

# High Resolution $T^2$ Association Tests of Complex Diseases Based on Family Data

Ruzong Fan<sup>1,2</sup>, Michael Knapp<sup>2</sup>, Matthias Wjst<sup>3</sup>, Caixia Zhao<sup>1</sup> and Momiao Xiong<sup>4</sup>

*Department of Statistics, <sup>1</sup>The Texas A&M University, 447 Blocker Building, College Station, Texas 77843-3143*

*Institute of Medical Biometry, Informatics and Epidemiology, <sup>2</sup>University of Bonn, Sigmund Freud Strasse 25, D-53105 Bonn, Germany*

*GSF - National Research Center for Environment and Health <sup>3</sup>Institute of Epidemiology, Ingolstaedter Landstrasse, 1 D-85764 Neuherberg, Germany*

*Human Genetics Center, <sup>4</sup>University of Texas - Houston, P.O. Box 20334, Houston, Texas 77225*

## Summary

This paper proposes family based Hotelling's  $T^2$  tests for high resolution linkage disequilibrium (LD) mapping or association studies of complex diseases. Assume that genotype data of multiple markers or haplotype blocks are available for a sample of nuclear families, in which some offspring are affected. Paired Hotelling's  $T^2$  test statistics are proposed for a high resolution association study using parents as controls for affected offspring, based on two coding methods: haplotype/allele coding and genotype coding. The paired Hotelling's  $T^2$  tests take not only the correlation between the haplotype blocks or markers into account, but also take the correlation within each parent-offspring pair into account. The method extends two sample Hotelling's  $T^2$  test statistics for population case control association studies, which are not valid for family data due to correlation of genetic data among family members. The validity of the proposed method is justified by rigorous mathematical and statistical proof under the large sample theory. The non-centrality parameter approximations of the test statistics are calculated for power and sample size calculations. From power comparison and type I error calculations, it is shown that the test statistic based on haplotype/allele coding is advantageous over the test statistic of genotype coding. Analysis using multiple markers may provide higher power than single marker analysis. If only one marker is utilized the power of the test statistic based on haplotype/allele coding is nearly identical to that of 1-TDT. Moreover, a permutation procedure is provided for data analysis. The method is applied to data from a German asthma family study. The results based on the paired Hotelling's  $T^2$  statistic tests confirm the previous findings. However, the paired Hotelling's  $T^2$  tests produce much smaller P-values than those of the previous study. The permutation tests produce similar results to those of the previous study; moreover, additional marker combinations are shown to be significant by permutation tests. The proposed paired Hotelling's  $T^2$  statistic tests are potentially powerful in mapping complex diseases. A SAS Macro, `Hotel_fam.sas`, has been written to implement the method for data analysis.

Keywords: linkage disequilibrium mapping, complex diseases

## Introduction

As an increasing density of markers/haplotypes is available for disease gene mapping, linkage disequilibrium (LD) mapping is becoming feasible for genome screen-

ing in disease-gene localization. LD mapping (or association study) can be based on either population or family data. In population based case control studies, allele frequencies of a marker are compared between cases and controls (Chapman & Wijsman, 1998; Kaplan & Martin, 2001; Nielsen & Weir, 2001; Olson & Wijsman, 1994). If the marker allele frequencies of cases

Correspondence to: Dr. Ruzong Fan, Tel. 979-845-3156, Fax 979-845-3144. E-mail: rfan@stat.tamu.edu

are significantly different from those of controls, it indicates that the marker is linked to the disease locus under certain assumptions of population history. When multiple bi-allelic markers such as single nucleotide polymorphisms (SNPs) are available, a generalized two sample Hotelling's  $T^2$  test statistic is proposed in Xiong *et al.* (2002) by simultaneously using all markers in analysis. If micro-satellites or haplotype block data are available, two sample Hotelling's  $T^2$  statistics are proposed by Fan & Knapp (2003) to test association, which extends the method of Xiong *et al.* (2002).

It is well-known that population based case control studies are prone to false positive due to inappropriate controls, which can occur if there is population admixture or stratification. Alternatively, the parents can be used as controls of affected offspring (Allen *et al.* 2003; Falk & Rubinstein, 1987; Ott, 1989; Schaid & Rowland, 1998; Spielman *et al.* 1993; Zhao *et al.* 2000). For parental controls, the methods proposed by Fan & Knapp (2003) and Xiong *et al.* (2002) are not valid, since cases and controls are correlated to each other. Not only does the correlation between the haplotype blocks or markers need to be taken into account, but also the correlation within each parent-offspring pair needs to be dealt with.

In this paper, high resolution paired Hotelling's  $T^2$  statistic tests are proposed to perform association studies based on multiple marker/haplotype data from parents and affected offspring. The method can be used to simultaneously analyze family data with multiple markers or haplotype blocks, which is based on two coding methods, genotype coding and haplotype/allele coding. The validity of the proposed method is justified by rigorous mathematical and statistical proof under the large sample theory. Type I error rates are calculated via simulations to evaluate the performance of the proposed test statistics. The non-centrality parameter approximations of the test statistics are calculated for power comparisons, (these are provided in Supplementary Information). A SAS Macro, Hotel.fam.sas, has been written to implement the method, and is available from the authors website <http://stat.tamu.edu/~rfan/software.html/>

As a practical example, the proposed methods are applied to data from a German asthma family study (Gohlke *et al.* 2004). The data was previously analyzed using the methods of Spielman *et al.* (1993) and Zhao

*et al.* (2000). The results based on the paired Hotelling's  $T^2$  statistic tests confirm the findings of Gohlke *et al.* (2004). However, the paired Hotelling's  $T^2$  tests produce much smaller P-values than those of Gohlke *et al.* (2004). The proposed paired Hotelling's  $T^2$  statistic tests are potentially powerful in mapping complex diseases.

## Methods

We assume that a disease locus  $D$  is located in a chromosome region. Suppose that the disease locus has two alleles  $D$  and  $d$ . Allele  $D$  is disease susceptible, and  $d$  is normal. Assume that the disease susceptible allele  $D$  has population frequency  $P_D$ , and the normal allele  $d$  has population frequency  $P_d$ . We start by introducing paired  $t$ -test statistics for an association study between a disease locus and a single bi-allelic marker, using parents as controls based on trio nuclear families. Then, the method is extended to paired Hotelling's  $T^2$  test statistics for high resolution association studies based on two coding methods: haplotype/allele coding and genotype coding, if data of multiple bi-allelic or multi-allelic markers or haplotype blocks are available.

*Paired  $t$ -Test Statistics of Single Bi-allelic Marker.* Suppose a bi-allelic marker  $H_1$  is typed in the region of disease locus  $D$ . Let  $H_{11}$  and  $H_{12}$  be the two alleles of marker  $H_1$ . Consider  $N$  trio families each consisting of an affected offspring and two parents. Assume that random mating/HWE is valid in the population. For the  $i$ -th family, let  $X_i^{(O)}$  be the number of allele  $H_{11}$  in the genotype of the offspring, i.e.,  $X_i^{(O)}$  takes values 2, 1, or 0 for offspring genotype  $H_{11}H_{11}$ ,  $H_{11}H_{12}$  or  $H_{12}H_{12}$ . Similarly, let  $Y_i^{(f)}$  be the number of allele  $H_{11}$  in the father's genotype; and let  $Y_i^{(m)}$  be the number of allele  $H_{11}$  in the mother's genotype. One may choose to use only one parent to construct  $t$ -test statistics. Suppose that only the fathers' genotypes are used in analysis. Let  $\bar{Y}^{(f)} = \sum_{i=1}^N Y_i^{(f)}/N$  be the mean of  $Y_i^{(f)}$ , and  $\bar{X}^{(O)} = \sum_{i=1}^N X_i^{(O)}/N$  be the mean of  $X_i^{(O)}$ , respectively. A paired  $t$ -test statistic can be defined as

$$t_f = \sqrt{N}(\bar{X}^{(O)} - \bar{Y}^{(f)})/\sqrt{S_f}, \quad (1)$$

where  $S_f = \sum_{i=1}^N [(X_i^{(O)} - Y_i^{(f)}) - (\bar{X}^{(O)} - \bar{Y}^{(f)})]^2 / (N - 1)$ . Notice that  $S_f$  is paired sample variance. Suppose that only one mother's genotypes are used in analysis. Let  $\bar{Y}^{(m)} = \sum_{i=1}^N Y_i^{(m)}/N$  be the mean of  $Y_i^{(m)}$ .

A paired *t*-test statistic can be defined as

$$t_m = \sqrt{N}(\bar{X}^{(O)} - \bar{Y}^{(m)}) / \sqrt{S_m}, \tag{2}$$

where  $S_m = \sum_{i=1}^N [(X_i^{(O)} - Y_i^{(m)}) - (\bar{X}^{(O)} - \bar{Y}^{(m)})]^2 / (N - 1)$ . Assume that the sample size *N* is sufficiently large that the large sample theory applies. Then the statistic *t<sub>f</sub>* (or *T<sub>m</sub>*) is asymptotically *t*-distributed.

To give an intuitive justification of the above paired *t<sub>f</sub>* (or *t<sub>m</sub>*) test statistic for an association study between disease locus *D* and marker *H<sub>1</sub>*, assume that the *N* families are independent. Then one may note that  $X_i^{(O)} - Y_i^{(f)}$  are independent variables,  $i = 1, \dots, N$ . However,  $X_i^{(O)}$  and  $Y_i^{(f)}$  are dependent variables within each trio family. In addition,  $\bar{X}^{(O)}$  should be close to  $\bar{Y}^{(f)}$  if the disease locus *D* is not associated with the marker *H<sub>1</sub>*. Thus,  $S_f/N$  is a valid sample variance of mean difference  $\bar{X}^{(O)} - \bar{Y}^{(f)}$ , and so the pair *t*-test statistic *t<sub>f</sub>* in (1) can be used for an association study between disease locus *D* and marker *H<sub>1</sub>*.

One may choose to use genotype information from both the father and mother in analysis. Let  $Y_i^{(P)} = [Y_i^{(f)} + Y_i^{(m)}] / 2$  be the average number of allele *H<sub>11</sub>* in the parents of *i*-th family, and  $\bar{Y}^{(P)} = \sum_{i=1}^N Y_i^{(P)} / N$  be the mean of  $Y_i^{(P)}$ . Then a paired *t*-test statistic can be defined as follows

$$t_p = \sqrt{N}(\bar{X}^{(O)} - \bar{Y}^{(P)}) / \sqrt{S_p}, \tag{3}$$

where  $S_p = \sum_{i=1}^N [(X_i^{(O)} - Y_i^{(P)}) - (\bar{X}^{(O)} - \bar{Y}^{(P)})]^2 / (N - 1)$ . For general nuclear families each consisting of two parents and multiple affected offspring, one may define  $X_i^{(O)}$  to be average number of allele *H<sub>11</sub>* in the affected offspring in the *i*-th family. That is,  $X_i^{(O)} = \sum_{j=1}^{n_i} X_{ij}^{(O)} / n_i$ , where  $n_i$  is the number of affected offspring, and  $X_{ij}^{(O)}$  is the number of allele *H<sub>11</sub>* in the *j*-th affected offspring. Then one may define paired the *t*-test statistics *t<sub>f</sub>*, *t<sub>m</sub>*, and *t<sub>p</sub>* in the same way as equations (1), (2), and (3).

One may choose to use one parent's genotype information in the analysis, in a sense that the father's genotype information is used for some families, and the mother's genotype information is used for the rest of the families. Then a paired *t*-test statistic can be built in a similar way as that of *t<sub>f</sub>* and *t<sub>m</sub>*. In practice, the information of one parent may be unavailable for some families. That is, only parent-offspring pairs are avail-

able. One may define the paired *t*-test statistic using the available parent-offspring genotypes, in a similar way as *t<sub>f</sub>* and *t<sub>m</sub>*. Sun *et al.* (1999) built a test statistic 1-TDT as follows: let *a* denote the number of parent-child pairs in which the parent's genotype possesses more alleles *H<sub>11</sub>* than the genotype of the child (i.e., parent *H<sub>11</sub>H<sub>11</sub>* and child *H<sub>11</sub>H<sub>12</sub>* or parent *H<sub>11</sub>H<sub>12</sub>* and child *H<sub>12</sub>H<sub>12</sub>*); and let *b* denote the number of parent-child pairs in which the parent's genotype possesses fewer alleles *H<sub>11</sub>* than the genotype of the child (i.e., parent *H<sub>11</sub>H<sub>12</sub>* and child *H<sub>11</sub>H<sub>11</sub>* or parent *H<sub>12</sub>H<sub>12</sub>* and child *H<sub>11</sub>H<sub>12</sub>*). The 1-TDT by Sun *et al.* (1999) is  $1 - TDT = (a - b)^2 / (a + b)$ . Suppose the parent in 1-TDT is the father. Using the notations *a* and *b* above, it is easy to see that  $S_f = \frac{1}{N-1} [a + b - (a - b)^2 / N]$  and  $\bar{X}^{(O)} - \bar{Y}^{(f)} = [a - b] / N$ . Therefore,  $t_f^2 = \frac{N-1}{N} \frac{(a-b)^2}{a+b-(a-b)^2/N}$ , which will give a similar result as that of 1-TDT as long as *N* is large enough. Allen *et al.* (2003) proposed likelihood methods to analyze data of parent-offspring pairs in the presence of informative missingness. To avoid theoretical subtlety, we mainly deal with nuclear families in this paper.

*Paired Hotelling's T<sup>2</sup> Test Statistics of Multiple Markers/Haplotype Blocks.* In the region of the disease locus *D*, assume that *J* haplotype blocks or markers *H<sub>1</sub>*, ..., *H<sub>J</sub>* are typed instead of only one bi-allelic marker. Let us denote the haplotypes/alleles of haplotype block/marker *H<sub>j</sub>* by *H<sub>j1</sub>*, ..., *H<sub>jn<sub>j</sub></sub>*. If *H<sub>j</sub>* is a haplotype block, *n<sub>j</sub>* denotes the number of its observed haplotypes. If *H<sub>j</sub>* is a marker, *n<sub>j</sub>* denotes the number of its alleles. Again, consider *N* trio families each consisting of an affected offspring and two parents. In addition, assume that random mating/HWE is valid in the population. We define coding vectors  $X_i^{(O)}$ ,  $Y_i^{(f)}$  and  $Y_i^{(m)}$  for the affected offspring, father and mother of the *i*-th trio family, respectively, by one of the following two ways (Fan & Knapp, 2003; Schaid, 1996, p. 430).

(i) *Haplotype/allele Coding.* For the offspring of the *i*-th family, let  $G_{ij}^{(O)}$  be his/her two haplotypes at block *H<sub>j</sub>* if *H<sub>j</sub>* is a haplotype block, or genotype at marker *H<sub>j</sub>* if *H<sub>j</sub>* is a marker. Define  $X_i^{(O)} = (z_{i11}^{(O)}, \dots, z_{i1(n_1-1)}^{(O)}, \dots, z_{ij1}^{(O)}, \dots, z_{ij(n_j-1)}^{(O)})^T$ , where  $z_{ijk}^{(O)}$  is the number of haplotypes/alleles *H<sub>jk</sub>* for the offspring of the *i*-th family, i.e.,

$$z_{ijk}^{(O)} = \begin{cases} 2 & \text{if } G_{ij}^{(O)} = H_{jk}H_{jk} \\ 1 & \text{if } G_{ij}^{(O)} = H_{jk}H_{jl}, l \neq k. \\ 0 & \text{else} \end{cases} \quad (4)$$

The dimension of  $X_i^{(O)}$  is  $(n_1 - 1) + \dots + (n_J - 1) = \sum_{j=1}^J n_j - J$ , which is usually smaller than dimension  $\sum_{j=1}^J n_j(n_j + 1)/2 - J$  of the following genotype coding method.

Assume that all markers  $H_j, j = 1, \dots, J$  are bi-allelic, e.g., SNPs. We notice that

$$X_i^{(O)} = (z_{i11}^{(O)}, \dots, z_{iJ1}^{(O)})^\tau,$$

and

$$z_{ij1}^{(O)} - 1 = \begin{cases} 1 & \text{if } G_{ij}^{(O)} = H_{j1}H_{j1} \\ 0 & \text{if } G_{ij}^{(O)} = H_{j1}H_{j2} \\ -1 & \text{if } G_{ij}^{(O)} = H_{j2}H_{j2} \end{cases}$$

Thus,  $z_{ij1}^{(O)} - 1$  is the same as the indicator variables  $X_{ij}$  defined on p. 1257, Xiong et al. (2002). Hence, haplotype/allele coding (4) generalizes the bi-allelic coding method of Xiong et al. (2002) to multiple allele markers or haplotype blocks.

(ii) *Genotype Coding*. Notice that  $G_{ij}^{(O)}$  can be one of  $n_j(n_j + 1)/2$  possible choices:  $n_j$  homozygous genotypes  $H_{jk}H_{jk}$ , and  $n_j(n_j - 1)/2$  heterozygous genotypes  $H_{jk}H_{jl}, k < l$ . Depending on the genotype, let us define an indicator vector  $X_{ij}^{(O)} = (x_{ij1}^{(O)}, \dots, x_{ij(n_j-1)}^{(O)}, x_{ij12}^{(O)}, \dots, x_{ij1n_j}^{(O)}, \dots, x_{ij(n_j-1)n_j}^{(O)})^\tau$ , where the indicator variables  $x_{ijk}^{(O)}, x_{ijkl}^{(O)}, k < l$ , are defined by

$$x_{ijk}^{(O)} = \begin{cases} 1 & \text{if } G_{ij}^{(O)} = H_{jk}H_{jk}, \\ 0 & \text{else} \end{cases} \quad (5)$$

$$x_{ijkl}^{(O)} = \begin{cases} 1 & \text{if } G_{ij}^{(O)} = H_{jk}H_{jl} \\ 0 & \text{else} \end{cases}$$

The dimension of  $X_{ij}^{(O)}$  is  $n_j(n_j + 1)/2 - 1$ , i.e., the total number  $n_j(n_j + 1)/2$  of genotypes of haplotype block  $H_j$  minus 1 to remove the redundancy. Let  $X_i^{(O)} = (X_{i1}^{(O)\tau}, \dots, X_{iJ}^{(O)\tau})^\tau$  be the combined genotype coding of the  $J$  haplotype blocks or markers  $H_1, \dots, H_J$ . The dimension of  $X_i^{(O)}$  is  $\sum_{j=1}^J n_j(n_j + 1)/2 - J$ .

For the father (or mother) of the  $i$ -th family, let  $G_{ij}^{(f)}$  (or  $G_{ij}^{(m)}$ ) be his (or her) two haplotypes if  $H_j$  is a haplotype block, or genotype at marker  $H_j$  if  $H_j$  is a marker. One may define a vector  $Y_i^{(f)}$  (or  $Y_i^{(m)}$ ) in the same way, based on either genotype coding or haplotype/allele coding method. Table 1 of Fan & Knapp (2003) gives

an example of genotype coding and haplotype/allele coding for a block with 3 haplotypes to illustrate the above two coding methods. Let  $Y_i^{(P)} = [Y_i^{(f)} + Y_i^{(m)}]/2$  be the average of parental coding vectors. Let  $\bar{X}^{(O)} = \sum_{i=1}^N X_i^{(O)}/N$  and  $\bar{Y}^{(P)} = \sum_{i=1}^N Y_i^{(P)}/N$  be the mean coding vectors of the offspring and parents, respectively. Similarly, let  $\bar{Y}^{(f)} = \sum_{i=1}^N Y_i^{(f)}/N$  and  $\bar{Y}^{(m)} = \sum_{i=1}^N Y_i^{(m)}/N$  be the mean coding vectors of the fathers and mothers, respectively.

Intuitively,  $\bar{X}^{(O)}$  and  $\bar{Y}^{(f)}$  (or  $\bar{Y}^{(m)}$  or  $\bar{Y}^{(P)}$ ) should be similar vectors if the disease locus  $D$  is not associated with haplotype blocks/markers  $H_j, j = 1, \dots, J$ . Actually, we prove in the following **Justification** and Appendix A that the expected value of  $\bar{X}^{(O)} - \bar{Y}^{(f)}$  (or  $\bar{X}^{(O)} - \bar{Y}^{(m)}$ , and so  $\bar{X}^{(O)} - \bar{Y}^{(P)}$ ) is a vector of 0s if there is no association. Hence, one may develop the test statistic based on the difference  $\bar{X}^{(O)} - \bar{Y}^{(f)}$  (or  $\bar{X}^{(O)} - \bar{Y}^{(m)}$  or  $\bar{X}^{(O)} - \bar{Y}^{(P)}$ ) to test for association between the disease locus  $D$  and haplotype blocks/markers  $H_j$ . To do this, one needs to consider the variance covariance matrix of  $\bar{X}^{(O)} - \bar{Y}^{(f)}$ . Again, assume that the  $N$  families are independent. Since the offspring's marker genotypes and the parents' marker genotypes are related to each other,  $\bar{X}^{(O)}$  and  $\bar{Y}^{(f)}$  are not independent. Moreover,  $X_i^{(O)}$  and  $Y_i^{(f)}$  are paired with each other for a trio nuclear family. Therefore, the extension of the paired  $t$ -test can be used to test the association between disease locus  $D$  and haplotype blocks/markers  $H_j$ . Define a paired-sample variance covariance matrix by

$$S_f = \frac{1}{N-1} \left[ \sum_{i=1}^N \left[ (X_i^{(O)} - Y_i^{(f)}) - (\bar{X}^{(O)} - \bar{Y}^{(f)}) \right] \right. \\ \left. \left[ (X_i^{(O)} - Y_i^{(f)}) - (\bar{X}^{(O)} - \bar{Y}^{(f)}) \right]^\tau \right] \\ = \frac{1}{N-1} \left[ \sum_{i=1}^N (X_i^{(O)} - \bar{X}^{(O)})(X_i^{(O)} - \bar{X}^{(O)})^\tau \right. \\ - \sum_{i=1}^N (X_i^{(O)} - \bar{X}^{(O)})(Y_i^{(f)} - \bar{Y}^{(f)})^\tau \\ - \sum_{i=1}^N (Y_i^{(f)} - \bar{Y}^{(f)})(X_i^{(O)} - \bar{X}^{(O)})^\tau \\ \left. + \sum_{i=1}^N (Y_i^{(f)} - \bar{Y}^{(f)})(Y_i^{(f)} - \bar{Y}^{(f)})^\tau \right].$$

**Table 1** One marker analysis of German asthma data at a 0.01 significance level. Abbreviation #Fam = No. of Families, P-perm = P-value of permutation test. For permutation test, the P-value is calculated based on B = 10<sup>6</sup> permutations.

Marker	Position	Gene	Control	# Fam	Statistic	df	P-value	P-perm
1609682	114686977	IL1_Alph	dad	89	$T_{Hf} = 6.97$	1	0.008	
16944	114775235	IL1_Beta	dad&mom	125	$T_{GP} = 9.55$	2	0.008	
			dad	125	$H_{Gf} = 11.97$	2	0.003	
1143623	114776197	IL1_Beta	dad&mom	126	$T_{GP} = 10.68$	2	0.005	
			dad	126	$T_{Hf} = 9.27$	1	0.002	
					$H_{Gf} = 11.58$	2	0.003	
1530549	114997823	IL1_Delt	dad&mom	114	$T_{GP} = 10.60$	2	0.005	
			dad	119	$H_{Gf} = 9.80$	2	0.008	
957200	115001704	IL1_Delt	dad&mom	119	$T_{GP} = 13.38$	2	0.001	
			mom	120	$T_{Gm} = 9.65$	2	0.008	
996878	115002503	IL1_Delt	dad&mom	110	$T_{GP} = 10.25$	2	0.006	
2515406	115002536	IL1_Delt	dad&mom	126	$T_{GP} = 10.54$	2	0.005	
			dad	126	$T_{Gf} = 10.56$	2	0.005	
996879	115002609	IL1_Delt	dad&mom	121	$T_{GP} = 9.80$	2	0.007	
			dad	124	$T_{Gf} = 9.77$	2	0.008	
2234678	115055457	IL1RN	dad&mom	116	$T_{GP} = 13.13$	2	0.001	
			dad	117	$T_{Hf} = 6.80$	1	0.009	
16065	115055524	IL1RN	dad&mom	116	$T_{GP} = 11.73$	2	0.003	
878972	115057605	IL1RN	dad&mom	121	$T_{GP} = 10.20$	2	0.006	
1794065	115059621	IL1RN	dad&mom	119	$T_{GP} = 11.25$	2	0.004	
392503	115064083	IL1RN	dad&mom	113	$T_{HP} = 7.06$	1	0.008	0.01
					$T_{GP} = 21.55$	2	$2 \times 10^{-5}$	0.003
			dad	116	$T_{Hf} = 6.90$	1	0.009	
					$T_{Gf} = 10.45$	2	0.005	
439154	115064289	IL1RN	dad	124	$T_{Hf} = 9.25$	1	0.002	
					$T_{Gf} = 9.34$	2	0.009	
1794067	115066272	IL1RN	dad&mom	106	$T_{HP} = 9.86$	2	0.007	
1794068	115066391	IL1RN	dad&mom	118	$T_{GP} = 11.74$	2	0.003	
419598	115067095	IL1RN	dad&mom	120	$T_{GP} = 11.61$	2	0.003	
446433	115067161	IL1RN	dad&mom	120	$T_{GP} = 11.61$	2	0.003	
442710	115067287	IL1RN	dad&mom	122	$T_{GP} = 11.97$	2	0.003	
408392	115067346	IL1RN	dad&mom	119	$T_{GP} = 10.91$	2	0.004	
447713	115067560	IL1RN	dad&mom	121	$T_{GP} = 11.60$	2	0.003	
128964	115067691	IL1RN	dad&mom	115	$T_{GP} = 11.76$	2	0.003	
0	115068005	IL1RN	dad&mom	101	$T_{GP} = 23.57$	9	0.005	
434792	115068358	IL1RN	dad&mom	121	$T_{GP} = 11.24$	2	0.004	
451578	115068445	IL1RN	dad&mom	122	$T_{GP} = 11.23$	2	0.004	
454078	115068515	IL1RN	dad&mom	88	$T_{HP} = 7.81$	1	0.005	0.009
					$T_{GP} = 21.81$	2	$2 \times 10^{-5}$	0.004
			dad	98	$T_{Hf} = 7.88$	1	0.005	
440286	115069357	IL1RN	dad&mom	119	$T_{GP} = 9.75$	2	0.008	
315951	115070310	IL1RN	dad	105	$T_{Gf} = 11.60$	2	0.003	

A paired Hotelling's  $T^2$  statistic can be defined as  $T_f^2 = N(\bar{X}^{(O)} - \bar{Y}^{(f)})^T S_f^{-1} (\bar{X}^{(O)} - \bar{Y}^{(f)})$  (Anderson, 1984; Hotelling, 1931), where  $(\bar{X}^{(O)} - \bar{Y}^{(f)})^T$  is transpose of  $(\bar{X}^{(O)} - \bar{Y}^{(f)})$ . Let us denote the above Hotelling's  $T^2$  statistic for haplotype/allele coding as  $T_{Hf}$ , and the Hotelling's  $T^2$  statistic for genotype coding as  $T_{Gf}$ . Assume that the sample size  $N$  is sufficiently large that the large sample theory applies. Under the null hypothesis of no association, the statistic

$T_{Hf}$  (or  $T_{Gf}$ ) is asymptotically distributed as central  $\chi^2$  with  $\sum_{j=1}^J n_j - J$  (or  $\sum_{j=1}^J n_j(n_j + 1)/2 - J$ ) degrees of freedom. Under the alternative hypothesis of association,  $T_{Hf}$  (or  $T_{Gf}$ ) is asymptotically distributed as non-central  $\chi^2$ . In the following, let us roughly show the above claims using  $T_{Hf}$  as an example. Let  $p = \sum_{j=1}^J n_j - J$ . Then  $[T_{Hf}/(N - 1)] [(N - p)/p]$  is asymptotically  $F$ -distributed with  $p$  and  $N - p$  degrees of freedom (Theorem 5.2.2, Anderson, 1984).

When the sample size  $N$  is sufficiently large,  $T_{Hf}$  is then asymptotically distributed as central  $\chi^2$  with  $p$  degrees of freedom.

In the above definition,  $\bar{Y}^{(f)}$  can be replaced by  $\bar{Y}^{(m)}$  or  $\bar{Y}^{(P)}$ . Based on the mother's genotypes, the Hotelling's  $T^2$  statistic for haplotype/allele coding is denoted as  $T_{Hm}$ , and the Hotelling's  $T^2$  statistic for genotype coding as  $T_{Gm}$ . Based on the average of codings of both parents' genotypes, the Hotelling's  $T^2$  statistic for haplotype/allele coding is denoted as  $T_{HP}$ , and the Hotelling's  $T^2$  statistic for genotype coding as  $T_{GP}$ . If only one marker/haplotype block  $H_1$  (or  $H_2$ ) is used, we denote the corresponding test statistics as  $T_{Hf1}$ ,  $T_{Gf1}$ ,  $T_{Hm1}$ ,  $H_{Gm1}$ ,  $T_{HP1}$ ,  $T_{GP1}$  (or  $T_{Hf2}$ ,  $T_{Gf2}$ ,  $T_{Hm2}$ ,  $H_{Gm2}$ ,  $T_{HP2}$ ,  $T_{GP2}$ ). One may want to notice that  $T_{Hf1}$  is equal to the square of the paired  $t$ -test statistic  $t_f$ , defined in (1), if only one bi-allele marker  $H_1$  is used in the analysis. Similarly,  $T_{Hm1}$  is equal to the square of  $t_m$  defined in (2), and  $T_{HP1}$  is equal to the square of  $t_p$  defined in (3), if only one bi-allele marker  $H_1$  is used in the analysis.

In practice, two situations can occur: (1) one parent is available for study and the other parent is missing; (2) both parents are available for study. If only one parent is available and the other is missing, one may use the available parent-offspring pair in the analysis. However, the test statistic can be problematic in the presence of informative missingness (Allen *et al.* 2003). In this paper, we deal with the perfect situation where both parents are available for study for each nuclear family.

For general nuclear families each consisting of two parents and multiple affected offspring, one may define  $X_i^{(O)} = \sum_{j=1}^{n_i} X_{ij}^{(O)} / n_i$ , where  $n_i$  is the number of affected offspring, and  $X_{ij}^{(O)}$  is the coding vector of the  $j$ -th affected offspring in the  $i$ -th family. Then one may define the paired Hotelling's  $T^2$  statistics  $T_f$ ,  $T_m$ , and  $T_p$  similarly, according to either haplotype/allele coding or genotype coding.

*Justification.* Let us denote  $A_O = (\text{Offspring is affected})$ . Assume that a trio family is ascertained through the affected offspring. The likelihood is the joint probability of the parental genotypes ( $G^{(f)}$ ,  $G^{(m)}$ ) and the offspring's genotype  $G^{(O)}$ , conditional on  $A_O$ , i.e.,  $P(G^{(f)}, G^{(m)}, G^{(O)} | A_O)$ . Here  $G^{(f)}$  is the father's genotype, and  $G^{(m)}$  is the mother's genotype. Under the null hypothesis of no association between the haplotype blocks or markers  $H_j$ ,  $j = 1, \dots, J$  and the disease locus  $D$ ,

i.e.,  $\Delta_{jk} = 0$  for all  $j$  and  $k$  (here  $\Delta_{jk} = P(H_{jk}D) - P(H_{jk})P_D$ , Appendix A), we show in Appendix A that the expectation  $E(\bar{X}^{(O)} - \bar{Y}^{(f)} | A_O) = E(\bar{X}^{(O)} - \bar{Y}^{(m)} | A_O) = E(\bar{X}^{(O)} - \bar{Y}^{(P)} | A_O) = 0$  for both the genotype coding and the haplotype/allele coding methods. Therefore,  $T_{Hf}$  and  $T_{Gf}$  (or  $T_{Hm}$  and  $T_{Gm}$ ) are valid statistics to test association between the disease locus  $D$  and blocks or markers  $H_1, \dots, H_J$ , since  $S_f/N$  (or  $S_m/N$ ) is the sample variance-covariance matrix of  $\bar{X}^{(O)} - \bar{Y}^{(f)}$  (or  $\bar{X}^{(O)} - \bar{Y}^{(m)}$ ). If the null hypothesis of no association is true, the marker genotypes of the two parents are independent. Therefore,  $T_{HP}$  and  $T_{GP}$  are valid statistics to test association between the disease locus  $D$  and blocks or markers  $H_1, \dots, H_J$ , and asymptotically distributed as central  $\chi^2$  under the null hypothesis.

*Population Stratification.* In the presence of population structure, association is prone to false positive. Let us allow for population stratification, in which there are  $\Gamma$  sub-populations with proportions  $\nu_1, \dots, \nu_\Gamma$  (Abecasis *et al.* 2000; Ewens & Spielman, 1995). In Appendix B, we justify that the Hotelling's  $T^2$  test statistics are valid to test association between the disease locus  $D$  and blocks or markers  $H_1, \dots, H_J$  in the presence of population stratification.

*Non-centrality Parameters.* The derivation of non-centrality parameters is provided in Supplementary Information.

*Permutation Test.* For each trio family, both haplotype/allele coding and genotyping coding are based on the observed pattern of allelic transmission. Under the null hypothesis of no association between the haplotypes or markers and the disease locus, the probability of the observed pattern of allelic transmission is 1/2, conditional on the parental genotypes. To see this, assume that one parent's genotype at marker  $H_1$  is  $H_{11}H_{12}$ , and the allele  $H_{11}$  is transmitted to the offspring. This happens with 1/2 probability. The other possibility is that the parent transmits allele  $H_{12}$  to the offspring. This also can happen with 1/2 probability. At each locus, replace the transmitted allele with the untransmitted one. This will lead to an unobserved pattern of allelic transmission, which may happen with 1/2 probability. Hence, each trio family corresponds to 2 events: the observed pattern of allelic transmission and the unobserved

pattern of allelic transmission. For a trio family, a random permutation can be constructed by replacing each observed pattern of allelic transmission by itself, or the the unobserved pattern of allelic transmission, with equal probability.

For a set of  $N$  families, there are  $2^N$  different permutations of the data. Suppose that  $B$  (say  $B = 10^6$ ) permutations are performed. For each permutation, the test statistics are calculated. Then the permutation P-value is the proportion of the  $B$  permutations which leads to an empirical test which is larger or equal to the test of the observed pattern of allelic transmission. Notice that the permutation depends on the genotypes of both parents and the offspring. Hence, permutation tests can only be performed for  $T_{GP}$  and  $T_{HP}$ , not for  $T_{Gf}$ ,  $T_{Hf}$ ,  $T_{Gm}$  and  $T_{Hm}$ .

## Results

### An Example

As an example, we analyzed the data from a German asthma family study (Gohlke *et al.* 2004). The dataset includes 127 trio families with 381 individuals. A detailed description of the dataset can be found in Gohlke *et al.* (2004), in which the data was previously analyzed using the methods of Spielman *et al.* (1993) and Zhao *et al.* (2000). Table 1 presents the results of a marker analysis at a 0.01 significance level. 28 markers show P-values  $\leq 0.01$ . Two SNPs, 392503 and 454078, produce the smallest P-value  $2 \times 10^{-5}$ . In Gohlke *et al.* (2004), Table 1, SNP 392503 produces a P-value 0.007, and SNP 454078 produces the smallest P-value 0.006. In Table 1 of the Supplementary Information, results are reported by using one marker analysis at a 0.05 significance level. The results of permutation tests for SNPs 392503 and 454078 are similar to the results of Gohlke *et al.* (2004).

By using adjacent markers, or every other adjacent marker, Table 2 presents results of two marker analysis at a 0.001 significance level. Here adjacent markers mean the two markers are next to each other, every other adjacent marker means that the two markers are separated by exactly one marker which is right in the centre of them. Marker pair 454078-380092 produces the smallest P-values ( $T_{Gm} = 34.24$ , P-value  $7 \times 10^{-7}$ ; and  $T_{GP} = 33.24$ , P-value  $1 \times 10^{-6}$ ). Marker pair 392503-439154

produces the next smallest P-value ( $T_{GP} = 28.65$ , P-value  $9 \times 10^{-6}$ ). Moreover, other marker pairs, such as 315934-392503, 315936-392503, 451578-454078, 434792-454078, 454078-973635, produce P-values around  $10^{-5}$ . Notice that each of these marker pairs contains either marker 392503 or 454078. Hence, the results are consistent with those of Table 1. Moreover, two marker analysis produces smaller P-values than one marker analysis. In Table 2 of Gohlke *et al.* (2004), some of the above marker pairs produce P-values around 0.01. In Table 2 of the Supplementary Information, results are reported by using two marker analysis at a 0.01 significance level. For the two marker pairs in Table 2 of Gohlke *et al.* (2004), the permutation tests provide similar P-values as those of Gohlke *et al.* (2004). However, marker pairs 996879-2234678, 440286-315952, 2234678-928940, and 434792-454078 produce small P-values in the permutation tests, which are not significant in Gohlke *et al.* (2004).

Table 3 presents the results of a three-marker analysis at a 0.001 significance level, except for some cases with permutation tests. Here the three markers can be exactly next to each other, or there is one marker which separates two of them. For instance, 2515406-996879-2234678 are three adjacent markers which are next to each other. However, 996878-996879-2234678 are three markers in which 996879 and 2234678 are next to each other, but 996878 and 996879 are separated by 2515406. Four three marker combinations, 315934-392503-439154, 451578-454078-380092, 434792-454078-380092, 454078-380092-973635, produce the smallest P-values, which are between  $10^{-7}$  and  $10^{-6}$ . Notice that each of these four combinations contains either marker 392503 or 454078. Four other combinations, 392503-439154-1794066, 454078-380092-440286, 996879-16065-928940, 996879-2234678-928940, produce small P-values which are between  $10^{-6}$  and  $10^{-5}$ . In Table 2 of Gohlke *et al.* (2004), combination 315934-392503-439154 has a P-value 0.004; and combinations 454078-973635-440286 and 451578-454078-973635 have P-values 0.002 and 0.004, respectively. The P-values based on our methods are between  $10^{-5}$  and  $10^{-4}$  for the two combinations 454078-973635-440286 and 451578-454078-973635 (Table 3). In Table 3 of the Supplementary Information, results are reported using three-marker analysis at a 0.01

**Table 2** Two marker analysis of German asthma data at a 0.001 significance level. Abbreviation **#Fam** = No. of Families, **P-perm** = P-value of permutation test. For permutation test, the P-value is calculated based on  $B = 10^6$  permutations.

Markers	Gene	Control	#Fam	Statistic	df	P-value	P-perm
996879-2234678	IL1.Beta-	dad&mom	112	$T_{Gf} = 23.84$	4	$9 \times 10^{-5}$	
		dad&mom	111	$T_{GP} = 24.16$	4	$7 \times 10^{-5}$	0.004
315934-392503	IL1RN	dad&mom	109	$T_{GP} = 27.90$	4	$1 \times 10^{-5}$	0.003
		dad	113	$T_{Gf} = 17.94$	4	0.001	
392503-439154	IL1RN	dad&mom	111	$T_{GP} = 28.65$	4	$\times 10^{-6}$	0.007
		dad	115	$T_{Gf} = 20.01$	4	$5 \times 10^{-4}$	
451578-454078	IL1RN	dad&mom	86	$T_{GP} = 25.55$	4	$4 \times 10^{-5}$	$3 \times 10^{-4}$
454078-380092	IL1RN	dad&mom	86	$T_{GP} = 33.24$	4	$1 \times 10^{-6}$	$4 \times 10^{-4}$
		dad	95	$T_{Hm} = 13.92$	2	0.001	
				$T_{Gm} = 34.24$	4	$7 \times 10^{-7}$	
973635-440286	IL1RN	dad	120	$T_{Hf} = 13.13$	2	0.001	
440286-315952	IL1RN	dad&mom	114	$T_{GP} = 24.58$	4	$\times 10^{-5}$	0.009
		dad	116	$T_{Gf} = 21.30$	4	$3 \times 10^{-4}$	
315951-3087271	IL1RN-2qterm?	dad	102	$T_{Gf} = 23.88$	4	$\times 10^{-5}$	
2515406-2234678	IL1.Delt-IL1RN	dad&mom	116	$T_{GP} = 21.99$	4	$2 \times 10^{-4}$	
996879-16065	IL1.Delt-IL1RN	dad&mom	111	$T_{GP} = 21.00$	4	$3 \times 10^{-4}$	
2234678-928940	IL1RN	dad&mom	112	$T_{GP} = 24.78$	4	$\times 10^{-5}$	0.001
315936-392503	IL1RN	dad&mom	100	$T_{GP} = 20.79$	4	$4 \times 10^{-4}$	
434792-454078	IL1RN	dad&mom	88	$T_{GP} = 25.55$	4	$4 \times 10^{-5}$	$3 \times 10^{-4}$
451578-380092	IL1RN	dad&mom	120	$T_{GP} = 21.69$	4	$2 \times 10^{-4}$	
		dad	121	$T_{Gf} = 18.86$	4	$8 \times 10^{-4}$	
		dad&mom	87	$T_{GP} = 25.38$	4	$4 \times 10^{-5}$	$5 \times 10^{-4}$
454078-973635	IL1RN	dad	96	$T_{Hf} = 19.72$	2	$5 \times 10^{-5}$	
				$T_{Gf} = 24.37$	4	$7 \times 10^{-5}$	
380092-440286	IL1RN	dad&mom	119	$T_{GP} = 18.14$	4	0.001	
440286-315951	IL1RN	dad	101	$T_{Gf} = 25.64$	4	$4 \times 10^{-5}$	
315952-3087271	IL1RN-2qterm?	dad&mom	115	$T_{GP} = 20.80$	4	$4 \times 10^{-4}$	
		dad	118	$T_{Gf} = 21.76$	4	$2 \times 10^{-4}$	

significance level. For the three combinations 315934-392503-439154, 454078-973635-440286 and 451578-454078-973635 in Table 2 of Gohlke *et al.* (2004), the permutation tests produce similar P-values. In addition, a few other combinations produce small P-values in the permutation tests, but are not significantly at the 0.01 significance level in Table 2 of Gohlke *et al.* (2004).

*Type I Errors.* Table 4 shows type I error rates at a significance level  $\alpha = 0.01$  using one marker  $H_1$  or two markers  $H_1$  and  $H_2$ . Three models are considered; in each model,  $H_1$  is a bi-allelic marker with equal allele frequency  $P(H_{11}) = P(H_{12}) = 0.50$ . In Model I, one marker  $H_1$  is used in analysis; 1-TDT is calculated based on the genotypes of fathers and offspring. In models II and III, two markers  $H_1$  and  $H_2$  are used in analysis. In Model II,  $r = 2, P(H_{21}) = 0.5, \Delta_{H_1, H_{21}} = 0.05$ . In Model

III,  $r = 4, P(H_{21}) = P(H_{22}) = P(H_{23}) = P(H_{24}) = 0.25, \Delta_{H_1, H_{21}} = P(H_{11} H_{21}) - P(H_{11})P(H_{21}) = 0.05, \Delta_{H_1, H_{22}} = P(H_{11} H_{22}) - P(H_{11})P(H_{22}) = 0.05, \Delta_{H_1, H_{23}} = P(H_{11} H_{23}) - P(H_{11})P(H_{23}) = -0.05, \Delta_{H_1, H_{24}} = P(H_{11} H_{24}) - P(H_{11})P(H_{24}) = -0.05$ . Each time, 10,000 simulated datasets are generated and each dataset contains  $N = 100$  or 250 or 500 trio families; a type I error rate is then calculated as the proportion of the 10,000 datasets whose empirical test statistics are greater than or equal to the cut-off point at the significance level  $\alpha = 0.01$ . The process is repeated 101 times. Thus, 101 type I error rates are calculated. The **Mean**, **Std Dev** (standard deviation), **Minimum** and **Maximum** of the 101 type I error rates are presented in Table 4.

From the results of Table 4, it is clear that  $T_{HP}$  has a lower type I error than that of  $T_{GP}$ , and  $T_{Hf}$  has a lower type I error than that of  $T_{Gf}$ . That is, the test statistic of the haplotype/allele coding method has a

**Table 3** Three marker analysis of German asthma data at a 0.001 significance level. Abbreviation **#Fam** = No. of Families, **P-perm** = P-value of permutation test. For permutation test, the P-value is calculated based on  $B = 10^6$  permutations

Markers	Control	# Fam	Statistic	df	P-value	P-perm
2515406-996879-2234678	dad&mom	112	$T_{GP} = 25.32$	6	$\times 10^{-4}$	
996879-2234678-16065	dad&mom	104	$T_{GP} = 23.63$	6	$\times 10^{-4}$	
2234678-16065-928940	dad&mom	104	$T_{GP} = 26.38$	6	$\times 10^{-4}$	
16065-928940-878972	dad&mom	107	$T_{GP} = 25.41$	6	$3 \times 10^{-4}$	
315936-315934-392503	dad&mom	100	$T_{GP} = 24.04$	6	$5 \times 10^{-4}$	
315934-392503-439154	dad&mom	107	$T_{GP} = 38.94$	6	$7 \times 10^{-7}$	0.003
			$T_{HP} = 11.85$	3	0.008	0.01
	dad	112	$T_{Gf} = 29.27$	6	$5 \times 10^{-5}$	
392503-439154-1794066	dad&mom	106	$T_{GP} = 35.99$	6	$3 \times 10^{-6}$	0.003
			$T_{HP} = 14.90$	3	0.002	0.003
	dad	111	$T_{Gf} = 24.41$	6	$4 \times 10^{-4}$	
	mom	111	$T_{Gm} = 31.02$	6	$3 \times 10^{-5}$	
439154-1794066-1794067	mom	111	$T_{Gm} = 23.03$	6	$8 \times 10^{-4}$	
434792-451578-454078	dad&mom	84	$T_{HP} = 26.88$	6	$2 \times 10^{-4}$	
451578-454078-380092	dad&mom	86	$T_{GP} = 38.52$	6	$9 \times 10^{-7}$	$3 \times 10^{-4}$
			$T_{HP} = 9.42$	3	0.02	0.03
	dad	95	$T_{Hf} = 16.77$	3	$8 \times 10^{-4}$	
			$T_{Gf} = 37.80$	6	$1 \times 10^{-6}$	
454078-380092-973635	dad&mom	85	$T_{GP} = 34.43$	6	$6 \times 10^{-6}$	$6 \times 10^{-4}$
			$T_{HP} = 11.39$	3	0.01	0.02
	dad	94	$T_{Hf} = 22.37$	3	$6 \times 10^{-5}$	
			$T_{Gf} = 38.53$	6	$9 \times 10^{-7}$	
973635-440286-315952	dad&mom	113	$T_{GP} = 26.15$	6	$2 \times 10^{-4}$	
	dad	115	$T_{Gf} = 23.43$	6	$7 \times 10^{-4}$	
440286-315952-315951	dad	98	$T_{Gf} = 27.86$	6	$1 \times 10^{-4}$	
315952-315951-3087271	dad	98	$T_{Gf} = 25.66$	6	$3 \times 10^{-4}$	
996878-996879-2234678	dad&mom	96	$T_{GP} = 26.31$	6	$2 \times 10^{-4}$	
996879-16065-928940	dad&mom	106	$T_{GP} = 35.31$	6	$\times 10^{-6}$	$7 \times 10^{-4}$
			$T_{HP} = 10.06$	3	0.02	0.02
	dad	113	$T_{Gf} = 21.74$	6	0.001	
2234678-928940-878972	dad&mom	108	$T_{GP} = 26.48$	6	$\times 10^{-4}$	
1794065-315934-392503	dad&mom	107	$T_{GP} = 27.44$	6	$1 \times 10^{-4}$	
315936-392503-439154	dad&mom	98	$T_{GP} = 30.40$	6	$3 \times 10^{-5}$	
598859-451578-454078	dad&mom	85	$T_{HP} = 18.96$	3	$3 \times 10^{-4}$	
			$T_{GP} = 27.66$	6	$1 \times 10^{-4}$	
434792-454078-380092	dad&mom	84	$T_{GP} = 38.95$	6	$7 \times 10^{-7}$	$2 \times 10^{-4}$
			$T_{HP} = 9.31$	3	0.03	0.02
	dad	93	$T_{Gf} = 33.25$	6	$9 \times 10^{-6}$	

lower type I error than the test statistic of the genotype coding method. Thus, the haplotype/allele coding method leads to more robust and reliable test statistics. In addition, the type I error rates of tests for both haplotype/allele coding and genotype coding are reasonable for models I and II when  $N \geq 100$ . The type I error rates of tests for haplotype/allele coding are reasonable for model III when  $N \geq 100$ . However, the type I error rates of tests for genotype coding are much higher than the nominal level 0.01 for model III when  $N = 100$ . As the number of trio families  $N \geq 250$ , the type I error rates of tests for both genotype coding and haplo-

type/allele coding are reasonable for model III. Notice that the degrees of freedom of tests  $T_{GP}$  and  $T_{HP}$  are 2, 4 and 11 for models I, II and III, respectively. Hence, the degrees of freedom of tests  $T_{GP}$  and  $T_{HP}$  are large for model III. When the degrees of freedom of tests are large, the asymptotic criteria can be problematic. In this case, a large sample is necessary to keep the type I error rates in a reasonable range.

*Power Calculation and Comparison.* To make power comparisons, we consider four genetic models: heterogeneous recessive, heterogeneous dominant,

**Table 3** (Continued)

Markers	Control	# Fam	Statistic	df	P-value	P-perm
451578-380092-973635	dad&mom	119	$T_{GP} = 22.29$	6	0.001	
	dad	120	$T_{Cf} = 22.00$	6	0.001	
454078-973635-440286	dad&mom	85	$T_{GP} = 27.42$	6	$1 \times 10^{-4}$	0.002
			$T_{HP} = 13.11$	3	0.004	0.007
	dad	94	$T_{Hf} = 22.80$	3	$5 \times 10^{-5}$	
			$T_{Cf} = 27.29$	6	$1 \times 10^{-4}$	
380092-440286-315952	dad&mom	114	$T_{GP} = 24.99$	6	$3 \times 10^{-4}$	
	dad	116	$T_{Cf} = 22.45$	6	0.001	
440286-315951-3087271	dad	100	$T_{Cf} = 26.05$	6	$2 \times 10^{-4}$	
315952-3087271-895496	dad&mom	68	$T_{GP} = 21.66$	6	0.001	
996878-2515406-2234678	dad&mom	100	$T_{GP} = 24.87$	6	$\times 10^{-4}$	
2515406-996879-16065	dad&mom	111	$T_{GP} = 22.42$	6	0.001	
996879-2234678-928940	dad&mom	108	$T_{GP} = 38.26$	6	$1 \times 10^{-6}$	$3 \times 10^{-4}$
			$T_{HP} = 13.81$	3	0.003	0.005
	dad	112	$T_{Cf} = 23.76$	6	$6 \times 10^{-4}$	
1-1794065-315934	dad&mom	101	$T_{GP} = 24.39$	6	$4 \times 10^{-4}$	
	dad	108	$T_{Cf} = 25.10$	6	$3 \times 10^{-4}$	
315934-392503-1794066	dad&mom	104	$T_{GP} = 25.42$	6	$\times 10^{-4}$	
392503-439154-1794067	dad&mom	95	$T_{GP} = 25.38$	6	$3 \times 10^{-4}$	
439154-1794066-1794068	dad&mom	112	$T_{GP} = 26.05$	6	$2 \times 10^{-4}$	
	mom	116	$T_{Gm} = 25.67$	6	$3 \times 10^{-4}$	
598859-434792-454078	dad&mom	84	$T_{HP} = 21.87$	3	$7 \times 10^{-5}$	
			$T_{GP} = 27.36$	6	$1 \times 10^{-4}$	
434792-451578-380092	dad&mom	117	$T_{GP} = 25.83$	6	$2 \times 10^{-4}$	
	dad	119	$T_{Hf} = 21.90$	6	0.001	
451578-454078-973635	dad&mom	85	$T_{GP} = 30.15$	6	$4 \times 10^{-5}$	$3 \times 10^{-4}$
			$T_{HP} = 11.67$	3	0.009	0.01
	dad	94	$T_{Hf} = 24.06$	3	$2 \times 10^{-5}$	
			$T_{Cf} = 29.95$	6	$4 \times 10^{-5}$	
454078-380092-440286	dad&mom	86	$T_{GP} = 33.78$	6	$7 \times 10^{-6}$	0.002
			$T_{HP} = 10.30$	3	0.02	0.02
	dad	95	$T_{Cf} = 34.58$	6	$5 \times 10^{-6}$	
973635-440286-315951	dad	100	$T_{Cf} = 24.78$	6	$4 \times 10^{-4}$	
440286-315952-3087271	dad&mom	111	$T_{GP} = 25.83$	6	$2 \times 10^{-4}$	
	dad	115	$T_{Cf} = 21.83$	6	0.001	
315951-3087271-724496	dad	101	$T_{Cf} = 24.03$	6	$5 \times 10^{-4}$	

additive and multiplicative. For optimistic models, Table 5 gives the penetrance probabilities taken from Nielsen *et al.* (1998) or Fan & Knapp (2003). For less optimistic models, Table 6 lists penetrance probabilities taken from Fan & Knapp (2003). Figure 1 shows power curves of  $T_{Hf1}$  (i.e., square of  $t_f$ ),  $T_{Cf1}$ , and 1-TDT at a significance level  $\alpha = 0.05$  using a bi-allele marker  $H_1$ , when  $P(H_{11}) = P(H_{12}) = 0.50$ ,  $P_D = 0.15$ ,  $N = 200$ , and  $\theta_{1D} = 1\text{cM}$  for the first set of parameters of the four genetic models of Table 5. Figure 2 shows power curves of  $T_{Hf1}$  (i.e., square of  $t_f$ ),  $T_{Cf1}$ , and 1-TDT at a significance level  $\alpha = 0.05$  using a bi-allele marker  $H_1$ , when  $P(H_{11}) = P(H_{12}) =$

$0.50$ ,  $P_D = 0.15$ ,  $N = 600$ , and  $\theta_{1D} = 1\text{cM}$  for the second set of parameters of the four genetic models of Table 6. From Figures 1 and 2, it is clear that the power of  $T_{Hf1}$  is nearly identical to that of 1-TDT. Generally,  $T_{Hf1}$  has higher power than that of  $T_{Cf1}$ , which is consistent with the result of Fan & Knapp (2003) for population case control studies. For the optimistic models in Table 5, the sample sizes required to achieve certain power levels are lower than those of the less optimistic models in Table 6.

Consider a situation where two markers  $H_1$  and  $H_2$  flank the disease locus  $D$ . Assume that a disease mutation  $D$  was introduced into the population  $T$  generations

**Table 4** Type I error rates at a significance level  $\alpha = 0.01$  using one marker  $H_1$  or two markers  $H_1$  and  $H_2$

Model	Test	df	# type I error rates	Mean	Std Dev	Minimum	Maximum
I N = 100	$T_{HP}$	1	101	0.012	0.001	0.009	0.014
	$T_{GP}$	2	101	0.015	0.001	0.013	0.018
	$T_{Hf}$	1	101	0.012	0.001	0.009	0.015
	$T_{Gf}$	2	101	0.013	0.001	0.010	0.016
	1-TDT	1	101	0.010	0.001	0.008	0.013
II N = 100	$T_{HP}$	2	101	0.013	0.001	0.010	0.016
	$T_{GP}$	4	101	0.019	0.001	0.016	0.023
	$T_{Hf}$	2	101	0.013	0.001	0.010	0.017
	$T_{Gf}$	4	101	0.016	0.001	0.013	0.018
II N = 250	$T_{HP}$	2	101	0.011	0.001	0.009	0.014
	$T_{GP}$	4	101	0.013	0.001	0.011	0.018
	$T_{Hf}$	2	101	0.011	0.001	0.009	0.014
	$T_{Gf}$	4	101	0.012	0.001	0.010	0.015
III N = 100	$T_{HP}$	4	101	0.016	0.001	0.012	0.019
	$T_{GP}$	11	101	0.050	0.002	0.044	0.058
	$T_{Hf}$	4	101	0.016	0.001	0.011	0.019
	$T_{Gf}$	11	101	0.033	0.002	0.028	0.039
III N = 250	$T_{HP}$	4	101	0.012	0.001	0.008	0.016
	$T_{GP}$	11	101	0.021	0.001	0.019	0.026
	$T_{Hf}$	4	101	0.012	0.001	0.010	0.015
	$T_{Gf}$	11	101	0.017	0.001	0.013	0.020
III N = 500	$T_{HP}$	4	101	0.011	0.001	0.008	0.015
	$T_{GP}$	11	101	0.015	0.001	0.012	0.018
	$T_{Hf}$	4	101	0.011	0.001	0.009	0.014
	$T_{Gf}$	11	101	0.013	0.001	0.010	0.016

**Table 5** First set of parameters of simulated genetic models

Model Type	$f_{DD}$	$f_{Dd}$	$f_{dd}$
Heterogeneous Recessive	1.00	0.05	0.05
Heterogeneous Dominant	1.00	0.95	0.05
Additive	1.00	0.50	0.0
Multiplicative	0.81	0.045	0.0025

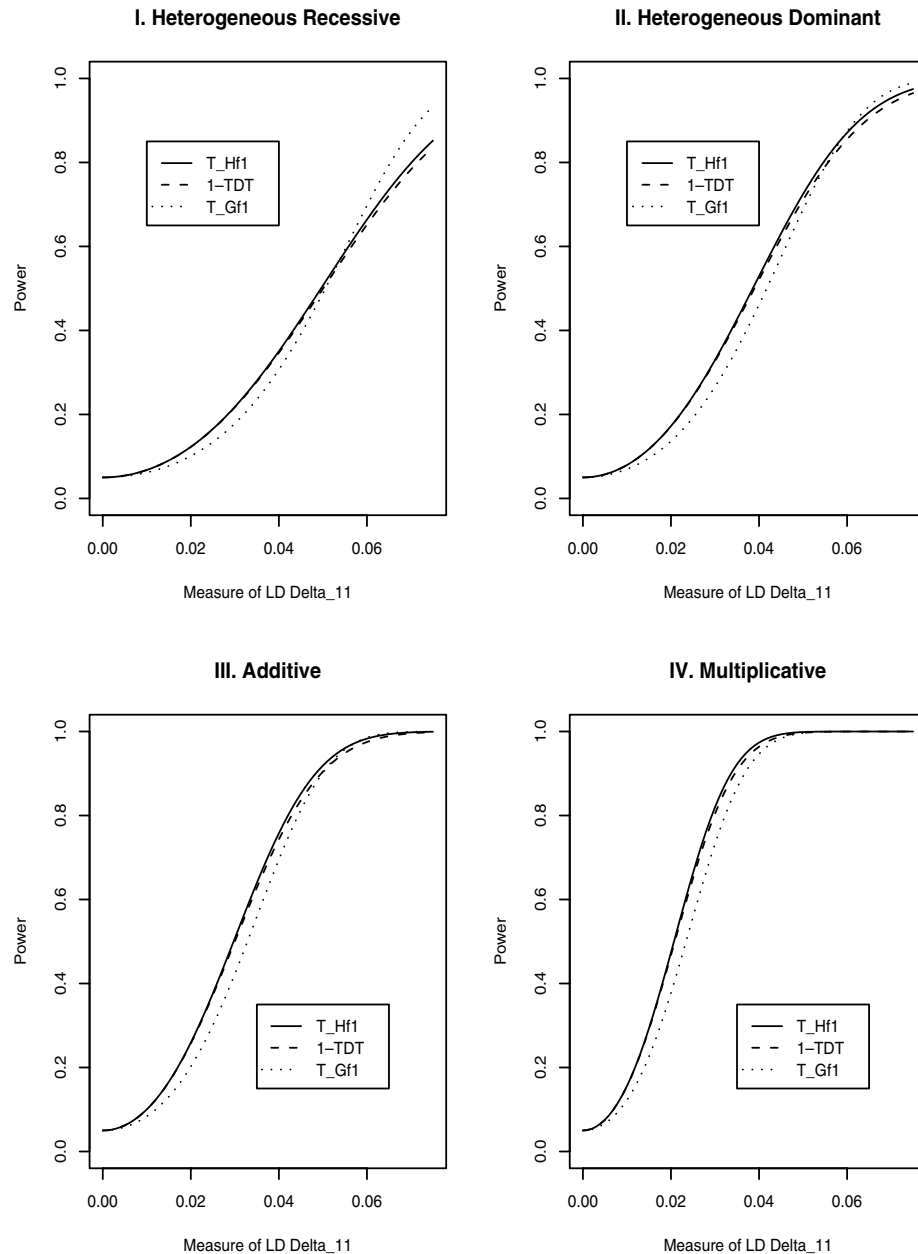
**Table 6** Second set of parameters of simulated genetic models

Model Type	$f_{DD}$	$f_{Dd}$	$f_{dd}$
Heterogeneous Recessive	0.16	0.04	0.04
Heterogeneous Dominant	0.08	0.08	0.02
Additive	0.108	0.0675	0.027
Multiplicative	0.12	0.06	0.03

ago, and was carried by allele  $H_{j1}$ . At the initial generation, the haplotype frequencies  $P(H_{j1}D)(0) = P_D$ , and  $P(H_{ji}D)(0) = 0, i = 2, \dots, n_j$ . Here  $H_{j1}D$  is haplotype  $DH_{j1}$ . Moreover,  $P(H_{j1}d)(0) = P(H_{j1}) - P_D$

and  $P(H_{ji}d)(0) = P(H_{ji}), i = 2, \dots, n_j$ . Let  $\theta_{jD}$  be the recombination fraction between marker  $H_j$  and disease locus  $D, j = 1, 2$ . Given a map distance  $\Omega_{jD}$  between marker  $H_j$  and disease locus  $D$ , the recombination fraction  $\theta_{jD}$  can be calculated by Haldane's map function  $\theta_{jD} = [1 - \exp(-2\Omega_{jD})]/2$  under the assumption of no interference. On average, the haplotype frequencies of the current generation can be approximately calculated by  $P(H_{ji}D)(T) = P(H_{ji}D)(0)e^{-T\theta_{jD}} + P_D P(H_{ji})(1 - e^{-T\theta_{jD}})$  and  $P(H_{ji}d)(T) = P(H_{ji}d)(0)e^{-T\theta_{jD}} + P_d P(H_{ji})(1 - e^{-T\theta_{jD}}), i = 1, \dots, n_j$ , where  $T$  indicates the generation.

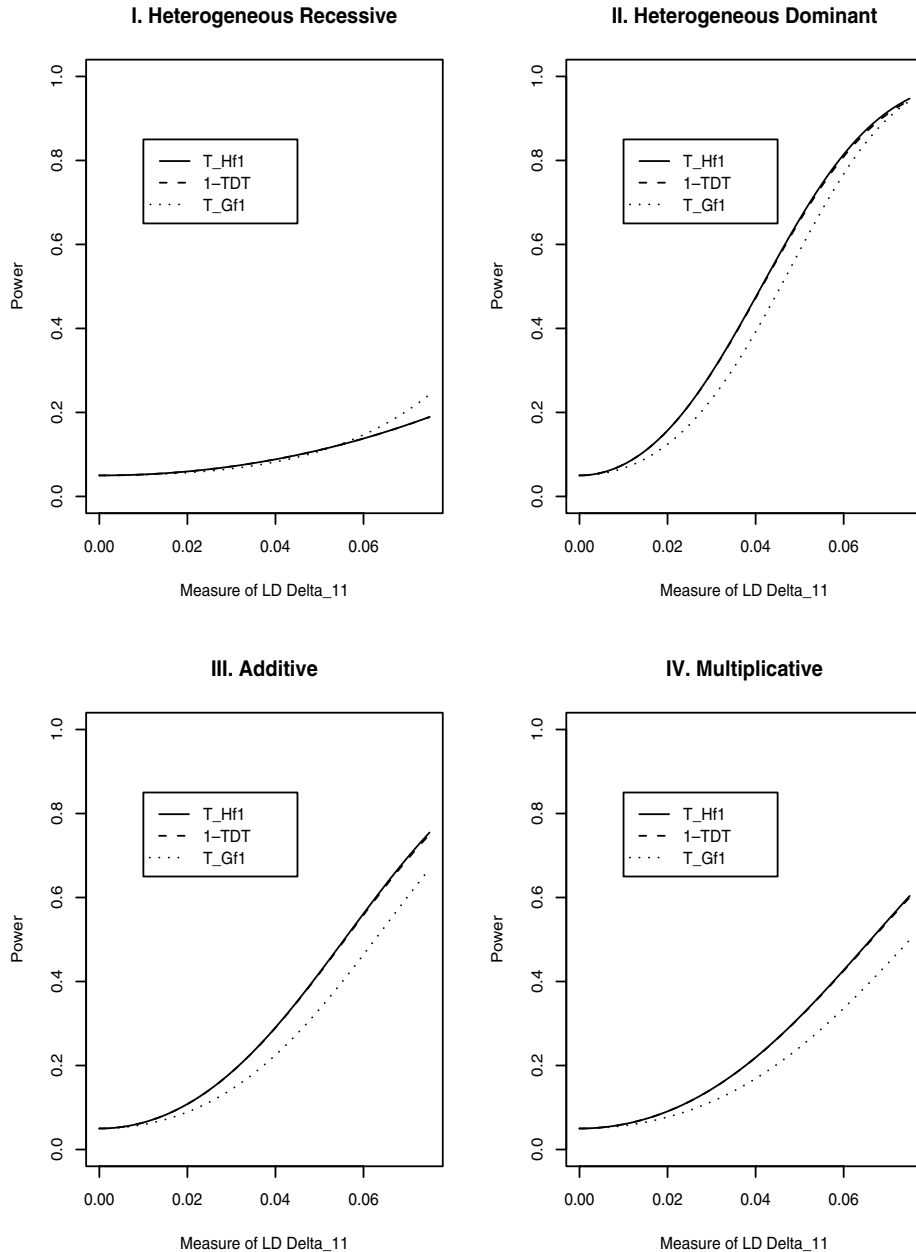
Assume that the disease susceptible allele  $D$  was carried by haplotype  $H_{11}H_{21}$  at the initial generation of the mutation. The initial haplotype frequencies  $P(H_{11}DH_{21})(0) = P_D$  and  $P(H_{1i}DH_{2s})(0) = 0, i = 1, \dots, n_1, s = 1, \dots, n_2$ , and  $(i, s) \neq (1, 1)$ . The other initial haplotype frequencies are  $P(H_{11}dH_{21})(0) = P(H_{11}H_{21}) - P_D$  and  $P(H_{1i}dH_{2s})(0) = P(H_{1i}H_{2s}), i = 1, \dots, n_1, s = 1, \dots, n_2$  and  $(i, s) \neq$



**Figure 1.** Power curves of  $T_{Hf1}$  (i.e., square of  $t_f$ ), 1-TDT and  $T_{Gf1}$  at a significance level  $\alpha = 0.05$  using a bi-allelic marker  $H_1$ , when  $P(H_{11}) = P(H_{12}) = 0.50$ ,  $P_D = 0.15$ ,  $N = 200$ , and  $\theta_{1D} = 1cM$  for the first set of parameters of the four genetic models of Table 5.  $\Delta_{11} = \Delta_{11} = P(H_{11}D) - P(H_{11})P_D$  is a measure of linkage disequilibrium between marker  $H_1$  and disease locus  $D$ .

(1, 1). On average, the haplotype frequencies of the current generation can be approximately calculated by  $P(H_{1i}DH_{2s})(T) = \Delta_{iD_s}(0)e^{-T(\theta_{1D} + \theta_{2D})} + P(H_{1i})\Delta_{2s}(0)e^{-T\theta_{2D}} + P(H_{2s})\Delta_{1i}(0)e^{-T\theta_{1D}} + P(H_{1i})P_DP(H_{2s})$  and  $P(H_{1i}dH_{2s})(T) = P(H_{1i}H_{2s}) - P(H_{1i}DH_{2s})(T)$ , where  $\Delta_{iD_s}(0) = P(H_{1i}DH_{2s})(0) -$

$P(H_{1i})\Delta_{2s}(0) - P(H_{2s})\Delta_{1i}(0) - P(H_{1i})P_DP(H_{2s})$  is a measure of the initial LD at the three loci for alleles  $H_{1i}$  and  $H_{2s}$ ,  $\Delta_{1i}(0) = P(H_{1i}D)(0) - P(H_{1i})P_D$  is a measure of the initial LD between allele  $H_{1i}$  and disease locus  $D$ , and  $\Delta_{2s}(0) = P(DH_{2s})(0) - P_DP(H_{2s})$  is a measure of the initial LD between

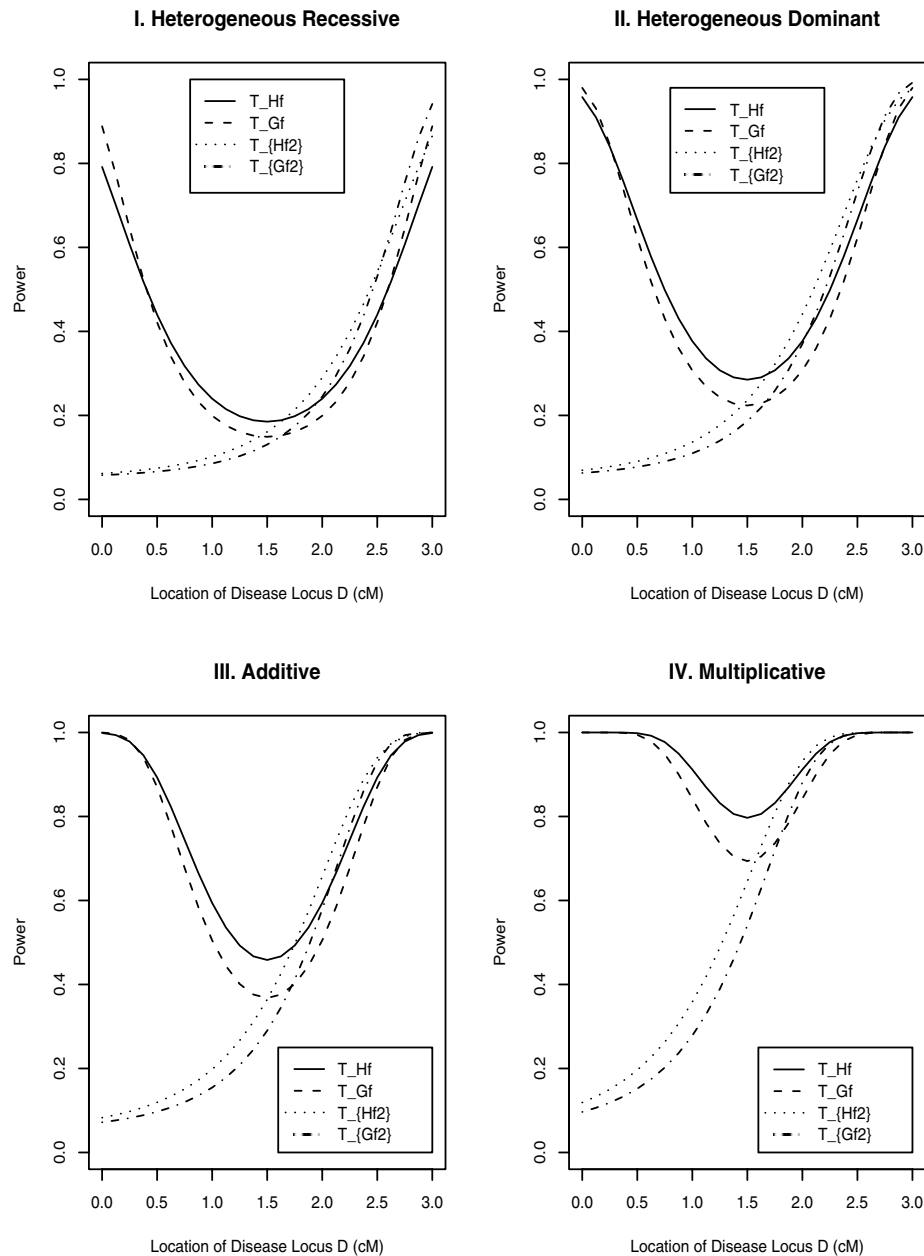


**Figure 2.** Power curves of  $T_{Hf1}$  (i.e., square of  $t_f$ ), 1-TDT and  $T_{Gf1}$  at a significance level  $\alpha = 0.05$  using a bi-allelic marker  $H_1$ , when  $P(H_{11}) = P(H_{12}) = 0.50$ ,  $P_D = 0.15$ ,  $N = 600$ , and  $\theta_{1D} = 1\text{cM}$  for the second set of parameters of the four genetic models of Table 6.  $\Delta_{11} = P(H_{11}D) - P(H_{11})P_D$  is a measure of linkage disequilibrium between marker  $H_1$  and disease locus  $D$ .

allele  $H_{2s}$  and disease locus  $D$  (Akey *et al.* 2001).

Assume that the distance between the two markers is 3cM. The marker  $H_1$  is located at position 0cM and the marker  $H_2$  is located at position 3cM. Since the location of disease locus  $D$  is unknown,

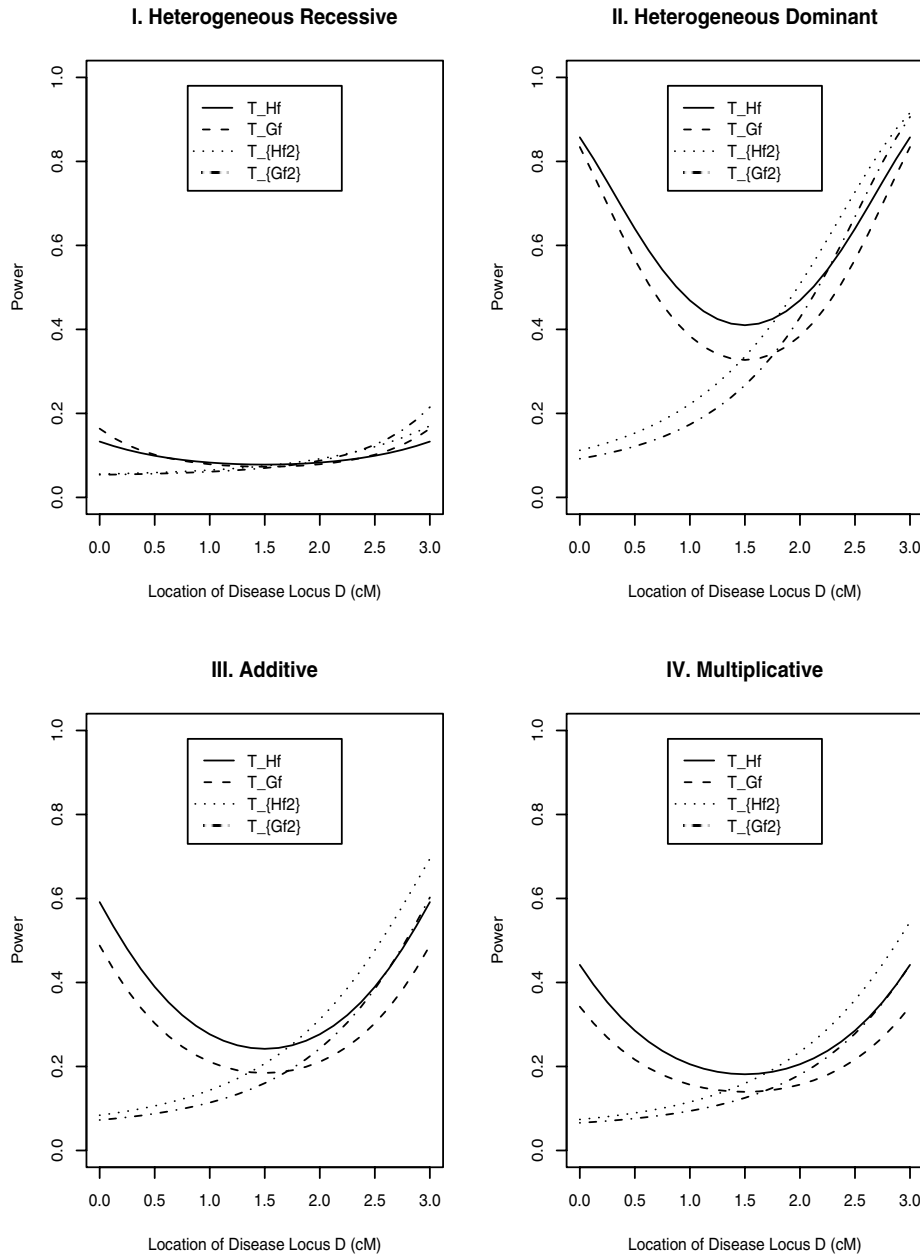
we assume that it is located in the interval between  $H_1$  and  $H_2$ . Figure 3 shows power curves of  $T_{Hf}$ ,  $T_{Gf}$ ,  $T_{Hf2}$  and  $T_{Gf2}$  at a significance level  $\alpha = 0.05$  using two bi-allele markers  $H_1$  and  $H_2$ , when  $P(H_{11}) = P(H_{21}) = 0.50$ ,  $P_D = 0.20$ ,  $T = 75$ ,  $\Delta_{H_1H_2} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.05$ ,



**Figure 3.** Power curves of  $T_{Hf}$ ,  $T_{Gf}$ ,  $T_{Hf2}$  and  $T_{Gf2}$  at a significance level  $\alpha = 0.05$  using two bi-allelic markers  $H_1$  and  $H_2$ , when  $P(H_{11}) = P(H_{21}) = 0.50$ ,  $P_D = 0.15$ ,  $T = 75$ ,  $\Delta_{H_{11}H_{21}} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.05$ , and  $N = 200$  for the first set of parameters of the four genetic models of Table 5. The marker  $H_1$  is located at position 0cM and the marker  $H_2$  is located at position 3cM.

and  $N = 200$  for the first set of parameters of the four genetic models of Table 5. Figure 4 shows power curves of  $T_{Hf}$ ,  $T_{Gf}$ ,  $T_{Hf2}$  and  $T_{Gf2}$  at a significance level  $\alpha = 0.05$  using two bi-allele markers  $H_1$  and  $H_2$ , when  $P(H_{11}) = P(H_{21}) = 0.50$ ,  $P_D = 0.20$ ,  $T = 50$ ,  $\Delta_{H_{11}H_{21}} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.05$ ,

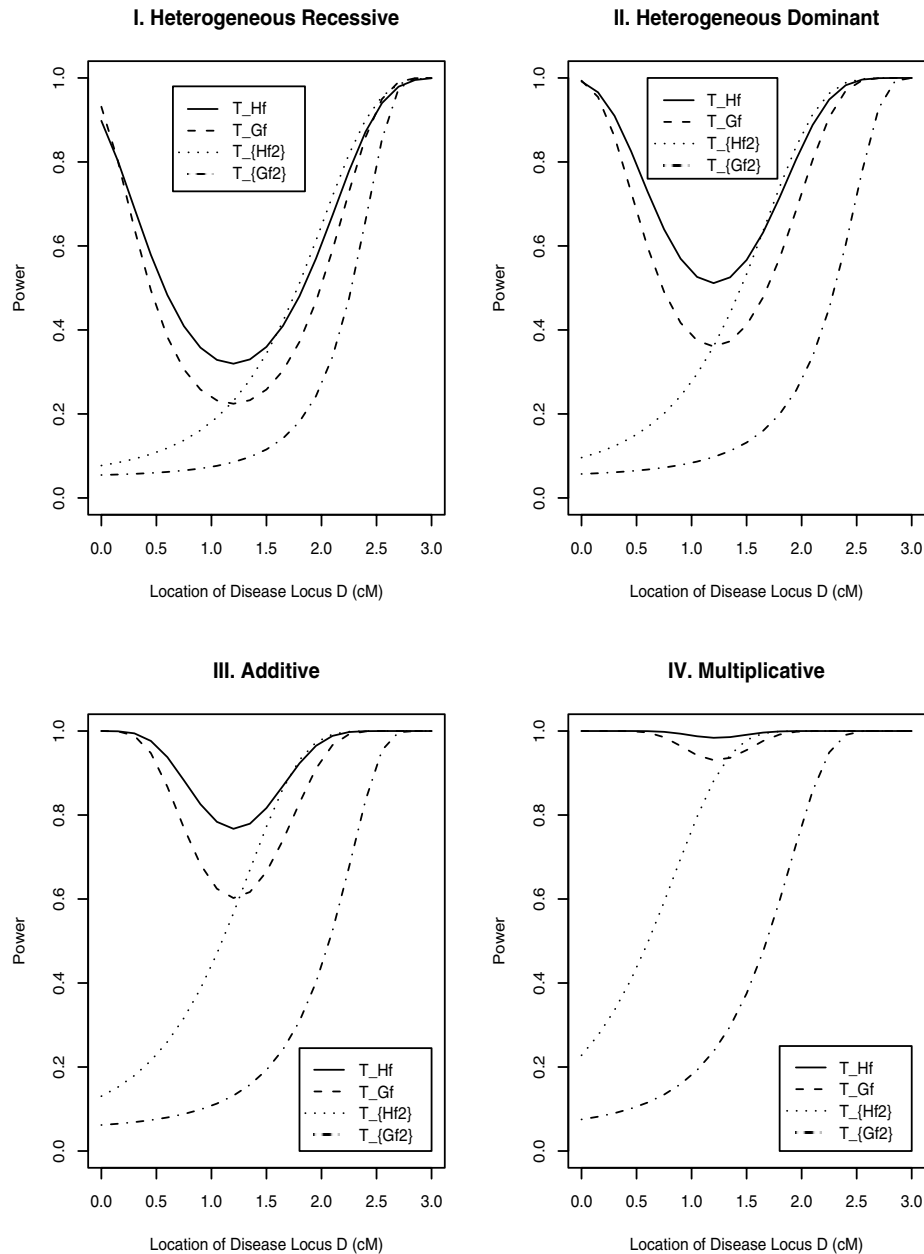
and  $N = 500$  for the second set of parameters of the four genetic models of Table 6.  $T_{Hf}$  (or  $T_{Hf2}$ ) is generally more powerful than  $T_{Gf}$  (or  $T_{Gf2}$ ), most likely because the degrees of freedom of test statistics of haplotype/allele coding are lower than those of genotype coding methods. When the disease locus  $D$



**Figure 4.** Power curves of  $T_{Hf}$ ,  $T_{Gf}$ ,  $T_{Hf2}$  and  $T_{Gf2}$  at a significance level  $\alpha = 0.05$  using two bi-allelic markers  $H_1$  and  $H_2$ , when  $P(H_{11}) = P(H_{21}) = 0.50$ ,  $P_D = 0.15$ ,  $T = 50$ ,  $\Delta_{H_1H_2} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.05$ , and  $N = 500$  for the second set of parameters of the four genetic models of Table 6. The marker  $H_1$  is located at position 0cM and the marker  $H_2$  is located at position 3cM.

is close to marker  $H_2$ , the power of  $T_{Hf}$  (or  $T_{Gf}$ ) is similar to that of  $T_{Hf2}$  (or  $T_{Gf2}$ ). However, the power of  $T_{Hf}$  (or  $T_{Gf}$ ) is much higher than that of  $T_{Hf2}$  (or  $T_{Gf2}$ ), if the disease locus  $D$  is far away from marker  $H_2$  (i.e., close to marker  $H_1$ ). Hence, it is advantageous to use two markers rather than one in the analysis.

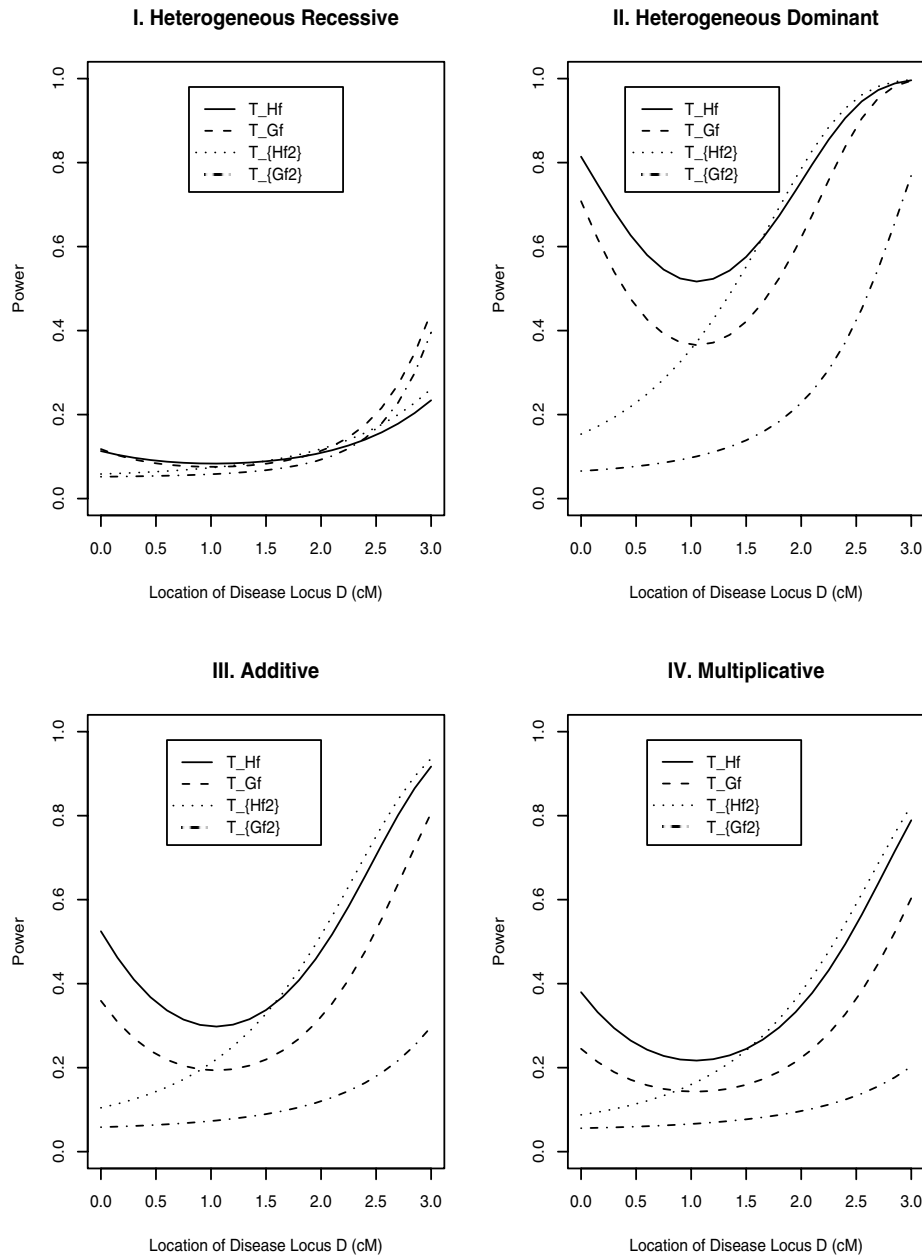
Instead of using two bi-allele markers  $H_1$  and  $H_2$  in the analysis, Figures 5 and 6 consider a situation where marker  $H_1$  is bi-allelic and marker  $H_2$  is quadri-allelic. The power curves of  $T_{Hf}$ ,  $T_{Gf}$ ,  $T_{Hf2}$  and  $T_{Gf2}$  are plotted in Figures 5 and 6 for the first set and the second set of parameters of the four genetic models of



**Figure 5.** Power curves of  $T_{Hf}$ ,  $T_{Gf}$ ,  $T_{Hf2}$  and  $T_{Gf2}$  at a significance level  $\alpha = 0.05$  using a bi-allelic marker  $H_1$  and a quadri-allelic marker  $H_2$ , when  $P(H_{11}) = 0.50$ ,  $P(H_{21}) = \dots = P(H_{24}) = 0.25$ ,  $P_D = 0.15$ ,  $T = 75$ ,  $\Delta_{H_{11}H_{21}} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.05$ ,  $\Delta_{H_{11}H_{22}} = P(H_{11}H_{22}) - P(H_{11})P(H_{22}) = 0.05$ ,  $\Delta_{H_{11}H_{23}} = P(H_{11}H_{23}) - P(H_{11})P(H_{23}) = -0.05$ ,  $\Delta_{H_{11}H_{24}} = P(H_{11}H_{24}) - P(H_{11})P(H_{24}) = -0.05$ , and  $N = 300$  for the first set of parameters of the four genetic models of Table 5. The marker  $H_1$  is located at position 0cM and the marker  $H_2$  is located at position 3cM.

Tables 5 and 6, respectively. The values of related parameters are given in the legends of the figures. Compared with Figures 3 and 4, Figures 5 and 6 show that  $T_{Hf}$  (or  $T_{Hf2}$ ) is generally much more

powerful than  $T_{Gf}$  (or  $T_{Gf2}$ ). The reason for this is that the degrees of freedom of test statistics of the genotype coding method increase a lot, if bi-allelic marker  $H_2$  is replaced by a quadri-allele marker.



**Figure 6.** Power curves of  $T_{Hf}$ ,  $T_{Gf}$ ,  $T_{Hf2}$  and  $T_{Gf2}$  at a significance level  $\alpha = 0.05$  using a bi-allelic marker  $H_1$  and a quadri-allelic marker  $H_2$ , when  $P(H_{11}) = 0.50$ ,  $P(H_{21}) = \dots = P(H_{24}) = 0.25$ ,  $P_D = 0.15$ ,  $T = 50$ ,  $\Delta_{H_{11}H_{21}} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.05$ ,  $\Delta_{H_{11}H_{22}} = P(H_{11}H_{22}) - P(H_{11})P(H_{22}) = 0.05$ ,  $\Delta_{H_{11}H_{23}} = P(H_{11}H_{23}) - P(H_{11})P(H_{23}) = -0.05$ ,  $\Delta_{H_{11}H_{24}} = P(H_{11}H_{24}) - P(H_{11})P(H_{24}) = -0.05$ , and  $N = 500$  for the second set of parameters of the four genetic models of Table 6. The marker  $H_1$  is located at position 0cM and the marker  $H_2$  is located at position 3cM.

However, the degrees of freedom of test statistics of the haplotype/allele coding method do not increase much. In fact, the degrees of freedom of  $T_{Gf}$  (or  $T_{Hf}$ ) is 4 (or 2) if both  $H_1$  and  $H_2$  are bi-allelic. If  $H_1$  is bi-allelic

and  $H_2$  is quadri-allelic, the degrees of freedom of  $T_{Gf}$  (or  $T_{Hf}$ ) increase to 11 (or 4). In addition, the power of  $T_{Gf}$  and  $T_{Hf}$  when the disease locus  $D$  is close to marker  $H_2$  is generally higher than that of  $T_{Gf}$

and  $T_{Hf}$  when the disease locus  $D$  is close to marker  $H_1$ .

## Discussion

Population based case control studies are prone to false positives due to the use of inappropriate controls. Alternatively, parents of affected offspring can be used as controls based on nuclear family data. For multiple marker genotype data, one may separately analyze the data one marker a time, which may not be a satisfactory strategy. It is better to simultaneously analyze the data to obtain a single result, instead of different results being obtained by separate analyses, marker by marker. Using parents as controls for affected offspring, paired Hotelling's  $T^2$  statistics are constructed in this paper using genotype data from multiple haplotype blocks or markers. The methods generalize the two sample Hotelling's  $T^2$  test statistics of population based case control association studies (Fan & Knapp 2003; Xiong *et al.* 2002). Two coding methods, haplotype/allele coding and genotype coding, are proposed to construct paired Hotelling's  $T^2$  test statistics. By power study, it is found that the test statistic based on haplotype/allele coding method is more powerful than the test statistic based on genotype coding method. Moreover, the type I error of the test statistic based on the haplotype/allele coding method is lower than the test statistic based on the genotype coding method. If only one marker is utilized, the power of the test statistic based on haplotype/allele coding is nearly identical to that of 1-TDT (Sun *et al.* 1999). In addition, analysis using multiple markers may provide greater power than a single marker analysis.

There are several advantages in performing the paired Hotelling's  $T^2$  tests  $T_H$  and  $T_G$  described above. First, the procedure to perform the tests is simple and straightforward, and is easy to implement by any statistical software such as SAS and Splus. Second, the methods can not only be used for bi-allelic markers, but also for multi-allelic markers or haplotype blocks. Third, the methods can be used in analyses for multiple markers or haplotype blocks. Fourth, regardless of whether the haplotype phase of multiple loci is known or unknown, the methods are feasible. Fifth, using multiple markers/haplotypes to perform Hotelling's  $T^2$  tests, may give a unique result, instead of performing analyses marker

by marker to get multiple results, which may complicate the interpretation of the study.

The proposed methods are applied to data of a German asthma family study (Gohlke *et al.* 2004). The results based on our paired Hotelling's  $T^2$  statistic tests confirm the findings of Gohlke *et al.* (2004). However, our methods produce much smaller P-values than those of Gohlke *et al.* (2004). Notice that the method of Zhao *et al.* (2000) is utilized in the previous two or three marker analyses of Gohlke *et al.* (2004). In Zhao *et al.* (2000), the authors compared their method with other approaches (Clayton, 1999; Clayton & Jones, 1999; Dudbridge *et al.* 2000; Lazzeroni & Lange, 1998; Sethuraman, 1997; Wilson, 1997; Sham, 1997), and concluded that their method is more powerful. In this article, we find that the proposed paired Hotelling's  $T^2$  statistic tests resulted in much smaller P-values than those of Zhao *et al.* (2000), and so the proposed approaches can be more powerful. One explanation for this is that our paired Hotelling's  $T^2$  test statistics take not only the correlation between the haplotype blocks or markers into account, but also take the correlation within each parent-offspring pair into account. Our methods are potentially useful in mapping complex diseases.

In Appendix A, we show that the expectation of the mean coding vector difference is zero under the null hypothesis for both the haplotype and the genotype coding methods, in a case of random mating/HWE in the population. In Appendix B, this assumption is somewhat relaxed by allowing sub-populations so that random mating/HWE is present in each of the sub-populations. In Tables 1–3, the vast majority of small P values are obtained with genotype coding. We would be very cautious to consider this a true positive finding. There is a subtle difference between haplotype and genotype coding. For the haplotype coding, the conditional expectation for a given parental mating type distribution is zero under the null, but this is not true for the genotype coding method. As an example, consider a single marker locus with two alleles and suppose that all parents are heterozygous at that marker locus. Then, in the case of no association, for haplotype coding we have  $Y_i^{(P)} = 1$  for all  $i$  and  $X_i^{(O)} = 2$  with probability  $1/4$ ,  $= 1$  with probability  $1/2$ , and  $= 0$  with probability  $1/4$ . But for genotype coding,  $Y_i^{(P)} = (0, 1)$  for all  $i$  and  $X_i^{(O)} = (1, 0)$  with probability  $1/4$ ,  $= (0, 1)$  with

probability 1/2, and  $=(0,0)$  with probability 1/4. Therefore,  $E(\bar{Y}^{(P)} - \bar{X}^{(O)})$  is different from zero. As a consequence, haplotype/allele coding seems to be more robust than genotype coding. From type I error rate evaluations, it is found that the genotype coding method tends to have increased type I error rates. If for some reason the distribution of parental mating types strongly deviates from the distribution expected under random mating/HWE, the expectation for the genotype coding method is different from zero even in the case of no association. A valid approach would be to perform permutation analysis by both the haplotype/allele and the genotype coding methods. Since the haplotype/allele coding method is more robust, the results of  $T_H$  are more credible. In the meantime, the results of  $T_G$  should be reported as additional information.

$T_{HP}$  and  $T_{GP}$  are valid statistics to test association between the disease locus  $D$  and blocks or markers  $H_1, \dots, H_j$ . If the null hypothesis of no association is true, the marker genotypes of the two parents are independent. Besides,  $T_{HP}$  and  $T_{GP}$  may have higher powers than  $T_{Hf}$  and  $T_{Gf}$  ( $T_{Hm}$  and  $T_{Gm}$ , Supplementary Information), if the marker genotypes of the two parents are independent under the alternative hypothesis. However, it would not be easy to justify that the marker genotypes of two parents are independent. Thus, the power approximations and non-centrality parameters of  $T_{HP}$  and  $T_{GP}$  derived in the Supplementary Information are problematic. For this reason we mainly focus our attention on using  $T_{Hf}$  and  $T_{Gf}$  ( $T_{Hm}$  and  $T_{Gm}$ ) in the power comparison. In practice,  $T_{HP}$  and  $T_{GP}$  can be used in data analysis.

Several issues need further investigation. First, we assume that there are no missing data. For practical genotype data, genotypic information may be missing at some markers for a portion of the sample (Allen *et al.* 2003). Then, the methods need to be updated to handle the missing data problem. Second, for late onset diseases in which parental controls are not available, normal sib controls can be used (Curtis, 1997; Spielman & Ewens, 1998). Paired Hotelling's  $T^2$  test statistics can be constructed using sib controls in the analysis with the concepts proposed in this paper. However, the power and sample size needed when sibs are used as controls are unknown. Third, it would be interesting to combine both population and family data in one analysis.

Fan & Knapp (2003) and Xiong *et al.* (2002) proposed two sample Hotelling's  $T^2$  test statistics for a population based case control study. One may want to combine the two-sample Hotelling's  $T^2$  statistics and the paired Hotelling's  $T^2$  statistics for one joint analysis, if the data consist of two parts: one part is the case control study with unrelated controls, and the other is the family data with parental/sib controls.

### Acknowledgment

We thank two anonymous reviewers for very detailed and thoughtful critiques, which make the paper more clear. R Fan was supported partially by a research fellowship from Alexander von Humboldt Foundation, Germany, and an International Research Travel Assistance Grant, Texas A&M University. M Knapp was supported by grant DFG Kn 378/1 (Project D1 of FOR 423) from the Deutsche Forschungsgemeinschaft. M Wjst was supported by project funded by the Deutsche Forschungsgemeinschaft DFG WI621/5-1, GSF FE 73922, the German National Genome Network UW S15T01. M Xiong was supported by NIH-NIAMS grant IP50-AR44888, NIH grant ES09912 and NIH grant HL74735. We thank Henning Gohlke and Margret Bahnweg for genotyping DNA samples in the German asthma IL1 project.

### Appendix A

Let  $f_{DD}, f_{Dd} = f_{dD}$  and  $f_{dd}$  be the probabilities that an individual with genotypes  $DD, Dd$  and  $dd$  is affected, respectively. Since allele  $D$  is disease susceptible, one may assume  $f_{DD} \geq f_{Dd} \geq f_{dd}$ . Let  $\bar{f}_{DD} = 1 - f_{DD}, \bar{f}_{Dd} = 1 - f_{Dd}$  and  $\bar{f}_{dd} = 1 - f_{dd}$ . For  $j = 1, \dots, J$ , denote the measures of LD between haplotype or allele  $H_{jk}$  of the haplotype block or marker  $H_j$  and the disease locus  $D$  of parent generation by  $\Delta_{jk} = P(H_{jk}D) - P(H_{jk})P_D, k = 1, \dots, n_j$ . Here  $P(H_{jk}D)$  is the frequency of parental haplotype  $H_{jk}D$ . Assume that one's affected status depends only on his/her own genotype at the disease locus. Let  $A_O$  indicate that the offspring is affected. The probability

$$\begin{aligned}
 P(A_O) &= f_{DD}P_D^2 + 2f_{Dd}P_DP_d + f_{dd}P_d^2 \\
 &= \sum_{s,t \in \{D,d\}} f_{st}P_sP_t. \tag{6}
 \end{aligned}$$

Notice that the parental haplotype frequencies  $P(H_{jk}D) = \Delta_{jk} + P(H_{jk})P_D, P(H_{jk}d) = -\Delta_{jk} +$

$P(H_{jk})P_d$ . Let  $\theta_{jD}$  be the recombination fraction between the haplotype block/marker  $H_j$  and disease locus  $D$ . The frequency of genotype  $H_{jk}H_{jl}$  in an affected offspring and a parent can be calculated as follows

$$\begin{aligned} a_{jkl} &= P[G_{ij}^{(O)} = H_{jk}H_{jl} | A_O] \\ &= P[G_{ij}^{(O)} = H_{jk}H_{jl}, A_O] / P(A_O) \\ &= (1 + 1_{(k \neq l)}) \\ &\quad \times \sum_{s,t \in \{D,d\}} f_{st} P_O(H_{jk}s) P_O(H_{jl}t) / P(A_O). \end{aligned} \quad (7)$$

$$\begin{aligned} \bar{a}_{jkl} &= P[G_{ij}^{(f)} = H_{jk}H_{jl} | A_O] \\ &= P[G_{ij}^{(m)} = H_{jk}H_{jl} | A_O] \\ &= (1 + 1_{(k \neq l)}) \sum_{s,t \in \{D,d\}} f_{st} [P(H_{jk}s)P(H_{jl})/2 \\ &\quad + P(H_{jk})P(H_{jl}s)/2] P_t / P(A_O), \end{aligned} \quad (8)$$

where  $P_O(H_{jk}s) = (1 - \theta_{jD})P(H_{jk}s) + \theta_{jD}P(H_{jk})P_s$  is the frequency of haplotype  $H_{jk}s$  of an offspring. Under the null hypothesis of no association between the haplotypes or markers  $H_j, j = 1, 2, \dots, J$ , and the disease locus  $D$ , i.e.,  $\Delta_{jk} = 0$  for all  $k$ , the haplotype frequencies are equal to the product of the allele frequencies, e.g.,  $P_O(H_{jk}s) = P(H_{jk})P_s, P_O(H_{jl}t) = P(H_{jl})P_t, P(H_{jk}s) = P(H_{jk})P_s$  and  $P(H_{jl}s) = P(H_{jl})P_s$ . From equations (7) and (8),  $a_{jkl} = \bar{a}_{jkl} = (1 + 1_{(k \neq l)})P(H_{jk})P(H_{jl})$ . Hence, the expectation  $E(\bar{X}^{(O)} - \bar{Y}^{(f)} | A_O) = E(\bar{X}^{(O)} - \bar{Y}^{(m)} | A_O) = E(\bar{X}^{(O)} - \bar{Y}^{(P)} | A_O) = 0$  for the genotype coding method. For the haplotype/allele coding method, equations (7) and (8) imply

$$\begin{aligned} E(z_{ijk}^{(O)} | A_O) &= 2a_{jkk} + \sum_{l \neq k} a_{jkl} \\ &= 2 \sum_{s,t \in \{D,d\}} f_{st} P_O(H_{jk}s) P_t / P(A_O) \end{aligned} \quad (9)$$

$$\begin{aligned} E(z_{ijk}^{(f)} | A_O) &= E(z_{ijk}^{(m)} | A_O) = 2\bar{a}_{jkk} + \sum_{l \neq k} \bar{a}_{jkl} \\ &= 2 \sum_{s,t \in \{D,d\}} f_{st} [P(H_{jk}s)/2 \\ &\quad + P(H_{jk})P_s/2] P_t / P(A_O). \end{aligned} \quad (10)$$

From equations (9) and (10), expectation  $E(z_{ijk}^{(O)} | A_O) - E(z_{ijk}^{(f)} | A_O) = 2P(H_{jk}) - 2P(H_{jk}) = 0$  for the haplo-

type/allele coding method, under the null hypothesis of no association between the haplotypes or markers  $H_j, j = 1, \dots, J$  and disease locus  $D$ . Similarly,  $E(\bar{X}^{(O)} - \bar{Y}^{(m)} | A_O) = E(\bar{X}^{(O)} - \bar{Y}^{(P)} | A_O) = 0$ .

## Appendix B

Assume that there is population stratification, Such that there are  $\Gamma$  sub-populations with proportions  $\nu_1, \dots, \nu_\Gamma$ . In each of the sub-populations, suppose that random mating/HWE is valid. In sub-population  $\gamma$ , let  $P^{(\gamma)}(H_{jk})$  be the frequency of haplotype or allele  $H_{jk}$  of the haplotype block or marker  $H_j$ ; and  $P_D^{(\gamma)}$  and  $P_d^{(\gamma)}$  be the frequencies of alleles  $D$  and  $d$ , respectively. Let us denote the measures of LD between haplotype or allele  $H_{jk}$  and the disease locus  $D$  of the parent generation by  $\Delta_{jk}^{(\gamma)} = P^{(\gamma)}(H_{jk}D) - P^{(\gamma)}(H_{jk})P_D^{(\gamma)}, k = 1, \dots, n_i$ . Here  $P^{(\gamma)}(H_{jk}D)$  is the frequency of parental haplotype  $H_{jk}D$  in the sub-population  $\gamma$ . Then the joint probability

$$\begin{aligned} P(A_O) &= \sum_{\gamma=1}^{\Gamma} P(A_O | \text{the family is drawn from} \\ &\quad \text{sub-population } \gamma) \nu_\gamma \\ &= \sum_{\gamma=1}^{\Gamma} \nu_\gamma \sum_{s,t \in \{D,d\}} f_{st} P_s^{(\gamma)} P_t^{(\gamma)}. \end{aligned} \quad (11)$$

The frequency of genotype  $H_{jk}H_{jl}$  in the affected offspring and parent can be calculated as follows

$$\begin{aligned} a_{jkl} &= P[G_{ij}^{(O)} = H_{jk}H_{jl} | A_O] \\ &= P[G_{ij}^{(O)} = H_{jk}H_{jl}, A_O] / P(A_O) \\ &= (1 + 1_{(k \neq l)}) \sum_{\gamma=1}^{\Gamma} \nu_\gamma \\ &\quad \times \sum_{s,t \in \{D,d\}} f_{st} P_O^{(\gamma)}(H_{jk}s) P_O^{(\gamma)}(H_{jl}t) / P(A_O) \end{aligned} \quad (12)$$

$$\begin{aligned} \bar{a}_{jkl} &= P[G_{ij}^{(f)} = H_{jk}H_{jl} | A_O] \\ &= P[G_{ij}^{(m)} = H_{jk}H_{jl} | A_O] \\ &= (1 + 1_{(k \neq l)}) \sum_{\gamma=1}^{\Gamma} \nu_\gamma \sum_{s,t \in \{D,d\}} \\ &\quad \frac{f_{st} [P^{(\gamma)}(H_{jk}s)P^{(\gamma)}(H_{jl}) + P^{(\gamma)}(H_{jk})P^{(\gamma)}(H_{jl}s)] P_t^{(\gamma)}}{2P(A_O)}, \end{aligned} \quad (13)$$

where  $P_O^{(\gamma)}(H_{jk}s) = (1 - \theta_{jD})P^{(\gamma)}(H_{jk}s) + \theta_{jD}P^{(\gamma)}(H_{jk})$   $P_s^{(\gamma)}$  is the frequency of haplotype  $H_{jk}s$  in an offspring of population  $\gamma$ . Under the null hypothesis of no association between the haplotypes or markers  $H_i, i = 1, 2, \dots, J$ , and the disease locus  $D$ , i.e.,  $\Delta_{ij}^{(\gamma)} = 0$  for all  $j$  and  $\gamma$ , the haplotype frequencies are equal to the product of the allele frequencies, e.g.,  $P_O^{(\gamma)}(H_{jk}s) = P^{(\gamma)}(H_{jk})P_s^{(\gamma)}$ ,  $P_O^{(\gamma)}(H_{jt}) = P^{(\gamma)}(H_{jt})P_t^{(\gamma)}$ ,  $P^{(\gamma)}(H_{jk}s) = P^{(\gamma)}(H_{jk})P_s^{(\gamma)}$  and  $P^{(\gamma)}(H_{jt}) = P^{(\gamma)}(H_{jt})P_t^{(\gamma)}$ . From equations (12) and (13),  $a_{jkl} = \bar{a}_{jkl} = (1 + 1_{(k \neq l)}) \sum_{\gamma=1}^{\Gamma} v_{\gamma} P^{(\gamma)}(H_{jk})P^{(\gamma)}(H_{jt}) \sum_{s,t \in \{D,d\}} f_{st} P_s^{(\gamma)} P_t^{(\gamma)} / P(A_O)$ . Hence, the expectation  $E(\bar{X}^{(O)} - \bar{Y}^{(f)} | A_O) = E(\bar{X}^{(O)} - \bar{Y}^{(m)} | A_O) = E(\bar{X}^{(O)} - \bar{Y}^{(P)} | A_O) = 0$  for the genotype coding method.

For the haplotype/allele coding method, equations (12) and (13) imply

$$E(z_{ijk}^{(O)} | A_O) = 2a_{jkk} + \sum_{l \neq k} a_{jkl} = 2 \sum_{\gamma=1}^{\Gamma} v_{\gamma} \sum_{s,t \in \{D,d\}} f_{st} P_O^{(\gamma)}(H_{jk}s) P_t^{(\gamma)} / P(A_O) \tag{14}$$

$$E(z_{ijk}^{(f)} | A_O) = E(z_{ijk}^{(m)} | A_O) = 2\bar{a}_{jkk} + \sum_{l \neq k} \bar{a}_{jkl} = 2 \sum_{\gamma=1}^{\Gamma} v_{\gamma} \sum_{s,t \in \{D,d\}} \frac{f_{st} [P^{(\gamma)}(H_{jk}s) + P^{(\gamma)}(H_{jk})P_s^{(\gamma)}] P_t^{(\gamma)}}{2P(A_O)} \tag{15}$$

From equations (14) and (15), expectation  $E(z_{ijk}^{(O)} | A_O) - E(z_{ijk}^{(f)} | A_O) = 0$  for the haplotype/allele coding method, under the null hypothesis of no association between the haplotypes or markers  $H_j, j = 1, \dots, J$  and disease locus  $D$ . Similarly,  $E(\bar{X}^{(O)} - \bar{Y}^{(m)} | A_O) = E(\bar{X}^{(O)} - \bar{Y}^{(P)} | A_O) = 0$ .

### References

Abecasis, G. R., Cardon, L. R. & Cookson, W. O. C. (2000) A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* **66**, 279–292.  
 Akey, J., Jin, L. & Xiong, M. M. (2001) Haplotype vs single marker linkage disequilibrium tests: what do we gain? *Eur J of Human Genetics* **9**, 291–300.

Allen, A. S., Rathouz, P. J. & Satten, G. A. (2003) Informative missingness in genetic association studies: case-parent designs. *Am J Hum Genet* **72**, 671–680.  
 Anderson, T. W. (1984) *An Introduction to Multivariate Statistical Analysis* 2nd edition. New York, Wiley.  
 Chapman, N. H. & Wijsman, E. M. (1998) Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. *Am J Hum Genet* **63**, 1872–1885.  
 Clayton, D. G. (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* **65**, 1170–1177.  
 Clayton, D. G. & Jones, H. (1999) Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet* **65**, 1161–1169.  
 Curtis, D. (1997) Use of siblings as controls in case-control association studies. *Ann Hum Genet* **61**, 319–333.  
 Dudbridge, F., Koeleman, B. P. C., Todd, J. A. & Clayton, D. G. (2000) Unbiased application of the transmission/disequilibrium test to multi-locus haplotypes. *Am J Hum Genet* **66**, 2009–2012.  
 Ewens, W. J. & Spielman, R. S. (1995) The transmission/disequilibrium test: history, subdivision, and admixture. *Am J of Hum Genet* **57**, 455–464.  
 Falk, C. T. & Rubinstein, P. (1987) Haplotype relative risk: an easy reliable way to construct a proper control sample for risk calculations. *Ann of Hum Genet* **51**, 227–233.  
 Fan, R. Z. & Knapp, M. (2003) Genome association studies of complex diseases by case-control designs. *Am J of Hum Genet* **72**, 850–868.  
 Gohlke, H., Illig, T., Bahnweg, M., Klopp, N. *et al.* (2004) Association of the interleukin 1 receptor antagonist gene. *Am Rev Res Crit Care Med* **169**, 1217–1223.  
 Hotelling, H. (1931) The generalization of student's ratio. *Ann Math Stat* **2**, 360–378.  
 Kaplan, N. & Martin, E. R. (2001) Power calculations for a general class of tests of linkage and association that use nuclear families with affected and unaffected sibs. *Theoretical Population Biology* **60**, 193–201.  
 Lazzeroni, L. C. & Lange, K. (1998) A conditional inference framework for extending the transmission/disequilibrium test. *Hum Hered* **48**, 67–81.  
 Nielsen, D. M., Ehm, M. G. & Weir, B. S. (1998) Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am J Hum Genet* **63**, 1531–1540.  
 Nielsen, D. M. & Weir, B. S. (2001) Association studies under general disease models. *Theoretical Population Biology* **60**, 253–263.  
 Olson, J. M. & Wijsman, E. M. (1994) Design and sample size considerations in the detection of linkage disequilibrium with a disease locus. *Am J Hum Genet* **55**, 574–580.  
 Ott, J. (1989) Statistical properties of the haplotype relative risk. *Genetic Epidemiology*, **6**, 127–130.

- Schaid, D. J. (1996) General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* **13**, 423–449.
- Schaid, D. J. & Rowland, C. (1998) Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. *Am J Hum Genet* **63**, 1492–1506.
- Sethuraman, B. (1997) *Topics in statistical genetics*. Ph.D diss, University of California, Berkeley.
- Sham, P. (1997) The transmission/disequilibrium tests for multiallelic loci. *Am J Hum Genet* **61**, 774–778.
- Spielman, R. S. & Ewens, W. J. (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* **62**, 450–458.
- Spielman, R. S., McGinnis, R. E. & Ewens, W. J. (1993) Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**, 506–516.
- Sun, F. Z., Flanders, W. D., Yang, Q. H., & Khoury, M. J. (1999) Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. *Am J of Epidemiology*, **150**, 97–104.
- Wilson, S. R. (1997) On extending the transmission/disequilibrium test (TDT). *Ann Hum Genet* **61**, 151–161.
- Xiong, M. M., Zhao, J. & Boerwinkle, E. (2002) Generalized  $T^2$  test for genome association studies. *Am J Hum Genet* **70**, 1257–1268.
- Zhao, H. Y., Zhang, S. L., Merikangas, K. R. *et al.* (2000) Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet* **67**, 936–946.

Received: 23 April 2004

Accepted: 20 August 2004