Journal of Human Genetics (2012) 57, 601–609

www.nature.com/jhg

ORIGINAL ARTICLE

A novel test of informative missingness using inconsistent linkage disequilibrium signals between case-parent triads and incomplete data

Chao-Yu Guo[1,2,3]

In general, multiple issues are examined before the analysis of genetic data such as Hardy–Weinberg Equilibrium and Mendelian errors. Although missing genotypes are commonly observed in genetic studies, potential bias due to informative missingness is usually overlooked. Therefore, the Test of Informative Missingness (TIM) was the first attempt to determine whether or not parental genotypes are missing informatively. The TIM is a useful tool for genetic data cleaning. For example, excluding single-nucleotide polymorphisms that appear to be missing informatively may further improve the quality of genetic data. Although the TIM has decent power, its performance is discernibly weaker when the minor allele/genotype introduces informative missingness. In an effort to avoid such reduced power, the newly proposed strategy detects informative missingness by comparing inconsistent linkage disequilibrium signals between intact case-parent triads and incomplete data. Computer simulations revealed that the new method was robust to population stratifications and more powerful than the TIM in most situations. In addition, the new method demonstrated decent power in the genome-wide association study, even if the most conservative correction for multiple testing was adopted.
Journal of Human Genetics (2012) **57,** 601–609; doi:10.1038/jhg.2012.78; published online 28 June 2012

Keywords: case-parent triads; GWAS; HRR; informative missingness; TDT

INTRODUCTION

Spurious associations due to population admixture could be a serious issue in genetic studies using unrelated subjects. To avoid false signals, the family-based approach, haplotype relative risk (HRR),[1] utilizes case-parent triads to detect linkage disequilibrium (LD) between a marker and a putative disease locus by comparing parental marker alleles transmitted to an affected offspring to those non-transmitted. Instead of treating transmitted and non-transmitted alleles as unrelated, the Transmission/Disequilibrium Test (TDT)[2] considered case-parent triads as matched data and examined whether or not heterozygous parents preferentially transmitted the specific allele to the affected offspring. The TDT is more powerful than the HRR, especially when population admixture is present. Therefore, the TDT is a popular study design for early onset diseases.

The greatest challenge in recruiting case-parent triads is that one or both parental genotypes may be unavailable due to declined participation, death, or other unexpected reasons. In the statistical analysis, both missing completely at random (MCAR) and missing at random (MAR) are ignorable.[3] If the events that lead to any particular value being missing are independent of both observed and unobserved parameters of interest, then the missing pattern is considered as MCAR. Given the observed data, if the missing mechanism does not depend on the unobserved data, then the missing pattern is MAR. The scenarios of MCAR and MAR could be confusing in the settings of a genetic study. For a single nucleotide polymorphism (SNP) with alleles A and C, there are three genotypes AA, AC and CC. In an admixed population, if the missing rates of the three genotypes are identical in all sub-populations, then the missing pattern is MACR. If the missing rates of the three genotypes are identical within each subgroup, but the missing rates differ across subgroups, then the missing pattern appears to be MAR. Two distinct types of missingness in genotype data should be noted due to different mechanisms. The first situation is that individuals may be unavailable due to death or non-participation. Therefore, there can be different missing rates for the offspring and their parents. As a result, informative missingness could occur solely in the parents, but not the offspring. The second situation is that the genotyping assay may have failed to deliver a 'call' at a particular locus for a particular specimen, even though the person was participating. The scenario may depend on the true genotype (hence be informative), but may not differ across individuals. As a result, informative missingness would exist in both the offspring and parents.

[1]Division of Biostatistics, Institute of Public Health, National Yang Ming University, Taipei, Taiwan, ROC; [2]Head and Neck Cancer Research Program, Cancer Research Center, National Yang Ming University, Taipei, Taiwan, ROC and [3]Genome Research Center, National Yang Ming University, Taipei, Taiwan, ROC
Correspondence: Professor CY Guo, Division of Biostatistics, Institute of Public Health, National Yang Ming University, No. 155, Sec. 2, Linong Street, Medical Building 2, Taipei 112, Taiwan, ROC.
E-mail: cyguo@ym.edu.tw
Received 16 March 2012; revised 26 May 2012; accepted 28 May 2012; published online 28 June 2012

In 1995, the estimated probability of transmission of certain alleles[4] was pointed out to be biased in the TDT using dyads (the affected offspring with only one parent), where only heterozygous parents and homozygous offspring contributed to the test. The 1-TDT[5] was free from such bias when parental genotypes were MCAR or MAR. In addition to the 1-TDT, the family-based association test by Rabinowitz and Laird[6] as well as several other strategies had been proposed[7–11] to accommodate incomplete triads. When parental genotypes were missing informatively, Allen *et al.*[12] and Chen[13] carried out valid tests to incorporate incomplete data. However, the two methods experienced substantially reduced statistical power when the underlying missing pattern was truly MCAR/MAR as discussed by Guo *et al.*[14] Although various scenarios had been well studied, all methods[5–14] focused on the missing pattern of parental genotypes and assumed that the offspring genotypes were MCAR/MAR. When the assumption of MCAR/MAR was violated among offspring genotypes, Guo[15] indicated that the TDT using only complete triads may still inflate the type-I error and/or reduce power due to ascertainment bias. This phenomenon suggests that if the missing pattern of offspring genotypes is not determined, a significant result of the TDT may not assure a true association, even if incomplete triads are excluded from the analysis. Therefore, missing mechanism is an important issue in analyzing genetic data.

The first attempt to determine whether or not parental genotypes are missing informatively was introduced by Guo *et al.*,[14] the Test of Informative Missingness (TIM), which compared the distribution of parental genotypes in triads with that of dyads, conditional on the genotypes of affected offspring. Differential distributions of parental genotypes in triads and dyads indicated that the missing pattern of parental genotypes was not ignorable. The TIM is a valuable tool for genetic data cleaning. A novel application for the TIM was to exclude SNPs that are missing informatively in a genome-wide association study (GWAS). In this way, fewer yet more reliable SNPs will be analyzed and this procedure may effectively reduce the excessive amount of false positives in the analysis. In the era of GWAS, one million SNPs are considered as the standard. SNPs with missing rates exceeding a specific threshold are now routinely excluded, because inclusions of SNPs with higher missing rates lead to too many significant results, which are thought to be false positives. This strict enforcement of nearly complete data raises some important issues, since excessive rates of significant results in a typical GWAS may be potentially caused by informative missingness.

Although the TIM demonstrates decent power, its performance is discernibly weaker when the minor allele/genotype introduces informative missingness. Insights of such reduced power could be comprehended by the following example. Assuming that the minor (major) allele is A (C) and the corresponding allele frequency is 0.3 (0.7). The frequencies of genotypes AA, AC and CC are 0.09, 0.42 and 0.49, respectively. If 10% of the subjects with the genotype AA are missing, but only 1% of the individuals with the genotype AC or CC are not available, then the missing pattern is informative. In a random sample of 10 000 subjects, one would expect that 90, 42 and 49 subjects are missing genotypes AA, AC and CC, respectively. In contrast, if the excessive missingness (10%) occurs in the major genotype CC, but only 1% of the individuals with the genotype AA or AC are absent, then one would expect a much larger number of individuals with missing genotypes, which results in a stronger signal for informative missingness. Since the TIM is conditional of the offspring genotypes, the size of each of the three offspring genotypes has an important role when comparing the distribution of parental genotypes in triads with that of dyads. It is worth noting that the offspring with the genotype AA is the minor group and their contribution to the test statistic of the TIM is less weighted. Hence, power of the TIM is considerably reduced under such circumstances.

In this article, a new strategy, which is not conditional on the genotypes of affected offspring, is proposed to avoid the weakness of the TIM method. This novel method extends the expectation-maximization algorithm-based HRR (EM-HRR),[11] which utilized all types of ascertained data (triads, dyads and monads; note that monads are affected offspring without any parent). A previous study[15] had revealed that when parental genotypes were missing informatively, inconsistent LD signals were frequently observed between the EM-HRR that included incomplete data and the HRR that used only complete triads. Since the Breslow-Day test[16,17] was designed to test homogeneity of multiple odds ratios, it detects inconsistent estimates of odds ratios from the EM-HRR and HRR. Therefore, the new test of informative missingness is named as TIMBD.

MATERIALS AND METHODS

Following the previous work,[14] let $M_k^{i,j}$ represents the observed sample size for each type of triad data. $k=$ '0', '1' or '2' denotes the total number of B_1 alleles transmitted to the offspring, and i, $j=$ '0', '1' or '2' denotes the total number of B_1 alleles for the father and mother, respectively. For example, $M_1^{0,1}$ represents the total number of triads where genotypes of the offspring, father and mother are B_1B_2, B_2B_2 and B_1B_2, respectively. Note that the superscript '*' indicates that the parental genotype is missing. For example, $M_k^{*,j}$ represents dyads with the missing father.

Let T_{B1}^s, N_{B1}^s, T_{B2}^s and N_{B2}^s denote the total number of transmitted alleles for B_1, non-transmitted alleles for B_1, transmitted alleles for B_2 and non-transmitted alleles for B_2, respectively, where the superscript s indicates the family types ($s=1$ for complete triads, $s=2$ for dyads and $s=3$ for monads).

The HRR is only applicable for complete triads, where $\mathrm{HRR}=(T_{B1}^1\times N_{B2}^1)/(N_{B1}^1\times T_{B2}^1)$. Unlike the original EM-HRR that utilizes both complete and incomplete data, the EM-HRR statistic in this article only includes dyads and monads such that $\text{EM-HRR}=((T_{B1}^2+T_{B1}^3)\times N_{B2}^2)/(N_{B1}^2\times(T_{B2}^2+T_{B2}^3))$. The EM-HRR statistic uses the same proportions estimated from the original EM-HRR and detailed calculations of the HRR and EM-HRR are displayed in Table 1.

Let $N_{\mathrm{HRR}}=T_{B1}^1+N_{B1}^1+T_{B2}^1+N_{B2}^1$ denotes the total number of alleles obtained from complete data and $N_{\mathrm{EM-HRR}}=T_{B1}^2+T_{B1}^3+N_{B1}^2+T_{B2}^2$

Table 1 The HRR and EM-HRR statistics for the TIMBD

	Triads		*Dyads/Monads*	
Alleles	*Transmitted*	*Non-transmitted*	*Transmitted*	*Non-transmitted*
B_1	T_{B1}^1	N_{B1}^1	$T_{B1}^2+T_{B1}^3$	N_{B1}^2
B_2	T_{B2}^1	N_{B2}^1	$T_{B2}^2+T_{B2}^3$	N_{B2}^2

Abbreviations: EM-HRR, expectation-maximization algorithm-based HRR; HRR, haplotype relative risk.

Note: $\mathrm{HRR}=(T_{B1}^1\times N_{B2}^1)/(N_{B1}^1\times T_{B2}^1)$ and $\text{EM-HRR}=((T_{B1}^2+T_{B1}^3)\times N_{B2}^2)/(N_{B1}^2\times(T_{B2}^2+T_{B2}^3))$, where

$T_{B1}^1=2M_2^{2,2}+2M_2^{2,1}+2M_2^{1,2}+2M_2^{1,1}+M_1^{1,2}+M_1^{2,1}+M_1^{2,0}+M_1^{0,2}+M_1^{1,1}+M_1^{1,0}+M_1^{0,1}$;

$N_{B1}^1=2M_2^{2,2}+2M_2^{2,1}+2M_2^{1,2}+2M_0^{1,1}+M_2^{2,1}+M_2^{1,2}+M_1^{2,0}+M_1^{0,2}+M_1^{1,1}+M_0^{1,0}+M_0^{0,1}$;

$T_{B2}^1=2M_0^{0,0}+2M_0^{0,1}+2M_0^{1,0}+2M_0^{1,1}+M_1^{1,2}+M_1^{2,1}+M_1^{2,0}+M_1^{0,2}+M_1^{1,1}+M_1^{1,0}+M_1^{0,1}$;

$N_{B2}^1=2M_1^{1,1}+2M_1^{1,0}+2M_1^{0,1}+2M_0^{0,0}+M_2^{2,1}+M_2^{1,2}+M_2^{2,0}+M_2^{0,2}+M_1^{1,1}+M_1^{1,0}+M_0^{0,1}$;

$T_{B1}^2=2M_2^{2,*}+2M_2^{*,2}+2M_2^{1,*}+2M_2^{*,1}+M_1^{2,*}+M_1^{*,2}+M_1^{1,*}+M_1^{*,1}+M_1^{0,*}+M_1^{*,0}$;

$N_{B1}^2=M_2^{2,*}+M_2^{*,2}+M_1^{2,*}+M_1^{*,2}+M_0^{1,*}+M_0^{*,1}+\{M_1^{1,*}+M_1^{*,1}\}_2$;

$T_{B2}^2=2M_0^{0,*}+2M_0^{*,0}+2M_0^{1,*}+2M_0^{*,1}+M_1^{2,*}+M_1^{*,2}+M_1^{1,*}+M_1^{*,1}+M_1^{0,*}+M_1^{*,0}$;

$N_{B2}^2=M_2^{1,*}+M_2^{*,1}+M_1^{0,*}+M_1^{*,0}+M_0^{0,*}+M_0^{*,0}+\{M_1^{1,*}+M_1^{*,1}\}_1$;

$T_{B1}^3=2M_2^{*,*}+M_1^{*,*}$; $T_{B2}^3=2M_0^{*,*}+M_1^{*,*}$

Note: $\{M_1^{1,*}+M_1^{*,1}\}_1$ is the EM algorithm estimate of the proportion of heterozygous parents ($M_1^{1,*}+M_1^{*,1}$) who transmitted the B_1 allele but not the other B_2 allele. Similarly, heterozygous parents who transmitted the B_2 allele but not the other B_1 allele were estimated by $\{M_1^{1,*}+M_1^{*,1}\}_2$.

$+T^3_{B2}+N^2_{B2}$ denotes the total number of alleles derived from incomplete data. Assuming the absence of genetic heterogeneity, the proofs (see Appendix A for details) indicate that parental genotypes are MCAR/MAR if and only if E(HRR) = E(EM-HRR). As a result, the Breslow-Day test is implemented to detect the inequality of the HRR and EM-HRR. Here, the Mantel-Haenszel Odds Ratio is defined as:

$$\mathrm{OR}_{\mathrm{MH}}=\frac{T^1_{B1}\times N^1_{B2}/N_{\mathrm{HRR}}+(T^2_{B1}+T^3_{B1})\times N^2_{B2}/N_{\text{EM-HRR}}}{N^1_{B1}\times T^1_{B2}/N_{\mathrm{HRR}}+N^2_{B1}\times(T^2_{B2}+T^3_{B2})/N_{\text{EM-HRR}}}.$$

The TIMBD is computed as:

$$\mathrm{TIMBD}=\frac{(T^1_{B1}-E_1)^2}{V_1}+\frac{((T^2_{B1}+T^3_{B1})-E_2)^2}{V_2}$$

(see Appendix B for details). Since the Breslow-Day test is available in many statistical packages, the TIMBD is not computing intensive. Under the null hypothesis of MCAR/MAR, the TIMBD has an asymptotic χ^2 distribution with one degree of freedom.

It is worth noting that both the HRR and EM-HRR are robust to population stratifications, even if allele frequencies in the sub-populations are extremely different. Hence, the TIMBD, which is based on the HRR and EM-HRR, is also robust to population admixture and remains a valid test under MAR.

Simulations

To provide fair comparisons, similar simulation schemes of the TIM[14] were adopted. Considering an SNP, simulations begin with the assumption that the population is under the Hardy–Weinberg Equilibrium. Let '*a*' and '*A*' denote the disease allele and normal allele, respectively. '*D*' means that an individual is diseased or affected. Let '*f*' denotes the probability of being affected when an individual carries 0 risk alleles (the phenocopy rate), and let '*K*' denotes the genotype relative risk. For a recessive disease model, the penetrance functions are $P(D|AA)=P(D|Aa)=f$ and $P(D|aa)=K\times f$, where $0\leqslant f\leqslant 1$ and $0\leqslant K\times f\leqslant 1$. The disease prevalence is determined by these probabilities and the risk allele frequency. Similarly, for a dominant disease model, $P(D|AA)=f$ and $P(D|Aa)=P(D|aa)=K\times f$. In addition, the confined additive model was also created as $P(D|AA)=f$; $P(D|Aa)=\min(K\times f,1)$; $P(D|aa)=\min(2\times K\times f,1)$. The affection status of each individual was determined according to these parameters.

Several disease allele frequencies as well as marker allele frequencies were examined. A range of possible values for the disequilibrium coefficient δ and recombination fraction θ were simulated. The frequencies of the disease and marker alleles, the disease model, the phenocopy rate and the penetrance rate are indicated in each table. According to these parameters, a general population was simulated where nuclear families have exactly one offspring. Parental genotypes under the Hardy–Weinberg Equilibrium were first simulated. Then based on the Mendelian law, offspring genotypes were then generated for each household. After genotypes were simulated for every triad, the disease status of the offspring was determined by the offspring genotype, the disease penetrance rate and the phenocopy rate. The next step was to create the missing data, where the parental genotypes as well as the offspring genotypes were assigned to be absent according to various missing rates, which were clearly indicated in the tables. The last step was to randomly select probands (triads, dyads and monads) from the simulated population.

In the second set of simulations, population stratifications were considered. The previous scheme[14] was adopted and two populations were sampled under the Hardy–Weinberg Equilibrium with expected samples sizes reflecting different disease allele frequencies in the two populations. For example, for a pure recessive model, if the disease allele frequencies of the two populations are 0.3 and 0.6, respectively, then 9% of the first and 36% of the second population would be affected and sampled. Therefore, one would expect 20 and 80% of the sample to come from the first and second populations, respectively. This is the ratio that one would observe in most samples with admixture. Because the disease allele frequencies are different in the two populations, the frequencies of the diseased individuals in the two samples are also different. The disease allele frequencies, the marker allele frequencies, the phenocopy rates and the penetrance rates for the two populations were indicated in the tables.

The simulations were repeated 10 000 (1000) times to examine type-I error (power) of several tests examined including the TIMBD. In general, parents of the affected offspring are difficult to recruit. Therefore, the missing rates ranged from 1 to 40% in computer simulations. Examples were the missing rates derived from the Framing Heart Study,[11] where the missing rate for systolic blood pressure was as high as 91% (247/271).

In this article, power under the GWAS scenario was also examined. Assuming that one million SNPs were tested, a much large sample size of 5000 triads was considered. In addition, the most stringent correction for multiple testing was adopted. Therefore, *P*-values that were smaller than the Bonferonni's adjusted $\alpha(5\times10^{-8})$ could be declared significant. A total of 10 000 repetitions were done for the GWAS scenario.

In Tables 2–7, the column marked 'TDT' reports results using the traditional TDT test on the subset of complete triads only. The column marked '1-TDT' uses both the complete triads and dyads. The column marked 'TIM' is the test of informative missingness[14] and the last column 'TIMBD' represents the new strategy proposed in this article. Allen *et al.*[12] commented that the original 1-TDT should not be used. Thus, the modified 1-TDT was used, but not the original 1-TDT, in computer simulations.

Table 2 Type-I error (%) of the TIMBD in a homogeneous population assuming MCAR

Model	*Theta*	*Delta*	*Missing rates*	*TDT*	*1-TDT*	*TIM*	*TIMBD*
Dominant	0.5	0	(0.3; 0.3; 0.3)	4.9	5.0	3.9	3.6
	0.5	0	(0.3; 0.1; 0.2)	5.0	5.0	4.1	3.3
	0.5	0	(0.2; 0.4; 0.1)	5.3	5.4	4.1	3.7
C. additive	0.5	0	(0.3; 0.3; 0.3)	4.9	4.8	4.1	3.7
	0.5	0	(0.3; 0.1; 0.2)	4.8	4.8	4.6	3.7
	0.5	0	(0.2; 0.4; 0.1)	5.1	5.0	4.2	3.9
Recessive	0.5	0	(0.3; 0.3; 0.3)	4.9	5.3	4.3	3.7
	0.5	0	(0.3; 0.1; 0.2)	5.2	4.9	3.8	3.2
	0.5	0	(0.2; 0.4; 0.1)	4.9	5.0	4.4	3.7
Dominant	0	0.14	(0.3; 0.3; 0.3)	53.3	68.8	3.8	3.4
	0	0.14	(0.3; 0.1; 0.2)	94.2	98.0	3.9	2.8
	0	0.14	(0.2; 0.4; 0.1)	52.2	68.2	4.0	3.3
C. additive	0	0.14	(0.3; 0.3; 0.3)	92.9	98.4	4.2	3.7
	0	0.14	(0.3; 0.1; 0.2)	99.9	100.0	4.3	3.2
	0	0.14	(0.2; 0.4; 0.1)	91.8	98.2	3.8	3.6
Recessive	0	0.14	(0.3; 0.3; 0.3)	24.0	31.7	4.4	4.0
	0	0.14	(0.3; 0.1; 0.2)	71.6	81.4	4.0	4.5
	0	0.14	(0.2; 0.4; 0.1)	23.7	31.1	4.4	4.0

Abbreviations: MCAR, missing completely at random; TDT, Transmission/Disequilibrium Test; TIM, Test of Informative Missingness.
Note: (1) The first, second and third numbers in the parenthesis are the missing rates for the father, mother and offspring, respectively. (2) Sample size = 500 families; the missing rates for the three genotypes B_1B_1, B_1B_2 and B_2B_2 are identical for each individual. Disease allele frequency = 0.3; minor marker allele frequency = 0.4; penetrance rate = 0.4; phenocopy rate = 0.2. (3) 'C. additive' is the confined additive model.

RESULTS

Type-I error

When the missing pattern was MCAR for any member of the triads, type-I errors of the TIMBD in a homogeneous population are displayed in Table 2. The disease and marker allele frequencies were 0.3 and 0.4, respectively. The disease penetrance and phenocopy rate were 0.4 and 0.2, respectively. Different disease and marker allele frequencies, penetrance rates and phenocopy rates yielded similar results, which were not shown in the tables. The underlying disease model was indicated in the first column. The second and third columns were the recombination fraction (θ) and disequilibrium

Table 3 Type-I error (%) of the TIMBD under population admixture with a moderate marker allele difference assuming MCAR/MAR

Model	*Theta*	*Delta*	*Missing rates 1*	*Missing rates 2*	*TDT*	*1-TDT*	*TIM*	*TIMBD*
Dominant	0.5	0	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	5.0	5.0	3.9	3.5
	0.5	0	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	5.0	5.2	3.8	3.6
	0.5	0	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	5.6	5.0	4.4	3.7
C. additive	0.5	0	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	5.6	4.9	4.2	3.9
	0.5	0	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	5.0	4.9	4.2	3.6
	0.5	0	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	4.7	4.6	4.4	3.7
Recessive	0.5	0	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	5.3	5.1	4.2	3.8
	0.5	0	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	5.0	4.6	4.4	3.8
	0.5	0	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	5.2	4.9	4.7	3.5
Dominant	0	0.1	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	23.0	31.6	4.2	3.2
	0	0.1	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	28.3	34.4	4.4	3.5
	0	0.1	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	26.9	32.5	4.6	3.8
C. additive	0	0.1	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	61.5	76.7	4.0	3.8
	0	0.1	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	73.4	82.9	4.4	3.5
	0	0.1	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	66.5	79.9	4.5	3.6
Recessive	0	0.1	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	19.5	25.7	4.6	4.1
	0	0.1	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	23.4	28.4	4.4	3.9
	0	0.1	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	20.1	26.5	4.3	3.7

Abbreviations: MAR, missing at random; MCAR, missing completely at random; TDT, Transmission/Disequilibrium Test; TIM, Test of Informative Missingness.
Note: (1) 'Missing rates 1' and 'Missing rates 2' are the missing rates for the first and second populations, respectively. (2) The first, second and third numbers in the parenthesis are missing rates for the father, mother and offspring, respectively. (3) The missing rates for the three genotypes B_1B_1, B_1B_2 and B_2B_2 are identical for each individual. (4) Sample size = 500 families; disease allele (minor marker allele) frequencies for the first and second populations are 0.2 and 0.6 (0.4 and 0.3), respectively.

Table 4 Type-I error (%) of the TIMBD under population admixture with an extreme marker allele difference assuming MCAR/MAR

Model	*Theta*	*Delta*	*Missing rates 1*	*Missing rates 2*	*TDT*	*1-TDT*	*TIM*	*TIMBD*
Dominant	0.5	0	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	4.8	4.9	4.3	2.5
	0.5	0	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	5.2	5.1	4.6	2.5
	0.5	0	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	5.1	5.1	8.9	2.5
C. additive	0.5	0	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	5.1	5.2	4.1	2.8
	0.5	0	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	5.0	4.8	4.7	2.7
	0.5	0	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	5.2	5.2	8.6	2.4
Recessive	0.5	0	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	4.8	4.5	4.6	2.4
	0.5	0	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	5.1	5.1	4.7	2.5
	0.5	0	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	4.9	5.2	9.3	2.6
Dominant	0	0.07	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	15.3	19.7	4.4	2.3
	0	0.07	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	18.8	22.7	4.1	2.1
	0	0.07	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	17.1	21.4	8.0	2.6
C. additive	0	0.07	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	41.5	55.0	4.6	2.8
	0	0.07	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	49.7	60.4	4.3	2.3
	0	0.07	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	44.4	57.1	8.7	2.6
Recessive	0	0.07	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	13.2	16.5	4.8	2.5
	0	0.07	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	15.4	17.4	4.3	2.4
	0	0.07	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	13.5	17.4	8.7	2.4

Abbreviations: MAR, missing at random; MCAR, missing completely at random; TDT, Transmission/Disequilibrium Test; TIM, Test of Informative Missingness.
Note: (1) 'Missing rates 1' and 'Missing rates 2' are the missing rates for the first and second populations, respectively. (2) The first, second and third numbers in the parenthesis are missing rates for the father, mother and offspring, respectively. (3) The missing rates for the three genotypes B_1B_1, B_1B_2 and B_2B_2 are identical for each individual. (4) Sample size = 500 families; disease allele (minor marker allele) frequencies for the first and second populations are 0.2 and 0.6 (0.6 and 0.2), respectively.

Table 5 Power (%) of the TIMBD assuming no linkage ($\theta = 0.5$) or association ($\delta = 0$)

Model	*Father*	*Mother*	*Offspring*	*TDT*	*1-TDT*	*TIM*	*TIMBD*
Dominant	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.15, 0.15, 0.15)	5.10	10.30	64.30	77.30
	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	16.20	5.90	64.30	80.10
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.15, 0.15)	6.20	9.70	30.60	88.20
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	8.90	3.60	34.50	86.00
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.1, 0.1, 0.1)	4.30	6.10	39.70	47.50
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	9.30	4.70	41.00	51.20
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.1, 0.1)	5.40	7.20	16.70	60.20
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	7.80	6.20	15.90	57.60
C. additive	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.15, 0.15, 0.15)	5.90	13.00	61.30	75.70
	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	16.20	5.50	66.70	81.00
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.15, 0.15)	4.70	9.60	31.30	85.60
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	7.20	3.40	32.30	86.20
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.1, 0.1, 0.1)	5.10	7.60	40.50	46.40
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	8.90	5.00	37.40	47.10
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.1, 0.1)	4.90	7.60	14.30	59.30
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	6.90	4.80	15.60	56.70
Recessive	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.15, 0.15, 0.15)	4.70	10.00	65.70	77.90
	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	15.70	5.60	64.00	79.90
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.15, 0.15)	4.50	8.70	34.80	89.50
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	8.70	5.90	31.80	86.30
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.1, 0.1, 0.1)	4.70	6.70	37.40	45.30
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	9.50	5.70	42.10	50.50
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.1, 0.1)	4.50	5.60	14.30	59.90
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	7.50	6.10	16.00	58.90

Abbreviations: TDT, Transmission/Disequilibrium Test; TIM, Test of Informative Missingness.
Note: (1) The three numbers in the parenthesis are the missing rates for genotypes B_1B_1, B_1B_2 and B_2B_2, respectively. (2) Sample size = 500 families; disease allele frequency = 0.3; minor marker allele frequency (B_1) = 0.4; penetrance rate = 0.4; phenocopy rate = 0.2.

coefficient (δ). The three missing rates for the father, mother and offspring were displayed, respectively, in the first, second and third number of the parenthesis in the fourth column. The three missing rates may be different. However, each of the three missing rates was identical for all genotypes, B_1B_1, B_1B_2 and B_2B_2, such that the missing patterns were considered as MCAR for each family member. When there were no linkage ($\theta = 0.5$) or association ($\delta = 0$), the TDT and 1-TDT showed the expected 5% chance of rejecting the null hypothesis in the upper nine rows of Table 2. When there is linkage ($\theta = 0$) and association ($\delta = 0.14$), power of the TDT and 1-TDT are displayed in the bottom nine rows of Table 2. Simulation results indicated that type-I errors of the TIM and TIMBD were less than the nominal level of 5%, regardless of the relationship between the marker and the disease alleles. Therefore, test statistics of the TIM and TIMBD were independent of the recombination fraction θ and disequilibrium coefficient δ.

When the parental and offspring genotypes were MCAR or MAR in an admixed population, type-I errors of the TIM and TIMBD were displayed in Tables 3 and 4. The disease penetrance and phenocopy rates were 0.4 and 0.2, respectively. This scenario implies that the genotype relative risk was 2. Higher or lower genotype relative risk yielded similar comparisons and the results were not shown. In Tables 3 and 4, the disease allele frequencies of the first and second populations were 0.2 and 0.6, respectively. Therefore, the degree of admixture was identical in Tables 3 and 4. However, in Tables 3 and 4), the minor marker allele frequencies for the first and second populations were 0.4 and 0.3 (0.6 and 0.2), respectively. Hence, the difference between the marker allele frequencies of the two populations was more extreme in Table 4 than that in Table 3. When the disease and marker allele frequencies were <0.2 or >0.6, the comparisons between the TIM and TIMBD were similar and the results are not shown.

Since the TDT and 1-TDT were robust to population stratifications, both methods demonstrated the expected 5% type-I errors when there was no linkage or association in the upper nine rows of Tables 3 and 4. Under the alternative hypothesis, the TDT showed the lowest power due to exclusions of dyads in the analysis in the bottom nine rows of Tables 3 and 4. Therefore, if the missing pattern was MCAR/MAR, then the 1-TDT was more powerful than the TDT for detecting LD, which matched previous reports by Sun *et al.*[5] and Guo *et al.*[11] Although the TIM performed well under population admixture and showed type-I errors $<5\%$ in Table 3, its type-I errors could be slightly inflated over 8% in rows 3, 6, 9, 12, 15 and 18 of Table 4. In both scenarios, type-I errors of the TIMBD did not exceed 5%, although it appeared conservative. Therefore, the simulation results revealed that the TIMBD was robust to population admixture, while the TIM may suffer slightly inflated type-I errors.

Notes of the Breslow-Day test[18] indicated its requirement of large sample sizes in stratums and behavior under 'small stratum' settings that introduced the conservative type-I error. In the simulations, the sample size of the EM-HRR (i.e., missing data stratum) was not large to reflect real life scenarios, where the proportion of missing data was not too high. Therefore, the type-I error of the TIMBD was slightly conservative. Note that the average marker allele frequency in Table 3 was higher than that in Table 4. As a result, the type-I error of the TIMBD decreased from Table 3 to Table 4. In other words, decreasing

Table 6 Power (%) of the TIMBD assuming linkage ($\theta = 0$) and association ($\delta = 0.1$)

Model	*Father*	*Mother*	*Offspring*	*TDT*	*1-TDT*	*TIM*	*TIMBD*
Dominant	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.15, 0.15, 0.15)	50.70	27.40	66.30	84.10
	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	82.00	61.80	66.30	86.10
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.15, 0.15)	50.80	76.60	38.90	86.60
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	29.70	53.50	38.50	85.60
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.1, 0.1, 0.1)	56.90	41.30	40.20	55.90
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	73.10	60.00	41.80	57.70
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.1, 0.1)	54.30	70.50	19.70	57.70
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	37.30	53.50	16.60	54.40
C. additive	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.15, 0.15, 0.15)	92.50	79.60	63.00	86.00
	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	98.90	95.70	64.10	89.40
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.15, 0.15)	92.30	98.50	35.90	85.80
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	76.80	92.40	35.60	85.40
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.1, 0.1, 0.1)	92.60	84.50	39.10	59.80
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	98.30	95.70	42.20	59.10
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.1, 0.1)	93.30	97.80	19.00	58.50
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	79.60	90.80	17.90	54.70
Recessive	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.15, 0.15, 0.15)	93.10	80.80	61.20	86.60
	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	98.90	94.50	68.50	91.20
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.15, 0.15)	92.70	98.50	36.80	85.50
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	71.70	90.90	36.20	86.40
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.1, 0.1, 0.1)	91.90	84.10	40.70	58.50
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	97.30	94.10	36.00	60.30
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.1, 0.1)	91.10	97.00	17.30	55.10
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	81.60	91.50	19.10	57.50

Abbreviations: TDT, Transmission/Disequilibrium Test; TIM, Test of Informative Missingness.
Note: (1) The three numbers in the parenthesis are the missing rates for genotypes B_1B_1, B_1B_2 and B_2B_2, respectively. (2) Sample size = 500 families; disease allele frequency = 0.3; minor marker allele frequency (B_1) = 0.4; penetrance rate = 0.4; phenocopy rate = 0.2.

marker allele frequencies introduced more conservative type-I errors of the TIMBD and such pattern matched the previous results.[18] Regardless, the TIMBD did not yield the inflated type-I error and remained a valid test, even if the sample sizes in some stratums were small.

Power

Simulation results displayed in Table 5 (no association ($\delta = 0$) or linkage ($\theta = 0.5$)) and in Table 6 (association ($\delta = 0.1$) and linkage ($\theta = 0$)) were circumstances under which genotypes of triads were missing informatively in a homogeneous population. The disease and marker allele frequencies were 0.3 and 0.4, respectively. The disease penetrance and phenocopy rate were 0.4 and 0.2, respectively. The following two scenarios were examined: (1) the odd rows in each disease model: informative missingness occurred solely in parents but not the offspring. One can see that the missing rate (15 or 10%) was identical for any offspring genotype; (2) the even rows in each disease model: informative missingness occurred in both offspring and parental genotypes.

In Table 5, the TDT using the subset of complete triads remained a valid test for LD under the first scenario (the odds rows), since the TDT revealed type-I errors approaching the 5% nominal level. However, the 1-TDT, which used both triads and dyads, showed the inflated type-I errors and the inflation increased with respect to the magnitude of informative missingness. Under the second scenario, the TDT and 1-TDT were no longer valid tests, but the 1-TDT was less inflated than the TDT. In either scenario, the TIMBD was consistently more powerful than the TIM and the difference was more discernible when informative missingness was introduced by the minor allele (B_1)/genotype (B_1B_1) (rows 3, 4, 7 and 8 in each disease model).

In Table 6, power of the 1-TDT was lower (higher) than the TDT, when the major (minor) genotype introduced the informative missingness. This fact suggested that including dyads in the analysis could either dampen or inflate power of the 1-TDT when the assumption of MCAR/MAR was violated, which matched the previous investigations.[19,20] The results revealed an important message that informative missingness could also prevent discoveries of putative disease genes.

The GWAS scenarios assuming no linkage or association were displayed in Table 7. The results were adjusted by the Bonferroni's correction for multiple testing (the adjusted $\alpha = 5 \times 10^{-8}$). The TIMBD demonstrated decent power in the GWAS scenarios. The results also revealed that the TDT could yield considerable false positives (the second row in each disease model), even if the correction for multiple testing was implemented. This phenomenon illustrated the relationship between excessive false positives and informative missingness in the GWAS analysis.

DISCUSSION

Unlike the TIM, which is conditional on the offspring genotypes, the novel strategy TIMBD detects informative missingness by inconsistent LD signals between the complete and incomplete data. Attributable to its family-based design, the TIMBD is robust to population stratifications and outperforms the TIM in most situations. The excessive false positives solely due to informative missingness were also observed in the GWAS scenarios. The TIMBD is applicable for general pedigrees,

Table 7 GWAS scenarios—Power (%) of the TIMBD after Bonferroni's correction for multiple testing ($\alpha = 10^{-8}$) assuming no linkage ($\theta = 0.5$) or association ($\delta = 0$)

Model	*Father*	*Mother*	*Offspring*	*TDT*	*1-TDT*	*TIM*	*TIMBD*
Dominant	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.15, 0.15, 0.15)	0.00	0.02	98.36	99.81
	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	0.15	0.00	99.02	99.94
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.15, 0.15)	0.00	0.00	47.63	100.00
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	0.01	0.00	51.46	99.99
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.1, 0.1, 0.1)	0.00	0.00	58.34	57.76
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	0.00	0.00	61.09	63.45
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.1, 0.1)	0.00	0.00	4.98	85.78
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	0.01	0.00	4.99	84.87
C. additive	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.15, 0.15, 0.15)	0.00	0.01	98.30	99.76
	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	0.16	0.00	99.06	99.88
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.15, 0.15)	0.00	0.00	47.40	100.00
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	0.00	0.00	51.30	100.00
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.1, 0.1, 0.1)	0.00	0.00	57.60	58.73
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	0.00	0.00	60.79	63.16
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.1, 0.1)	0.00	0.00	4.70	86.28
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	0.00	0.00	5.01	84.70
Recessive	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.15, 0.15, 0.15)	0.00	0.02	98.20	99.80
	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	(0.01, 0.01, 0.15)	0.17	0.00	98.85	99.91
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.15, 0.15)	0.00	0.00	47.84	99.99
	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	(0.15, 0.01, 0.01)	0.00	0.00	51.77	100.00
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.1, 0.1, 0.1)	0.00	0.00	57.95	57.73
	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	(0.01, 0.01, 0.1)	0.00	0.00	61.17	63.49
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.1, 0.1)	0.00	0.00	4.52	86.16
	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	(0.1, 0.01, 0.01)	0.00	0.00	5.24	84.34

Abbreviations: TDT, Transmission/Disequilibrium Test; TIM, Test of Informative Missingness.
Note: (1) The three numbers in the parenthesis are the missing rates for genotypes B_1B_1, B_1B_2 and B_2B_2, respectively. (2) Sample size = 5000 families; disease allele frequency = 0.3; minor marker allele frequency (B_1) = 0.4; penetrance rate = 0.4; phenocopy rate = 0.2.

when independent triads, dyads and monads are identified from the independent pedigrees (see Supplementary data for the application in SAS/STAT software, SAS Institute inc., Cary, NC, USA).

In addition to non-random genotyping failure, which introduces informative missingness in both the offspring and parents, informative missingness may occur due to death or refusal to participate related to the outcome. One example to consider is asthma,[21,22] which could be diagnosed in both children and adults. The other plausible scenario is informative missingness in the parents, but not the offspring, as seen in age-dependent diseases, such as cancer,[23] Parkinson's disease,[24] diabetes[25] and cardiovascular diseases.[26,27] Same as the TIM, the limitation of the TIMBD is that it could not detect informative missingness that exists solely in the offspring, but not the parents, which could be classified as ascertainment bias. However, the TIMBD could be the foundation and/or step stones for considering ascertainment bias in genetic studies, since it could determine whether or not parental genotypes are missing informatively.

It is worth noting that the HRR and EM-HRR are based on the 2×2 contingency tables, hence the TIMBD could be easily extended into the logistic regression framework to adopt the Breslow-Day test in the logit model. In this way, the TIMBD could adjust for covariates related to missingness and ensures a valid test under various conditions of MAR as discussed previously.[14]

ACKNOWLEDGEMENTS

This work is supported by the research grant 'NSC-99-2314-B-006-053' awarded by National Science Council, Taiwan. It is also partially supported by a grant from the Ministry of Education, Aim for the Top University Plan. I appreciate much the insightful and valuable comments raised by the two reviewers that substantially improved this article.

1 Falk, C. T. & Rubinstein, P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* **51,** 227–233 (1987).
2 Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the sinsulin gene region and insulin dependent diabetes mellitus. *Am. J. Hum. Genet.* **52,** 506–516 (1993).
3 Little, R. J. A. & Rubin, D. B. *Statistical Analysis with Missing Data*. 2nd edn. (Wiley, New York, 2002).
4 Curtis, D. R. & Sham, P. C. A note on the application of the transmission disequilibrium test when a parent is missing. *Am. J. Hum. Genet.* **56,** 811–812 (1995).
5 Sun, F., Flanders, W., Yang, Q. & Khoury, J. Transmission Disequilibrium Test (TDT) with only one parent is available: The 1-TDT. *Am. J. Epidemiol.* **150,** 97–104 (1999).
6 Rabinowitz, D. & Laird, N. M. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.* **504,** 227–233 (2000).
7 Goring, H. H. & Terwilliger, J. D. Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am. J. Hum. Genet.* **66,** 1310–1327 (2000).
8 Clayton, D. A generalization of the transmission/disequilibrium test for uncertain haplotype transmission. *Am. J. Hum. Genet.* **65,** 1170–1177 (1999).
9 Weinberg, C. R. Allowing for missing parents in genetic studies of case-parents triads. *Am. J. Hum. Genet.* **64,** 1186–1193 (1999).
10 Gordon, D., Haynes, C., Johnnidis, C., Patel, S. B., Bowcock, A. M. & Ott, J. A. transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. *Eur. J. Hum. Genet.* **12,** 752–761 (2004).
11 Guo, C. Y., Destefano, A. L., Lunetta, K. L., Dupuis, J. & Cupples, L. A. Expectation Maximization Algorithm Based Haplotype Relative Risk (EM-HRR): test of linkage disequilibrium using incomplete case-parents trios. *Hum. Hered.* **59,** 125–135 (2005).

12 Allen, A. S., Rathouz, P. J. & Satten, G. A. Informative missingness in genetic association studies: case-parent designs. *Am. J. Hum. Genet.* **72,** 671–680 (2003).
13 Chen, Y. H. New approach to association testing in case-parent designs under informative parental missingness. *Genet. Epidemiol.* **27,** 131–140 (2004).
14 Guo, C. Y., Cupples, L. A. & Yang, Q. Testing informative missingness in genetic studies using case–parent triads. *Eur. J. Hum. Genet.* **16,** 992–1001 (2008).
15 Guo, C. Y. The impact of complex informative missingness on the validity of the transmission/disequilibrium test (TDT). *BMC Proc.* **1(Suppl 1), S26 (2007).**
16 Breslow, N. E. & Day, N. E. *Statistical Methods in Cancer Research, Volume II: The Design and Analysis of Cohort Studies* (Oxford University Press, USA, 1994).
17 Breslow, N. E. & Day, N. E. *Statistical Methods in Cancer Research, Volume I: The Analysis of Case-Control Studies* (Oxford University Press, Inc., New York, 1993).
18 Paul, S. R. & Donner, A. Small sample performance of tests of homogeneity of odds ratios in K 2 × 2 table. *Stat. Med.* **11,** 159–165 (1992).
19 Guo, C. Y., Cui, J. & Cupples, L. A. Impact of non-ignorable missingness on genetic tests of linkage and/or association using case-parents trios. *BMC Genet.* **6(Suppl 1), S90 (2005).**
20 Guo, C. Y. Validity of the transmission/disequilibrium test (TDT) under impact of complex informative missingness. *BMC Proc.* **1(Suppl 1), S26 (2007).**
21 Weiss, S. T. & Silverman, E. K. Pro: Genome-Wide Association Studies (GWAS) in Asthma. *Am. J. Respir. Crit. Care Med.* **184,** 631–633 (2011).
22 Adcock, I. M. & Barnes, P. J. Con: genome-wide association studies have not been useful in understanding asthma. *Am. J. Respir. Crit. Care Med.* **184,** 633–636 (2011).
23 Barrett, J. H., Iles, M. M., Harland, M., Taylor, J. C., Aitken, J. F., Andresen, P. A. *et al.* Genome-wide association study identifies three new melanoma susceptibility loci. *Nat. Genet.* **43,** 1108–1113 (2011).
24 Simón-Sánchez, J., van Hilten, J. J., van de Warrenburg, B., Post, B., Berendse, H. W., Arepalli, S. *et al.* Genome-wide association study confirms extant PD risk loci among the Dutch. *Eur. J. Hum. Genet.* **19,** 655–661 (2011).
25 Kooner, J. S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P. *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat. Genet.* **43,** 984–989 (2011).
26 Newton-Cheh, C., Guo, C. Y., Wang, T. J., O'donnell, C. J., Levy, D. & Larson, M. G. Genome-wide association study of electrocardiographic and heart rate variability traits: the Framingham Heart Study. *BMC Med. Genet.* **8,** S7 (2007).
27 Cupples, L. A., Arruda, H. T., Benjamin, E. J., D'Agostino, Sr R.B., Demissie, S., Destefano, A. L. *et al.* The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Med. Genet.* **8,** S1 (2007).
28 Paul, S. R. & Donner, A. A comparison of tests of tests of homogeneity of odds ratios in K 2 × 2 table. *Stat. Med.* **8,** 1455–1468 (1989).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

APPENDIX A

Let $P_k^{i,j}$ represents the theoretical probability for each type of triad data. $k =$ '0', '1' or '2' denotes the total number of B_1 alleles transmitted to the offspring, and $i, j =$ '0', '1' or '2' denotes the total number of B_1 alleles for the father and mother, respectively.

Assuming the absence of heterogeneity, when parental genotypes were incomplete, Guo *et al.*[11] applied the EM algorithm to estimate the proportion of heterozygous parents $(M_1^{1,*} + M_1^{*,1})$ who transmitted the B_1 allele but not the other B_2 allele, which was denoted by $\{M_1^{1,*} + M_1^{*,1}\}_1$. Similarly, heterozygous parents who transmitted the B_2 allele but not the other B_1 allele were denoted by $\{M_1^{1,*} + M_1^{*,1}\}_2$. The EM-HRR by Guo *et al.*[11] avoided biased results warned by Curtis and Sham.[4] An important feature of the HRR is that genotypes of affected offspring are always present (assuming no genotyping failure) due to the ascertainment criteria, which collects an affected individual first and then seeks his/her parents. Note that only N_{B1}^2 and N_{B2}^2 are involved with the EM estimates. Since transmitted alleles can be inferred unambiguously, these alleles do not require EM algorithm estimates. In addition, both N_{B1}^3 and N_{B2}^3 are defined as 0, because none of the parental genotypes are present in monads to infer which alleles are not transmitted.

Assuming the absence of heterogeneity, under the null hypothesis of MCAR/MAR, one would expect that the EM estimates of the non-transmitted allele from the one heterozygous parent $(M_1^{1,*} + M_1^{*,1})$ will be unbiased. Therefore, the following four equations (1–4) will hold.

$$E\left(\frac{T_{B1}^1}{N_{HRR}}\right) = E\left(\frac{T_{B1}^2 + T_{B1}^3}{N_{EM\text{-}HRR}}\right) \quad (1)$$

$$E\left(\frac{T_{B2}^1}{N_{HRR}}\right) = E\left(\frac{T_{B2}^2 + T_{B2}^3}{N_{EM\text{-}HRR}}\right) \quad (2)$$

$$E\left(\frac{N_{B1}^1}{N_{HRR}}\right) = E\left(\frac{N_{B1}^2 + N_{B1}^3}{N_{EM\text{-}HRR}}\right) \quad (3)$$

$$E\left(\frac{N_{B2}^1}{N_{HRR}}\right) = E\left(\frac{N_{B2}^2 + N_{B2}^3}{N_{EM\text{-}HRR}}\right) \quad (4)$$

In this case, the EM-HRR is expected to yield an identical LD signal (i.e., odds ratio (OR)) to that of the HRR. When there is no linkage (i.e., recombination fraction $\theta = 0.5$) or association (that is, disequilibrium coefficient $\delta = 0$), $E(HRR) = E(EM\text{-}HRR) = 1$. If there is linkage ($\theta \neq 0.5$) and association ($\delta \neq 0$), then $E(HRR) = E(EM\text{-}HRR) \neq 1$.

Under the alternative hypothesis of informative missingness, equations (1–4) may be violated. As a result, expectations of the HRR and EM-HRR are dissimilar, $E(HRR) \neq E(EM\text{-}HRR)$, regardless of the LD information.

Following notations in Table 1 and Appendix of Guo *et al.*,[14] let PT_{B1}^s, PN_{B1}^s, PT_{B2}^s and PN_{B2}^s denote the conditional probability of a parent transmitted B_1 alleles, non-transmitted B_1 alleles, transmitted B_2 alleles and non-transmitted B_2 alleles, respectively, from type s families, where $s = 1$ indicates complete triads, 2 for dyads and 3 denotes monads.

Details of the conditional probability of transmitting and non-transmitting a specific marker allele for all three types of families are displayed in the following:

$$PT_{B1}^1 = 2P_2^{2,2} + 2P_2^{2,1} + 2P_2^{1,2} + 2P_2^{1,1} + P_1^{1,2} + P_1^{2,1} + P_1^{2,0} + P_1^{0,2} + P_1^{1,1} + P_1^{1,0} + P_1^{0,1}$$

$$PN_{B1}^1 = 2P_2^{2,2} + 2P_1^{2,1} + 2P_1^{1,2} + 2P_0^{1,1} + P_2^{2,1} + P_2^{1,2} + P_1^{2,0} + P_1^{0,2} + P_1^{1,1} + P_0^{1,0} + P_0^{0,1}$$

$$PT_{B2}^1 = 2P_0^{0,0} + 2P_0^{0,1} + 2P_0^{1,0} + 2P_0^{1,1} + P_1^{1,2} + P_1^{2,1} + P_1^{2,0} + P_1^{0,2} + P_1^{1,1} + P_1^{1,0} + P_1^{0,1}$$

$$PN_{B2}^1 = 2P_2^{1,1} + 2P_1^{1,0} + 2P_1^{0,1} + 2P_0^{0,0} + P_2^{2,1} + P_2^{1,2} + P_1^{2,0} + P_1^{0,2} + P_1^{1,1} + P_0^{1,0} + P_0^{0,1}$$

$$PT_{B1}^2 = 2P_2^{2,*} + 2P_2^{*,2} + 2P_2^{1,*} + 2P_2^{*,1} + P_1^{2,*} + P_1^{*,2} + P_1^{1,*} + P_1^{*,1} + P_1^{0,*} + P_1^{*,0}$$

$$PN_{B1}^2 = P_2^{2,*} + P_2^{*,2} + P_1^{2,*} + P_1^{*,2} + P_0^{1,*} + P_0^{*,1} + P\{M_1^{1,*} + M_1^{*,1}\}_2$$

$$PT_{B2}^2 = 2P_0^{0,*} + 2P_0^{*,0} + 2P_0^{1,*} + 2P_0^{*,1} + P_1^{2,*} + P_1^{*,2} + P_1^{1,*} + P_1^{*,1} + P_1^{0,*} + P_1^{*,0}$$

$$PN_{B2}^2 = P_2^{1,*} + P_2^{*,1} + P_1^{0,*} + P_1^{*,0} + P_0^{0,*} + P_0^{*,0} + P\{M_1^{1,*} + M_1^{*,1}\}_1$$

$$PT_{B1}^3 = 2P_2^{*,*} + P_1^{*,*}$$

$$PT_{B2}^3 = 2P_0^{*,*} + P_1^{*,*}$$

The HRR using complete triads is defined as $HRR = (PT_{B1}^1 \times PN_{B2}^1)/(PN_{B1}^1 \times PT_{B2}^1)$ and the EM-HRR using dyads and monads is defined as $EM\text{-}HRR = ((PT_{B1}^2 + PT_{B1}^3) \times PN_{B2}^2)/(PN_{B1}^2 \times (PT_{B2}^2 + PT_{B2}^3))$. It is straightforward to show that, under the null hypothesis of MCAR/MAR ($P_{o11} = P_{o12} = P_{o22} = P_o$, $P_{f11} = P_{f12} = P_{f22} = P_f$, and $P_{m11} = P_{m12} = P_{m22} = P_m$), HRR is identical to EM-HRR regardless of linkage (θ) and association information (δ).

If E(HRR) = E(EM-HRR), then the EM estimators are unbiased so that one would expect the offspring and parental genotypes to be MCAR/MAR. Because the underlying distribution/parameters are the same in the complete triads and incomplete data, which forces missing rates for different genotypes to be identical. Therefore, the probability of the Breslow-Day test showing a significant difference between the HRR and EM-HRR is expected to be the predetermined significance level.

APPENDIX B

Note that $E_1 = E(T^1_{B1} \mid OR_{MH})$ and $E_2 = E(T^2_{B1} + T^3_{B1} \mid OR_{MH})$ are the expected values of (T^1_{B1}) and $(T^2_{B1} + T^3_{B1})$ given OR_{MH}, $V_1 = Var(T^1_{B1} \mid OR_{MH})$ and $V_2 = Var((T^2_{B1} + T^3_{B1}) \mid OR_{MH})$ are estimators of the variance of (T^1_{B1}) and $(T^2_{B1} + T^3_{B1})$ given the value of OR_{MH} and conditional on the value of $t_1 = T^1_{B1} + N^1_{B1}$ and $t_2 = (T^2_{B1} + T^3_{B1}) + (N^2_{B1})$ (see Paul and Donner[28] for details). $E(T^1_{B1} \mid OR_{MH})$ can be found by solving the following quadratic equation:

$$\frac{E_1 \times (N^1_{B1} + N^1_{B2} - t_1 + E_1)}{(t_1 - E_1)(N^1_{B1} + N^1_{B2} - E_1)} = OR_{MH} \tag{B.1}$$

equation (B.1) takes the unique root in the interval $\max[0, t_1 - (N^1_{B1} + N^1_{B2})] \leqslant E_1 \leqslant \min[t_1, (T^1_{B1} + T^1_{B2})]$.

V_1 is then obtained as

$$V_1 = \left(\frac{1}{E_1} + \frac{1}{t_1 - E_1} + \frac{1}{(T^1_{B1} + T^1_{B2}) - E_1} + \frac{1}{(N^1_{B1} + N^1_{B2}) - t_1 + E_1}\right)^{-1}.$$

Similarly, E_2 can be found by solving the following quadratic equation:

$$\frac{E_2 \times (N^2_{B1} + N^2_{B2} - t_2 + E_2)}{(t_2 - E_2)(N^2_{B1} + N^2_{B2} - E_2)} = OR_{MH} \tag{B.2}$$

equation (B.2) takes the unique root in the interval $\max[0, t_2 - (N^2_{B1} + N^2_{B2})] \leqslant E_2 \leqslant \min\{t_2, [(T^2_{B1} + T^3_{B1}) + (T^2_{B2} + T^3_{B2})]\}$.

V_2 is then obtained as

$$V_2 = \left(\frac{1}{E_2} + \frac{1}{t_2 - E_2} + \frac{1}{[(T^2_{B1} + T^3_{B1}) + (T^2_{B2} + T^3_{B2})] - E_2} + \frac{1}{(N^2_{B1} + N^2_{B2}) - t_2 + E_2}\right)^{-1}.$$

Equations (B.1 and B.2) involve finding the root of two quadratic equations, followed by the usual summation to obtain the test statistic.