European Journal of Human Genetics (2008) 16, 992–1001

www.nature.com/ejhg

ARTICLE

Testing informative missingness in genetic studies using case–parent triads

Chao-Yu Guo*[,1,2,3], Laura Adrienne Cupples[4] and Qiong Yang[4]

[1]*Clinical Research Program, Children's Hospital Boston, Boston, MA, USA;* [2]*Program in Genomics, Department of Medicine, Children's Hospital Boston, Boston, MA, USA;* [3]*Department of Pediatrics, Harvard Medical School, Boston, MA, USA;* [4]*Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA*

In genetic studies, the transmission/disequilibrium test (TDT) using case–parent triads has gained popularity attributable to its robustness to population admixture. Several extensions have been proposed to accommodate incomplete triads. Some strategies assume that parental genotypes are missing completely at random (MCAR) to insure an unbiased conclusion and some methods allow parental genotypes to be missing informatively, resulting in reduced power when the missing data pattern is indeed MCAR. However, these tests assumed that offspring genotypes were MCAR. Recently, Guo indicated that when offspring genotypes were missing informatively, an occurrence that can be considered as ascertainment bias, inflated type-I error and/or reduced power may occur using the TDT when incomplete triads are excluded. In an effort to avoid an erroneous conclusion, we propose a strategy called testing informative missingness (TIM) that compares conditional distributions of parental genotypes among complete triads and incomplete data with only one parent to examine the missing data pattern. Through computer simulations, TIM has decent power to detect informative missingness and is robust to population admixture. In addition, we illustrate TIM with an application to the Framingham Heart Study.
European Journal of Human Genetics (2008) **16,** 992–1001; doi:10.1038/ejhg.2008.38; published online 12 March 2008

Keywords: transmission disequilibrium test; family-based study; informative missingness; triads; dyads

Introduction

Using unrelated subjects in a case–control study is a popular design for testing association between genetic markers and phenotypes. Spurious association may occur due to migration, nonrandom mating or population admixture. In order to avoid spurious evidence of association, Falk and Rubinstein[1] proposed the Haplotype Relative Risk (HRR), which uses case–parent triads, as a method to test linkage disequilibrium (LD) between a marker and a putative disease locus. The HRR compares parental marker alleles transmitted to an affected offspring to those not transmitted as a test for association. When population admixture is present, HRR is conservative resulting in reduced power for testing associations. Spielman *et al*[2] suggested the transmission/disequilibrium test (TDT) to adjust for population admixture using a matched study design. TDT examines if heterozygous parents preferentially transmit certain alleles to an affected offspring.

Although family-based triads are robust to population admixture, the collection of parental genotypes is often difficult because of death or refusal to participate. Family-based association tests, such as the HRR and TDT, are generally not applicable when parental genotypes are not complete. Curtis and Sham[3] showed that the estimate of the probability of transmission of certain alleles is biased in the TDT when one parent is missing, and only heterozygous parents and homozygous offspring contribute to

*Correspondence: Professor C-Y Guo, Clinical Research Program and Program in Genomics, Children's Hospital Boston and Harvard Medical School, Address: 300 Longwood Avenue, Boston, MA 02115, USA.
Tel: +1 617 355 0685 or 1 617 919 4798; Fax: +1 617 355 2312;
E-mail: chao-yu.guo@childrens.harvard.edu
Received 7 April 2007; revised 23 October 2007; accepted 1 February 2008; published online 12 March 2008

the test. Assuming that parental genotypes are missing completely at random (MCAR), such bias is avoided by the 1-TDT test (TDT with only one parent) proposed by Sun *et al*,[4] a test that uses genotypes of the affected offspring and the one available parent, but excludes affected offspring where both the offspring and the parent are heterozygous. In addition, several other strategies with the same assumption have been proposed to accommodate incomplete triads.[5–7]

Allowing for informative missingness of parental genotypes, Allen *et al*[8] and Chen[9] proposed valid tests incorporating incomplete triads. However, their strategies are less powerful when the missing pattern was indeed MAR or MCAR. For example, in Chen's Table 4,[9] the power of the 1-d.f. score statistic is less than that of TDT using intact triads only for a common (rare) allele under the dominant (recessive) disease model. So is the 2-d.f. score statistic for both rare and common variant alleles under the multiplicative inheritance. This means that the inclusion of dyads (incomplete triads with only one parental genotype) reduces the power of the score test in these cases.

Regardless of different missing data patterns among parental genotypes, the above methods assumed that offspring genotypes were MCAR. Recently, Guo[10] derived the conditional distribution of ascertained triads that allows informative missingness for offspring genotypes, as well as their parental genotypes, and evaluated several tests under such scenarios. Guo[10] indicated that when offspring genotypes were missing informatively, a circumstance that can be considered as ascertainment bias, inflated type-I error and/or reduced power may occur using the TDT excluding incomplete triads. Therefore, if the missing data pattern for offspring genotypes is not confirmed to be MCAR, a significant result from the TDT using only intact triads does not assure true association between the marker and a putative disease locus.

In an effort to assure a valid conclusion, we introduce a new test called Testing Informative Missingness (TIM) to determine whether the missing data pattern in ascertained triads is informative or not.

Statistical method

We derived the conditional distribution of ascertained triads and dyads (Table 1) in the Appendix 1.[11] Note that $P_k^{i,j}$ and $M_k^{i,j}$ represent the theoretical probability and observed counts for each type of triad data. $k =$ '0', '1' or '2' represents the total number of B_1 alleles transmitted to the offspring, and $i, j =$ '0', '1' or '2' represents the ordered total number of B_1 alleles for fathers and mothers, respectively. Note that we use the superscript '*' to denote that the parental genotype is missing.

Based on Table 1, we calculated the conditional distribution of parental genotypes among triads and dyads, displayed in Table 2. Under the null hypothesis of MCAR, conditional on offspring genotypes, the distribution of parental genotypes among triads and dyads are identical. Therefore, a logistic regression approach can be implemented to test for informative missingness of parental genotypes.

Let the outcome variable Y be 1, if the parental genotype is from a complete triad and $Y = 0$, if the parental genotype is from a dyad. Parents of a triad contribute two independent observations and the available parent among dyads contributes one observation. By choosing genotype B_1B_1 to be the reference group, let the first (second) dummy variable of parental genotype be $D_1 = 1$ ($D_2 = 1$) if the parental genotype is B_1B_2 (B_2B_2); otherwise $D_1 = 0$ ($D_2 = 0$). Similarly, let the first (second) dummy variable of offspring genotype $G_1 = 1$ ($G_2 = 1$) if the offspring genotype is B_1B_2 (B_2B_2); otherwise $G_1 = 0 (G_2 = 0)$. As a result, conditional on the affected offspring genotypes, the distribution of parental genotypes among triads can be compared to that of dyads by the logistic model as

$$\begin{aligned}\text{Logit}(P) &= \text{Log}\left(\frac{P}{1-P}\right)\\ &= \beta_0 + \beta_{D_1}D_1 + \beta_{D_2}D_2 + \beta_{G_1}G_1 + \beta_{G_2}G_2,\end{aligned}$$

where

$$\begin{aligned}P &= \Pr(Y = 1|D_1, D_2, G_1, G_2)\\ &= \frac{e^{\beta_0+\beta_{D_1}D_1+\beta_{D_2}D_2+\beta_{G_1}G_1+\beta_{G_2}G_2}}{1+e^{\beta_0+\beta_{D_1}D_1+\beta_{D_2}D_2+\beta_{G_1}G_1+\beta_{G_2}G_2}}\end{aligned}$$

Under the null hypothesis that the genotypes of ascertained triads and dyads are MCAR, the null hypothesis of TIM is $\beta_{D_1} = \beta_{D_2} = 0|G_1, G_2$, which states that the distributions of parental genotypes are identical among triads and dyads controlling for offspring genotypes.

Scenarios under missing at random

Little and Rubin[12] indicated that MCAR means that the cause of missingness is unrelated to the items and the observed values from a random subsample of the sampled value. Missing at random (MAR) means that the probability of a missing value for an outcome depends on the observed responses of other covariates, but given these, it does not depend on the missing value itself. Within subgroups formed by the observed covariates on which the missingness depends, the data are MCAR. Therefore, scenarios under MAR are also considered as the null distribution, since it becomes MCAR by adjusting for available covariates related to the missing data mechanism.

For example, suppose you are modeling weight (Y) as a function of sex (X). Some respondents would not disclose their weight, so you are missing some values for Y. One sex may be less likely to disclose its weight. That is, the probability that Y is missing depends only on the value of X. Such data are MAR and weight conditional on sex ($Y|X$)

Table 1 Conditional distribution of ascertained triads and dyads

Types of data	*Affected child*	*Father*	*Mother*	*Probability*	*Observation*
Type 1: Both parents available	B_1B_1	B_1B_1	B_1B_1	$P_2^{2,2} = \mu^2 \times (1\text{-}P_{f11}) \times (1\text{-}P_{m11}) \times (1\text{-}P_{o11})$	$M_2^{2,2}$
		B_1B_1	B_1B_2	$P_2^{1,2} = \mu\nu \times (1\text{-}P_{f11}) \times (1\text{-}P_{m12}) \times (1\text{-}P_{o11})$	$M_2^{2,1}$
		B_1B_2	B_1B_1	$P_2^{2,1} = \mu\nu \times (1\text{-}P_{f12}) \times (1\text{-}P_{m11}) \times (1\text{-}P_{o11})$	$M_2^{1,2}$
		B_1B_2	B_1B_2	$P_2^{1,1} = \nu^2 \times (1\text{-}P_{f12}) \times (1\text{-}P_{m12}) \times (1\text{-}P_{o11})$	$M_2^{1,1}$
	B_1B_2	B_1B_1	B_1B_2	$P_1^{1,2} = \mu\zeta \times (1\text{-}P_{f11}) \times (1\text{-}P_{m12}) \times (1\text{-}P_{o12})$	$M_1^{1,2}$
		B_1B_2	B_1B_1	$P_1^{2,1} = \mu\zeta \times (1\text{-}P_{f12}) \times (1\text{-}P_{m11}) \times (1\text{-}P_{o12})$	$M_1^{2,1}$
		B_1B_1	B_2B_2	$P_1^{0,2} = \mu\tau \times (1\text{-}P_{f11}) \times (1\text{-}P_{m22}) \times (1\text{-}P_{o12})$	$M_1^{2,0}$
		B_2B_2	B_1B_1	$P_1^{2,0} = \mu\tau \times (1\text{-}P_{f22}) \times (1\text{-}P_{m11}) \times (1\text{-}P_{o12})$	$M_1^{0,2}$
		B_1B_2	B_1B_2	$P_1^{1,1} = 2\nu\zeta \times (1\text{-}P_{f12}) \times (1\text{-}P_{m12}) \times (1\text{-}P_{o12})$	$M_1^{1,1}$
		B_1B_2	B_2B_2	$P_1^{0,1} = \nu\tau \times (1\text{-}P_{f12}) \times (1\text{-}P_{m22}) \times (1\text{-}P_{o12})$	$M_1^{1,0}$
		B_2B_2	B_1B_2	$P_1^{1,0} = \nu\tau \times (1\text{-}P_{f22}) \times (1\text{-}P_{m12}) \times (1\text{-}P_{o12})$	$M_1^{0,1}$
	B_2B_2	B_1B_2	B_1B_2	$P_0^{1,1} = \zeta^2 \times (1\text{-}P_{f12}) \times (1\text{-}P_{m12}) \times (1\text{-}P_{o22})$	$M_0^{1,1}$
		B_1B_2	B_2B_2	$P_0^{0,1} = \zeta\tau \times (1\text{-}P_{f12}) \times (1\text{-}P_{m22}) \times (1\text{-}P_{o22})$	$M_0^{1,0}$
		B_2B_2	B_1B_2	$P_0^{1,0} = (1\text{-}P_{f22}) \times (1\text{-}P_{m12}) \times (1\text{-}P_{o22})$	$M_0^{0,1}$
		B_2B_2	B_2B_2	$P_0^{0,0} = \tau^2 \times (1\text{-}P_{f22}) \times (1\text{-}P_{m22}) \times (1\text{-}P_{o22})$	$M_0^{0,0}$
	Total 1			$\text{Sum1} = \sum_{i,j,k} P_i^{j,k}$	N_{triads}
Type 2: One parent available	B_1B_1	B_1B_1		$P_2^{2,*} = (\mu^2+\mu\nu) \times [(1\text{-}P_{f11})P_{m11}] \times (1\text{-}P_{o12})$	$M_2^{2,*}$
			B_1B_1	$P_2^{*,2} = (\mu^2+\mu\nu) \times [(1\text{-}P_{m11})P_{f12}] \times (1\text{-}P_{o12})$	$M_2^{*,2}$
		B_1B_2		$P_2^{1,*} = (\nu^2+\mu\nu) \times [(1\text{-}P_{f12})P_{m12}] \times (1\text{-}P_{o12})$	$M_2^{1,*}$
			B_1B_2	$P_2^{*,1} = (\nu^2+\mu\nu) \times [(1\text{-}P_{m12})P_{f11}] \times (1\text{-}P_{o12})$	$M_2^{*,1}$
	B_1B_2	B_1B_1		$P_1^{2,*} = (\mu\zeta+\mu\tau) \times [(1\text{-}P_{f11})P_{m12}] \times (1\text{-}P_{o12})$	$M_1^{2,*}$
			B_1B_1	$P_1^{*,2} = (\mu\zeta+\mu\tau) \times [(1\text{-}P_{m11})P_{f22}] \times (1\text{-}P_{o12})$	$M_1^{*,2}$
		B_1B_2		$P_1^{1,*} = (\mu\zeta+2\nu\zeta+\nu\tau) \times [(1\text{-}P_{f12})P_{m11}] \times (1\text{-}P_{o12})$	$M_1^{1,*}$
			B_1B_2	$P_1^{*,1} = (\mu\zeta+2\nu\zeta+\nu\tau) \times [(1\text{-}P_{m12})P_{f12}] \times (1\text{-}P_{o12})$	$M_1^{*,1}$
		B_2B_2		$P_1^{0,*} = (\mu\tau+\nu\tau) \times [(1\text{-}P_{f22})P_{m11}] \times (1\text{-}P_{o12})$	$M_1^{0,*}$
			B_2B_2	$P_1^{*,0} = (\mu\tau+\nu\tau) \times [(1\text{-}P_{m22})P_{f12}] \times (1\text{-}P_{o12})$	$M_1^{*,0}$
	B_2B_2	B_1B_2		$P_0^{1,*} = (\zeta^2+\zeta\tau) \times [(1\text{-}P_{f12})P_{m12}] \times (1\text{-}P_{o22})$	$M_0^{1,*}$
			B_1B_2	$P_0^{*,1} = (\zeta^2+\zeta\tau) \times [(1\text{-}P_{m12})P_{f22}] \times (1\text{-}P_{o22})$	$M_0^{*,1}$
		B_2B_2		$P_0^{0,*} = (\tau^2+\zeta\tau) \times [(1\text{-}P_{f22})P_{m12}] \times (1\text{-}P_{o22})$	$M_0^{0,*}$
			B_2B_2	$P_0^{0,*} = (\tau^2+\zeta\tau) \times [(1\text{-}P_{m22})P_{f12}] \times (1\text{-}P_{o22})$	$M_0^{*,0}$
	Total 2			$\text{Sum2} = \sum_{i,j} P_i^{j,*} + \sum_{i,k} P_i^{*,k}$	N_{dyads}

Table 2 Conditional distribution of parental genotypes among triads and dyads

Offspring genotype	*Parental genotype*	*Triads*	*Dyads*
B_1B_1	B_1B_1	$p_2^{2,2}+\frac{p_2^{1,2}+p_2^{2,1}}{2}$	$p_2^{2,*}+p_2^{*,2}$
	B_1B_2	$p_2^{1,1}+\frac{p_2^{1,2}+p_2^{2,1}}{2}$	$p_2^{1,*}+p_2^{*,1}$
B_1B_2	B_1B_1	$\frac{p_1^{1,2}+p_1^{2,1}+p_1^{0,2}+p_1^{2,0}}{2}$	$p_1^{2,*}+p_1^{*,2}$
	B_1B_2	$p_1^{1,1}+\frac{p_1^{1,2}+p_1^{2,1}+p_1^{0,1}+p_1^{1,0}}{2}$	$p_1^{1,*}+p_1^{*,1}$
	B_2B_2	$\frac{p_1^{0,1}+p_1^{1,0}+p_1^{0,2}+p_1^{2,0}}{2}$	$p_1^{0,*}+p_1^{*,0}$
B_2B_2	B_1B_2	$p_0^{1,1}+\frac{p_0^{0,1}+p_0^{1,0}}{2}$	$p_0^{1,*}+p_0^{*,1}$
	B_2B_2	$p_0^{0,0}+\frac{p_0^{0,1}+p_0^{1,0}}{2}$	$p_0^{0,*}+p_0^{*,0}$
Total		Sum1	Sum2

is MCAR. Therefore, the data can be considered as MCAR within subgroups formed by the observed items (covariates) on which the missingness depends. Here, let the covariates be X_1, $X_2,\ldots,X_K$ and the logistic model is

$$\text{Logit}(P)=\beta_0+\beta_{D_1}D_1+\beta_{D_2}D_2+\beta_{G_1}G_1+\beta_{G_2}G_2+\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k$$

The null hypothesis of MCAR is $\beta_{D_1}=\beta_{D_2}=0|G_1, G_2, X_1, X_2,\ldots,X_K$. Therefore, even when the missing data mechanism is MAR but not MCAR, the TIM does not reject the null hypothesis when covariates related to missingness X_1, $X_2,\ldots,X_K$ are taken into account.

Simulations

We first assumed that the population is free from population stratification. Let '*a*' and '*A*' denote the disease and normal allele. Let *D* denote that an individual is diseased. Let *f* denote the probability of being affected when an individual carries 0 risk alleles (the phenocopy rate), and let *K* denote the genotype relative risk (GRR). For a recessive disease model, the penetrance functions are $P(D|AA)=P(D|Aa)=f$ and $P(D|aa)=K\times f$, where $0\leq f\leq 1$ and $0\leq K\times f\leq 1$. The disease prevalence is determined by these probabilities and the risk allele frequency. Similarly, for a dominant disease model, $P(D|AA)=f$ and $P(D|Aa)=P(D|aa)=K\times f$. For an additive disease model, $P(D|AA)=f$; $P(D|Aa)=K\times f$; $P(D|aa)=2\times K\times f-1(K>0.5)$. We considered additive, recessive, and dominant disease models in our simulations and the affection status of each individual is determined according to these parameters.

We simulated a general population where nuclear families have exactly one offspring. We randomly assigned each offspring, father, and mother to be missing according to various probabilities indicated in each table. Therefore, only a proportion of families with an affected offspring were eligible for the study and a total of 500 families were sampled.

Several disease allele (denoted *a*) and marker allele (denoted B_1) frequencies were examined. A range of possible values for the disequilibrium coefficient $\delta=p(aB_1)-p(a)p(B_1)$ and recombination fraction θ were simulated. In the tables and figures displayed, the frequencies of the disease and marker alleles, the disease model, phenocopy rate, and the penetrance are indicated in each table.

We repeated the simulation 10 000 (1000) times to examine type-I error (power) of several tests examined including the TIM. Under the null hypothesis of MCAR, the fraction of times that the test statistic exceeds the critical value, defined by the asymptotic distribution of the statistic, is the type-I error. The power of each test is the proportion of test statistics in the total number of simulations exceeding the critical value under the alternative hypothesis. Type-I error and power of several LD tests were also evaluated under the various patterns of parental genotype missingness.

In a second set of simulations we introduced population stratification by sampling two populations with expected samples sizes reflecting different disease frequencies in the subpopulations. 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, we will expect 20% and 80% of the sample to come from the first and second populations, respectively. This is the ratio we would observe in most admixed samples. Because the disease allele frequencies are different in the two populations, the frequencies of diseased individuals in the two samples are also different.

In Tables 3–6, the column marked 'TDT' reports results using the traditional TDT[2] test on the subset of complete triads only. The columns marked '1-TDT' and 'EM-HRR' (expectation maximization algorithm based haplotype relative risk)[7] use both the complete triads and dyads. The column marked 'TIM' is the test of informative missingness.

Results

When genotypes of ascertained offspring and parents are MCAR, the type-I errors of TIM in a homogeneous population are displayed in Table 3. Both the disease and marker allele frequencies are 0.3. The disease penetrance and phenocopy rate are 0.4 and 0.2, respectively. Different disease and marker allele frequencies, penetrance, and phenocopy rates yielded similar results, not shown here. The underlying disease model (dominant, additive, or recessive) is indicated in the first column. The second and third columns are the recombination fraction θ and disequi-

Table 3 Type-I error (%) of TIM at $\alpha = 0.05$ based on 10 000 replicates

Model	θ	δ	*Missing rates*	*TDT*	*1-TDT*	*EM-HRR*	*TIM*
Dominant	0.5	0	(0.3; 0.3; 0.3)	5.2	4.9	5.4	3.9
	0.5	0	(0.3; 0.1; 0.2)	4.9	4.9	5.0	4.0
	0.5	0	(0.2; 0.4; 0.1)	4.8	4.9	5.5	3.9
Additive	0.5	0	(0.3; 0.3; 0.3)	4.9	4.7	4.9	4.0
	0.5	0	(0.3; 0.1; 0.2)	5.3	5.1	5.6	4.4
	0.5	0	(0.2; 0.4; 0.1)	5.2	5.3	5.5	4.1
Recessive	0.5	0	(0.3; 0.3; 0.3)	4.9	5.3	6.1	4.0
	0.5	0	(0.3; 0.1; 0.2)	5.2	5.0	5.4	4.0
	0.5	0	(0.2; 0.4; 0.1)	4.8	4.8	5.2	4.1
Dominant	0	0.05	(0.3; 0.3; 0.3)	11.8	15.2	17.4	4.2
	0	0.05	(0.3; 0.1; 0.2)	26.1	31.5	34.8	4.7
	0	0.05	(0.2; 0.4; 0.1)	11.7	14.8	17.3	4.3
Additive	0	0.05	(0.3; 0.3; 0.3)	24.7	32.9	37.4	4.0
	0	0.05	(0.3; 0.1; 0.2)	48.7	58.0	61.6	4.1
	0	0.05	(0.2; 0.4; 0.1)	24.1	34.1	37.2	3.9
Recessive	0	0.05	(0.3; 0.3; 0.3)	7.8	9.1	9.8	4.2
	0	0.05	(0.3; 0.1; 0.2)	16.5	19.2	20.8	4.1
	0	0.05	(0.2; 0.4; 0.1)	7.4	8.6	9.5	4.1

Abbreviations: EM-HRR, expectation maximization algorithm based haplotype relative risk; MCAR, missing completely at random; TDT, transmission/disequilibrium test; TIM, testing informative missingness. Sample size = 500 families; both the disease and minor marker allele frequencies are 0.3; penetrance rate = 0.4; phenocopy rate = 0.2.
The first, second, and third number in parenthesis are missing rates for fathers, mothers, and offspring, respectively. The missing rates for all three genotypes B_1B_1, B_1B_2 and B_2B_2 are identical (MCAR).
TDT uses complete trios only; 1-TDT, EM-HRR, and TIM use complete trios and dyads.

librium coefficient δ. The three missing rates for fathers, mothers, and offspring are displayed in the first, second, and third number of the parenthesis in the fourth column. The three missing rates may differ. However, each of the three missing rates is identical for all genotypes B_1B_1, B_1B_2, and B_2B_2, such that the missing patterns are considered as MCAR. When there is no linkage ($\theta = 0.5$) and no association ($\delta = 0$), the TDT, 1-TDT, and EM-HRR have expected 5% chance of rejecting the null hypothesis. When there is linkage ($\theta = 0$) and association ($\delta = 0.05$), the power of TDT, 1-TDT, and EM-HRR are displayed in the bottom rows of Table 3. One can see that TIM has expected 5% type-I error regardless of the relationship between the marker and the disease alleles (independent of values of θ and δ).

When genotypes of ascertained offspring and parents are MCAR in an admixed population, the type-I error of TIM is displayed in Table 4. The disease allele (minor marker allele) frequency for the first and second populations are 0.2 and 0.6 (0.4 and 0.3), respectively. The disease penetrance and phenocopy rates are 0.4 and 0.2, respectively. Since TDT, 1-TDT, and EM-HRR are robust to population stratification, all tests have expected 5% error rates when there is no linkage ($\theta = 0.5$) and no association ($\delta = 0$). Under the alternative hypothesis with linkage ($\theta = 0$) and association ($\delta = 0.05$), TDT has the lowest power due to the exclusion of dyads in the analysis. Therefore, the 1-TDT is more powerful than TDT and EM-HRR has the highest power for detecting linkage and association, matching previous reports.[4,7] Since TIM has expected 5% type-1 error in the extreme scenarios we simulated, TIM is also robust to population admixture.

Simulation results displayed in Tables 5 and 6 are circumstances under which genotypes of trios are missing informatively in a homogeneous population. The disease and marker allele frequencies are 0.3 and 0.4, respectively. The disease penetrance and phenocopy rate are 0.4 and 0.2, respectively. Within each disease model, we first display four scenarios where informative missingness of genotypes occurred in parents only and genotypes of ascertained offspring are MCAR (20% missing rates for all genotypes). Secondly, we introduced informative missingness for genotypes of ascertained offspring as well as their parents and the results are displayed in rows 5–8 within each disease model.

In Table 5, when there is no association ($\delta = 0$) and no linkage ($\theta = 0.5$) and offspring genotypes are MCAR (rows 1–4 in each disease model), the TDT using the subset of complete triads remains a valid test for linkage and association with expected 5% type-I error. The 1-TDT and EM-HRR, using both triads and dyads, had inflated type-I errors and the inflation increased with respect to magnitude and pattern of informative missingness. TIM has better power to detect informative missingness in the more frequent parental genotypes (B_2B_2) than the less frequent genotypes (B_1B_1). When offspring genotypes are missing informatively (rows 5–8 in each disease model), by excluding dyads from the analysis, TDT is no longer valid for testing linkage and association. However, incorporation of dyads and monads reduced such biases.[10] The simulation results suggest that the TIM maintains good power when offspring genotypes are also missing informatively.

In Table 6, when there was association ($\delta = 0.05$) and linkage ($\theta = 0$), the power of the 1-TDT and EM-HRR can be lower or higher than TDT, suggesting that incorporating dyads either dampened or inflated the power of those tests when the MCAR assumption was violated, matching the investigations by Guo *et al*.[10,13] In the scenarios we examined, the TIM has decent power to detect informative missingness and its performance is closely related to the missing data pattern. Power of TIM was not confounded by linkage and association between the disease locus and the marker.

Application to the Framingham Heart Study

The Framingham Heart Study began in 1948 with the enrollment of 5209 men and women.[14,15] In 1971, 5124

Table 4 Type-I error (%) of TIM at $\alpha = 0.05$ under population admixture based on 10 000 replicates

Model	θ	δ	*Missing rates 1*	*Missing rates 2*	*TDT*	*1-TDT*	*EM-HRR*	*TIM*
Dominant	0.5	0	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	5.0	5.0	5.3	3.9
	0.5	0	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	5.0	5.2	5.5	3.8
	0.5	0	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	5.6	5.0	5.9	4.4
Additive	0.5	0	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	5.6	4.9	5.2	4.2
	0.5	0	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	5.0	4.9	5.2	4.2
	0.5	0	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	4.7	4.6	5.3	4.4
Recessive	0.5	0	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	5.3	5.1	5.5	4.2
	0.5	0	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	5.0	4.6	5.1	4.4
	0.5	0	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	5.2	4.9	5.3	4.7
Dominant	0	0.05	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	11.1	13.7	15.2	4.2
	0	0.05	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	12.4	14.4	15.5	4.3
	0	0.05	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	12.6	14.5	13.4	4.7
Additive	0	0.05	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	22.7	29.7	33.3	3.9
	0	0.05	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	27.1	33.7	36.4	4.2
	0	0.05	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	25.8	32.0	31.1	4.7
Recessive	0	0.05	(0.3; 0.3; 0.3)	(0.3; 0.3; 0.3)	7.6	9.0	9.9	4.4
	0	0.05	(0.3; 0.1; 0.2)	(0.3; 0.1; 0.2)	8.1	9.0	9.7	4.1
	0	0.05	(0.3; 0.1; 0.2)	(0.2; 0.4; 0.1)	8.4	9.5	8.9	4.3

Abbreviations: EM-HRR, expectation maximization algorithm based haplotype relative risk; MCAR, missing completely at random; TDT, transmission/disequilibrium test; TIM, testing informative missingness.
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; penetrance rate = 0.4; phenocopy rate = 0.2.
Missing rates 1 and 2 denote missing parameters for the first and second population, respectively.
The first, second, and third number in parenthesis are missing rates for fathers, mothers, and offspring, respectively. The missing rates for all three genotypes B_1B_1, B_1B_2, and B_2B_2.are identical (MCAR).
TDT uses complete trios only; 1-TDT, EM-HRR, and TIM use complete trios and dyads.

men and women were enrolled into the Framingham Offspring Study, which included the offspring (and their spouses) of the original cohort. Offspring participants underwent examinations approximately every 4 years; the design and methodology have been previously described.[16,17] The sample analyzed was comprised of Framingham Offspring Study participants who attended the sixth examination cycle between 1995 and 1998 and the apolipoprotein E (apoE) genotypes of the first generation cohort. The Framingham Heart Study protocol is approved by the Boston Medical Center Institutional Review Board and all participants provided written informed consent.

Ordovas *et al*[18] reported evidence for association of the apoE isoform with elevated total cholesterol (TC) levels in the Framingham Heart Study. Jarvik *et al*[19] addressed the possible influence of apoE genotype on age-related changes in TC from a male twin longitudinal study. Several studies of unrelated subjects also reported association between the apoE gene and TC.[20–24]

Because genotypes of the first generation cohort were collected nearly 40 years after the initialization of the study, it has been questioned whether the missing data pattern of parental genotypes of the Framingham Heart Study was affected by potential survival bias. Therefore, we applied TIM to the relation of elevated total cholesterol and APOE genotype. The apoE gene has three common alleles, which are apoE2, apoE3, and apoE4 and genotype frequencies are 0, 0.089, 0.021, 0.642, 0.223, and 0.025 for E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, and E4/E4, respectively. We adopted a similar approach implemented by Guo *et al*[25] to combine the rare allele apoE2 (associated with low TC) with the major allele apoE3 to compare those with at least one apoE4 allele (associated with high TC) to those without any. Of the 3532 participants attending the sixth offspring examination, there were 1041 individuals with at least one parental APOE genotype. Among the 1044, there were 472 with elevated total cholesterol (greater than 200 mg per 100 ml), deriving from 427 independent nuclear families. Therefore, there were 229 dyads and 198 triads included in the analysis.

In Table 7, we display the distribution of parental genotypes among dyads and triads by offspring genotypes. The logistic regression of missing status yielded a *P*-value of 0.8624 for parental genotype adjusting for offspring genotype. Therefore, there was no statistically significant evidence of informative missingness of parental APOE genotypes in the families with offspring who have elevated total cholesterol at the Framingham Heart Study.

Table 5 Power (%) of TIM at $\alpha = 0.05$ with no linkage ($\theta = 0.5$) and no association ($\delta = 0$) based on 1000 replicates

Model	*Father's missing rates*	*Mother's missing rates*	*Offspring's missing rates*	*TDT*	*1-TDT*	*EM-HRR*	*TIM*
Dominant	(0.0, 0.2)	(0.0, 0.2)	(0.2, 0.2, 0.2)	5.2	14.0	20.4	90.2
	(0.2, 0, 0)	(0.2, 0, 0)	(0.2, 0.2, 0.2)	5.1	15.3	19.5	65.7
	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	(0.2, 0.2, 0.2)	3.9	38.0	48.1	91.8
	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	(0.2, 0.2, 0.2)	5.2	37.4	53.6	59.8
	(0, 0, 0.2)	(0, 0, 0.2)	(0, 0, 0.2)	25.9	6.5	5.9	93.3
	(0.2, 0, 0)	(0.2, 0, 0)	(0.2, 0, 0)	12.9	5.3	6.5	64.6
	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	65.9	10.7	13.8	94.4
	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	30.5	3.7	4.7	66.5
Additive	(0, 0, 0.2)	(0, 0, 0.2)	(0.2, 0.2, 0.2)	4.5	14.8	21.4	91.5
	(0.2, 0, 0)	(0.2, 0, 0)	(0.2, 0.2, 0.2)	4.4	13.0	19.2	64.8
	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	(0.2, 0.2, 0.2)	5.1	39.8	48.3	89.7
	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	(0.2, 0.2, 0.2)	4.7	34.4	49.9	61.0
	(0, 0, 0.2)	(0, 0, 0.2)	(0, 0, 0.2)	27.5	6.3	5.8	92.0
	(0.2, 0, 0)	(0.2, 0, 0)	(0.2, 0, 0)	13.9	5.4	5.8	65.6
	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	69.8	12.2	14.6	95.7
	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	30.2	5.1	6.9	63.3
Recessive	(0, 0, 0.2)	(0, 0, 0.2)	(0.2, 0.2, 0.2)	6.4	15.9	21.7	91.1
	(0.2, 0, 0)	(0.2, 0, 0)	(0.2, 0.2, 0.2)	5.5	14.8	21.2	67.0
	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	(0.2, 0.2, 0.2)	4.6	38.8	48.6	90.0
	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	(0.2, 0.2, 0.2)	5.2	33.5	49.5	62.0
	(0, 0, 0.2)	(0, 0, 0.2)	(0, 0, 0.2)	25.4	6.9	6.0	91.5
	(0.2, 0, 0)	(0.2, 0, 0)	(0.2, 0, 0)	12.2	3.7	5.2	66.7
	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	65.8	11.4	15.2	94.8
	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	29.9	4.1	5.9	66.0

Abbreviations: EM-HRR, expectation maximization algorithm based haplotype relative risk; TDT, transmission/disequilibrium test; TIM, testing informative missingness.
Sample size = 500 families; disease allele frequency = 0.3; marker allele B_1 frequency = 0.4; penetrance rate = 0.4; phenocopy rate = 0.2.
The three numbers in the parenthesis are missing rates for the B_1B_1, B_1B_2, and B_2B_2 genotype, respectively.
TDT uses complete trios only; 1-TDT, EM-HRR, and TIM use complete trios and dyads.

Discussion

For the case–parent triads design, the TDT[2] cannot include families with incomplete parental genotypes. Approaches such as the 1-TDT[4] and EM-HRR[7] were designed to include such families due to missingness related to a disease in ascertainment, not to genotyping failure. They may be more powerful than the TDT[2] but are valid only if the missingness is not informative, that is, missingness independent of the underlying genotype (MAR). Although approaches proposed by Allen *et al*[8] and Chen[9] can include incomplete triads and are valid under informative missingness, they may not be as powerful as 1-TDT[4] and EM-HRR[7] when the missing data pattern is truly MCAR. Regardless of different missing data patterns among parental genotypes, the above methods assumed that offspring genotypes were MCAR. Recently, Guo[10] indicated that when offspring genotypes were missing informatively, a circumstance that can arise from ascertainment bias, inflated type-I error and/or reduced power may occur using the TDT excluding incomplete triads.

The purpose of this work is to provide a test for informative missingness in the context of case–parent triad designs for genetic linkage and/or association studies, in an effort to avoid a biased conclusion. Our approach compares the parental genotype distribution in triads to that of dyads conditional on the genotypes of affected offspring. Differential parental genotype distributions in triads and dyads indicate that parental genotypes are missing informatively. We have shown, through theoretical derivations and computer simulations, that TIM is not affected by linkage (θ) or association (δ). It provides expected 5% type-I error at $\alpha = 0.05$ level under MCAR and is robust to population admixture. Simulation results suggest that TIM has adequate power to test informative missingness in moderately sized sample. In the logistic regression framework, TIM remains a valid test under MAR by conditioning on available covariates X_1, X_2,…,X_K related to missingness.

Given a significant TIM result, assuming that informative missingness exists only in parental genotypes due to, for example, a late onset fatal disease such as cardiovascular disease, Allen *et al*[8] and Chen's[9] strategies are recommended to incorporate dyads. Otherwise, the 1-TDT[4] and EM-HRR[7] are appropriate and may provide higher power. However, one should be aware of the basic assumption of absence of ascertainment bias in all TDT/family-based association tests. If TIM is significant for an early onset fatal disease, one should be aware that none of existing methods yields a valid result, including the TDT with only complete trios, as illustrated and discussed by Guo.[10]

Table 6 Power (%) of TIM at $\alpha = 0.05$ with linkage ($\theta = 0$) and association ($\delta = 0.05$) based on 1000 replicates

Model	*Father's missing rates*	*Mother's missing rates*	*Offspring's missing rates*	*TDT*	*1-TDT*	*EM-HRR*	*TIM*
Dominant	(0.0, 0.2)	(0.0, 0.2)	(0.2, 0.2, 0.2)	17.1	4.5	4.5	90.5
	(0.2, 0.0)	(0.2, 0.0)	(0.2, 0.2, 0.2)	17.3	48.5	58.0	64.9
	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	(0.2, 0.2, 0.2)	14.3	10.2	13.1	91.7
	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	(0.2, 0.2, 0.2)	18.7	75.7	85.7	63.4
	(0.0, 0.2)	(0.0, 0.2)	(0.0, 0.2)	65.6	29.6	28.0	91.7
	(0.2, 0.0)	(0.2, 0.0)	(0.2, 0.0)	4.6	16.3	22.6	68.9
	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	89.1	36.5	43.7	94.9
	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	7.8	17.0	24.0	66.6
Additive	(0.0, 0.2)	(0.0, 0.2)	(0.2, 0.2, 0.2)	39.5	14.5	12.3	90.8
	(0.2, 0.0)	(0.2, 0.0)	(0.2, 0.2, 0.2)	42.6	76.3	83.7	69.1
	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	(0.2, 0.2, 0.2)	34.8	3.9	5.3	92.5
	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	(0.2, 0.2, 0.2)	15.2	75.7	85.7	64.6
	(0.0, 0.2)	(0.0, 0.2)	(0.0, 0.2)	84.1	54.0	51.3	93.7
	(0.2, 0.0)	(0.2, 0.0)	(0.2, 0.0)	11.9	37.4	46.7	69.1
	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	97.3	62.4	68.7	95.1
	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	6.7	17.4	24.3	67.0
Recessive	(0.0, 0.2)	(0.0, 0.2)	(0.2, 0.2, 0.2)	8.5	6.0	7.5	89.4
	(0.2, 0.0)	(0.2, 0.0)	(0.2, 0.2, 0.2)	10.1	33.5	41.8	66.3
	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	(0.2, 0.2, 0.2)	7.7	19.9	26.0	90.3
	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	(0.2, 0.2, 0.2)	18.1	75.5	86.4	62.3
	(0.0, 0.2)	(0.0, 0.2)	(0.0, 0.2)	49.2	17.2	15.5	91.7
	(0.2, 0.0)	(0.2, 0.0)	(0.2, 0.0)	5.6	9.0	12.7	68.0
	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	(0.05, 0.05, 0.4)	83.9	27.7	32.1	93.6
	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	(0.4, 0.05, 0.05)	6.5	15.6	23.7	69.1

Abbreviations: EM-HRR, expectation maximization algorithm based haplotype relative risk; TDT, transmission/disequilibrium test; TIM, testing informative missingness.
Sample size = 500 families; disease allele frequency = 0.3; marker allele B1 frequency = 0.4; penetrance rate = 0.4; phenocopy rate = 0.2.
The three numbers in the parenthesis are missing rates for the B_1B_1, B_1B_2, and B_2B_2 genotype, respectively.
TDT uses complete trios only; 1-TDT, EM-HRR, and TIM use complete trios and dyads.

Table 7 Distribution of APOE genotypes

Offspring genotype	*Parental genotype*	*Triads*	*Dyads*
33	33	154	129
	34	24	21
34	33	20	21
	34	22	18
	44	3	3
44	34	6	5
	44	0	1
Total		229	198

The proposed test TIM is developed for case–parent triads designs. However, it is readily applicable to other designs consisting of parents and affected offspring by selecting triads and dyads from the data. But this may not be the most powerful approach due to deletion of sibling information. Therefore, our future work will extend TIM to consider more general pedigrees. Many recently published genome-wide association studies (GWAs) are case–control designs, but there are also family-based GWAs with available genotyping on parent–offspring trios, such as the Framingham Heart Study and International Multi-Center ADHD Genetics Project (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap). It is likely that more family-based GWA studies using DNA already collected in the past emerge in the near future. The advantage of parent–offspring design over case–control design is that the former design is immune for population admixture, and it tests for linkage as well as association. Therefore, our method may be useful for these designs to further reduce false positives due to informative missingness in the modern era of genome-wide studies.

Acknowledgements

This work was supported by, in part, National Heart, Lung and Blood Institute's Framingham Heart Study (Contract No. N01-HC-25195). We acknowledge the support of the Genomics Program at Children's Hospital Boston. We thank anonymous reviewers and the editor for their insightful comments and suggestions.

References

1 Falk CT, Rubinstein P: Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* 1987; **51**: 227–233.
2 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the sinsulin gene region and insulin dependent diabetes mellitus. *Am J Hum Genet* 1993; **52**: 506–516.

3 Curtis DR, Sham PC: A note on the application of the transmission disequilibrium test when a parent is missing. *Am J Hum Genet* 1995; **56**: 811–812.
4 Sun F, Flanders W, Yang Q, Khoury J: Transmission disequilibrium test (TDT) with only one parent is available: the 1-TDT. *Am J Epidemiol* 1999; **150**: 97–104.
5 Clayton D: A generalization of the transmission/disequilibrium test for uncertain haplotype transmission. *Am J Hum Genet* 1999; **65**: 1170–1177.
6 Weinberg CR: Allowing for missing parents in genetic studies of case-parents triads. *Am J Hum Genet* 1999; **64**: 1186–1193.
7 Guo CY, Destefano AL, Lunetta KL, Dupuis J, Cupples LA: Expectation maximization algorithm based haplotype relative risk (EM-HRR): test of linkage disequilibrium using incomplete case-parents trios. *Hum Hered* 2005; **59**: 125–135.
8 Allen AS, Rathouz PJ, Satten GA: Informative missingness in genetic association studies: case-parent designs. *Am J Hum Genet* 2003; **72**: 671–680.
9 Chen YH: New approach to association testing in case-parent designs under informative parental missingness. *Genet Epidemiol* 2004; **27**: 131–140.
10 Guo CY: Validity of the transmission/disequilibrium test (TDT) under impact of complex informative missingness. *BMC Proc* 2007; **1** (Suppl 1): S26.
11 Ott J: Statistical properties of the haplotype relative risk. *Genet Epidemiol* 1989; **6**: 127–130.
12 Little RJA, Rubin DB: *Statistical Analysis with Missing Data*. New York: Wiley, 1987.
13 Guo CY, Cui J, Cupples LA: Impact of non-ignorable missingness on genetic tests of linkage and/or association using case-parents trios. *BMC Genet* 2005; **6** (Suppl 1): S90.
14 Dawber TR, Meadors GF, Moore FE: Epidemiologic approaches to heart disease: the Framingham study. *Am J Public Health* 1951; **41**: 279–286.
15 Dawber TR, Kannel WB, Lyell LP: An approach to longitudinal studies in a community: the Framingham heart study. *Ann NY Acad Sci* 1963; **107**: 539–556.
16 Feinleib M, Kannel WB, Garrison RJ, McNamara PM, Castelli WP: The Framingham offspring study. Design and preliminary data. *Prev Med* 1975; **4**: 518–525.
17 Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP: An investigation of coronary heart disease in families. The Framingham offspring study. *Am J Epidemiol* 1979; **110**: 281–290.
18 Ordovas JM, Litwack-Klein L, Wilson PW, Schaefer MM, Schaefer EJ: Apolipoprotein E isoform phenotyping methodology and population frequency with identification of apoE1 and apoE5 isoforms. *J Lipid Res* 1987; **28**: 371–380.
19 Jarvik GP, Austin MA, Fabsitz RR *et al*: Genetic influences on age-related change in total cholesterol, low density lipoprotein-cholesterol, and triglyceride levels: longitudinal apolipoprotein E genotype effects. *Genet Epidemiol* 1994; **11**: 375–384.
20 Jarvik GP, Goode EL, Austin MA *et al*: Evidence that the apolipoprotein E-genotype effects on lipid levels can change with age in males: a longitudinal analysis. *Am J Hum Genet* 1997; **61**: 171–181.
21 Kallio MJ, Salmenpera L, Siimes MA, Perheentupa J, Gylling H, Miettinen TA: The apolipoprotein E phenotype has a strong influence on tracking of serum cholesterol and lipoprotein levels in children: a follow-up study from birth to the age of 11 years. *Pediatr Res* 1998; **43**: 381–385.
22 Fulton JE, Dai S, Grunbaum JA, Boerwinkle E, Labarthe DR: Apolipoprotein E affects serial changes in total and low-density lipoprotein cholesterol in adolescent girls: Project HeartBeat!. *Metabolism* 1999; **48**: 285–290.
23 Srinivasan SR, Ehnholm C, Elkasabany A, Berenson G: Influence of apolipoprotein E polymorphism on serum lipids and lipoprotein changes from childhood to adulthood: the Bogalusa Heart Study. *Atherosclerosis* 1999; **143**: 435–443.
24 Hak AE, Witteman JC, Hugens W *et al*: The increase in cholesterol with menopause is associated with the apolipoprotein E genotype. A population-based longitudinal study. *Atherosclerosis* 2004; **175**: 169–176.
25 Guo CY, Lunetta KL, DeStefano AL, Ordovas JM, Cupples LA: Informative transmission disequilibirum test (i-TDT): combined linkage and association mapping that includes unaffected offspring as well as affected offspring. *Genet Epidemiol* 2007; **31**: 115–133.

Appendix 1

Distribution of ascertained triads and dyads

First, we assume that the data consists of genotypes of bi-allelic markers such as a single nucleotide polymorphism. Therefore, there are exactly two alleles, B_1 and B_2, at the marker locus. We first derive the distribution of complete triads as the following: Let G_o, G_f, G_m be the offspring's, father's, and mother's genotypes, respectively. Let G_{of} and G_{om} be the allele of offspring inherited from the father and mother, respectively. Then, G_o, when it is heterozygous, really represents a set of two possible pairs of values, ($G_{of}=B_1$, $G_{om}=B_2$) or ($G_{of}=B_2$, $G_{om}=B_1$). Let I_f, I_m, and I_o be binary indicator functions for father, mother, and offspring having missing genotype information. For example, $I_f=1$ if the father's genotype is missing and 0 otherwise.

Here, we do not consider imprinting and the four possible joint probabilities of a given parental genotype and the probability of transmitting a given allele to the offspring from that parent, all conditional on offspring affected status are:

$$\mu = \Pr(G_f = (B_1B_1) \& G_{of} = (B_1) \text{ or } G_m = (B_1B_1) \& G_{om} = (B_1) | \text{affected offspring})$$

$$\upsilon = \Pr(G_f = (B_1B_2) \& G_{of} = (B_1) \text{ or } G_m = (B_1B_2) \& G_{om} = (B_1) | \text{affected offspring})$$

$$\zeta = \Pr(G_f = (B_1B_2) \& G_{of} = (B_2) \text{ or } G_m = (B_1B_2) \& G_{om} = (B_2) | \text{affected offspring})$$

$$\tau = \Pr(G_f = (B_2B_2) \& G_{of} = (B_2) \text{ or } G_m = (B_2B_2) \& G_{om} = (B_2) | \text{affected offspring})$$

When the disease model is recessive, Ott[11] (Table 2) showed that $\mu=(s+\delta/r)s$, $\upsilon=(s+\delta/r)\ (1-s)-\theta\delta/r$, $\zeta=(1-s-\delta/r)s+\theta\delta/r$ and $\tau=(1-s-\delta/r)\ (1-s)$, where '$r$' is the allele frequency of the recessive disease allele 'a', and 's' is the allele frequency of marker allele 'B_1'. The parameter θ denotes the recombination fraction, and $\delta=p(aB_1)-p(a)p(B_1)$ denotes the disequilibrium coefficient between the marker and the disease locus. The conditional probabilities under the dominant or additive disease model can be derived similarly.

Assuming random mating and no missing parental genotype in the population, the probability of ascertaining a triad with the father, mother, and affected offspring's

genotypes being B_1B_1, B_1B_2, and B_1B_2, respectively, is $\Pr(G_f = (B_1B_1); G_m = (B_1B_2); G_o = (B_1B_2)|\text{affected offspring}) = \mu \times \zeta$.

However, it is unrealistic to assume the completeness of parental genotypes when collecting sample. For example, parental genotypes may not be available due to death from the disease under study (ie, missing pattern of parental genotypes is related to the disease under study or informative missingness) or due to random refusal of participation (MCAR). Allowing for differential missing rates for offspring, fathers, and mothers, let P_{o11}, P_{o12}, and P_{o22} denote missing rates for offspring with B_1B_1, B_1B_2, and B_2B_2 genotypes, respectively. Similarly, let P_{f11}, P_{f12}, and P_{f22} (P_{m11}, P_{m12}, and P_{m22}) denote missing rates for father (mother) with B_1B_1, B_1B_2, and B_2B_2 genotypes, respectively. Note that we do not assume any pattern for the nine missing parameters, ie, missingness of a given parent's genotype can be dependent or independent of the other parent's and/or offspring's genotype.

Take the above missing parameters into consideration, the conditional probability of ascertaining a complete triad with the father, mother, and affected offspring's genotypes being B_1B_1, B_1B_2, and B_1B_2, respectively, is $\Pr(I_f = 0\& G_f = (B_1B_1);\ I_m = 0\&G_m = (B_1B_2);\ I_o = 0\&G_o = (B_1B_2)|\text{affected offspring}) = \mu \times \zeta \times (1-P_{f11}) \times (1-P_{m12}) \times (1-P_{o12})$. The rest probabilities can be derived in the same manner and are displayed in Table 1.

Null distribution of TIM under MCAR

If genotypes are MCAR, then the probability of missing a subject is independent of the subject's genotype, ie, $P_{o11} = P_{o12} = P_{o22} = P_o$, $P_{f11} = P_{f12} = P_{f22} = P_f$, $P_{m11} = P_{m12} = P_{m22} = P_m$. Note that P_o, P_f, and P_m are three parameters and need not to be identical. In addition, $\sum_{i,j,k} P_i^{j,k} = (1-P_f)\times(1-P_m)\times(1-P_o)$ and $\sum_{i,j} P_i^{j,*} = [(1-P_f)P_m + P_f(1-P_m)]\times(1-P_o)$. Under the null hypothesis of MCAR and conditional on genotypes of affected offspring B_1B_1, the proportion of parents with B_1B_1 genotypes among triads

$$\left[\left(P_2^{2,2} + \frac{P_2^{1,2} + P_2^{2,1}}{2}\right) \Big/ \sum_{i,j,k} P_i^{j,k}\right]$$

is equivalent to that of dyads

$$\left[\left(P_2^{2,*} + P_2^{*,2}\right) \Big/ \sum_{i,j} P_i^{j,*}\right],$$

which are both equivalent to $\mu \times (\mu + \nu)$. The rest conditional null distributions can be derived in a similar manner.