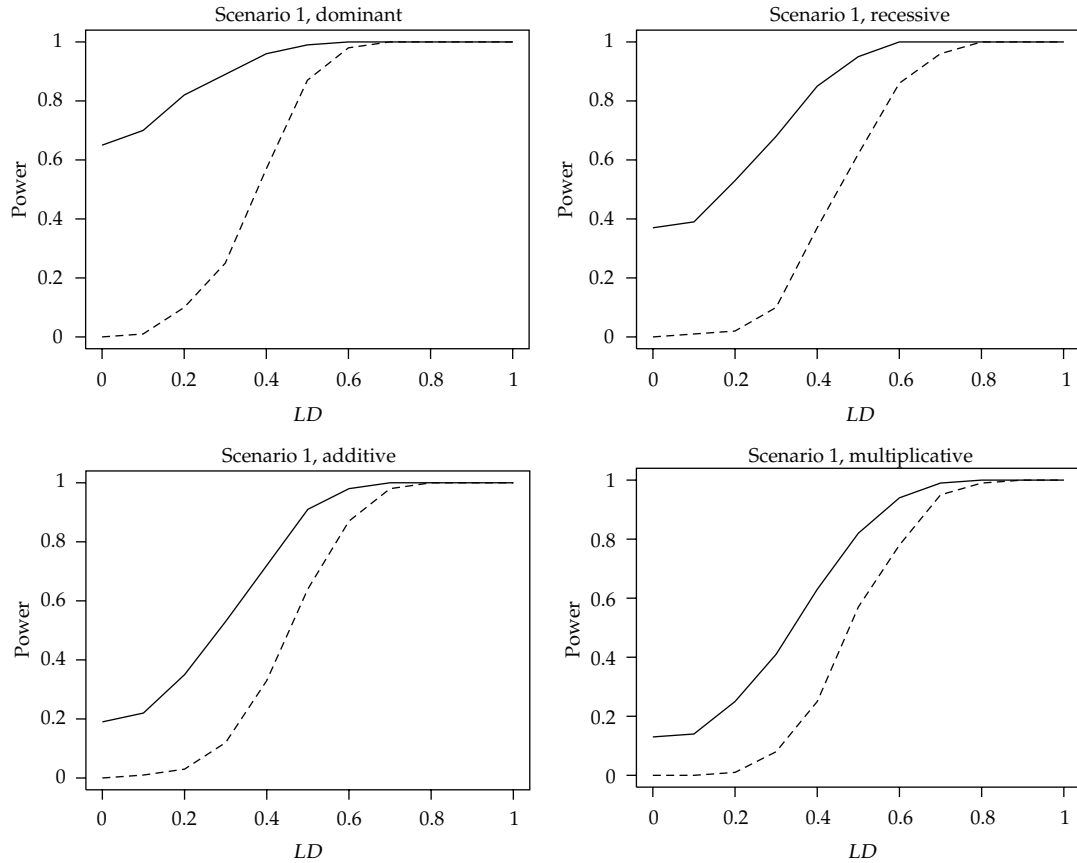
Model	$\Delta = 0.1$		$\Delta = 0.5$		$\Delta = 0.9$	
	RC-TDT	RC-TDH	RC-TDT	RC-TDH	RC-TDT	RC-TDH
D1	0.13	0.87	0.68	0.97	0.99	1.00
D2	0.08	0.41	0.61	0.73	0.97	0.98
A1	0.10	0.43	0.65	0.77	0.97	0.97
A2	0.09	0.16	0.56	0.59	0.98	0.96
R1	0.21	0.86	0.99	1.00	1.00	1.00
R2	0.15	0.40	0.98	0.99	1.00	1.00

D (dominant), R (recessive), A (additive);  $f_{DD}$ : 1 (1.0), 2 (0.5); with type-I error rate .05 based on 500 independent replicates of 150 nuclear families.  $\Delta$  is the measurement for linkage disequilibrium. When  $\Delta = 0$ , there is no linkage disequilibrium. In this simulation study, all the parental marker genotypes are missing.

**Table 9:** Powers of the RC-TDT and RC-TDH in simulation study 2.

Scenario	$\Delta$	Dominant		Recessive		Additive		Multiplicative	
		RC-TDT	RC-TDH	RC-TDT	RC-TDH	RC-TDT	RC-TDH	RC-TDT	RC-TDH
S1	0.0	0.00	0.65	0.00	0.37	0.00	0.19	0.00	0.13
	0.2	0.10	0.82	0.02	0.53	0.03	0.35	0.01	0.25
	0.4	0.57	0.96	0.37	0.85	0.33	0.72	0.25	0.63
	0.6	0.98	1.00	0.86	1.00	0.87	0.98	0.78	0.94
	0.8	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00
	1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S5	0.0	0.00	0.17	0.00	0.04	0.00	0.01	0.00	0.01
	0.2	0.02	0.27	0.00	0.09	0.01	0.10	0.00	0.06
	0.4	0.25	0.62	0.04	0.29	0.16	0.43	0.06	0.24
	0.6	0.79	0.93	0.22	0.53	0.65	0.81	0.40	0.62
	0.8	0.99	1.00	0.61	0.89	0.96	0.99	0.81	0.94
	1.0	1.00	1.00	0.89	0.99	1.00	1.00	0.97	0.99

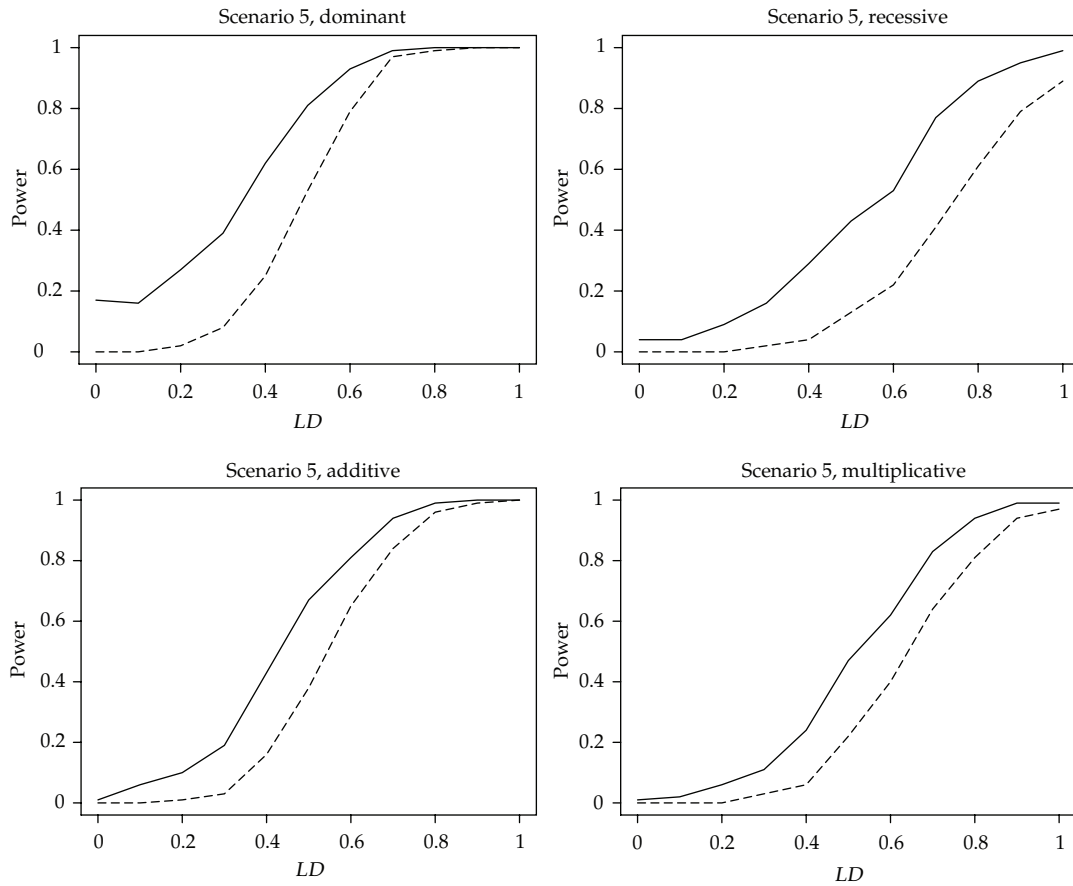
In this simulation, we used 50% families with available parental marker genotypes and 50% families without parental marker genotypes.



**Figure 1:** Power of RC-TDH (solid) and RC-TDT (dashed) in Table 5 (scenario 1). This figure is based on scenario 1:  $\theta = 0.01$ ,  $p_D = 0.1$ ,  $p_A = 0.5$ ,  $f_{DD} = 1$  and  $f_{dd} = 0.01$ . The type I error rate is 0.001 based on 500 independent replicates of 200 nuclear families, 50% of which without parental information. Every family contains 3 sibs and at least one is affected.  $LD$  is the measurement for linkage disequilibrium as defined by  $\Delta$  in Section 4.1. When  $LD = 0$ , there is no linkage disequilibrium.

parental marker genotype data and extends the RC-TDT of Knapp [10, 11]. The proposed new approach was validated by simulation studies and GAW14 data sets, and the results indicate that the new approach might improve the power of family-based linkage analysis for a broad range of  $LD$ . Moreover, the simulation studies also indicate that the systematic power advantage of the RC-TDH test over the RC-TDT holds regardless of the underlying genetic models (e.g., recessive, dominant, additive, multiplicative).

Similar to RC-TDT, the new approach can utilize the missing parental information that can be reconstructed from the child genotypes, especially including some families with genotype-concordant or phenotype-concordant sibs. In addition, the proposed test is a sibship-oriented method which does not require specification of the underlying genetic model; it naturally uses the multiple siblings by considering the sibship as a whole. The second part of the RC-TDH statistic, the THT part of the test statistic, is based on information from IBD. This is quite obvious in the situation of affected sib-pairs, where the THT is essentially equivalent to the so-called mean test [4, 13].



**Figure 2:** Power of RC-TDH (solid) and RC-TDT (dashed) in Table 5 (scenario 5). This figure is based on scenario 5:  $\theta = 0.01$ ,  $p_D = 0.1$ ,  $p_A = 0.5$ ,  $f_{DD} = 0.5$  and  $f_{da} = 0.01$ . The type I error rate is 0.001 based on 500 independent replicates of 200 nuclear families, 50% of which without parental information. Every family contains 3 sibs and at least one is affected.  $LD$  is the measurement for linkage disequilibrium as defined by  $\Delta$  in Section 4.1. When  $LD = 0$ , there is no linkage disequilibrium.

Many other linkage analysis tests such as the tests implemented by Genehunter have relatively low power with respect to TDT or TDH when  $LD$  is present. In reality, some degree of  $LD$  is often present particularly when we use dense genetic markers (e.g., SNPs) along the genome because they are available at increasingly cheaper cost, and these dense markers are already very affordable. With a large number of dense genetic markers, some markers may be expected to fall into the  $LD$  block of the causal variants. When using these affordable dense markers along the genome or candidate gene regions, we believe that RC-TDH will have better chance of success than the classical IBD-based linkage methods in detecting linkage signals along the genome.

As high density SNP arrays become increasingly affordable to researchers, genome-wide linkage studies are becoming common. Our TDH test has simple closed form test statistics which is computationally easy in addition to good overall power across a broad range of  $LD$ . Thus the proposed method would be potentially useful for genomewide linkage analysis. In contrast, likelihood ratio test for mixture likelihood is generally computationally

intensive [5, 17]. Many existing linkage tests and algorithms such as the likelihood ratio test discussed in Lo et al. [5] would be too computationally intensive for genomewide studies or when the number of genotyped markers is large.

It is possible to further extend the method to be applicable to markers with more than two alleles, which would be of great interest in studying haplotypes of multiple loci. However, our proposed tests are already applicable to the commonly used biallelic markers; for instance, the widely used single nucleotide polymorphisms (SNPs) are convenient biallelic markers.

## Appendix

### A. Computational Details for the RC-TDH Test

When there are no parents who have been typed, the conditional probability has been derived in equation (A.6) of Knapp [10]. When only one parent has been typed as  $AB$ , the same constraint for reconstruction applies, thus (A.6) of Knapp [10] also works. Next we derive the conditional probability when only one parent has been typed as  $AA$ . The case of when only one parent has been typed as  $BB$  is obvious due to symmetry between  $A$  and  $B$ .

#### A.1. One Parental Genotype Has Been Typed as $AA$

Note that the family index  $i$  has been dropped in the following formula.

Only one parental genotype has been typed, which is  $AA$ , but the genotype of the missing parent can be reconstructed as  $AB$ , if there is at least one child with genotype  $AB$  and at least one child with genotype  $AA$ . Here, the condition  $R$  is  $N^{AB} > 0$  and  $N^{AA} > 0$ . To calculate the conditional distribution of  $T$ , we first calculate the probability of satisfying the constraint for reconstruction,  $R$ :

$$\begin{aligned}
 P_{H_0}(R) &= P_{H_0}(N^{AA} > 0 \text{ and } N^{AB} > 0) \\
 &= 1 - P_{H_0}(N^{AA} = 0) - P_{H_0}(N^{AB} = 0) + P_{H_0}(N^{AA} = 0 \text{ and } N^{AB} = 0) \quad (\text{A.1}) \\
 &= 1 - 2\left(\frac{1}{2}\right)^{n_c}.
 \end{aligned}$$

Then we calculate the joint probability of  $T$  and  $R$ :

$$\begin{aligned}
 P_{H_0}(\{T = c\} \cap R) &= P_{H_0}(T = c \cap N^{AA} > 0 \cap N^{AB} > 0) \\
 &= P_{H_0}(T = c) - P_{H_0}(T = c \cap (N^{AA} = 0 \cup N^{AB} = 0)) \\
 &= P_{H_0}(T = c) - P_{H_0}(T = c \cap N^{AA} = 0) - P_{H_0}(T = c \cap N^{AB} = 0) + 0 \\
 &= \binom{n_a}{c - n_a} \left(\frac{1}{2}\right)^{n_a} - P_{H_0}(T = c \cap N^{AA} = 0) - P_{H_0}(T = c \cap N^{AB} = 0). \quad (\text{A.2})
 \end{aligned}$$

There are three cases for the calculation:

$$\text{case 1: } c = n_a, P_{H_0}(\{T = c\} \cap R) = \binom{n_a}{c-n_a} (1/2)^{n_a} - (1/2)^{n_c},$$

$$\text{case 2: } n_a < c < 2n_a, P_{H_0}(\{T = c\} \cap R) = \binom{n_a}{c-n_a} (1/2)^{n_a},$$

$$\text{case 3: } c = 2n_a, P_{H_0}(\{T = c\} \cap R) = \binom{n_a}{c-n_a} (1/2)^{n_a} - (1/2)^{n_c}.$$

Therefore the distribution of  $T$  conditioned on  $R$  is

$$P_{H_0}(T = c | R) = \begin{cases} \frac{\binom{n_a}{c-n_a} (1/2)^{n_a} - (1/2)^{n_c}}{1 - 2(1/2)^{n_c}}, & c = n_a, \\ \frac{\binom{n_a}{c-n_a} (1/2)^{n_a}}{1 - 2(1/2)^{n_c}}, & n_a < c < 2n_a, \\ \frac{\binom{n_a}{c-n_a} (1/2)^{n_a} - (1/2)^{n_c}}{1 - 2(1/2)^{n_c}}, & c = 2n_a. \end{cases} \quad (\text{A.3})$$

### ***A.2. At Least One Parental Genotype Is Missing and Cannot Be Reconstructed, but the Condition for the S-TDT Is Satisfied***

In a sibship 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  $AA$  is  $r$  and the number of sibs who are of genotype  $AB$  is  $s$ . Let  $x$  be the number of  $AA$  sibs and let  $y$  be the number of  $AB$  sibs who are classified as affected. As discussed in Spielman and Ewens [9], given the totals  $r, s, a, u$ , and  $t$ , the numbers  $x, y$  can be regarded as two entries in a  $2 \times 3$  contingency table with marginal totals  $a, u, r, s$ , and  $t - r - s$ . Therefore, the distribution of  $T = 2x + y$  can be obtained by the generalized hypergeometric distribution [18, page 47]. More specifically, we have

$$P(T = c) = \sum_{i=\max(c-2a, c-2r, 0)}^{\min(s, a, c)} \frac{\binom{r}{(c-i)/2} \cdot \binom{s}{i} \cdot \binom{t-r-s}{a-((c+i)/2)}}{\binom{t}{a}}, \quad 1 \leq c \leq \min(2r + s, 2a). \quad (\text{A.4})$$

More formulas of parental marker genotype reconstruction probabilities under various missing genotypes types and constraints, as well as detailed derivations of these formulas, can be found in Han [16].

### **Acknowledgments**

This research was partially supported by a Stony Wold-Herbert Foundation grant, the MPD Research Consortium Project Grant (1P01 CA108671), and the New York University Cancer Center Supporting Grant (2P30 CA16087) and by the NYU NIEHS Center Grant (5P30 ES00260). The research of JH was carried out as part of her Ph.D. dissertation work at New York University.

## References

- [1] Y. Shao, "Linkage Analysis," in *Encyclopedia of Quantitative Risk Analysis and Assessment*, John Wiley & Sons, Hoboken, NJ, USA, 2008.
- [2] J. Ott, *Analysis of Human Genetic Linkage*, Johns Hopkins University, 3rd edition, 1999.
- [3] E. S. Lander and N. J. Schork, "Genetic dissection of complex traits," *Science*, vol. 265, no. 5181, pp. 2037–2048, 1994.
- [4] Y. Shao, "Adjustment for transmission heterogeneity in mapping complex genetic diseases using mixture models and score tests," *Proceeding of the American Statistical Association*, pp. 383–393, 2005.
- [5] S. H. Lo, X. Liu, and Y. Shao, "A marginal likelihood model for family-based data," *Annals of Human Genetics*, vol. 67, no. 4, pp. 357–366, 2003.
- [6] R. S. Spielman, R. E. McGinnis, and W. J. Ewens, "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)," *American Journal of Human Genetics*, vol. 52, no. 3, pp. 506–516, 1993.
- [7] H. Zhao, "Family-based association studies," *Statistical Methods in Medical Research*, vol. 9, no. 6, pp. 563–587, 2000.
- [8] W. J. Ewens and R. S. Spielman, "The transmission/disequilibrium test," in *Handbook of Statistical Genetics*, D. J. Balding, M. Bishop, and C. Cannings, Eds., John Wiley & Sons, 2nd edition, 2003.
- [9] R. S. Spielman and W. J. Ewens, "A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test," *American Journal of Human Genetics*, vol. 62, no. 2, pp. 450–458, 1998.
- [10] M. Knapp, "The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test," *American Journal of Human Genetics*, vol. 64, no. 3, pp. 861–870, 1999.
- [11] M. Knapp, "Using exact P values to compare the power between the reconstruction-combined transmission/disequilibrium test and the sib transmission/disequilibrium test," *American Journal of Human Genetics*, vol. 65, no. 4, pp. 1208–1210, 1999.
- [12] D. Curtis, "Use of siblings as controls in case-control association studies," *Annals of Human Genetics*, vol. 61, no. 4, pp. 319–333, 1997.
- [13] J. Huang and Y. Jiang, "Linkage detection adaptive to linkage disequilibrium: the disequilibrium maximum-likelihood-binomial test for affected-sibship data," *American Journal of Human Genetics*, vol. 65, no. 6, pp. 1741–1759, 1999.
- [14] W. C. Blackwelder and R. C. Elston, "A comparison of sib-pair linkage tests for disease susceptibility loci," *Genetic Epidemiology*, vol. 2, no. 1, pp. 85–97, 1985.
- [15] M. Boehnke and C. D. Langefeld, "Genetic association mapping based on discordant sib pairs: the discordant-alleles test," *American Journal of Human Genetics*, vol. 62, no. 4, pp. 950–961, 1998.
- [16] J. Han, *Family-based linkage analysis allowing for missing parental information [Ph.D. thesis]*, New York University, 2005.
- [17] X. Liu and Y. Shao, "Asymptotics for likelihood ratio tests under loss of identifiability," *The Annals of Statistics*, vol. 31, no. 3, pp. 807–832, 2003.
- [18] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, John Wiley & Sons, New York, NY, USA, 3rd edition, 1968.



# Hindawi

Submit your manuscripts at  
<http://www.hindawi.com>

