OPEN ACCESS Freely available online

PLOS ONE

Handling Missing Data in Transmission Disequilibrium Test in Nuclear Families with One Affected Offspring

Gulhan Bourget*

Department of Mathematics, California State University, Fullerton, California, United States of America

Abstract

The Transmission Disequilibrium Test (TDT) compares frequencies of transmission of two alleles from heterozygote parents to an affected offspring. This test requires all genotypes to be known from all members of the nuclear families. However, obtaining all genotypes in a study might not be possible for some families, in which case, a data set results in missing genotypes. There are many techniques of handling missing genotypes in parents but only a few in offspring. The robust TDT (rTDT) is one of the methods that handles missing genotypes for all members of nuclear families [with one affected offspring]. Even though all family members can be imputed, the rTDT is a conservative test with low power. We propose a new method, Mendelian Inheritance TDT (MITDT-ONE), that controls type I error and has high power. The MITDT-ONE uses Mendelian Inheritance properties, and takes population frequencies of the disease allele and marker allele into account in the rTDT method. One of the advantages of using the MITDT-ONE is that the MITDT-ONE can identify additional significant genes that are not found by the rTDT. We demonstrate the performances of both tests along with Sib-TDT (S-TDT) in Monte Carlo simulation studies. Moreover, we apply our method to the type 1 diabetes data from the Warren families in the United Kingdom to identify significant genes that are related to type 1 diabetes.

Citation: Bourget G (2012) Handling Missing Data in Transmission Disequilibrium Test in Nuclear Families with One Affected Offspring. PLoS ONE 7(10): e46100. doi:10.1371/journal.pone.0046100

Editor: Zhaoxia Yu, University of California, Irvine, United States of America

Received February 27, 2012; **Accepted** August 28, 2012; **Published** October 8, 2012

Funding: The author has no funding or support to report.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: galpargu@fullerton.edu

Introduction

The Transmission Disequilibrium Test (TDT) is the most widely used family-based test for linkage disequilibrium [1], [2]. It was first introduced to handle one affected offspring in a nuclear family, and was later extended to two or more affected offspring, and to multi-allelic markers as well. The TDT is a test for linkage in the presence of linkage disequilibrium [1], [2].

The TDT compares frequencies of the transmission of two alleles from heterozygote parents to an affected offspring. The TDT requires complete genotypes from parents and offspring. However, sometimes genotypes may not be available. If genotypes of parents are missing, including only complete cases [3], [4], [5], [6], [7], or reconstructing missing parental genotypes by assuming a missing at random (MAR) model [8] have been suggested as common approaches in practice. However, if parental genotypes are missing due to his genotype at the locus of interest, then the informatively missing model is more appropriate than the MAR model [9]. Also, including only complete families and families with only one parent missing in informatively missing parent(s) [3], [6], [7], [10] reconstructing parental genotypes from their affected offspring [2], or from affected and unaffected siblings (Reconstruction-Combined TDT) [4], [11], or completely ignoring parental genotypes and comparing frequencies of genotypes of unaffected and affected offspring (S-TDT) [12], [13],[14], [15], or combining different data sets from families with parental genotypes and from families with missing parental genotype data but whose siblings' genotypes are unaffected (C-TDT) [12] has been also proposed as alternative approaches.

The robust TDT (rTDT) was proposed to handle any missing genotypes in a nuclear family with one affected offspring and bi-allelic marker [16]. The rTDT does not assume any missing model, and defines an interval estimate of TDT by considering all possible completions of missing genotypes. Sebastiani et al. [16] claimed that rTDT has more power than TDT. The simulation study was not performed, and the claim of having more power than TDT was shown mathematically for a specific missing pattern for each family [16]. That is, they assumed that missing families have the same form: the genotype of one parent is missing, the other parent has a heterozygous genotype, and the affected child has homozygous genotype [see Discussion section for more details]. This specific missing pattern for each family is not a reasonable assumption in practice. Alpargu (Bourget) [17] defined the rTDT for *two* affected offspring, and showed in simulation studies that rTDT was too conservative, and had low power. Because of its poor performance, the Mendelian Inheritance-Transmission Disequilibrium Test (MI-TDT), which takes population frequencies of the disease allele (p) and marker allele (m) into account in rTDT, was proposed [17]. The MI-TDT performed better than rTDT by controlling type I error rates and having high power. Since, MI-TDT outperformed rTDT, in this paper we propose the Mendel Inheritance-Transmission Disequilibrium Test (MITDT-ONE) for *one* affected offspring. The MITDT-ONE considers p and m in rTDT. The simulation study replicating real life scenarios such as different missing models and different genetic models shows that MITDT-ONE outperforms rTDT by providing better control of type I error rates and producing higher power.

Methods

We demonstrate the features of rTDT and MITDT-ONE with an example. We assume that we have genotypes of nuclear families with one affected offspring, and bi-allelic markers with alleles 1 and 2. In a given data set, there are (1,1), (1,2), or (2,2) complete genotypes or (0,0) missing genotypes. For each family, there are three genotypes with the first two genotypes for parents and the last genotype for offspring (e.g., (1,2)(1,1)(1,2)). If at least one of the genotypes is unknown, then the data is called incomplete. Otherwise it is called complete. Hence, a whole data set has two parts for a given marker: complete and incomplete trio genotypes.

The TDT considers transmission from heterozygote parents (h) to affected offspring. Let u be the number of h that transmit allele 1 to an affected offspring, and v be the number of h that transmit allele 2 to an affected offspring. Then, the TDT statistic for complete data

$$\chi^2_{\mathrm{TDT}} = \frac{(u-v)^2}{u+v} \qquad (1)$$

tests linkage (θ) between a disease and a marker locus in the presence of linkage disequilibrium ($\delta>0$ or $\delta<0$) [1]. Under the null hypothesis of no linkage ($\theta=0.5$), χ^2_{TDT} follows a central chi-square distribution with 1 degree of freedom (df).

We construct interval estimates of MITDT-ONE and rTDT as follows: (1) compute maximum and minimum increments in u and v by considering all possible admissible completions of missing genotypes (u_{inc} (v_{inc}) for maximum increments of u (v)), (2) find population frequencies of disease allele (p), and marker allele (m), and finally, (3) compute maximum and minimum values of u and v (u_{max} and u_{min} for u and v_{max} and v_{min} for v). While all three steps are involved in MITDT-ONE, rTDT does not require step (2). This is the only important difference between two methods. However, MITDT-ONE requires the value of p, which is difficult to know in some diseases. We can overcome the knowledge of p by assuming $p \approx m$ because McGinnis (1998) [18] showed that TDT is able to detect linkage, and its power exceeds 0.5 only when δ is close to its most positive value $\delta_{\max}$ (see the definition of $\delta_{\max}$ in the following section) when $\delta>0$, and allele frequencies m and p are similar in magnitude at marker and disease locus.

For complete families, let us assume that we have 50 heterozygote parents (h_c) in which 35 of them transmit allele 1 (u_c), and 15 of them transmit allele 2 (v_c). Using (1), we compute $\chi^2_{\mathrm{TDT}} = (35-15)^2/50 = 8$. The chi-square distribution with 1 df at 5% nominal level is 3.84. Based on only complete cases, we reject the null hypothesis of no linkage at 5% nominal level. Now, assume $m=0.25$ and $p=0.05$ with two missing families as in Table 1.

The first step of imputing missing cases involves only possible admissible completions. The MITDT-ONE and rTDT (as does TDT) consider families with at least one heterozygote parent. For example, if the incomplete case is (1,1)(0,0)(1,2), we do not consider the completion (1,1)(2,2)(1,2) because both parents have homozygous genotypes. Moreover, in family 2 above, (1,2)(1,1)(2,2) is not a possible admissible completion because the only possible completions for offspring are (1,1) or (1,2). All possible admissible genotypes are defined in Table 2.

Table 1. Two missing cases.

Family	Parents	Children
1	(0,0)(1,2)	(0,0)
2	(1,2)(1,1)	(0,0)

doi:10.1371/journal.pone.0046100.t001

Under the null hypothesis ($\theta=0.5$), heterozygote parent transmits allele 1 but not allele 2 to an affected offspring with probability $\theta_1 = (m+\delta/p)(1-m)-(\theta\delta)/p$, and the same parent transmits allele 2 but not allele 1 to an affected offspring with probability $\theta_2 = (1-m-\delta/p)m+(\theta\delta)/p$, where δ is the coefficient of disequilibrium, m is the frequency of the marker allele 1, and p is the population relative frequency of disease allele [19]. The χ^2 statistic compares the number of transmissions with probabilities θ_1 and θ_2. It can be shown that these probabilities are the same under the null hypothesis. Thus, the expected number of transmissions are the same. Thus, $\mathrm{E}(u)=\mathrm{E}(v)$. However, the probabilities are different when there is linkage, and hence the number of transmissions are different. This means that the χ^2 statistic is related to the parameters δ,m,p, and θ.

All these families have equal probabilities of being considered under the null hypothesis of no linkage. However, MITDT-ONE and rTDT consider increments in u (u_{inc}) and v (v_{inc}). The exact maximum and minimum values of TDT in (1) are attained by rTDT. The interval estimate of rTDT is [5.45,9.98]. While the minimum value is attained when $u=35$ and $v=18$ (scenarios 7 and 9), the maximum value is attained when $u=38$ and $v=15$ (scenarios 5 and 8). The interval estimate of MITDT-ONE is [7.61,8.27] with the same completion of the families as rTDT.

Both tests use the same admissible cases and consider lower limits to identify significant genes. Both methods reject the null hypothesis of no linkage at 5% nominal level in the above example. The interval estimate of MITDT-ONE is always contained in the interval estimate of rTDT (see in Construction of the MITDT-ONE and rTDT for more details). It is important to note that MITDT-ONE and rTDT have the same minimum values for u and v but differ at maximum values of u and v. Therefore, MITDT-ONE will never have less power than rTDT. Since the MITDT-ONE has more power and controls type I error rates better, we suggest using the MITDT-ONE test instead of rTDT test.

Construction of the MITDT-ONE and rTDT

There are 17 admissible missing cases in a nuclear family with one affected offspring (Table 3). Sebastiani et al. [16] proposed an interval estimate of rTDT for one affected offspring. They proceeded in the following way: χ^2_{TDT} in (1) is a monotone convex function on a closed domain. Thus, it achieves its maximum and

Table 2. Admissible cases.

Family	Scenario	Parent	Children	u_{inc}	v_{inc}
1	1	(1,1)(1,2)	(1,1)	1	0
	2	(1,1)(1,2)	(1,2)	0	1
	3	(2,2)(1,2)	(1,2)	1	0
	4	(2,2)(1,2)	(2,2)	0	1
	5	(1,2)(1,2)	(1,1)	2	0
	6	(1,2)(1,2)	(1,2)	1	1
	7	(1,2)(1,2)	(2,2)	0	2
2	8	(1,2)(1,1)	(1,1)	1	0
	9	(1,2)(1,1)	(1,2)	0	1

doi:10.1371/journal.pone.0046100.t002

Table 3. Number of missing cases in a family with one affected offspring.

	Offspring Genotype					
Case	**Parental Genotype**	**(0,0)**	**(1,1)**	**(1,2)**	**(2,2)**	**Total**
1	(0,0) (0,0)	+	+	+	+	4
2	(0,0) (1,1)	+	+	+	−	3
3	(0,0) (1,2)	+	+	+	+	4
4	(0,0) (2,2)	+	−	+	+	3
5	(1,1) (1,2)	+	⋆	⋆	⋆	1
6	(1,2) (1,2)	+	⋆	⋆	⋆	1
7	(1,2) (2,2)	+	⋆	⋆	⋆	1
Total number of admissible incomplete trios						17

The symbols +, −, and ⋆ denote possible incomplete, impossible incomplete, and complete cases, respectively.

doi:10.1371/journal.pone.0046100.t003

minimum values at one of its extreme points. The maximum and minimum values of u and v were considered to define the maximum and minimum values of χ^2_{TDT}. First, all possible admissible completions were identified (Tables 4 and 5), and then the maximum and minimum increments in u and v (Table 6) were defined as

$$u_{\mathrm{inc}} = 2n_1 + 2n_2 + n_3 + n_5 + n_6 + 2n_8 + 2n_9 + n_{10} + n_{12} + n_{13} + n_{15} + 2n_{16} + n_{17}$$

$$v_{\mathrm{inc}} = 2n_1 + n_3 + 2n_4 + n_5 + n_7 + 2n_8 + n_{10} + 2n_{11} + n_{12} + n_{14} + n_{15} + 2n_{16} + n_{17},$$

where n_k $(k = 1,2,\ldots,17)$ is the number of missing families in case k. The maximum and minimum values of u and v were defined as

$$u_{\min} = u_{\mathrm{c}} + n_9, \qquad u_{\max} = u_{\mathrm{c}} + u_{\mathrm{inc}}$$

$$v_{\min} = v_{\mathrm{c}} + n_{11}, \qquad v_{\max} = v_{\mathrm{c}} + v_{\mathrm{inc}}, \tag{2}$$

where $u_c(v_c)$ is the number of h that transmit allele 1 (2) to affected offspring in complete data set. And finally, the interval estimate $[\chi^2_{\min}, \chi^2_{\max}]$ of rTDT was defined as

1. If $u_{\min} \geq v_{\max}$, then

$$\chi^2_{\min} = \chi^2(u_{\min}, v_{\max}) \leq \chi^2_{\mathrm{TDT}} \leq \chi^2(u_{\max}, v_{\min}) = \chi^2_{\max}.$$

2. If $u_{\max} \leq v_{\min}$, then

$$\chi^2_{\min} = \chi^2(u_{\max}, v_{\min}) \leq \chi^2_{\mathrm{TDT}} \leq \chi^2(u_{\min}, v_{\max}) = \chi^2_{\max}.$$

Table 4. List of admissible completions for cases 1–8.

Case	Incomplete Genotypes		Admissible Completions		Increments	
k	**Parents**	**Offspring**	**Parents**	**Offspring**	u_{inc}	v_{inc}
1	(0, 0)(0, 0)	(0, 0)	(1,1) (1,1)	(1,1)	0	0
			(1,1) (1,2)	(1,1)	1	0
			(1,1) (1,2)	(1,2)	0	1
			(1,1) (2,2)	(1,2)	0	0
			(1,2) (1,2)	(1,1)	2	0
			(1,2) (1,2)	(1,2)	1	1
			(1,2) (1,2)	(2,2)	0	2
			(2,2) (1,2)	(1,2)	1	0
			(2,2) (1,2)	(2,2)	0	1
			(2,2) (2,2)	(2,2)	0	0
2	(0, 0)(0, 0)	(1, 1)	(1,1) (1,1)	(1,1)	0	0
			(1,1) (1,2)	(1,1)	1	0
			(1,2) (1,2)	(1,1)	2	0
3	(0, 0)(0, 0)	(1, 2)	(1,1) (1,2)	(1,2)	0	1
			(1,1) (2,2)	(1,2)	0	0
			(1,2) (1,2)	(1,2)	1	1
			(1,2) (2,2)	(1,2)	1	0
4	(0, 0)(0, 0)	(2, 2)	(1,2) (1,2)	(2,2)	0	2
			(1,2) (2,2)	(2,2)	0	1
			(2,2) (2,2)	(2,2)	0	0
5	(0, 0)(1, 1)	(0, 0)	(1,1) (1,1)	(1,1)	0	0
			(1,2) (1,1)	(1,1)	1	0
			(1,2) (1,1)	(1,2)	0	1
			(2,2) (1,1)	(1,2)	0	0
6	(0, 0)(1, 1)	(1, 1)	(1,1) (1,1)	(1,1)	0	0
			(1,2) (1,1)	(1,1)	1	0
7	(0, 0)(1, 1)	(1, 2)	(1,2) (1,1)	(1,2)	0	1
			(2,2) (1,1)	(1,2)	0	0
8	(0.0) (1,2)	(0,0)	(1,1) (1,2)	(1,1)	1	0
			(1,1) (1,2)	(1,2)	0	1
			(1,2) (1,2)	(1,1)	2	0
			(1,2) (1,2)	(1,2)	1	1
			(1,2) (1,2)	(2,2)	0	2
			(2,2) (1,2)	(1,2)	1	0
			(2,2) (1,2)	(2,2)	0	1

doi:10.1371/journal.pone.0046100.t004

3. In all other cases:

$$\chi^2_{\min} = 0 \leq \chi^2_{\mathrm{TDT}} \leq \max\{\chi^2(u_{\max} v_{\min}), \chi^2(u_{\min}, v_{\max})\} = \chi^2_{\max}.$$

The value of $\chi^2_{\min}$ ($\chi^2_{\max}$) makes a decision against (conforming) the null hypothesis. If χ^2_{TDT} for complete data (i.e., missing data are ignored) and $\chi^2_{\min}$ reach the conclusion of the alternative hypothesis (i.e., significant genes), and $\chi^2_{\min} \leq \chi^2_{\mathrm{TDT}}$, then rTDT affirms significant genes of complete data. Similarly, the value of $\chi^2_{\max}$ ratifies the insignificant genes if χ^2_{TDT} and $\chi^2_{\max}$ cannot reject

Table 5. List of admissible completions for cases 9–17.

Case	Incomplete Genotypes		Admissible Completions		Increments	
k	Parents	Offspring	Parents	Offspring	u_{inc}	v_{inc}
9	(0, 0)(1, 2)	(1, 1)	(1,1) (1,2)	(1,1)	1	0
			(1,2) (1,2)	(1,1)	2	0
10	(0, 0)(1, 2)	(1, 2)	(1,1) (1,2)	(1,2)	0	1
			(1,2) (1,2)	(1,2)	1	1
			(2,2) (1,2)	(1,2)	1	0
11	(0, 0)(1, 2)	(2, 2)	(1,2) (1,2)	(2,2)	0	2
			(2,2) (1,2)	(2,2)	0	1
12	(0, 0)(2, 2)	(0, 0)	(1,1) (2,2)	(1,2)	0	0
			(1,2) (2,2)	(1,2)	1	0
			(1,2) (2,2)	(2,2)	0	1
			(2,2) (2,2)	(2,2)	0	0
13	(0, 0)(2, 2)	(1, 2)	(1,1) (2,2)	(1,2)	0	0
			(1,2) (2,2)	(1,2)	1	0
			(2,2) (2,2)	(1,2)	0	0
14	(0, 0)(2, 2)	(2, 2)	(1,2) (2,2)	(2,2)	0	1
			(2,2) (2,2)	(2,2)	0	0
15	(1, 1)(1, 2)	(0, 0)	(1,1) (1,2)	(1,1)	1	0
			(1,1) (1,2)	(1,2)	0	1
16	(1, 2)(1, 2)	(0, 0)	(1,2) (1,2)	(1,1)	2	0
			(1,2) (1,2)	(1,2)	1	1
			(1,2) (1,2)	(2,2)	0	2
17	(1, 2)(2, 2)	(0, 0)	(1,2) (2,2)	(1,2)	1	0
			(1,2) (2,2)	(2,2)	0	1

doi:10.1371/journal.pone.0046100.t005

the null hypothesis, and $\chi^2_{TDT} \le \chi^2_{max}$. In all other scenarios, rTDT cannot verify any conclusions of complete data.

Sebastiani et al. [16] did not run any simulation study to demonstrate the performance of rTDT. They theoretically showed that if all missing families are in case 9, which is not a reasonable assumption in practice, then rTDT has higher power than the classical χ^2_{TDT}. Since the power of TDT depends on linkage disequilibrium (δ), and relative frequencies of marker allele (m) and disease allele (p) [20], we ran simulation studies to take into account different realistic disease models and missing models, involving m and p. The simulation results show that rTDT overestimates the values of u_{max} (v_{max}) (results are not shown), and hence becomes a conservative test with low power. Since u_{min} (v_{min}) does not involve u_{inc} (v_{inc}), we decided to scale down u_{inc} (v_{inc}) to have a smaller value of u_{max} (v_{max}) for MITDT-ONE. One way to achieve this goal is to involve m and p in scaling. These parameters appear together in maximum linkage disequilibrium $\delta_{max} = \min\{(1-m)p,(1-p)m\}$ when linkage disequilibrium is positive $\delta > 0$, and $\delta_{max} = \min\{mp,(1-m)(1-p)\}$ when linkage disequilibrium is negative ($\delta < 0$) [18]. We scale u_{inc} (v_{inc}) with $(1-m)p$ and $(1-p)m$ when $\delta > 0$, and define u^*_{max} (v^*_{max}) for MITDT-ONE as the average of these values. That is,

$$u^*_{max} = u_c + u^*_{inc}, \qquad v^*_{max} = v_c + v^*_{inc}, \tag{3}$$

where

$$u^*_{inc} = \frac{(1-m)\cdot p\cdot u_{inc} + (1-p)\cdot m\cdot u_{inc}}{2}. \tag{4}$$

Similarly, we can define v^*_{inc} by replacing in (4) u_{inc} with v_{inc}.

Since TDT provides better power when linkage disequilibrium is at its maximum (δ_{max}) for $\delta > 0$, and $m \approx p$ [18], we can reformulate (4) for real sample data as

$$u^*_{inc} = (1-m)\cdot m\cdot u_{inc}, \qquad v^*_{inc} = (1-m)\cdot m\cdot v_{inc} \tag{5}$$

The lowest values of the interval estimates of rTDT and MITDT-ONE find significant genes when they are actually not. The way the interval estimate for MITDT-ONE constructed guarantees that its lowest interval estimate (χ^2_2) is always larger than the lowest interval estimate of rTDT (χ^2_1). This fact can be shown theoretically in the following way: let us assume $u_{min} \ge v_{max}$ (the other two conditions in (31) can be shown similarly). Since $v^*_{max} < v_{max}$, we have

$$\chi^2_1 = \frac{(u_{min} - v_{max})^2}{u_{min} + v_{max}} < \frac{(u_{min} - v_{max})^2}{u_{min} + v^\star_{max}} < \frac{(u_{min} - v^\star_{max})^2}{u_{min} + v^\star_{max}} = \chi^2_2. \tag{6}$$

We claimed that rTDT is a conservative test. We have observed this through simulation study but not theoretically. The reason rTDT becomes conservative is that the value of χ^2_1, in general, falls below the value of chi-square distribution with 1 df at α nominal level (for example, when $\alpha = 0.05$, this value is 3.84).

Results

Simulation

We replicated the simulation study in [17] for one affected offspring. Let us assume a bi-allelic marker with alleles 1 and 2 which is linked to a bi-allelic disease locus with disease-predisposing allele D and non-predisposing allele d. The penetrance for DD,Dd and dd genotypes are α,β and γ, respectively, with $0 \le \alpha,\gamma,\beta \le 1$, and the population frequencies for the marker with disease locus haplotype for 1D, 1d, 2D and 2d are c_1,c_2,c_3 and c_4, respectively, where $c_1 + c_2 + c_3 + c_4 = 1$. The population relative frequency of disease allele D is p $(= c_1 + c_3)$. The frequencies of the marker alleles 1 and 2 are $m(= c_1 + c_2)$ and $1-m(= c_3 + c_4)$, respectively. The recombination fraction between the disease and marker locus is θ, and the coefficient of disequilibrium is $\delta(= c_1c_4 - c_2c_3)$. The probability of a heterozygote parent transmitting marker allele 1 to a particular affected child [18] is defined as

$$P_t = 0.5 + \underbrace{(1-2\theta)}_{L_t}\underbrace{\left[\frac{c_1c_4 - c_2c_3}{H}\right]}_{M_t} \tag{7}$$

$$\times \underbrace{\left[p^2\left(\frac{\alpha^2-\beta^2}{4}\right) + 2p(1-p)\left(\frac{(\alpha+\beta)^2-(\beta+\gamma)^2}{16}\right) + (1-p)^2\left(\frac{\beta^2-\gamma^2}{4}\right)\right]}_{R_t} \tag{8}$$

$$= 0.5 + L_tM_tR_t,$$

where

Table 6. Admissible increments of u and v.

Case	Parents	Offspring	Increment (i,j)						Min.	Max.
			(0,0)	(1,0)	(2,0)	(1,1)	(0,1)	(0,2)	Inc	Inc
1	(0,0) (0,0)	(0,0)	+	+	+	+	+	+	0	2
2		(1,1)	+	+	+	−	−	−	0	2
3		(1,2)	+	+	−	+	+	−	0	2
4		(2,2)	+	−	−	−	+	+	0	2
5	(0,0) (1,1)	(0,0)	+	+	−	−	+	−	0	1
6		(1,1)	+	+	−	−	−	−	0	1
7		(1,2)	+	−	−	−	+	−	0	1
8	(0,0) (1,2)	(0,0)	−	+	+	+	+	+	1	2
9		(1,1)	−	+	+	−	−	−	1	2
10		(1,2)	−	+	−	+	+	−	1	2
11		(2,2)	−	−	−	−	+	+	1	2
12	(0,0) (2,2)	(0,0)	+	+	−	−	+	−	0	1
13		(1,2)	+	+	−	−	−	−	0	1
14		(2,2)	+	−	−	−	+	−	0	1
15	(1,1) (1,2)	(0,0)	−	+	−	−	+	−	1	1
16	(1,2) (1,2)	(0,0)	−	−	+	+	−	+	2	2
17	(1,2) (2,2)	(0,0)	−	+	−	−	+	−	1	1

In (i,j) , i and j represent the increments in u and v, respectively. The plus (minus) sign indicates that the increment is plausible (not plausible). The last two columns show the maximum and minimum increments in each cases.

doi:10.1371/journal.pone.0046100.t006

$$H = 2(c_1c_4 + c_2c_3)\left[p^2\left(\frac{\alpha+\beta}{2}\right)^2 + \frac{1}{2}p(1-p)\left(\frac{\alpha+2\beta+\gamma}{2}\right)^2 + (1-p)^2\left(\frac{\beta+\gamma}{2}\right)^2\right] + 2c_1c_3[p^2\alpha^2 + \frac{1}{2}p(1-p)(\alpha+\beta)^2 + (1-p)^2\beta^2] + 2c_2c_4[p^2\beta^2 + \frac{1}{2}p(1-p)(\beta+\gamma)^2 + (1-p)^2\gamma^2].$$

Our simulation study demonstrates realistic complex disease models. We generated 5,000 data sets for four different missing models and three genetics models (additive, dominant and recessive). In each simulation, we generated 100 families and each family consisted of one affected and one unaffected offspring, and 50 heterozygote fathers and 50 heterozygote mothers. In disease models, the probabilities of an affected child given the homozygosity (DD), heterozygosity (Dd), and absence of the disease alleles (dd) are defined as α,β, and γ, respectively. The values of these parameters were as $\alpha=0.8,\gamma=0.025$ for dominant ($\alpha=\beta$), additive ($\beta=(\alpha+\gamma)/2$), and recessive models ($\beta=\gamma$). In missing models, we consider (1) Missing Completely at Random (MCAR) for all genotypes, (2) informative missing for parental genotypes and MCAR for offspring genotypes, (3) informative missing for all genotypes, and (4) MCAR for parental genotypes and informative missing for offspring genotypes. A model is called "informatively missing" if at least two of the P_{11},P_{12},P_{22} are not equal, where $P_{i11},P_{i12},P_{i22},(i=\text{father }(f),\text{ mother }(m),\text{offspring }(o))$ are missing rates for f, m, and o with (1,1), (1,2) and (2,2) genotypes, respectively. In Table 7, the first column k/l denotes the missing patterns ($k=1,2,3,4$) and missing rates ($l=1,\ldots 6$).

The performances of the methods were demonstrated by validity and power analysis. The S-TDT, which ignores genotypes of the parents and compares frequencies of the affected and unaffected offspring [see 14 for the computation of S-TDT], was included to compare our methods with one of the widely used family based methods. Since S-TDT completely ignores parental genotypes and requires unaffected offspring genotypes from these families, and also assumes affected offspring genotypes are available, none of the missing mechanism models were taken into account. It means that the type I error rates for S-TDT are all the same whatever the missing mechanism models are for a given δ value.

In validity and power analysis tables, the TDT ignores missing cases and considers only complete cases, S-TDT ignores parental genotypes and considers only genotypes of affected and unaffected offspring of all 100 families (genotypes are all known), and MITDT and rTDT use all 100 families after construction of all possible admissible genotypes.

The most positive value of linkage disequilibrium is defined as $\delta_{\max}=\min\{(1-m)p,(1-p)m\}$ when $\delta>0$, and the most negative value of linkage disequilibrium is defined as $\delta_{\max}=\min\{\text{mp},(1-p)(1-m)\}$ when $\delta<0$. Since type I error rate and power results for $\delta_{\max}$ $(\frac{1}{2}\delta_{\max})$ when $\delta>0$ at m are equal to type I error rate and power results for $\delta_{\max}$ $(\frac{1}{2}\delta_{\max})$ when $\delta<0$ at $1-m$, we only consider the values of δ when $\delta>0$. In the presence of positive linkage disequilibrium ($\delta>0$), the null hypotheses are there is no linkage ($\theta=0.5$) in validity analysis, and there is a complete linkage ($\theta=0$) in power analysis. The values of δ were chosen as moderate $(\frac{1}{2}\delta_{\max})$ and maximum $(\delta_{\max})$ with $m=0.25$ and $p=0.05$.

Table 7. Missing model (MM) and missing rates (MR).

MM/MR	Missing Rates		
	Father	Mother	Offspring
	$(P_{f11},P_{f12},P_{f22})$	$(P_{m11},P_{m12},P_{m22})$	$(P_{o11},P_{o12},P_{o22})$
1/1	(0.10,0.10,0.10)	(0.10,0.10,0.10)	(0.10,0.10,0.10)
2/1	(0.05, 0.05, 0.10)	(0.05, 0.05, 0.10)	(0.10,0.10,0.10)
2/2	(0.05, 0.075, 0.10)	(0.05, 0.075, 0.10)	(0.10,0.10,0.10)
2/3	(0.05, 0.10, 0.10)	(0.05, 0.10, 0.10)	(0.10,0.10,0.10)
2/4	(0.10, 0.05, 0.05)	(0.10, 0.05, 0.05)	(0.10,0.10,0.10)
2/5	(0.10, 0.075, 0.05)	(0.10, 0.075, 0.05)	(0.10,0.10,0.10)
2/6	(0.10, 0.10, 0.05)	(0.10, 0.10, 0.05)	(0.10,0.10,0.10)
3/1	(0.05, 0.05, 0.10)	(0.05, 0.05, 0.10)	(0.05, 0.05, 0.10)
3/2	(0.05, 0.075, 0.10)	(0.05, 0.075, 0.10)	(0.05, 0.075, 0.10)
3/3	(0.05, 0.10, 0.10)	(0.05, 0.10, 0.10)	(0.05, 0.10, 0.10)
3/4	(0.10, 0.05, 0.05)	(0.10, 0.05, 0.05)	(0.10, 0.05, 0.05)
3/5	(0.10, 0.075, 0.05)	(0.10, 0.075, 0.05)	(0.10, 0.075, 0.05)
3/6	(0.10, 0.10, 0.05)	(0.10, 0.10, 0.05)	(0.10, 0.10,0.05)
4/1	(0.10, 0.10, 0.10)	(0.10, 0.10, 0.10)	(0.05, 0.05, 0.10)
4/2	(0.10, 0.10, 0.10)	(0.10, 0.10, 0.10)	(0.05, 0.075, 0.10)
4/3	(0.10, 0.10, 0.10)	(0.10, 0.10, 0.10)	(0.05, 0.10, 0.10)
4/4	((0.10, 0.10, 0.10)	(0.10, 0.10, 0.10)	(0.10, 0.05, 0.05)
4/5	(0.10, 0.10, 0.10)	(0.10, 0.10, 0.10)	(0.10, 0.075, 0.05)
4/6	(0.10, 0.10, 0.10)	(0.10, 0.10, 0.10)	(0.10, 0.10,0.05)

Missing models:(1) Missing Completely at Random (MCAR) for all genotypes, (2) informative missing for parental genotypes and MCAR for offspring genotypes, (3) informative missing for all genotypes, and (4) MCAR for parental genotypes and informative missing for offspring genotypes. P_{i11},P_{i12}, and P_{i22} $(i=f,m,o)$ denote missing rates for father (f), mother (m), and offspring (o) with (1,1), (1,2), and (2,2) genotypes, respectively.

doi:10.1371/journal.pone.0046100.t007

Validity Analysis

When $\theta=0.5$, the probability that an informative parent transmits marker allele 1 to a particular affected child (P_t) becomes 0.5 because L_t is zero in (8). That is, the value of δ in M_t and the disease model in R_t are not involved in validity analysis. It means that type I error rates are the same for every disease model.

All testing procedures (TDT, MITDT-ONE, rTDT) except S-TDT were valid tests at 1% and 5% significance levels (Tables 8 and 9). Since TDT, MITDT-ONE and rTDT takes also information about genotypes of parents into account as opposed to S-TDT, this information had a positive impact on the sizes of the tests. Since S-TDT had inflated type I errors, we excluded its performance in power analysis. Overall, MITDT-ONE outperformed rTDT by providing type I error rates close to the corresponding significance levels. The rTDT was the conservative test. Actually, this was the main reason for us to propose a new test that controls type I error rates better. The results in Tables 8 and 9 show that the MITDT-ONE achieved this goal. Since MITDT-ONE (and rTDT) does not assume any specific missing models, we suggest that MITDT-ONE should be preferred over some widely used family based testing procedures.

Power Analysis

In power analysis, the null hypothesis is that there is a complete linkage ($\theta=0$). When $\theta=0$, the probability of an informative parent transmitting marker allele 1 to a particular affected child (P_t) becomes greater than or equal to 0.5 because L_t,M_t, and R_t contribute to the value of P_t. It means information from linkage disequilibrium and α, β and γ (parameters of disease model) have positive effect on power. This theoretical fact was also observed through simulation studies in Tables 10, 11, 12, 13, 14, 15. The pattern of power for all disease models, missing rates, missing models, and strength of linkage disequilibrium were the same for different significance levels (1% and 5%). However, the power values were better at 5% significance level than at 1% significance level.

Table 8. Type I error rates at 1% significance level under the null hypothesis of $\theta=0.5$.

δ	GM	MM/MR	S-TDT	TDT	rTDT	MI-TDT
$\frac{1}{2}\delta_{\max}$	D, A, R	1/1	0.024	0.009	0	0.007
		2/1	0.024	0.009	0	0.007
		2/2		0.01	0	0.007
		2/3		0.009	0	0.007
		2/4		0.009	0	0.006
		2/5		0.009	0	0.007
		2/6		0.009	0	0.007
		3/1	0.024	0.01	0	0.007
		3/2		0.011	0	0.007
		3/3		0.01	0	0.007
		3/4		0.009	0	0.007
		3/5		0.01	0	0.008
		3/6		0.01	0	0.009
		4/1	0.024	0.01	0	0.007
		4/2		0.01	0	0.007
		4/3		0.01	0	0.007
		4/4		0.01	0	0.009
		4/5		0.01	0	0.009
		4/6		0.01	0	0.009
$\delta_{\max}$	D, A, R	1/1	0.022	0.007	0	0.006
		2/1	0.022	0.007	0	0.004
		2/2		0.007	0	0.005
		2/3		0.007	0	0.006
		2/4		0.008	0	0.004
		2/5		0.008	0	0.004
		2/6		0.008	0	0.005
		3/1	0.022	0.009	0	0.005
		3/2		0.009	0	0.005
		3/3		0.008	0	0.006
		3/4		0.008	0	0.005
		3/5		0.009	0	0.006
		3/6		0.008	0	0.007
		4/1	0.022	0.008	0	0.006
		4/2		0.008	0	0.006
		4/3		0.008	0	0.006
		4/4		0.008	0	0.007
		4/5		0.008	0	0.007
		4/6		0.008	0	0.007

doi:10.1371/journal.pone.0046100.t008

Table 9. Type I error rates at 5% significance level under the null hypothesis of $\theta = 0.5$.

δ	GM	MM/MR	S-TDT	TDT	rTDT	MI-TDT
$\frac{1}{2}\delta_{\max}$	D, A, R	1/1	0.104	0.05	0	0.04
		2/1	0.104	0.048	0	0.036
		2/2		0.048	0	0.037
		2/3		0.05	0	0.04
		2/4		0.047	0.001	0.034
		2/5		0.047	0.001	0.034
		2/6		0.05	0.001	0.037
		3/1	0.104	0.052	0.001	0.035
		3/2		0.052	0	0.036
		3/3		0.052	0	0.041
		3/4		0.048	0.003	0.036
		3/5		0.053	0.002	0.04
		3/6		0.053	0.001	0.042
		4/1	0.104	0.053	0.001	0.039
		4/2		0.052	0	0.04
		4/3		0.052	0	0.041
		4/4		0.053	0.001	0.042
		4/5		0.053	0	0.043
		4/6		0.053	0	0.045
$\delta_{\max}$	D, A, R	1/1	0.102	0.045	0	0.036
		2/1	0.102	0.043	0	0.031
		2/2		0.043	0	0.034
		2/3		0.045	0	0.036
		2/4		0.042	0.001	0.028
		2/5		0.042	0.001	0.03
		2/6		0.045	0.001	0.034
		3/1	0.102	0.045	0	0.032
		3/2		0.046	0	0.034
		3/3		0.047	0	0.037
		3/4		0.041	0.003	0.033
		3/5		0.047	0.002	0.037
		3/6		0.045	0.001	0.038
		4/1	0.102	0.046	0	0.036
		4/2		0.046	0	0.037
		4/3		0.047	0	0.037
		4/4		0.045	0	0.038
		4/5		0.046	0	0.039
		4/6		0.047	0	0.039

In column 2, D, A, and R represent dominant, additive, and recessive genetic models (GM), respectively.

doi:10.1371/journal.pone.0046100.t009

Table 10. Power values at 1% significance level when alternative hypothesis is $\theta = 0$.

δ	GM	MM/MR	S-TDT	TDT	rTDT	MI-TDT
$\frac{1}{2}\delta_{\max}$	Dominant	1/1	0.291	0.769	0.072	0.829
		2/1	0.291	0.782	0.102	0.833
		2/2		0.785	0.098	0.835
		2/3		0.769	0.072	0.829
		2/4		0.78	0.18	0.815
		2/5		0.782	0.175	0.819
		2/6		0.767	0.145	0.823
		3/1	0.291	0.764	0.136	0.815
		3/2		0.776	0.111	0.826
		3/3		0.769	0.072	0.829
		3/4		0.795	0.449	0.83
		3/5		0.812	0.385	0.838
		3/6		0.778	0.294	0.833
		4/1	0.291	0.761	0.094	0.821
		4/2		0.762	0.08	0.823
		4/3		0.769	0.072	0.829
		4/4		0.769	0.298	0.826
		4/5		0.77	0.266	0.829
		4/6		0.777	0.244	0.835
	Additive	1/1	0.257	0.72	0.053	0.788
		2/1	0.257	0.729	0.077	0.792
		2/2		0.735	0.074	0.795
		2/3		0.72	0.053	0.788
		2/4		0.724	0.143	0.772
		2/5		0.731	0.139	0.774
		2/6		0.716	0.113	0.781
		3/1	0.257	0.714	0.105	0.77
		3/2		0.726	0.083	0.786
		3/3		0.72	0.053	0.788
		3/4		0.751	0.383	0.788
		3/5		0.767	0.327	0.795
		3/6		0.731	0.248	0.793
		4/1	0.257	0.713	0.068	0.778
		4/2		0.715	0.059	0.781
		4/3		0.72	0.053	0.788
		4/4		0.721	0.253	0.785
		4/5		0.724	0.218	0.788
		4/6		0.729	0.197	0.795

doi:10.1371/journal.pone.0046100.t010

When the linkage disequilibrium was at its moderate level ($\frac{1}{2}\delta_{\max}$), dominant models had the highest power following by additive and recessive models. While the power of MITDT-ONE ranged between 0.73 (0.94) and 0.84 (0.89), the power of rTDT ranged between 0.042 (0.17) and 0.45 (0.68) when $\alpha = 1\%$ (5%). When linkage disequilibrium was at its maximum ($\delta_{\max}$), all testing procedures lacked power because the value of P_t in (8) was close to 0.5 (this value was exactly 0.5 in validity analysis). When $\delta = \delta_{\max}$, recessive models had the highest power, following by additive and dominant models, which was a reserve observation for $\delta_{\max}$. Over all, MITDT-ONE was the only method that provided the highest power at any significance level.

Real Data: U.K. Warren Family

We illustrate the robustness of the MITDT-ONE for type 1 diabetes at insulin dependent diabetes mellitus 2 locus (IDDM2) on chromosome 11p15. At our request, Neil Walker of the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes

Table 11. Power analysis continues.

δ	GM	MM/MR	S-TDT	TDT	rTDT	MI-TDT
$\frac{1}{2}\delta_{\max}$	Recessive	1/1	0.231	0.68	0.042	0.756
		2/1	0.231	0.688	0.064	0.755
		2/2		0.693	0.061	0.761
		2/2		0.68	0.042	0.756
		2/2		0.686	0.119	0.734
		2/2		0.69	0.117	0.737
		2/2		0.676	0.097	0.746
		3/1	0.231	0.671	0.089	0.731
		3/2		0.682	0.068	0.748
		3/3		0.68	0.042	0.756
		¾		0.714	0.342	0.753
		3/5		0.728	0.287	0.76
		3/6		0.693	0.207	0.761
		4/1	0.231	0.673	0.055	0.744
		4/2		0.676	0.048	0.747
		4/3		0.68	0.042	0.756
		4/4		0.683	0.218	0.753
		4/5		0.686	0.186	0.756
		4/6		0.69	0.167	0.764
$\delta_{\max}$	Dominant	1/1	0.021	0.009	0	0.007
		2/1	0.021	0.009	0	0.006
		2/2		0.009	0	0.006
		2/3		0.009	0	0.007
		2/4		0.009	0	0.005
		2/5		0.009	0	0.005
		2/6		0.009	0	0.007
		3/1	0.021	0.01	0	0.007
		3/2		0.01	0	0.007
		3/3		0.009	0	0.007
		¾		0.009	0	0.007
		3/5		0.01	0	0.007
		3/6		0.009	0	0.007
		4/1	0.021	0.009	0	0.007
		4/2		0.009	0	0.007
		4/3		0.009	0	0.007
		4/4		0.01	0	0.008
		4/5		0.009	0	0.008
		4/6		0.009	0	0.008

doi:10.1371/journal.pone.0046100.t011

Table 12. Power analysis continues.

δ	GM	MM/MR	S-TDT	TDT	rTDT	MI-TDT
$\delta_{\max}$	Additive	1/1	0.015	0.012	0	0.016
		2/1	0.015	0.009	0	0.014
		2/2		0.01	0	0.014
		2/3		0.012	0	0.016
		2/4		0.01	0	0.011
		2/5		0.01	0	0.011
		2/6		0.012	0	0.014
		3/1	0.015	0.009	0	0.012
		3/2		0.01	0	0.013
		3/3		0.012	0	0.016
		¾		0.011	0.001	0.014
		3/5		0.013	0	0.015
		3/6		0.013	0	0.016
		4/1	0.015	0.012	0	0.015
		4/2		0.012	0	0.015
		4/3		0.012	0	0.016
		4/4		0.012	0	0.017
		4/5		0.012	0	0.018
		4/6		0.013	0	0.018
	Recessive	1/1	0.014	0.014	0	0.021
		2/1	0.014	0.012	0	0.017
		2/2		0.013	0	0.019
		2/3		0.014	0	0.021
		2/4		0.013	0	0.015
		2/5		0.013	0	0.015
		2/6		0.013	0	0.018
		3/1	0.014	0.012	0	0.014
		3/2		0.013	0	0.017
		3/3		0.014	0	0.021
		¾		0.013	0.001	0.017
		3/5		0.016	0.001	0.019
		3/6		0.015	0	0.021
		4/1	0.014	0.012	0	0.02
		4/2		0.012	0	0.02
		4/3		0.014	0	0.021
		4/4		0.014	0	0.022
		4/5		0.014	0	0.023
		4/6		0.015	0	0.023

doi:10.1371/journal.pone.0046100.t012

and Inflammation Laboratory (JDRF/WT DIL) compiled data from 475 families with *two* affected offspring from the U.K. Warren Families for 52 SNPs. This data set was analyzed by [17] to demonstrate the method of MI-TDT for two affected offspring. The author of [21] used extensive logistic regression studies on the same data set, and identified −23 *Hph*I, +1,140A/C, +1428 *Fok*I, and VNTR as significant SNPs. The same SNPs as in [21] and six more were also identified by [17].

We considered the same U.K. Warren Families but chose the first affected child from each family to have only *one* affected offspring to demonstrate the performance of MITDT-ONE and rTDT. For the MITDT-ONE, we need to know frequencies of marker allele 1 (m) and disease allele (p) for each SNP. The values of m were provided to us along with the data set, except two (VNTR (DIL967) and TH micro' Z (DIL950)), but not the values of p. McGinnis (1998) [18] showed that TDT was able to detect linkage and its power exceeded 0.5 only when δ was close to $\delta_{\max}$ and allele frequencies m and p were similar in magnitude at the marker and disease locus. Therefore, we chose optimal values for p by assuming $m=p$.

The percentage of missing genotypes ranged from low (4% for DIL977) to high (52% for DIL997). Table 16 reports 18 significant

Table 13. Power values at 5% significance level when alternative hypothesis is $\theta = 0$.

δ	GM	MM/MR	S-TDT	TDT	rTDT	MI-TDT
$\frac{1}{2}\delta_{\max}$	Dominant	1/1	0.571	0.911	0.249	0.941
		2/1	0.571	0.919	0.295	0.942
		2/2		0.919	0.287	0.943
		2/3		0.911	0.249	0.941
		2/4		0.919	0.484	0.938
		2/5		0.92	0.469	0.938
		2/6		0.913	0.402	0.937
		3/1	0.571	0.912	0.366	0.935
		3/2		0.916	0.314	0.939
		3/3		0.911	0.249	0.941
		¾		0.928	0.734	0.941
		3/5		0.933	0.67	0.944
		3/6		0.916	0.553	0.941
		4/1	0.571	0.909	0.309	0.936
		4/2		0.911	0.274	0.937
		4/3		0.911	0.249	0.941
		4/4		0.911	0.57	0.939
		4/5		0.913	0.535	0.94
		4/6		0.914	0.507	0.944
	Additive	1/1	0.524	0.885	0.198	0.92
		2/1	0.524	0.894	0.243	0.92
		2/2		0.893	0.234	0.922
		2/3		0.885	0.198	0.92
		2/4		0.894	0.421	0.914
		2/5		0.894	0.407	0.916
		2/6		0.885	0.343	0.914
		3/1	0.524	0.883	0.313	0.91
		3/2		0.89	0.262	0.916
		3/3		0.885	0.198	0.92
		3/4		0.904	0.681	0.918
		3/5		0.909	0.608	0.924
		3/6		0.89	0.489	0.918
		4/1	0.524	0.881	0.254	0.914
		4/2		0.883	0.221	0.916
		4/3		0.885	0.198	0.92
		4/4		0.885	0.504	0.917
		4/5		0.886	0.468	0.919
		4/6		0.888	0.44	0.923

doi:10.1371/journal.pone.0046100.t013

Table 14. Power analysis continues.

δ	GM	MM/MR	S-TDT	TDT	rTDT	MI-TDT
$\frac{1}{2}\delta_{\max}$	Recessive	1/1	0.498	0.862	0.168	0.905
		2/1	0.498	0.873	0.205	0.907
		2/2		0.873	0.198	0.909
		2/3		0.862	0.168	0.905
		2/4		0.873	0.378	0.899
		2/5		0.872	0.365	0.9
		2/6		0.863	0.302	0.898
		3/1	0.498	0.86	0.272	0.895
		3/2		0.87	0.221	0.901
		3/3		0.862	0.168	0.905
		¾		0.885	0.633	0.902
		3/5		0.891	0.562	0.908
		3/6		0.869	0.448	0.904
		4/1	0.498	0.858	0.22	0.899
		4/2		0.86	0.19	0.901
		4/2		0.862	0.168	0.905
		4/2		0.863	0.459	0.902
		4/2		0.865	0.425	0.904
		4/2		0.866	0.395	0.909
$\delta_{\max}$	Dominant	1/1	0.104	0.04	0	0.038
		2/1	0.104	0.043	0	0.034
		2/2		0.042	0	0.034
		2/3		0.04	0	0.038
		2/4		0.042	0	0.032
		2/5		0.042	0	0.03
		2/6		0.04	0	0.037
		3/1	0.104	0.045	0	0.032
		3/2		0.046	0	0.033
		3/3		0.042	0	0.039
		¾		0.041	0.004	0.033
		3/5		0.044	0.002	0.037
		3/6		0.042	0.001	0.039
		4/1	0.104	0.042	0	0.039
		4/2		0.043	0	0.039
		4/3		0.042	0	0.039
		4/4		0.042	0.001	0.04
		4/5		0.043	0.001	0.039
		4/6		0.042	0	0.041

doi:10.1371/journal.pone.0046100.t014

SNPs out of 52 at 5% significance level for complete genotypes. Since we tested 52 SNPs, we applied Bonferroni multiple testing procedure at 0.05% significance level or 99.95% confidence level, and identified seven significant SNPs (underlined p-values). Since percentage of missing genotypes ranged from small to high, one should be cautious to declare significant SNPs when missing genotypes are ignored. Since DIL950 was insignificant for complete data, we dropped it from the computation of MITDT-ONE and rTDT. DIL967 was significant for complete data but its marker allele were not provided to us. Since we did not have any knowledge about the value of m, and did not want to assign any preferential value, we considered equal frequencies for $m = 0.5$ and $1 - m = 0.5$.

The MITDT-ONE and rTDT could verify if the significant SNPs for complete data are also significant when missing genotypes are taken into account. However, if either method could not reach significant result as in complete case, it does not mean that these SNPs are insignificant. It simply means that both methods reach an inconclusive decision. Moreover, the number of significant SNPs could be smaller when either test is employed, compared to the number of significant SNPs for complete data.

Table 15. Power analysis continues.

δ	GM	MM/MR	S-TDT	TDT	rTDT	MI-TDT
δ_{max}	Additive	1/1	0.08	0.055	0	0.07
		2/1	0.08	0.054	0	0.061
		2/2		0.056	0	0.065
		2/3		0.055	0	0.07
		2/4		0.053	0.002	0.055
		2/5		0.055	0.002	0.058
		2/6		0.055	0.001	0.062
		3/1	0.08	0.052	0	0.055
		3/2		0.056	0	0.06
		3/3		0.055	0	0.07
		¾		0.056	0.007	0.059
		3/5		0.065	0.005	0.069
		3/6		0.059	0.003	0.068
		4/1	0.08	0.055	0	0.065
		4/2		0.055	0	0.066
		4/3		0.055	0	0.07
		4/4		0.057	0.003	0.07
		4/5		0.058	0.003	0.071
		4/6		0.058	0.002	0.075
	Recessive	1/1	0.077	0.066	0	0.084
		2/1	0.077	0.064	0	0.078
		2/2		0.066	0	0.081
		2/3		0.066	0	0.084
		2/4		0.063	0.002	0.07
		2/5		0.065	0.002	0.072
		2/6		0.065	0.001	0.075
		3/1	0.077	0.061	0.001	0.068
		3/2		0.065	0	0.075
		3/3		0.066	0	0.083
		¾		0.067	0.008	0.074
		3/5		0.078	0.006	0.083
		3/6		0.07	0.004	0.083
		4/1	0.077	0.064	0.001	0.077
		4/2		0.065	0	0.079
		4/2		0.066	0	0.083
		4/2		0.067	0.004	0.084
		4/2		0.068	0.003	0.086
		4/2		0.07	0.002	0.09

doi:10.1371/journal.pone.0046100.t015

Out of 18 significant SNPs in complete cases, MITDT-ONE (rTDT) verified seven (three) to be significant (Table 17). The MITDT-ONE as well as rTDT found 23 *Hph*I, +1428 *Fok*I, and VNTR as significant SNPs as in [21] and [17]. Furthermore, MITDT-ONE identified four more same SNPs in [17] as significant; hence, we suggest researchers to investigate these SNPs as possible casual variant genes.

Discussion

Sebastiani et al. [16] proposed to handle missing genotypes of parents or offspring in a nuclear family with one affected offspring.

Table 16. Type I Diabetes (IDDM): The significant SNPs for complete data.

SNP	Variant	%	χ_c^2	$P(\chi^2 > \chi_c^2)$
DIL997	C/T	52	4	0.0455003
DIL996	C/T	28	4.2631579	0.0389475
DIL989	C/T	26	4.7407407	0.0294564
DIL985	C/T	42	6.1084337	0.0134538
DIL984	G/A	22	3.8571429	0.0495346
DIL977	G/A	4	17.386831	0.0000305
DIL976	T/G	36	10.940828	0.0009407
DIL975	C/T	30	11.571429	0.0006697
DIL974	A/C	30	16.568966	0.0000469
DIL973	T/C	16	14.069767	0.0001762
DIL971	G/C	20	10.971429	0.0009253
DIL969	A/T	6	23.027397	$15.97\text{x}10^{-5}$
DIL967	VNTR	6	21.300341	$3.93\text{x}10^{-6}$
DIL965	T/C	20	14.901478	0.0001133
DIL963	A/C	22	10.414286	0.0012504
DIL954	C/T	36	6.2857143	0.0121715
DIL3872	C/G	18	7.4745763	0.0062576
DIL2048	C/T	12	3.7815126	0.0518218

The third, fourth, and fifth columns show the percentages of missing data, the TDT statistics for complete data, and uncorrected p-values at 5% significance level. The significance SNPs are shown by underlined p-values for Bonferroni at 0.05% significance level.

doi:10.1371/journal.pone.0046100.t016

However, rTDT produces a conservative test and lacks power. Hence, we proposed MITDT-ONE to correct the problems of rTDT. The MITDT-ONE takes population frequencies of marker allele m and disease allele p into account in the rTDT method. With these m and p values, we restrict the domain of rTDT to have much better estimates for the maximum values of u and v.

The minimum values of the interval estimates of MITDT-ONE and rTDT make a decision against the null hypothesis of no linkage. One of the advantages of using MITDT-ONE is that significance results achieved by complete data is ratified when the minimum value of the interval estimate is smaller than the value of TDT for complete data. The other advantage of our method is that it allows researchers to implement our method to any missing rates. As discussed in the introduction, many studies deal with missing genotypes in parents but not in offspring. Moreover, these methods assume some missing mechanism (e.g., MAR) to recover parental genotypes. Thus, another strength of MITDT-ONE is that it does not assume any missing model but simply considers the Mendelian Inheritance property to define all possible admissible genotypes in parents or offspring. Also, MITDT-ONE and rTDT become classical TDT when $u_{inc} = v_{inc} = 0$.

In the construction of MITDT-ONE, we consider cases where all genotypes of family members are missing (Case 1). It is intuitive that since these families do not have any information they should be ignored from the study. We suggest that these families be omitted from the data if only one SNP is studied. However, if more than one SNP are studied then we suggest keeping them in the computation of MITDT-ONE to have same number of families for each SNP.

Table 17. The Significant SNPs at 5% significance level when the MITDT-ONE is applied.

SNP	Variant	Name	dbSNP	χ_c^2	$\chi_{\min}^2$	$P(\chi^2 > \chi_c^2)$	$P(\chi^2 > \chi_{\min}^2)$
DIL977*	G/A	+1,428 *Fok*I	rs3842756	17.39	16.82	0.0000305	0.0000412
DIL973	T/C	+1,127 *Pst*I	rs3842752	14.07	8.00	0.0001762	0.0046696
DIL971	G/C	+805 *Dra*III	rs3842748	10.98	6.13	0.0009253	0.0133311
DIL969*	A/T	−23 *Hph*I	rs689	23.03	21.44	15.97×10^{-5}	36.48×10^{-5}
DIL967*	VNTR	VNTR	-	21.30	18.27	3.93×10^{-6}	0.0000192
DIL965	T/C	−2,221 *Msp*I	rs3842729	14.90	7.97	0.0001133	0.0047659
DIL963	A/C	−2,733A/C	rs3842727	10.41	4.68	0.0012504	0.0306104

The third and fourth columns show the name of the SNP defined in Barratt et al. (2004) and the SNP database, respectively. The fifth and sixth columns show the statistics for complete and incomplete data when MITDT-ONE is applied, respectively. The seventh and eight columns show the type I errors of the columns fifth and sixth, respectively.

*are SNPs found in association by using rTDT.

doi:10.1371/journal.pone.0046100.t017

In summary, simulation studies show that MITDT-ONE controls type I error rates very well and produces high power when degree of linkage disequilibrium is mild.

More than one offspring: rTDT for two affected offspring was proposed by [17]. However, it was a conservative test and had low power. Hence, Alpargu [17] proposed MI-TDT to remedy the problems. With the motivation of Alpargu [17], we proposed MITDT-ONE. Both MITDT-ONE and MI-TDT correct the problems arising from rTDT. Theoretically, it is possible to propose our method for families with at least three and more affected offspring. However, the computation will be tedious because the number of missing cases increases as the number of affected offspring increases. Moreover, in the linkage studies it is very rare to have more than two affected offspring.

Multiple alleles: We proposed MITDT-ONE for bi-allelic cases. However, it is possible to extend to multi-allelic cases. We consider two approaches that have been used in practice [22,23]. In the first approach, all alleles except the allele of interest are grouped as allele 2, and the MITDT-ONE for bi-allelic case is applied [22]. In the second approach, if we have q alleles, then for each allele, the first approach is applied to obtain q MITDT-ONE statistics, then the largest MITDT-ONE is chosen as the test statistic [23] to make a decision about significant gene.

Acknowledgments

We thank the members of the DNA resource team and Neil Walker of Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory (JDRF/WT DIL) for sample and data services (http://www-gene.cimr.cam.ac.uk/todd)). The author thanks the two referees for their valuable comments that helped improved the quality of the article.

Author Contributions

Analyzed the data: GB. Wrote the paper: GB.

References

1. 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: 506516.
2. Spielman RS, Ewens WJ (1996) The tdt and other family-based tests for linkage disequi-librium and association. Am J Hum Genet 59: 983989.
3. Clayton D (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65: 11701177.
4. Knapp M (1999) The transmission/disequilibrium test and parental-genotype reconstruc-tion: the reconstruction-combined transmission/disequilibrium test. Am J Hum Genet 64: 861870.
5. Knapp M (1999) A note on power approximations for the transmission/disequilibrium test. Am J Hum Genet 64: 11771185.
6. Weinberg CR (1999) Allowing for missing parents in genetic studies of case-parent triads. Am J Hum Genet 64: 11861193.
7. Cervino ACL, Hill AVS (2000) Comparison of tests for association and linkage in incom-plete families. Am J Hum Genet 67: 120–132.
8. Little RJA, Rubin DB (2002) Statistical Analysis With Missing Data. Chichester: John Wiley.
9. Allen AS, Rathouz PJ, Satten GA (2003) Informative massiveness in genetic association studies: case-parent designs. Am J Hum Genet 72: 671680.
10. Chen YH (2004) New approach to association testing in case-parent designs under infor-mative parental missingness. Genetic Epidemiology 27: 131–140.
11. Boenhnke M, Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62: 950–961.
12. Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62: 450458.
13. Horvath S, Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 63: 18861897.
14. Monks SA, Kaplan NL, Weir BS (1998) A comparative study of sibship tests of linkage and/or association. Am J Hum Genet 63: 1507–1516.
15. Martin ER, Monks SA, Warren LL, Kaplan NL (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am J Hum Genet 67: 146–154.
16. Sebastiani P, Abad-Grau MM, Alpargu G, Ramoni M (2004) Robust transmission/disequilibrium test for incomplete family genotypes. Genetics 168(4): 2329–2337.
17. Alpargu G (2011) Allowing for missing genotypes in any members of the nuclear families in transmission disequilibrium test. Computational Statistics and Data Analysis 55: 1236–1249.
18. McGinnis RE (1998) Hidden linkage: a comparison of the affected sib pair (asp) test and transmission/disequilibrium test (tdt). Ann Hum Genet 62: 159179.
19. Ott J (1989) Statistical properties of the haplotype relative risk. Genet Epidemiol 6: 127130.
20. Abecasis GR, Cookson WO, Cardon LR (2000) Pedigree tests of transmission disequilib-rium. Euro J Huts Genet 8: 545–551.
21. Barratt BJ, Payne F, Lowe CE, Hermann R, Healy BC, et al. (2004) Remapping the insulin gene/iddm2 locus in type 1 diabetes. Diabetes 53(7): 1884–1889.
22. Schaid DJ (1996) General score tests for associations of genetic markers with diease using cases and their parents. Genet Epidemiol 13: 423–449.
23. Ewens WJ, Spielman RS (1997) Disease associations and the transmission/disequilibrium test. In Dracopoli NC (ed) Current protocols in human genetics. Supl 15, pp. 1.12.1–1.12.13. New York: Wiley.