European Journal of Human Genetics (2005) 13, 1007–1012

www.nature.com/ejhg

ARTICLE

Combining the case–control methodology with the small size transmission/disequilibrium test for multiallelic markers

Wei Guo[1] and Wing K Fung*[,1]

[1]*Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China*

Case–control studies compare marker-allele distributions in affected and unaffected individuals, and significant results may be due to linkage but can also simply reflect population structure. To test for linkage after obtaining a significant case–control finding, within-family analysis can be performed. In a transmission/disequilibrium test (TDT), genotypes of cases are compared to those of their parents to explore whether a specific allele, or marker, at a locus of interest is transmitted to a greater degree than Mendelian inheritance would warrant. For multiallelic markers, several authors have proposed extensions to the TDT. In this article, we propose a TDT test, utilizing the available information of a case–control study in the grouping of alleles for multiallelic markers, and thereby increase the statistical power of a TDT test with a small sample size.
European Journal of Human Genetics (2005) **13,** 1007–1012. doi:10.1038/sj.ejhg.5201453;
published online 15 June 2005

Keywords: linkage disequilibrium (LD); case–control; transmission/disequilibrium test (TDT); multiallelic marker

Introduction

The transmission/disequilibrium test (TDT)[1] is a powerful method for testing linkage between a marker and the disease gene in the presence of association.[2–4] Case–control studies compare marker-allele distributions in affected and unaffected individuals; when a significant result is obtained, it may be due to linkage or population structure. To test for linkage after obtaining a significant case–control result, within-family tests can be performed. In a TDT test, genotypes of cases are compared to those of their parents to explore whether a specific allele, or marker, at a locus of interest is transmitted to a greater degree than Mendelian inheritance would warrant. In order to integrate all available information and thereby increase the statistical power, Nagelkerke *et al*[5] proposed a new TDT statistic by combining a TDT test and a case–control result using the generalized logistic regression.

For multiallelic markers, there are a number of extensions to the TDT after its debut.[6–10] In this article, we develop a new approach by combining a small size TDT test and a case–control result for multiallelic markers. As a measure of association, linkage disequilibrium (LD) has a great effect on the power of the TDT test, so we explore the LD sign (positive or negative) using the case–control data, then use the biallelic TDT statistic computed for one allele subset including alleles with the same LD sign *versus* all other alleles combined. We find from a simulation study that it is possible to increase the power of the TDT test when the sample size is relatively small.

In the following sections, we review some existing TDT tests for multiallelic markers. Then we introduce our new approach, involving the combination of a TDT test and a case–control result. Next, we describe the steps of our proposed simulation study, and based on the simulation results we evaluate the performance of the test statistics. Finally, a few concluding remarks are given in the discussion section.

*Correspondence: Professor WK Fung, Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China. Tel: +852 2859 1988; Fax: +852 2858 9041;
E-mail: wingfung@hku.hk
Received 14 February 2005; revised 17 May 2005; accepted 18 May 2005; published online 15 June 2005

Method

We consider a random mating population in which Hardy–Weinberg equilibrium is assumed. Suppose a biallelic disease locus has alleles D and d. Consider a multiallelic marker with L alleles, $M_1, \ldots, M_L$. Let the allele frequencies of M_i and D be p_i and q, respectively; and the frequencies of haplotype M_iD and $M_i d$ be h_{i1} and h_{i2}, respectively. The LD between the marker allele M_i and the disease allele D is given as $\Delta_{(i)} = h_{i1} - p_i q$. Suppose the penetrances of the disease given genotypes DD, Dd and dd are f_2, f_1 and f_0, respectively, and $Pr(A) = f_2 q^2 + 2f_1 q(1-q) + f_0(1-q)^2$ represents the prevalence of the disease in the population. Let θ be the recombination fraction between the marker locus and the disease locus. For $i = 1, \ldots, L$, $j = 1, \ldots, L$, let n_{ij} denote the number of those parents who transmit the M_i allele but not the M_j allele to their affected children. Let $n_{i\cdot} = \sum_{j \neq i} n_{ij}$ denote the number of heterozygous parents who transmit the M_i allele, and let $n_{\cdot i} = \sum_{j \neq i} n_{ij}$ denote the number of heterozygous parents who have the M_i allele but do not transmit it. Homozygous parents are not included in the sample, as they are noninformative for the transmission tests. For the allele pair M_i and M_i^c, the biallelic TDT statistic is given by

$$TDT_{(i)} = \frac{(n_{i\cdot} - n_{\cdot i})^2}{(n_{i\cdot} + n_{\cdot i})}, \tag{1}$$

which asymptotically follows a χ^2 distribution with one degree of freedom under the null hypothesis of no linkage.

Existing test statistics

For multiallelic markers, there are a number of extensions to the TDT test. For example, the generalized TDT (GTDT) statistic[9] is proposed as,

$$GTDT = \boldsymbol{d}' \boldsymbol{V}^{-1} \boldsymbol{d}, \tag{2}$$

where $\boldsymbol{d}' = (d_1, d_2, \ldots, d_{L-i\cdot})$, $d_i = n_{i\cdot} - n_{\cdot i}$, and $\boldsymbol{V}$ is the estimate of the variance and covariance matrix. A simpler test statistic, Tmhet,[7] is given as

$$Tmhet = \frac{L-1}{L} \sum_{i=1}^{L} \frac{(n_{i\cdot} - n_{\cdot i})^2}{(n_{i\cdot} + n_{\cdot i})}. \tag{3}$$

Under the null hypothesis, both the GTDT and Tmhet statistics follow asymptotically a χ^2 distribution with $L-1$ degrees of freedom, and both reduce to the biallelic TDT statistic when $L = 2$.

The maximal TDT statistic, maxTDT,[9] is defined as

$$\max TDT = \max_i TDT_{(i)}. \tag{4}$$

Since the exact and asymptotic distributions of the maxTDT are not available, the critical value at a given level of significance α is determined using the simulation method proposed by Kaplan *et al.*[11]

Combining TDT with case–control (ccTDT)

When a case–control study is carried out first and a TDT study is carried out subsequently within the same population to corroborate case–control findings independently, it is possible to combine the TDT test and a case–control comparison in order to integrate all available information, particularly when the sample size of the TDT test is not large enough to detect linkage for multiallelic markers. We try to find the LD sign of each marker allele through the prior case–control analysis so that the biallelic TDT test can be constructed through combining alleles with the same LD signs.

Suppose for simplicity m unrelated cases and m controls are sampled randomly from the population. For the multiallelic marker, let t_{1i} and t_{2i}, $i = 1, \ldots, L$, be the numbers of allele i in cases and controls, respectively. The usual χ^2 test for allele i is

$$\chi^2_{(i)} = \frac{4m(t_{1i} - t_{2i})^2}{t_{\cdot i}(4m - t_{\cdot i})}, \tag{5}$$

where $t_{\cdot i} = t_{1i} + t_{2i}$, is the number of allele i in both cases and controls. The frequencies of the marker allele i in the case and control groups can be calculated as follows

$$\begin{aligned} &Pr(M_i|\text{case}) \\ &= Pr(M_i M_i^c|\text{case})/2 + Pr(M_i M_i|\text{case}) \\ &= \{[f_2 Pr(M_i D/M_i^c D) + f_1 Pr(M_i D/M_i^c d) \\ &+ f_1 Pr(M_i d/M_i^c D) + f_0 Pr(M_i d/M_i^c d)]/2 \\ &+ [f_2 Pr(M_i D/M_i D) + f_1 Pr(M_i D/M_i d) \\ &+ f_0 Pr(M_i d/M_i d)]\}/Pr(A) \\ &= p_i + \Delta_{(i)}[(f_2 - f_1)q + (f_1 - f_0)(1-q)]/Pr(A), \end{aligned} \tag{6}$$

and similarly, $Pr(M_i|\text{control}) = p_i - \Delta_{(i)}\ [(f_2 - f_1)q + (f_1 - f_0)(1-q)]/(1 - Pr(A))$. So the expectation of the difference of allele numbers,

$$\begin{aligned} E\{t_{1i} - t_{2i}\} &= 2m[Pr(M_i|\text{case}) - Pr(M_i|\text{control})] \\ &= 2m\Delta_{(i)}[(f_2 - f_1)q + (f_1 - f_0)(1-q)]/ \\ &\quad [Pr(A)(1 - Pr(A))], \end{aligned} \tag{7}$$

is determined by the LD measure $\Delta_{(i)}$, and it is possible to use this difference to estimate the signs of the LD. Naturally, we might expect to construct a more powerful test by grouping those alleles with positive (or negative) signs of $\Delta_{(i)}$ as a single allele in the biallelic TDT test. The details of our TDT test combining with the case–control information, ccTDT, are given below:

Step 1: Group all L marker alleles according to the signs of their LD estimates, and suppose $\Lambda^+ = \{j: t_{1j} - t_{2j} \geq 0\}$, and $\Lambda^- = \{j:\ t_{1j} - t_{2j} < 0\}$;

Step 2: For any allele subset $G = \{i_1, \ldots, i_g\}$, where $1 \leq g \leq L$, we can calculate the biallelic case–control χ^2 test statistic which regards the allele sets $\{i_1, \ldots, i_g\}$ and $\{i_1, \ldots, i_g\}^c$ as two single alleles,

$$\chi^2_{(i_1, \ldots i_g)} = \frac{4m\left(\sum_{k=1}^{g} t_{1i_k} - \sum_{k=1}^{g} t_{2i_k}\right)^2}{\left(\sum_{k=1}^{g} t_{\cdot i_k}\right)\left(4m - \sum_{k=1}^{g} t_{\cdot i_k}\right)}; \tag{8}$$

Step 3: Take the subset G that maximizes the biallelic χ^2 test statistic for all subsets of Λ^+ and Λ^- as follows,

that is

$$\chi^2_G = \max\{\chi^2_G : G \subseteq \Lambda^+, \quad \text{or} \quad G \subseteq \Lambda^-\}.$$

According to the subset G obtained by the case–control data, the usual biallelic TDT test can be constructed as

$$ccTDT = \frac{[\sum_{k=1}^{g}(n_{i_k\cdot} - n_{\cdot i_k})]^2}{\sum_{k=1}^{g}(n_{i_k\cdot} + n_{\cdot i_k}) - 2\sum_{r=1}^{g}\sum_{s=r+1}^{g}(n_{i_r i_s} + n_{i_s i_r})}, \quad (9)$$

which follows a χ^2 distribution with one degree of freedom under the null hypothesis of no linkage, because the subset G is determined independently by the case–control data before the TDT test is constructed.

Simulation

Simulation design

In this section, we make a power comparison between the GTDT, Tmhet, maxTDT and ccTDT tests at the $\alpha = 0.05$ significance level. Three disease models of inheritance are considered: (1) recessive model $f_2 = 1$, $f_1 = f_0 = 0$; (2) additive model $f_2 = 1$, $f_1 = 0.5$, $f_0 = 0$; (3) dominant model $f_2 = f_1 = 1$, $f_0 = 0$. Suppose D is the disease allele with population frequency $q = 0.01$, and d is the normal allele.

Since the LD is the most important factor affecting the power of the association and linkage tests for the multi-allelic marker, similar to the population design of Kaplan *et al*[11] we design the simulated populations according to the LD situation and the association index I^* where

$$I^* = \sum_{i=1}^{L} \frac{[Pr(M_i|\text{case}) - Pr(M_i|\text{control})]^2}{Pr(M_i|\text{case}) + Pr(M_i|\text{control})} \quad (10)$$

is based on the theory of testing for the equality of two independent multinomial distributions.[12]

In the first simulation study, we aim to investigate the performance of the test statistics under different LD situations. In all, 200 cases, 200 controls and 100 trios are taken independently from an identical population. A six-allelic marker with equal frequency $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$ is taken. As shown in Table 1, we consider six LD modes as (1) one positive peak; (2) two positive peaks; (3) one positive peak and one negative peak; (4) two positive peaks and one negative

Table 1 LD situations for a six-allelic marker in simulated populations under recessive model

				$\Delta_{(i)}$			I^*
Population	*i* = 1	*i* = 2	*i* = 3	*i* = 4	*i* = 5	*i* = 6	*Recessive*
1	*0.001614*	−0.000323	−0.000323	−0.000323	−0.000323	−0.000323	0.07
2	*0.000925*	*0.000925*	−0.000462	−0.000462	−0.000462	−0.000462	0.07
3	*0.001025*	0.000050	0.000050	*−0.001025*	−0.000050	−0.000050	0.07
4	*0.000627*	*0.000627*	0.000004	*−0.001061*	−0.000097	−0.000097	0.07
5	*0.000744*	*0.000744*	0.000050	*−0.000744*	*−0.000744*	−0.000050	0.07
6	0.000613	0.000613	0.000613	−0.000613	−0.000613	−0.000613	0.07

Italic values are the peaks of the linkage disequilibrium at L marker alleles.

Table 2 Conditional marker allele distributions and LD situations for a seven-allelic marker in simulated populations under recessive model

					Allele				I^*
Population		*i* = 1	*i* = 2	*i* = 3	*i* = 4	*i* = 5	*i* = 6	*i* = 7	*Recessive*
7	*Unimodal*								
	$Pr(M_i\|D)$	*0.7836*	0.0361	0.0361	0.0361	0.0361	0.0361	0.0361	0.07
	$Pr(M_i\|d)$	*0.6100*	0.0650	0.0650	0.0650	0.0650	0.0650	0.0650	
	$\Delta_{(i)}$	*0.0017*	−0.0003	−0.0003	−0.0003	−0.0003	−0.0003	−0.0003	
	$f(M_i)$	*0.6117*	0.0647	0.0647	0.0647	0.0647	0.0647	0.0647	
8	*Bimodal*								
	$Pr(M_i\|D)$	*0.2055*	*0.2055*	0.1180	0.1180	0.1180	0.1180	0.1169	0.07
	$Pr(M_i\|d)$	*0.3000*	*0.3000*	0.0800	0.0800	0.0800	0.0800	0.0800	
	$\Delta_{(i)}$	*−0.0009*	*−0.0009*	0.0004	0.0004	0.0004	0.0004	0.0004	
	$f(Mi)$	*0.2991*	*0.2991*	0.0804	0.0804	0.0804	0.0804	0.0804	
9	*Uniform*								
	$Pr(M_i\|D)$	0.2052	0.2052	0.2052	0.0961	0.0961	0.0961	0.0961	0.07
	$Pr(M_i\|d)$	0.1422	0.1422	0.1422	0.1433	0.1433	0.1433	0.1433	
	$\Delta_{(i)}$	0.0006	0.0006	0.0006	−0.0005	−0.0005	−0.0005	−0.0005	
	$f(M_i)$	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	0.1429	

Italic values are the peaks of the linkage disequilibrium at L marker alleles.

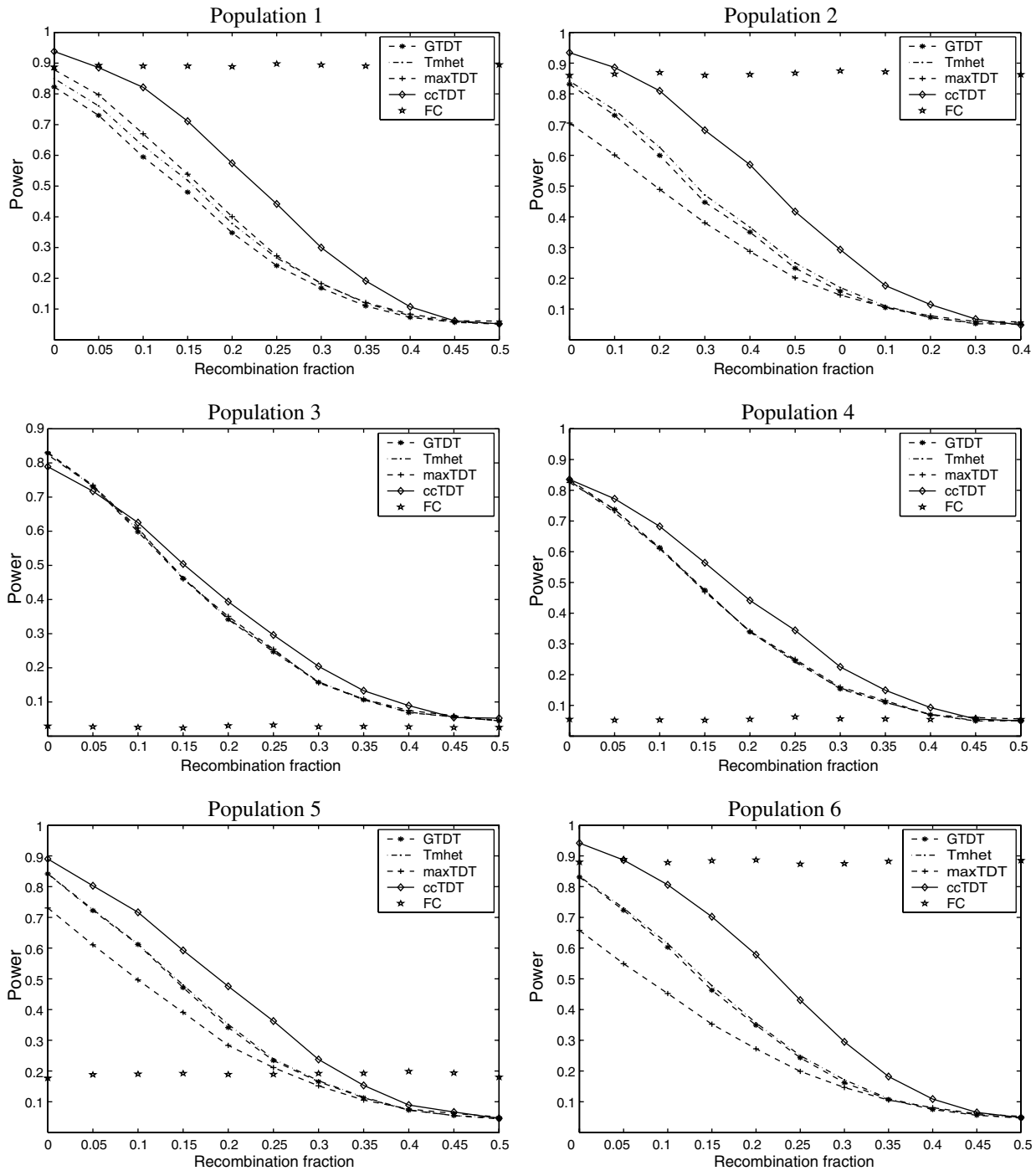

Figure 1 Power comparisons of the GTDT, Tmhet, maxTDT and ccTDT tests for populations 1–6 given in Table 1 under the recessive model. FC is the frequency of replicates in which the 'correct' subset of positively associated markers is identified. The number of the cases, controls and trios are 200, 200 and 100, respectively. The power is based on 5000 replications.

peak; (5) two positive peaks and two negative peaks; (6) equal LD magnitude ($|\Delta_{(i)}|$) of six LD between the disease allele and each marker allele, where the peaks are in italics. In this simulation study, we also present the frequency of replicates in which the 'correct' subset of positively associated markers (FC) is identified.

In the second simulation study, we investigate the effect of the sample size of the case–control comparison and

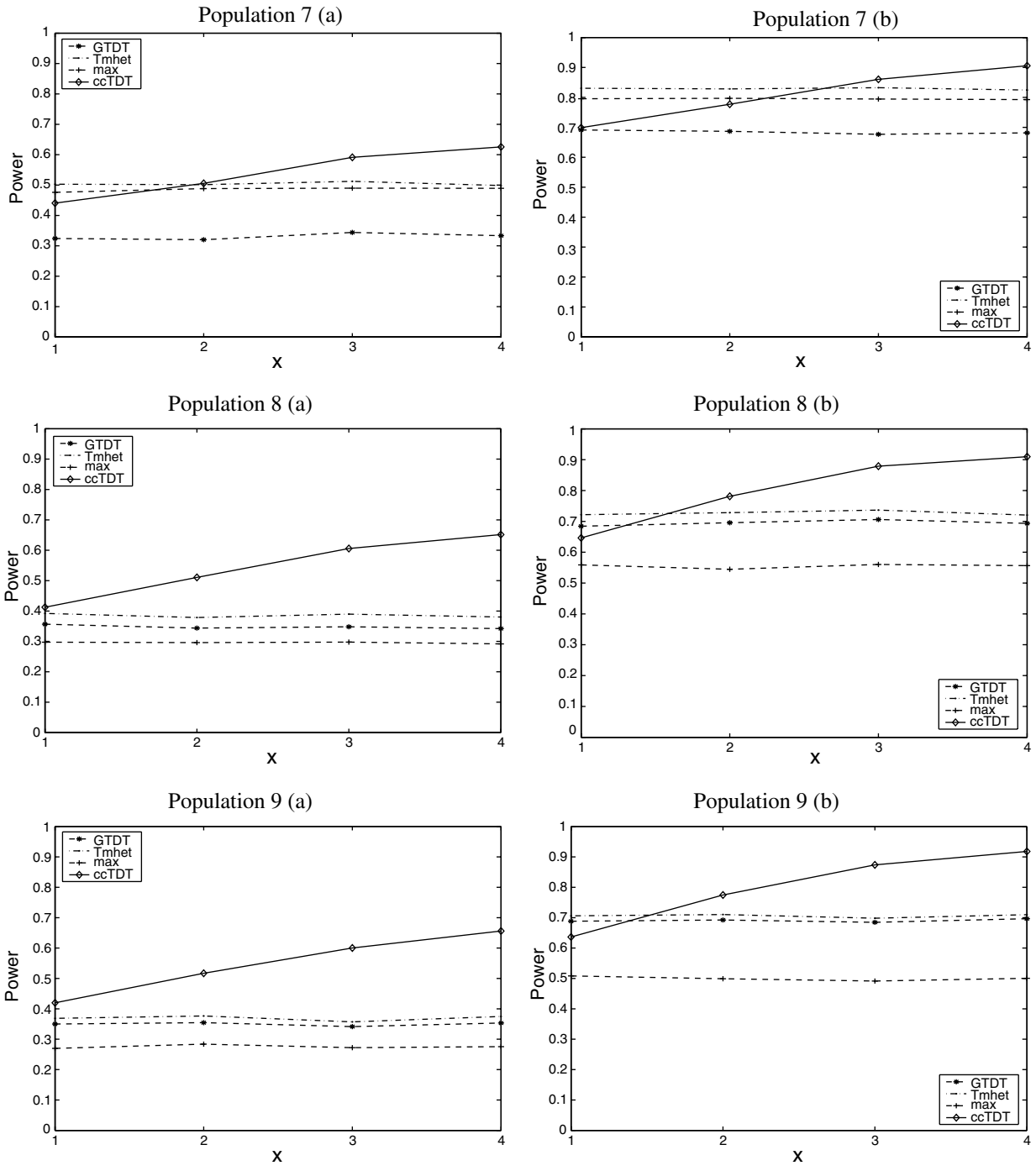

Figure 2 Power comparisons of the GTDT, Tmhet, maxTDT and ccTDT tests for populations 7–9 given in Table 2 under the recessive model for $\theta = 0.05$. The equal sizes of the case–control are 50, 100, 200 and 400, which are, respectively, denoted as 1, 2, 3 and 4 in the *x*-axis. The numbers of trios are (**a**) 50 and (**b**) 100. The power is based on 5000 replications.

the TDT trios, where the equal sizes of cases and controls are taken as 50, 100, 200 or 400, and the number of trios used in the TDT tests is 50 or 100 when the recombination fraction is fixed at 0.05. As shown in Table 2, we consider the unimodal, bimodal and uniform conditional marker allele distributions in these three populations, which are analogous to the population design given by Kaplan *et al.*[11]

Simulation procedure

The steps of the simulation study are given below:

1. Specify (a) the frequencies of L marker alleles $M_1, \ldots, M_L$ and the disease allele D, $p_1, \ldots, p_L$ and q, (b) the coefficients of LD between the marker allele M_i and the disease allele D, $\Delta_{(i)}$, $i=1, \ldots, L$ in a random mating population.
2. Sample. the genotype data of m cases and m controls, then look for the allele subset G based on the case–control result.
3. Sample N case–parents trios by the multinomial distribution based on the transmission probabilities according to Kaplan *et al*[11] and obtain the values of the four test statistics: GTDT, Tmhet, maxTDT and ccTDT.
4. For each of the test statistics, reject H_0 if the statistic is larger than its asymptotic or simulated critical value. The simulated critical value of the maxTDT test is obtained by 5000 replications.
5. Repeat steps 1–4, 5000 times.

Simulation results

In Figure 1, we demonstrate the size and power comparison of the six LD situations for a six-allelic marker locus under the recessive model. When there is no linkage $\theta=0.5$, all four tests control the size $\alpha=0.05$ well. When there is linkage between the marker and the disease gene, for all six populations, the ccTDT test is more powerful than the other three TDT tests. We should note that for populations 1, 2 and 6, the frequencies of the replicates in which the 'correct' subsets of positively associated markers (FC) are identified are about 90%, while for populations 3–5 the frequencies are much lower. Nevertheless, in almost 100% of the replicates the positive and negative LD peaks are classified correctly for the latter populations (results omitted). All tests achieve the highest power under the recessive model, but each performs similar for both the additive and the dominant models (results not shown). For the TDT tests on multiallelic markers, the GTDT and Tmhet are found to have a similar performance and achieve a higher power when several alleles are more or less equally associated.

In Figure 2, we investigate the effect of the size of the case–control sample, that is, 50, 100, 200 and 400, to the small trio size (a) 50 and (b) 100. Since the case–control samples are not used by the GTDT, Tmhet and maxTDT tests, their power curves in Figure 2 are almost horizontal (they are not entirely horizontal due to simulation variation). For populations 8 and 9, the ccTDT test has a better performance than the other tests when the numbers of cases and controls are larger than 50 and 100 and the number of trios are 50 and 100, respectively. However, for population 7, the ccTDT test does not outperform the others until the numbers of cases and controls increase to 200. As the sample size of the case–control increases, the power of the ccTDT test increases as well (Figure 2).

Discussion

In this article we investigate an extension of the TDT test, utilizing the available information of a case–control study in order to increase the statistical power of a TDT test with a small-sized sample. As shown by several investigators, the TDT test is valid for linkage detection when LD exists, and the power of the TDT test depends on the magnitude of LD.[6,8,11] Based on this property, we develop a new test under Hardy–Weinberg equilibrium, ccTDT, which uses the biallelic TDT statistic computed for one allele subset including alleles with the same sign of LD *versus* all others combined based on the information of a case–control sample. A nice property of ccTDT is that its asymptotic distribution is known and the critical value can be easily determined, which is not the case for maxTDT.

The simulation findings demonstrate that the ccTDT performs well for the small-sized trios, but also performs more powerfully as the size of the case–control sample increases. The test is expected to have a good power, especially for bimodal and uniform models.

Acknowledgements

We thank two referees and David Wilmshurst for helpful comments that improved the presentation of the paper. This project is partly supported by the research Grant NSF 10329102 of China.

References

1 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. *Am J Hum Genet* 1993; **52**: 506–516.
2 Ewens WJ, Spielman RS: The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* 1995; **57**: 455–464.
3 Harley JB, Moser KL, Neas BR: Logistic transmission modeling of simulated data. *Genet Epidemiol* 1995; **12**: 607–612.
4 Lazzeroni LC, Lange K: A conditional inference framework for extending the transmission/disequilibrium test. *Hum Hered* 1998; **48**: 67–81.
5 Nagelkerke NJ, Hoebee B, Teunis P, Kimman TG: Combining the transmission disequilibrium test and case–control methodology using generalized logistic regression. *Eur J Hum Genet* 2004; **12**: 964–970.
6 Bickeboller H, Clerget-Darpoux F: Statistical properties of the allelic and genotypic transmission/disequilibrium test for multi-allelic markers. *Genet Epidemiol* 1995; **12**: 865–870.
7 Spielman RS, Ewens WJ: The TDT and other family-based tests for linkage diseqilibrium and association. *Am J Hum Genet* 1996; **59**: 983–989.
8 Sham PC, Curtis D: An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet* 1995; **59**: 97–105.
9 Schaid DJ: General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* 1996; **13**: 423–450.
10 Cleves MA, Olson JM, Jacobs KB: Exact transmission-disequilibrium tests with multiallelic markers. *Genet Epidemiol* 1997; **14**: 337–347.
11 Kaplan NL, Martin ER, Weir BS: Power studies for the transmission/disequilibrium test with multiple alleles. *Am J Hum Genet* 1997; **60**: 691–702.
12 Bishop YMM, Feinberg SE, Holland PW: *Discrete multivariate analysis: theory and practice*. Cambridge, MA: MIT Press, 1975.