Sibship T^2 association tests of complex diseases for tightly linked markers

Ruzong Fan[1]* *and Michael Knapp*[2]

[1]Department of Statistics, The Texas A&M University, 447 Blocker Building, College Station, Texas 77843-3143, USA
[2]Institute of Medical Biometry, Informatics and Epidemiology, University of Bonn, Sigmund Freud Strasse 25, D-53105 Bonn, Germany
* *Correspondence to*: Tel: +1 979 845 3156; Fax: +1 979 845 3144; E-mail: rfan@stat.tamu.edu

Date received (in revised form): 28th April 2005

Abstract

For population case-control association studies, the false-positive rates can be high due to inappropriate controls, which can occur if there is population admixture or stratification. Moreover, it is not always clear how to choose appropriate controls. Alternatively, the parents or normal sibs can be used as controls of affected sibs. For late-onset complex diseases, parental data are not usually available. One way to study late-onset disorders is to perform sib-pair or sibship analyses. This paper proposes sibship-based Hotelling's T^2 test statistics for high-resolution linkage disequilibrium mapping of complex diseases. For a sample of sibships, suppose that each sibship consists of at least one affected sib and at least one normal sib. Assume that genotype data of multiple tightly linked markers/haplotypes are available for each individual in the sample. Paired Hotelling's T^2 test statistics are proposed for high-resolution association studies using normal sibs as controls for affected sibs, based on two coding methods: 'haplotype/allele coding' and 'genotype coding'. The paired Hotelling's T^2 tests take into account not only the correlation among the markers, but also take the correlation within each sib-pair. The validity of the proposed method is justified by rigorous mathematical and statistical proofs under the large sample theory. The non-centrality parameter approximations of the test statistics are calculated for power and sample size calculations. By carrying out power and simulation studies, it was found that the non-centrality parameter approximations of the test statistics were accurate. By power and type I error analysis, the test statistics based on the 'haplotype/allele coding' method were found to be advantageous in comparison to the test statistics based on the 'genotype coding' method. The test statistics based on multiple markers can have higher power than those based on a single marker. The test statistics can be applied not only for bi-allelic markers, but also for multi-allelic markers. In addition, the test statistics can be applied to analyse the genetic data of multiple markers which contain double heterozygotes — that is, unknown linkage phase data. An SAS macro, Hotel_sibs.sas, is written to implement the method for data analysis.

Keywords: *linkage disequilibrium mapping, complex diseases*

Introduction

In recent years, there has been great interest in the research of association studies of complex diseases.[1−6] By association studies, we mean linkage disequilibrium (LD) mapping of genetic traits. For population case-control studies, the marker allele frequency in cases can be compared with that of controls using χ^2 test statistics.[7−11] If there is association between one marker and the trait locus, it is expected that the χ^2 tests would lead to significant results. Essentially, this method can be applied to analyse the data for one marker at a time. For multiple markers, the linkage phase may be unknown,[12] and the method cannot be applied simultaneously to analyse the data of multiple markers which contain double heterozygotes. With the development of dense maps such as single nucleotide polymorphisms (SNPs), haplotype maps and high-resolution micro-satellites in the human genome, enormous amounts of genetic data on human chromosomes are becoming available.[13−15] It is interesting when building appropriate models and useful algorithms in association mapping of complex diseases to have the ability to use multiple markers/haplotypes simultaneously.

For tightly linked genetic markers, one may perform association studies of complex diseases based on the Hotelling's T^2 test statistics.[16] For population case-control data, Xiong *et al*. proposed two sample Hotelling's T^2 test statistics to analyse genotype data of multiple bi-allelic markers such as SNPs;[17] in addition, logistic regression models were proposed.[2,18] To analyse the multi-allelic micro-satellite or haplotype data, Fan and Knapp extended Xiong *et al*. method using two coding methods — 'haplotype/allele coding' and 'genotype coding'.[19] For the genetic data of nuclear families or parent–offspring pairs, paired Hotelling's T^2 test statistics were proposed, in order to perform association studies based on multiple markers/haplotypes.[20]

For late-onset complex diseases, parental data are usually not available. One way to study late-onset disorders is to perform sib-pair or sibship analyses.[21,22] This paper proposes sibship-based paired Hotelling's T^2 test statistics for high-resolution LD mapping of complex diseases. For a sample of sibships, suppose that each sibship consists of at least one affected sib and at least one normal sib. Assume that genotype data for multiple markers are available for each individual in the sample. Paired Hotelling's T^2 test statistics are proposed for high-resolution association studies, using normal sibs as controls for affected sibs. The paired Hotelling's T^2 tests not only take the correlation among the markers into account, but also the correlation within each sib-pair. The validity of the proposed method is justified by rigorous mathematical and statistical proofs under the large sample theory. The non-centrality parameter approximations of the test statistics are calculated for power calculations and comparisons; these are included in the section: Supplementary information: Non-centrality parameters. Type I error rates are calculated by simulations to evaluate the performance of the proposed test statistics. In the section: Supplementary information: Simulation study, the results from the simulation study are presented, to show that the non-centrality parameter approximations of the test statistics are accurate. An SAS macro, Hotel_sibs.sas, was written to implement the method and can be downloaded from the authors' website (http://www.stat.tamu.edu/~rfan/software.html/).

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.

Paired Hotelling's T^2 test statistics

In the region of the disease locus D, assume that J tightly linked markers $H_1, \ldots, H_J$ are typed. By tightly linked, we mean that the markers are so close to each other that the recombination fractions among markers are 0. Let us denote the alleles of marker H_j by $H_{j1}, \ldots, H_{jn_j}$, where n_j denotes the number of its alleles. Here, markers can be micro-satellites or di-allelic markers such as SNPs or haplotypes. If $H_1, \ldots, H_J$ are phase-known haplotypes, the methods developed in this paper are still valid, since the haplotypes can be treated as markers; but the related terminology needs to be changed accordingly. Usually, haplotypes consist of phase-unknown markers; in these cases, we prefer to analyse the genotype marker data directly, instead of estimating the haplotypes first and then analysing the haplotype data. The method developed in this paper can be used to analyse phase-unknown genotype data directly. Consider N sib-pairs, each consisting of an affected sibling and a normal sibling. We define coding vectors $X_i^{(A)}$ and $Y_i^{(U)}$ for the affected sibling and normal sibling of the i-th sib-pair, respectively, by one of the following two ways.[19,20]

(i) *Haplotype/allele coding*: For the affected sibling of the i-th sib-pair, let $G_{ij}^{(A)}$ be his/her genotype at marker H_j. Define $X_i^{(A)} = (z_{i11}^{(A)}, \ldots, z_{i1(n_1-1)}^{(A)}, , \ldots, z_{iJ1}^{(A)}, \ldots, z_{iJ(n_J-1)}^{(A)})^\tau$, where $z_{ijk}^{(A)}$ is the number of alleles H_{jk} for the affected sibling of the i-th sib-pair — that is,

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

Here and hereafter, the superscript τ denotes the transposition of a matrix or a vector. The dimension of $X_i^{(A)}$ is $(n_1 - 1) + \cdots + (n_J - 1) = \sum_{j=1}^J n_j - J$, which is usually smaller than dimension $\sum_{j=1}^J n_j(n_j + 1)/2 - J$ of the following genotype coding method.

(ii) *Genotype coding*: Note that $G_{ij}^{(A)}$ 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 $X_{ij}^{(A)} = (x_{ij1}^{(A)}, \ldots, x_{ij(n_j-1)}^{(A)}, x_{ij12}^{(A)}, \ldots, x_{ij1n_j}^{(A)}, \ldots, x_{ij(n_j-1)n_j}^{(A)})^\tau$. Here, $x_{ijk}^{(A)}$ is the indicator variable of genotype $H_{jk}H_{jk}$ defined by $x_{ijk}^{(A)} = \begin{cases} 1 & \text{if } G_{ij}^{(A)} = H_{jk}H_{jk} \\ 0 & \text{else} \end{cases}$; and $x_{ijkl}^{(A)}, k < l$ is the indicator variable of genotype $H_{jk}H_{jl}$ defined by $x_{ijkl}^{(A)} = \begin{cases} 1 & \text{if} G_{ij}^{(A)} = H_{jk}H_{jl} \\ 0 & \text{else} \end{cases}$. The dimension of $X_{ij}^{(A)}$ is $n_j(n_j + 1)/2 - 1$ — that is, the total number $n_j(n_j + 1)/2$ of genotypes of marker H_j minus 1 to remove the redundancy. Let $X_i^{(A)} = (X_{i1}^{(A)^\tau}, \ldots, X_{iJ}^{(A)^\tau})^\tau$ be the combined genotype coding of the J markers $H_1, \ldots H_J$. The dimension of $X_i^{(A)}$ is $\sum_{j=1}^J n_j(n_j + 1)/2 - J$.

For the unaffected sibling of the i-th sib-pair, let $G_{ij}^{(U)}$ be his/her genotype at marker H_j. One may define a vector $Y_i^{(U)}$ in the same way, based on either the 'genotype coding' or 'haplotype/allele coding' method. Table 1 in reference 19 gives an example of 'genotype coding' and 'haplotype/allele coding' for a marker with three alleles, to illustrate the above two coding methods.

Let $\bar{X}^{(A)} = \sum_{i=1}^N X_i^{(A)}/N$ and $\bar{Y}^{(U)} = \sum_{i=1}^N Y_i^{(U)}/N$ be average coding vectors of affected and unaffected siblings, respectively. Intuitively, $\bar{X}^{(A)}$ and $\bar{Y}^{(U)}$ should be similar vectors if the disease locus D is not associated with markers H_j, $j = 1, \ldots, J$. In the Appendix we prove that the expected value of $\bar{X}^{(A)} - \bar{Y}^{(U)}$ is 0 if there is no association. Hence, one may build a test statistic based on the difference $\bar{X}^{(A)} - \bar{Y}^{(U)}$ to test the association between disease locus D and markers H_j. To do

Table 1. Type I error rates of $N = 200$ or 300 sib-pairs at a significance level $\alpha = 0.01$ using one marker, H_1, or two markers, H_1 and H_2. In model I, one bi-allelic marker H_1 is used, $P(H_{11}) = P(H_{12}) = 0.50$. In model II, two bi-allelic markers H_1 and H_2 are used, $P(H_{ij}) = 0.5$, $i, j = 1, 2$, $\Delta_{H_{11}H_{21}} = 0.05$. In model III, one quadric-allelic marker H_1 is used, $P(H_{21}) = P(H_{22}) = 0.35$, $P(H_{23}) = P(H_{24}) = 0.15$. Abbreviations: df = degrees of freedom; Std Dev = standard deviation.

Model	Test	df	# type I error rates	Mean	Std Dev	Minimum	Maximum
I	T_H	1	100	0.010808	0.0014264	0.0066	0.0140
N=200	T_G	2	100	0.011240	0.0013923	0.0082	0.0152
II	T_H	2	100	0.011286	0.0014717	0.0070	0.0146
N=200	T_G	4	100	0.012352	0.0015899	0.0088	0.0160
III	T_H	3	100	0.011660	0.0014348	0.0078	0.0146
N=200	T_G	9	100	0.014352	0.0018710	0.0102	0.0196
III	T_H	3	100	0.011186	0.0015669	0.0074	0.0160
N=300	T_G	9	100	0.013076	0.0017027	0.0084	0.0176

this, one needs to consider the variance–covariance matrix of $\bar{X}^{(A)} - \bar{Y}^{(U)}$. Since siblings' marker genotypes are related to each other, $\bar{X}^{(A)}$ and $\bar{Y}^{(U)}$ are not independent. Moreover, $X_i^{(A)}$ and $Y_i^{(U)}$ are paired with each other in a sib-pair. Therefore, paired T^2 test statistics can be used to test the association between disease locus D and markers H_j as follows. Define a paired-sample variance–covariance matrix by

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

A paired Hotelling's T^2 statistic can be defined as $T^2 = N(\bar{X}^{(A)} - \bar{Y}^{(U)})^\tau S^{-1}(\bar{X}^{(A)} - \bar{Y}^{(U)})$.[16,23] Let us denote the above Hotelling's T^2 statistic for 'haplotype/allele coding' as T_H, and the Hotelling's T^2 statistic for 'genotype coding' as T_G. 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_H (or T_G) is asymptotically distributed as central χ^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_H (or T_G) is asymptotically distributed as non-central χ^2. For power calculation and comparison, the non-centrality parameter of statistic T_H or T_G can be derived under the alternative hypothesis of association.

For general sibships each containing at least one affected sibling and at least one normal sibling, the Hotelling's T^2 test statistics T_H and T_G above can be generalised as follows. Assume that N sibships are available. In the i-th sibship, assume that n_i siblings are affected and m_i siblings are normal. Let $\bar{X}_i^{(A)}$ and $\bar{Y}_i^{(U)}$ be average coding vectors of affected and normal siblings, respectively. To be precise, let $X_{ij}^{(A)}, j = 1, \cdots, n_i$ be the coding vectors of the affected siblings of the i-th sibship. Then, $\bar{X}_i^{(A)} = \sum_{j=1}^{n_i} X_{ij}^{(A)}/n_i$; $\bar{Y}_i^{(U)}$ is defined, accordingly. Utilising $\bar{X}_i^{(A)}$ to replace $X_i^{(A)}$ and $\bar{Y}_i^{(U)}$ to replace $Y_i^{(U)}$ in the above paragraph and defining $\bar{X}^{(A)} = \sum_{i=1}^{N} \bar{X}_i^{(A)}/N$ and $\bar{Y}^{(U)} = \sum_{i=1}^{N} \bar{Y}_i^{(U)}/N$, we may define the related Hotelling's T^2 test statistics T_H and T_G.

Non-centrality parameters

The derivation of non-centrality parameters of sib-pairs is provided in the section Supplementary information: Non-centrality parameters.

Results

Type I errors

Tables 1, 2 and 3 show type I error rates of test statistics T_H and T_G 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 model I, one marker H_1 is used in analysis: H_1 is a bi-allelic marker with equal allele frequency $P(H_{11}) = P(H_{12}) = 0.50$. In model II, two bi-allelic markers H_1 and H_2 are used in analysis, where $P(H_{ij}) = 0.5$, $i, j = 1, 2$,

Table 2. Type I error rates of $N = 200$ or 300 sibships at a significance level $\alpha = 0.01$ using one marker, H_1, or two markers, H_1 and H_2. The number of sib-pairs is equal to $N/2$; in each sib-pair, one sibling is affected and the other is normal. The number of sibships of size 3 is $N/2$; in each of $N/4$ sibships of size 3, one is affected and the other two are normal; in the remaining $N/4$ sibships of size 3, two are affected and the other one is normal. The other parameters of each model are the same as those of Table 1. Abbreviations: df = degrees of freedom; Std Dev = standard deviation.

Model	Test	df	# type I error rates	Mean	Std Dev	Minimum	Maximum
I	T_H	1	100	0.010642	0.0014406	0.0062	0.0134
N=200	T_G	2	100	0.011278	0.0015023	0.0076	0.0154
II	T_H	2	100	0.011096	0.0014418	0.0078	0.0154
N=200	T_G	4	100	0.012138	0.0014825	0.0082	0.0154
III	T_H	3	100	0.011536	0.0014156	0.070	0.0158
N=200	T_G	9	100	0.014202	0.0016562	0.096	0.0182
III	T_H	3	100	0.011098	0.0015214	0.0076	0.0152
N=300	T_G	9	100	0.012790	0.0016883	0.0086	0.0186

$\Delta_{H_{11}H_{21}} = 0.05$. In model III, one marker H_1 is used in analysis, where H_1 is a quadri-allelic marker with allele frequencies $P(H_{21}) = P(H_{22}) = 0.35$, $P(H_{23}) = P(H_{24}) = 0.15$.

Each time, 5,000 simulated datasets are generated and each dataset contains $N = 200$ or 300 sibships under the assumption that there is no association between the marker(s) and the disease locus; a type I error rate is then calculated as the proportion of the 5,000 datasets for which the empirical test statistics are greater than, or equal to, the cut-off point at the significance level $\alpha = 0.01$. The process is repeated 100 times. Thus, 100 type I error rates are calculated. The mean, standard deviation, minimum and maximum of the 100 type I error rates are presented in the entries of Tables 1, 2 and 3. Since the disease locus is not associated with the marker(s), the empirical test statistics which are greater than or equal to the cut-off point at the significance level $\alpha = 0.01$ are treated as false positives. Thus, the type I error rates of Tables 1, 2 and 3 are empirical results.

Table 3. Type I error rates of $N = 200$ or 300 sibships at a significance level $\alpha = 0.01$ using one marker, H_1, or two markers, H_1 and H_2. The number of sib-pairs is equal to $N/2$; the number of sibships of size 3 is $N/5$; and the number of sibships of size 4 is $3N/10$. In each sib-pair, one sibling is affected and the other is normal. In each of $N/10$ sibships of size 3, one is affected and the other two are normal; in the remaining $N/10$ sibships of size 3, two are affected and the other is normal. In each of $N/10$ sibships of size 4, one is affected and the other three are normal; in each of $N/10$ sibships of size 4, two are affected and the other two are normal; in the remaining $N/10$ sibships of size 4, three are affected and the other one is normal. The other parameters of each model are the same as those of Table 1. Abbreviations: df = degrees of freedom; Std Dev = standard deviation.

Model	Test	df	# type I error rates	Mean	Std Dev	Minimum	Maximum
I	T_H	1	100	0.010670	0.0014040	0.0072	0.0136
N=200	T_G	2	100	0.011156	0.0015397	0.0066	0.0142
II	T_H	2	100	0.011218	0.0014678	0.0078	0.0166
N=200	T_G	4	100	0.012304	0.0011921	0.0092	0.0156
III	T_H	3	100	0.011518	0.0014639	0.0082	0.015
N=200	T_G	9	100	0.014356	0.0015381	0.0102	0.018
III	T_H	3	100	0.011228	0.0013312	0.0078	0.0160
N=300	T_G	9	100	0.012544	0.0015203	0.0086	0.0182

In Table 1, only sib-pairs are used in the calculations. In each sib-pair, one sibling is affected and the other one is normal. In Table 2, combinations of both sib-pairs and sibships of size 3 are used: the number of sib-pairs is equal to $N/2$; the number of sibships of size 3 is $N/2$; in each of $N/4$ sibships of size 3, one is affected and the other two are normal; in the remaining $N/4$ sibships of size 3, two are affected and the other one is normal. In Table 3, combinations of sib-pairs and sibships of sizes 3 and 4 are used: the number of sib-pairs is equal to $N/2$; the number of sibships of size 3 is $N/5$; and the number of sibships of size 4 is $3N/10$; in each of $N/10$ sibships of size 3, one is affected and the other two are normal; in the remaining $N/10$ sibships of size 3, two are affected and the other one is normal; in each of $N/10$ sibships of size 4, one is affected and the other three are normal; in each of $N/10$ sibships of size 4, two are affected and the other two are normal; in the remaining $N/10$ sibships of size 4, three are affected and the other one is normal.

From the results presented in Tables 1, 2 and 3, it is clear that T_H has a lower type I error than T_G. That is, the test statistic of the 'haplotype/allele coding' method has a lower type I error than the test statistic of the 'genotype coding' method. The 'haplotype/allele coding' method leads to more robust and reliable test statistics. The type I error rates of the test statistic of the 'haplotype/allele coding' method are reasonable for models I, II and III when $N = 200$. In addition, the type I error rates of the test statistic of the 'genotype coding' method are reasonable for models I and II when $N = 200$. The type I error rates of the test statistic for the 'genotype coding' method are slightly higher than the nominal level 0.01 for model III when $N = 200$ and become lower when $N = 300$. Note that the number of degrees of freedom for tests T_G and T_H is 3 and 9, respectively, for model III. Hence, the number of degrees of freedom for test T_G is large for model III. When the number of degrees of freedom for tests is 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.

The results are similar in Tables 1, 2 and 3. Thus, the type I error rates are little affected by the varying structure of the sibships. The reason for this is that we basically take averages of the coding vectors for sibships whose size is larger than 2.

Power calculation and comparison

To make power comparisons, we consider four genetic models: heterogeneous recessive, heterogeneous dominant, additive and multiplicative. For optimistic models, Table 4 gives penetrance probabilities taken from Nielsen *et al.* or Fan and Knapp.[11,19] For less optimistic models, Table 5 lists penetrance probabilities taken from Fan and Knapp.[19] For $j = 1, \ldots, J$, let us denote the measures of LD between allele H_{jk} of the marker H_j and the disease locus D by $\Delta_{jk} = P(H_{jk}D) - P(H_{jk})P_D$, $k = 1, \ldots, n_i$. Here, $P(H_{jk}D)$ is the frequency of haplotype $H_{jk}D$, and $P(H_{jk})$ is the frequency of allele H_{jk}. For two bi-allelic markers H_1 and H_2, let $\Delta_{H_1H_2} = P(H_{11}H_{21}) - P(H_{11})P(H_{21})$ be the measure of LD between the two markers, where $P(H_{11}H_{21})$ is the frequency of haplotype $H_{11}H_{21}$. Assume that the two markers H_1 and H_2 flank the disease locus D in the order H_1DH_2. Let $\Delta_{1D2} = P(H_{11}DH_{21}) - P(H_{11})\Delta_{21} - P_D\Delta_{H_1H_2} - P(H_{21})\Delta_{11} - P(H_{11})P_DP(H_{21})$ be the measure of the third order LD.[24] Here, $P(H_{11}DH_{21})$ is the frequency of haplotype $H_{11}DH_{21}$.

Table 4. 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

Figure 1 shows power curves of T_H and T_G against the measure of LD Δ_{11} at a significance level $\alpha = 0.05$ using two bi-allelic marker H_1 and H_2, when $P(H_{i1}) = P(H_{i2}) = 0.50$, $i = 1, 2$, $P_D = 0.15$ and $N = 200$ sib-pairs for the first set of parameters of the four genetic models of Table 4. The power curves of T_{H1} and T_{G1} are calculated based on one marker H_1. In the graphs, Delta_11 = Δ_{11}; the other parameters are given in the legend of the Figure. Figure 2 shows power curves of T_H and T_G against the measure of LD Δ_{11} at a significance level $\alpha = 0.05$ using two bi-allelic marker H_1 and H_2, when $P(H_{i1}) = P(H_{i2}) = 0.50$, $i = 1, 2$, $P_D = 0.15$ and $N = 600$ sib-pairs for the second set of parameters of the four genetic models listed in Table 5. Similarly to Figure 1, the power curves of T_{H1} and T_{G1} are calculated based on one marker H_1. The other parameters are the same as those of Figure 1.

From Figures 1 and 2, it is clear that T_H generally has a higher power than that of T_G. This is consistent with the results of Fan and Knapp for population case-control studies and Fan *et al.* for nuclear family data.[19,20] This is most likely due to the large number of degrees of freedom of the test statistic T_G. The power of T_H (or T_G) based on two markers H_1 and H_2 is generally higher than that of T_{H1} (or T_{G1}), which is only based on one marker H_1. Hence, it is advantageous to use two markers rather than one marker in the analysis. This observation can be generalised — that is, it is

Table 5. 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

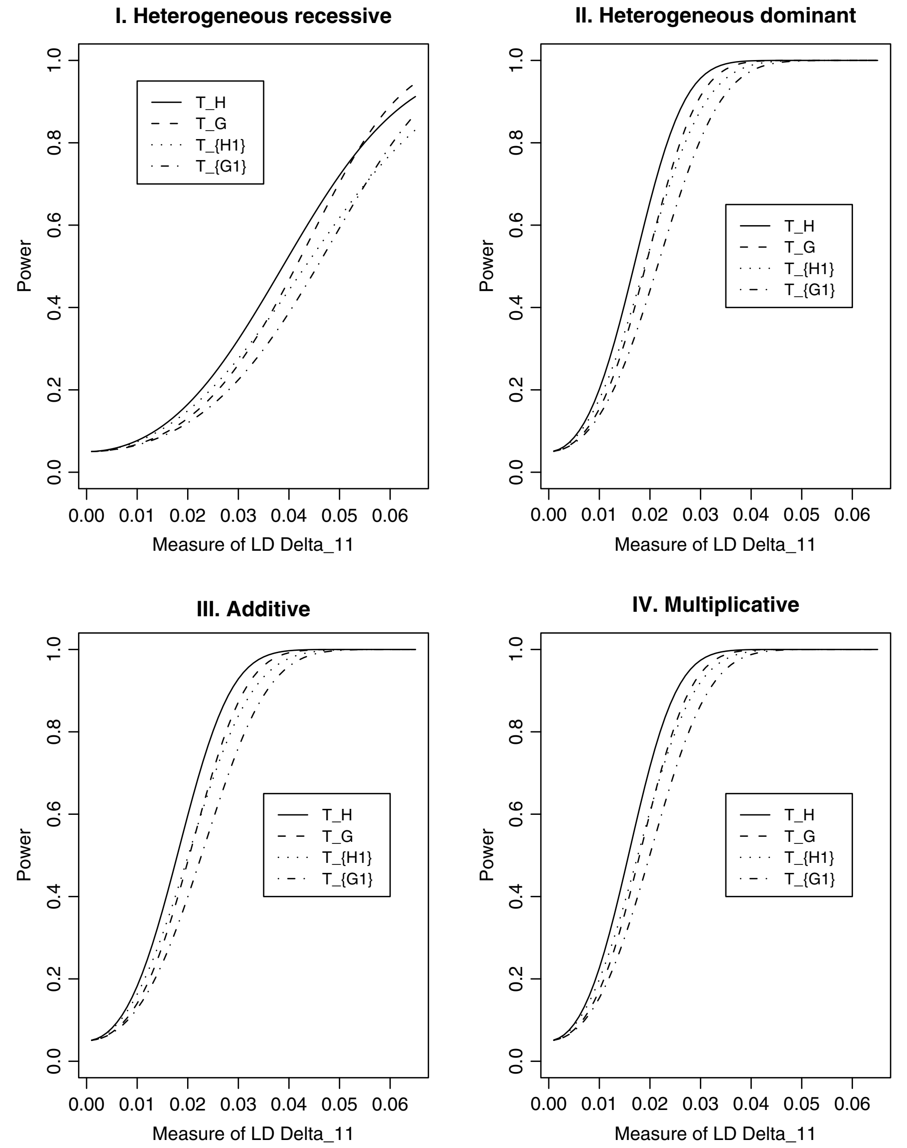

Figure 1. Power curves of T_H and T_G at a significance level $\alpha = 0.05$, using two bi-allelic markers H_1 and H_2, when $P(H_{i1}) = P(H_{i2}) = 0.50$, $i = 1{,}2$, $P_D = 0.15$, and $N = 200$ sib-pairs for the first set of parameters of the four genetic models of Table 4. The power curves of T_{H1} and T_{G1} are calculated based on one marker H_1. In the graphs, Delta_11 $= \Delta_{11} = P(H_{11}D) - P(H_{11})P_D$ is a measure of linkage disequilibrium (LD) between marker H_1 and disease locus D; in addition, the other parameters are given by $\Delta_{21} = P(H_{21}D) - P(H_{21})P_D = \Delta_{11}$, $\Delta_{H_1H_2} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.05$, and $\Delta_{1D2} = P(H_{11}DH_{21}) - P(H_{11})\Delta_{21} - P_D\Delta_{H_1H_2} - P(H_{21})\Delta_{11} - P(H_{11})P_DP(H_{21}) = \Delta_{11}/3$.

advantageous to use multiple tightly linked markers in analysis. Note that the number of degrees of freedom of test statistic T_G can increase rapidly as the number of markers increases. This is particularly true when multi-allelic markers are used in analysis; but the number of degrees of freedom of T_H only increases by one if one more bi-allelic marker is added to the analysis. Thus, T_H has the advantage of high power when multiple markers are used; in addition, the number of degrees of freedom of T_H would be not very large. For optimistic models in Table 4, the sample sizes required to achieve certain power levels are lower than those of the less optimistic models in Table 5.

Not only can the test statistics T_H and T_G be applied to analyse the genetic data of the bi-allelic markers, but they can also be applied to analyse the genetic data of the multi-allelic markers. Figure 3 shows the power curves of T_H and T_G against the measure of LD Δ_{11} at a significance level $\alpha = 0.05$ using a quadri-allelic marker H_1, when $P(H_{11}) = P(H_{12}) = 0.35$,

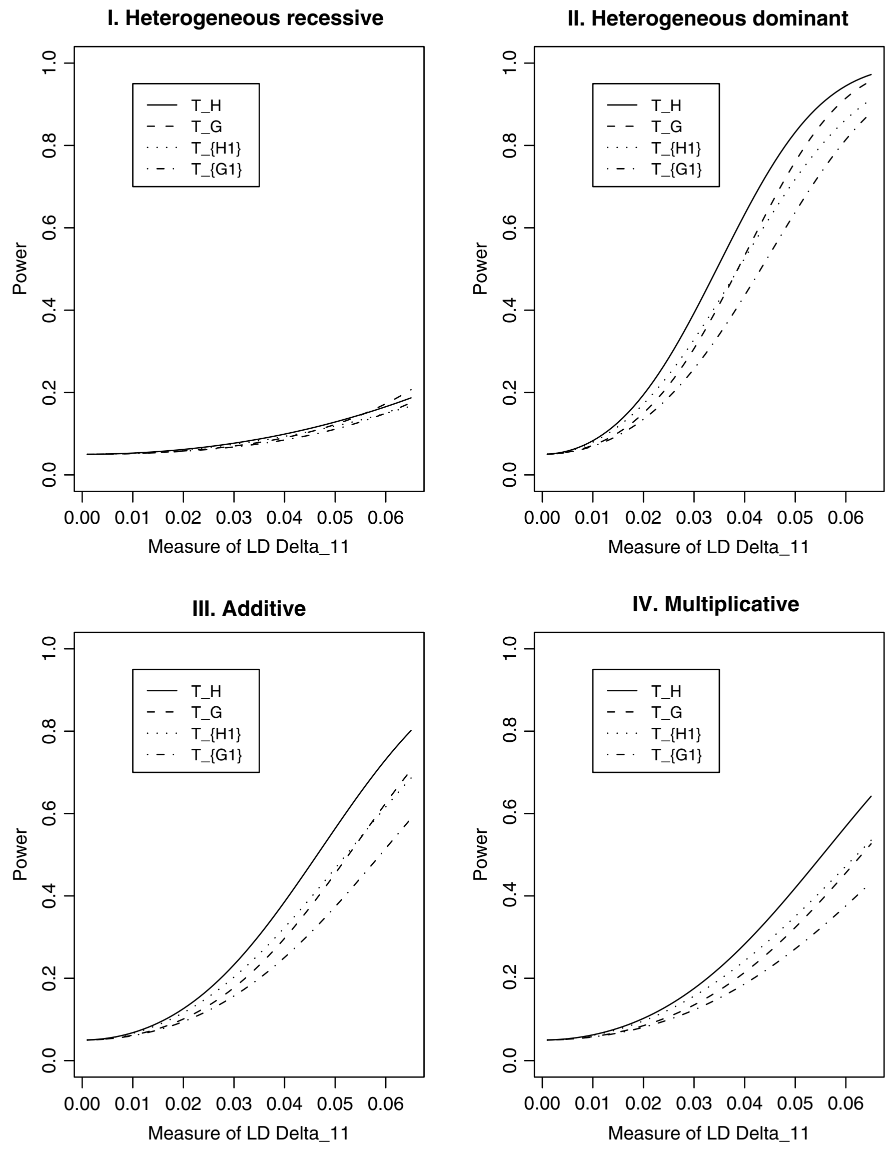

Figure 2. Power curves of T_H and T_G at a significance level $\alpha = 0.05$, using two bi-allelic markers H_1 and H_2, when $P(H_{i1}) = P(H_{i2}) = 0.50$, $i = 1,2$, $P_D = 0.15$ and $N = 600$ sib-pairs for the second set of parameters of the four genetic models of Table 5. The power curves of T_{H1} and T_{G1} are calculated based on one marker H_1. In the graphs, Delta_11 $= \Delta_{11} = P(H_{11}D) - P(H_{11})P_D$ is a measure of linkage disequilibrium (LD) between marker H_1 and disease locus D; in addition, the other parameters are given by $\Delta_{21} = P(H_{21}D) - P(H_{21})P_D = \Delta_{11}$, $\Delta_{H_1H_2} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.05$, and $\Delta_{1D2} = P(H_{11}DH_{21}) - P(H_{11})\Delta_{21} - P_D\Delta_{H_1H_2} - P(H_{21})\Delta_{11} - P(H_{11})P_DP(H_{21}) = \Delta_{11}/3$.

$P(H_{13}) = P(H_{14}) = 0.15$, $P_D = 0.15$ and $N = 200$ sib-pairs for the first set of parameters of the four genetic models of Table 4. The other parameters are given in the legend of the Figure. Figure 4 shows power curves of T_H and T_G at a significance level $\alpha = 0.05$ using a quadri-allelic marker H_1, when $P(H_{11}) = P(H_{12}) = 0.35$, $P(H_{13}) = P(H_{14}) = 0.15$, $P_D = 0.15$ and $N = 600$ sib-pairs for the second set of parameters of the four genetic models of Table 5. Similarly to Figures 1 and 2, T_H generally has a higher power than that of T_G.

In addition to the power curves of T_H and T_G, which are based on sib-pair data, Figures 3 and 4 show the simulated power curves of ST_H and ST_G, which are based on sibships of varying structures. In Figure 3, combinations of both sib-pairs and sibships of size 3 are used to calculate the simulated power curves of ST_H and ST_G: the number of sib-pairs is equal to $N/2 = 100$; the number of sibships of size 3 is $N/2 = 100$; in each of $N/4 = 50$ sibships of size 3, one is affected and the other two are normal; in the remaining

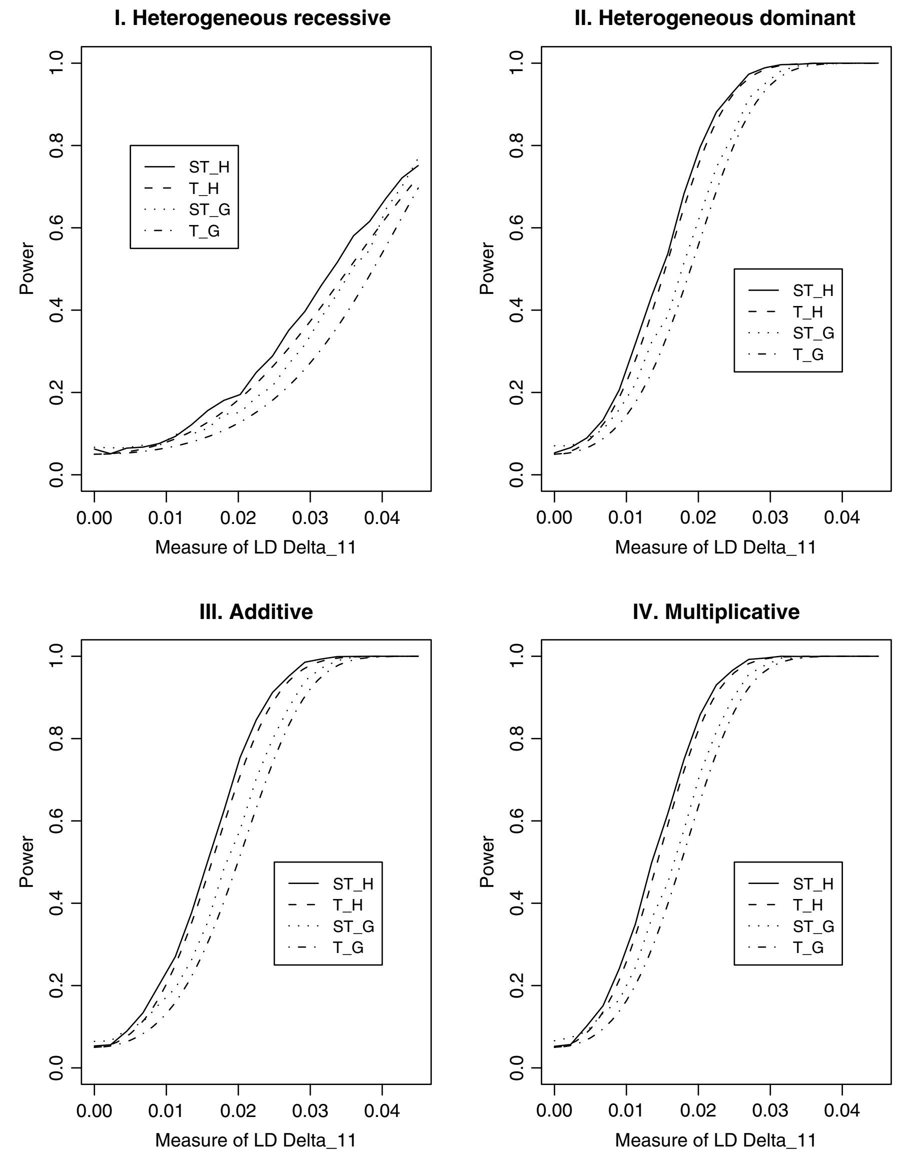

Figure 3. Power curves of T_H and T_G at a significance level $\alpha = 0.05$ using a quadric-allelic marker H_1, when $P(H_{11}) = P(H_{12}) = 0.35$, $P(H_{13}) = P(H_{14}) = 0.15$ $P_D = 0.15$ and $N = 200$ sib-pairs for the first set of parameters of the four genetic models of Table 4. Delta_11 $= \Delta_{11} = P(H_{11}D) - P(H_{11})P_D$ is a measure of linkage disequilibrium (LD) between marker H_1 and disease locus D. In addition, $\Delta_{12} = -\Delta_{11}$, $\Delta_{13} = -\Delta_{14} = \Delta_{11}/2$. The simulated power curves of ST_H and ST_G are calculated using combinations of both sib-pairs and sibships of size 3: the number of sib-pairs is equal to $N/2 = 100$; the number of sibships of size 3 is $N/2 = 100$; in each of $N/4 = 50$ sibships of size 3, one is affected and the other two are normal; in the remaining $N/4 = 50$ sibships of size 3, two are affected and the other one is normal.

$N/4 = 50$ sibships of size 3, two are affected and the other one is normal. In Figure 4, combinations of sib-pairs and sibships of sizes 3 and 4 are used to calculate the simulated power curves of ST_H and ST_G: the number of sib-pairs is equal to $N/2 = 300$; the number of sibships of size 3 is $N/5 = 120$; and the number of sibships of size 4 is $3N/10 = 180$; in each of $N/10 = 60$ sibships of size 3, one is affected and the other two are normal; in the remaining $N/10 = 60$ sibships of size 3, two are affected and the other one is normal; in each of $N/10 = 60$ sibships of size 4, one is affected and the other three are normal; in each of $N/10 = 60$ sibships of size 4, two are affected and the other two are normal; in the remaining $N/10 = 60$ sibships of size 4, three are affected and the other one is normal.

To calculate the simulated power curves ST_H and ST_G, the interval (0, 0.045) of the LD measure Δ_{11} of LD is

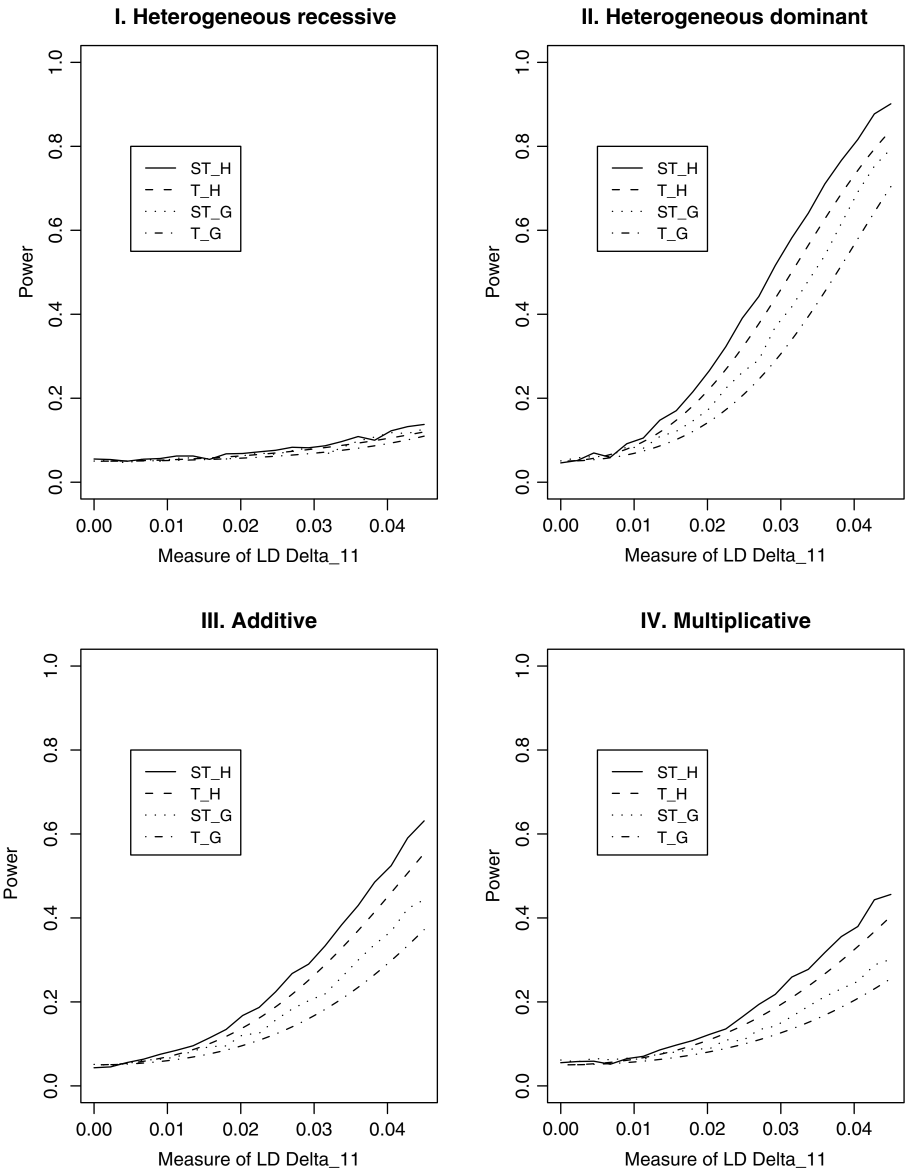

Figure 4. Power curves of T_H and T_G at a significance level $\alpha = 0.05$ using a quadric-allelic marker H_1, when $P(H_{11}) = P(H_{12}) = 0.35$, $P(H_{13}) = P(H_{14}) = 0.15$, $P_D = 0.15$ and $N = 600$ sib-pairs for the second set of parameters of the four genetic models of Table 5. Delta_11 $= \Delta_{11} = P(H_{11}D) - P(H_{11})P_D$ is a measure of linkage disequilibrium (LD) between marker H_1 and disease locus D. In addition, $\Delta_{12} = -\Delta_{11}$, $\Delta_{13} = -\Delta_{14} = \Delta_{11}/2$. The simulated power curves of ST_H and ST_G are calculated using combinations of both sib-pairs and sibships of size 3 and sibships of size 4; the number of sib-pairs is equal to $N/2 = 300$; the number of sibships of size 3 is $N/2 = 120$; and the number of sibships of size 4 is $3N/10 = 180$; in each of $N/10 = 60$ sibships of size 3, one is affected and the other two are normal; in the remaining $N/10 = 60$ sibships of size 3, two are affected and the other one is normal; in each of $N/10 = 60$ sibships of size 4, one is affected and the other three are normal; in each of $N/10 = 60$ sibships of size 4, two are affected and the other two are normal; in the remaining $N/10 = 60$ sibships of size 4, three are affected and the other one is normal.

uniformly divided into 20 subintervals in Figures 3 and 4. Correspondingly, the 20 subintervals lead to 21 endpoints. For each endpoint, there is a set of parameters for each power curve. Using the set of parameters, 2,500 datasets are simulated for each endpoint. For each dataset, the empirical statistics T_H and T_G were calculated. The simulated power is the proportion of the 2,500 simulated datasets for which the empirical statistic is larger than the cut-off point of the corresponding χ^2-distribution at a 0.05 significance level.

From Figures 3 and 4, it can be seen that the simulated power ST_H is generally higher than the power of T_H, and the simulated power ST_G is generally higher than the power of

T_G. Intuitively, sibships of large size contain more information than that of a sib-pair. The test statistics T_H and T_G can accurately capture the information contained in sibships of large size. Moreover, it can also be seen in Tables 1, 2 and 3 that the type I error is not inflated by including sibships of varying structure.

Simulation study

To evaluate the accuracy of the non-centrality parameter approximations, we performed simulations for the power curves in Figures 1, 2, 3 and 4. The results are presented in the section: Supplementary information: Simulation study. It can be seen that the approximations are excellent.

Discussion

The goal of this study was to develop sibship-based Hotelling's T^2 test statistics for high-resolution association mapping of complex diseases. This extends our previous research of paired Hotelling's T^2 test statistics of nuclear family data or parent–offspring pairs.[20] For late-onset complex diseases, parental data are usually not available. This motivated us to perform sib-pair or sibship analyses to study late-onset disorders. Based an two coding methods—'haplotype/allele coding' and 'genotype coding'—paired Hotelling's T^2 test statistics T_H and T_G are proposed for high-resolution association studies, using normal sibs as controls for affected sibs. The test statistics can be applied to any number of markers, which can be either bi-allelic or multi-allelic. After power calculation and comparison, it was found that it is advantageous to use two markers rather than one marker in the analysis. This observation can be generalised — that is, it is advantageous to use multiple tightly linked markers in analysis. The test statistic T_H based on the 'haplotype/allele coding' method is generally more powerful than the test statistic T_G based on the 'genotype coding' method. This is most likely due to the large number of degrees of freedom of T_G. Moreover, the type I error rates of the test statistic T_H are lower than those of test statistic T_G.

For population case-control association studies, false-positive rates can be high due to inappropriate controls, which can occur if there is population admixture or stratification.[25] Moreover, it is not always clear how to choose the appropriate controls. Alternatively, the parents or normal sibs can be used as controls of affected sibs.[22,26–29] For parental/sibling controls, the methods proposed by Fan and Knapp[19] and Xiong *et al.*[17] are not valid, since cases and controls are correlated with each other. The two sample Hotelling's T^2 test statistics only take into account the correlation among markers.[17,19] For sibship data, not only the correlation among the markers but also the correlation within each sib-pair needs to be taken into account. The paired Hotelling's T^2 test statistics T_H and T_G developed in this paper correctly take both the correlation among the markers and the correlation within each sib-pair into account. The proposed method is potentially useful in association mapping of late-onset complex diseases.

Cordell and Clayton[2] and Chapman *et al.*[18] proposed logistic regression models for population-based case control studies or family studies. Both our proposed method and the logistic regression models can be used in association studies of multi-locus marker data. One advantage of the logistic regression models is that it is easy to add covariates to model the environmental effects, in addition to the genetic effects; however, it is not clear how to incorporate the environmental effects into our Hotelling's T^2 test statistics. While we are able to calculate the non-centrality parameters for our T^2 test statistics for power and sample size calculations, it is not clear if one might get similar results for the logistic regression models. In the study by Cordell and Clayton,[2] the authors mainly discuss the analysis of SNP data and only briefly describe a way to analyse the multi-allelic markers data. We feel that more investigations are necessary in order for multi-allelic markers data to be used in the logistic regression models. By contrast, our proposed T^2 can be used to analyse either bi-allelic or multi-allelic marker data, or both simultaneously. Moreover, more investigations are needed to make power comparisons of the two methods.

In Figures 3 and 4, we show that the power of test statistics T_H and T_G based on combinations of sibships of varying structures are generally higher than the power of the test statistics based on sib-pairs. This is because the test statistics T_H and T_G use the average coding vectors for sibships whose sizes are larger than 2. This averaging strategy does not affect the mean of the coding vectors $\bar{X}^{(A)}$ and $\bar{Y}^{(U)}$, but it will lead to a variance–covariance matrix *S*, which increases the test statistics. Moreover, it can be seen from Tables 1, 2 and 3 that the type I error is not inflated by including sibships of varying structure. Although the proposed test statistics benefit from this, it is unlikely that they are optimal. One way would be to use weighted sibships in constructing test statistics. In this paper, 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.[26] As a result, the methods used here need to be updated to address the problem of missing data. Another issue is that it is not clear how to combine population data, the nuclear family data and sibship data in one single analysis. In practice, the three types of genetic data can be available. They can be analysed separately, but it would be preferable to combine them in a unified analysis, which may lead to higher power. These issues needs more in-depth investigation.

Acknowledgments

M. Knapp was supported by grant KN 370/1-1 (Project D1 of FOR 423) from the Deutsche Forschungsgemeinschaft. R. Fan was supported by the National Science Foundation Grant DMS-0505025.

Appendix

Consider a sib-pair in which one sibling is affected and the other is unaffected/normal. For convenience, assume that the first sibling is affected and the second sibling is normal. Let us denote $A_1 =$ *(the first sibling is affected)*, $U_2 =$ *(the second sibling is unaffected)*. Let f_{DD}, $f_{Dd} = f_{dD}$ and f_{dd} be the probabilities that an individual with genotypes *DD*, *Dd* and *dd* is affected with the disease, respectively. Since allele *D* is disease susceptible, one may assume that $f_{DD} \geq f_{Dd} \geq f_{dd}$. Let $\bar{f}_{DD} = 1 - f_{DD}$, $\bar{f}_{Dd} = 1 - f_{Dd}$ and $\bar{f}_{dd} = 1 - f_{dd}$. Denote the disease prevalence in population by $A = f_{DD}P_D^2 + 2f_{Dd}P_DP_d + f_{dd}P_d^2$, and $\bar{A} = \bar{f}_{DD}P_D^2 + 2\bar{f}_{Dd}P_DP_d + \bar{f}_{dd}P_d^2 = 1 - A$. Assume that the affected status of an individual depends only on his/her own genotype at the disease locus. Let us denote the event *(i IBD) = the sib-pair share i gene identical by descent (IBD) at the disease locus D*. Then the joint probability

$$\begin{aligned}P(A_1, U_2) &= P(A_1, U_2|2 \quad \text{IBD})/4 + P(A_1, U_2|1 \quad \text{IBD})/2 \\ &\quad + P(A_1, U_2|0 \quad \text{IBD})/4 \\ &= \frac{1}{4}\left[\sum_{s,t\in\{D,d\}} f_{st}\bar{f}_{st}P_sP_t + 2\sum_{s,t,q\in\{D,d\}} f_{st}\bar{f}_{tq}P_sP_tP_q \right. \\ &\quad \left. + \sum_{s,t,q,r\in\{D,d\}} f_{st}\bar{f}_{qr}P_sP_tP_qP_r\right] \\ &= \frac{1}{4}\left[\sum_{s,t\in\{D,d\}} f_{st}\bar{f}_{st}P_sP_t + 2\sum_{s,t,q\in\{D,d\}} f_{st}\bar{f}_{tq}P_sP_tP_q + A\bar{A}\right], \end{aligned} \tag{1}$$

where *s*, *t*, *q*, *r* take values of disease allele *D* and *d*. To calculate the above equations, we consider the three partitions (2 IBD), (1 IBD) and (0 IBD). These three partitions have probabilities 1/4, 1/2 and 1/4, respectively. Conditional on each partition, the corresponding conditional probabilities are then calculated. The frequency of homozygous genotype $H_{jk}H_{jk}$ in an affected sibling is given by:

$$\begin{aligned}a_{jkk} &= P[G_{ij}^{(A)} = H_{jk}H_{jk}|A_1, U_2] \\ &= P[G_{ij}^{(A)} = H_{jk}H_{jk}, A_1, U_2, (2\text{ IBD})\cup(1\text{ IBD})\cup(0\text{ IBD})]/ \\ &\quad P(A_1, U_2) \\ &= \left[\frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}\bar{f}_{st}P(H_{jk}s)P(H_{jk}t)\right. \\ &\quad + \frac{1}{2}\sum_{s,t,q\in\{D,d\}} f_{st}\bar{f}_{tq}P(H_{jk}t)P(H_{jk}s)P_q \\ &\quad \left. + \frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}P(H_{jk}s)P(H_{jk}t)\bar{A}\right]/P(A_1, U_2). \end{aligned} \tag{2}$$

Similarly, the frequency of homozygous genotype $H_{jk}H_{jk}$ in an unaffected sibling is given by:

$$\begin{aligned}\bar{a}_{jkk} &= P[G_{ij}^{(U)} = H_{jk}H_{jk}|A_1, U_2] = P[G_{ij}^{(U)} = H_{jk}H_{jk}, A_1, U_2, \\ &\qquad (2\text{ IBD})\cup(1\text{ IBD})\cup(0\text{ IBD})]/P(A_1, U_2) \\ &= \left[\frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st}f_{st}P(H_{jk}s)P(H_{jk}t)\right. \\ &\quad + \frac{1}{2}\sum_{s,t,q\in\{D,d\}} \bar{f}_{st}f_{tq}P(H_{jk}t)P(H_{jk}s)P_q \\ &\quad \left. + \frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st}P(H_{jk}s)P(H_{jk}t)A\right]/P(A_1, U_2). \end{aligned} \tag{3}$$

Note that $\bar{a}_{jkk}$ can be calculated by the formula for a_{jkk} by substituting f_{st} with $\bar{f}_{st}$ and vice versa. Note that the haplotype frequencies $P(H_{jk}D) = \Delta_{jk} + P(H_{jk})P_D$, $P(H_{jk}d) = -\Delta_{jk} + P(H_{jk})P_d$. Under the null hypothesis of no association between the markers H_i, $i = 1, 2, \ldots, J$, and the disease locus *D* — that is, $\Delta_{ij} = 0$ for all *j*, the haplotype frequencies are equal to the product of allele frequencies; for example, $P(H_{jk}D) = P(H_{jk})P_D$ and $P(H_{jk}d) = P(H_{jk})P_d$. From equations (4) and (5), $a_{jkk} = \bar{a}_{jkk} = P(H_{jk})^2$.

Similarly, the frequency of the heterozygous genotype $H_{jk}H_{jl}$, $k \neq l$, in an affected sibling can be calculated as follows:

$$\begin{aligned}a_{jkl} &= P[G_{ij}^{(A)} = H_{jk}H_{jl}|A_1, U_2] = P[G_{ij}^{(A)} = H_{jk}H_{jl}, A_1, U_2, \\ &\qquad (2\text{ IBD})\cup(1\text{ IBD})\cup(0\text{ IBD})]/P(A_1, U_2) \\ &= \left[\frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}\bar{f}_{st}(P(H_{jk}s)P(H_{jl}t) + P(H_{jk}t)P(H_{jl}s))\right. \\ &\quad + \frac{1}{2}\sum_{s,t,q\in\{D,d\}} f_{st}\bar{f}_{tq}(P(H_{jk}t)P(H_{jl}s) + P(H_{jk}s)P(H_{jl}t))P_q \\ &\quad \left. + \frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}(P(H_{jk}s)P(H_{jl}t) + P(H_{jk}t)P(H_{jl}s))\bar{A}\right]/P(A_1, U_2). \end{aligned} \tag{4}$$

The frequency of the heterozygous genotype $H_{jk}H_{jl}$, $k \neq l$, in an unaffected sibling can be calculated as follows:

$$\begin{aligned}\bar{a}_{jkl} &= P[G_{ij}^{(U)} = H_{jk}H_{jl}|A_1, U_2] = P[G_{ij}^{(U)} = H_{jk}H_{jl}, A_1, U_2, \\ &\qquad (2\text{ IBD})\cup(1\text{ IBD})\cup(0\text{ IBD})]/P(A_1, U_2) \\ &= \left[\frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st}f_{st}(P(H_{jk}s)P(H_{jl}t) + P(H_{jk}t)P(H_{jl}s))\right. \\ &\quad + \frac{1}{2}\sum_{s,t,q\in\{D,d\}} \bar{f}_{st}f_{tq}(P(H_{jk}t)P(H_{jl}s) + P(H_{jk}s)P(H_{jl}t))P_q \\ &\quad \left. + \frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st}(P(H_{jk}s)P(H_{jl}t) + P(H_{jk}t)P(H_{jl}s))A\right]/P(A_1, U_2). \end{aligned} \tag{5}$$

Note that $\bar{a}_{jkl}$ can be calculated by the formula for a_{jkl} by substituting f_{st} using $\bar{f}_{st}$ and vice versa. Under the null hypothesis of no association between the markers H_i, $i = 1, 2, \ldots, J$, and the disease locus D — that is, $\Delta_{ij} = 0$ for all j, the haplotype frequencies are equal to the product of the allele frequencies; for example, $P(H_{jk}D) = P(H_{jk})P_D$, $P(H_{jk}d) = P(H_{jk})P_d$, $P(H_{jl}D) = P(H_{jl})P_D$ and $P(H_{jl}d) = P(H_{jl})P_d$. From equations (4) and (5), $a_{jkl} = \bar{a}_{jkl} = 2P(H_{jk})P(H_{jl})$. Therefore, the expectation $E(\bar{X}^{(A)} - \bar{Y}^{(U)}|A_1, U_2) = 0$ for the 'genotype coding' method.

For the 'haplotype/allele coding' method, equations (2), (3), (4) and (5) imply

$$\begin{aligned} \mathrm{E}(z_{ijk}^{(A)}|A_1, U_2) = 2a_{jkk} + \sum_{l\neq k} a_{jkl}, \mathrm{E}(z_{ijk}^{(U)}|A_1, U_2) \\ = 2\bar{a}_{jkk} + \sum_{l\neq k} \bar{a}_{jkl}. \end{aligned} \tag{6}$$

From equation (6), expectation $\mathrm{E}(z_{ijk}^{(A)} - z_{ijk}^{(U)}|A_1, U_2) = 2P(H_{jk}) - 2P(H_{jk}) = 0$ by 'haplotype/allele coding' method, under the null hypothesis of no association between the markers H_i, $j = 1, \ldots, J$ and disease locus D.

References

1. Botstein, D. and Risch, N. (2003), 'Discovering genotypes underlying human phenotypes: Past successes for Mendelian disease, future approaches for complex disease', *Nat. Genet.* Vol. 33(Suppl.), pp. 228–237.
2. Cordell, H.J. and Clayton, D.G. (2002), 'A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to HLA in type 1 diabetes', *Am. J. Hum. Genet.* Vol. 70, pp. 124–141.
3. Rannala, B. and Reeve, J.P. (2001), 'High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence', *Am. J. Hum. Genet.* Vol. 69, pp. 159–178, p. 672.
4. Risch, N. (2001), 'Implications of multilocus inheritance for gene-disease association studies', *Theor. Popul. Biol.* Vol. 60, pp. 215–220.
5. Risch, N. and Merikangas, K. (1996), 'The future of genetic studies of complex human diseases', *Science* Vol. 273, pp. 1516–1517.
6. Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993), 'Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM)', *Am. J. Hum. Genet.* Vol. 52, pp. 506–516.
7. Chapman, N.H. and Wijsman, E.M. (1998), 'Genome screens using linkage disequilibrium tests: Optimal marker characteristics and feasibility', *Am. J. Hum. Genet.* Vol. 63, pp. 1872–1885.
8. Olson, J.M. and Wijsman, E.M. (1994), 'Design and sample size considerations in the detection of linkage disequilibrium with a disease locus', *Am. J. Hum. Genet.* Vol. 55, pp. 574–580.
9. Kaplan, N. and 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', *Theor. Popul. Biol.* Vol. 60, pp. 193–201.
10. Kaplan, N. and Morris, R. (2001), 'Issues concerning association studies for fine mapping a susceptibility gene for a complex disease', *Genet. Epidemiol.* Vol. 20, pp. 432–457.
11. Nielsen, D.M., Ehm, M.G. and Weir, B.S. (1998), 'Detecting marker-disease association by testing for Hardy–Weinberg disequilibrium at a marker locus', *Am. J. Hum. Genet.* Vol. 63, pp. 1531–1540.
12. Ott, J. (1999), Analysis of human genetic linkage, 3rd edition, Johns Hopkins University Press, Baltimore and London.
13. The International HapMap Consortium (2003), 'The International HapMap Project', *Nature* Vol. 426, pp. 789–796.
14. The International SNP Map Working Group (2001), 'A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms', *Nature* Vol. 409, pp. 928–933.
15. Kong, A., Gudbjartsson, D.F., Sainz, J. *et al.* (2002), 'A high resolution recombination map of the human genome', *Nat. Genet.* Vol. 31, pp. 241–247.
16. Hotelling, H. (1931), 'The generalization of Student's ratio', *Ann. Math. Stat.* Vol. 2, pp. 360–378.
17. Xiong, M.M., Zhao, J. and Boerwinkle, E. (2002), 'Generalized T^2 test for genome association studies', *Am. J. Hum. Genet.* Vol. 70, pp. 1257–1268.
18. Chapman, J.M., Cooper, J.D., Todd, J. and Clayton, D. (2003), 'Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of the statistical power', *Hum. Hered.* Vol. 56, pp. 18–31.
19. Fan, R.Z. and Knapp, M. (2003), 'Genome association studies of complex diseases by case-control designs', *Am. J. Hum. Genet.* Vol. 72, pp. 850–868.
20. Fan, R.Z., Knapp, M., Wjst, M. *et al.* (2005), 'High resolution T^2 association tests of complex diseases based on family data', *Ann. Hum. Genet.* Vol. 69, pp. 187–208.
21. Loukola, A., Chadha, M., Penn, S.G. *et al.* (2004), 'Comprehensive evaluation of the association between prostate cancer and genotype/haplotypes in CYP17A1, CYP3A4, and SRD5A2', *Eur. J. Hum. Genet.* Vol. 12, pp. 321–332.
22. Spielman, R.S. and Ewens, W.J. (1998), 'A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test', *Am. J. Hum. Genet.* Vol. 62, pp. 450–458.
23. Anderson, T.W. (1984), An introduction to multivariate statistical analysis, 2nd edition Wiley, New York.
24. Thomson, G. and Baur, M.P. (1984), 'Third order linkage disequilibrium', *Tissue Antigens* Vol. 24, pp. 250–255.
25. Ewens, W.J. and Spielman, R.S. (1995), 'The transmission/disequilibrium test: History, subdivision, and admixture', *Am. J. Hum. Genet.* Vol. 57, pp. 455–464.
26. Allen, A.S., Rathouz, P.J. and Satten, G.A. (2003), 'Informative missingness in genetic association studies: Case-parent designs', *Am. J. Hum. Genet.* Vol. 72, pp. 671–680.
27. Curtis, D. (1997), 'Use of siblings as controls in case-control association studies', *Ann. Hum. Genet.* Vol. 61, pp. 319–333.
28. Falk, C.T. and Rubinstein, P. (1987), 'Haplotype relative risk: An easy reliable way to construct a proper control sample for risk calculations', *Ann. Hum. Genet.* Vol. 51, pp. 227–233.
29. Zhao, H.Y., Zhang, S.L., Merikangas, K.R. *et al.* (2000), 'Transmission/disequilibrium tests using multiple tightly linked markers', *Am. J. Hum. Genet.* Vol. 67, pp. 936–946.

Supplementary information: Non-centrality parameters

Consider N sib-pairs, each consisting of an affected sibling and a normal sibling. For convenience, assume that the first sibling is affected and the second sibling is normal in each sib-pair. Let us denote $A_1 =$ *(the first sibling is affected)*, $U_2 =$ *(the second sibling is unaffected)*. For 'haplotype/allele coding', the coding vector of the affected sibling in the i-th sib-pair is $X_i^{(A)} = (z_{i11}^{(A)}, \ldots, z_{i1(n_1-1)}^{(A)}, \ldots, z_{iJ1}^{(A)}, \ldots, z_{iJ(n_J-1)}^{(A)})^\tau$. Similarly, $Y_i^{(U)} = (z_{i11}^{(U)}, \ldots, z_{i1(n_1-1)}^{(U)}, \ldots, z_{iJ1}^{(U)}, \ldots, z_{iJ(n_J-1)}^{(U)})^\tau$ is the coding vector of the normal sibling. Denote the variance–covariance matrix of $X_i^{(A)} - Y_i^{(U)}$ by $\Sigma_{hap} = \mathrm{Var}(X_i^{(A)} - Y_i^{(U)}|A_1, U_2) = \mathrm{Var}(X_i^{(A)}|A_1, U_2) - \mathrm{Cov}(X_i^{(A)}, Y_i^{(U)}|A_1, U_2) - \mathrm{Cov}(Y_i^{(U)}, X_i^{(A)}|A_1, U_2) + \mathrm{Var}(Y_i^{(U)}|A_1, U_2)$. The elements of the above variance–covariance matrices are given in Appendices A, B, and C: $\mathrm{Var}(X_i^{(A)}|A_1, U_2)$ and

$\mathrm{Var}(Y_i^{(U)}|A_1, U_2)$ in Appendix A, and $\mathrm{Cov}(X_i^{(A)}, Y_i^{(U)}|A_1, U_2)$ in Appendices B and C. Using quantities of $\mathrm{E}(z_{ijk}^{(A)}|A_1, U_2)$ and $\mathrm{E}(z_{ijk}^{(U)}|A_1, U_2)$ in the Appendix to the manuscript, $\mathrm{E}(X_i^{(A)} - Y_i^{(U)}|A_1, U_2)$ can be calculated. The non-centrality parameter λ_H of Hotelling's statistics T_H is given by $\lambda_H = N\mathrm{E}(X_i^{(A)} - Y_i^{(U)}|A_1, U_2)^\tau[\Sigma_{hap}]^{-1}\mathrm{E}(X_i^{(A)} - Y_i^{(U)}| A_1, U_2)$.

For the 'genotype coding' method, the coding vector of the affected sibling in the *i*-th sib-pair is $X_{ij}^{(A)} = (x_{ij1}^{(A)}, \ldots, x_{ij(n_j-1)}^{(A)}, x_{ij12}^{(A)}, \ldots, x_{ij1n_j}^{(A)}, \ldots, x_{ij(n_j-1)n_j}^{(A)})^\tau$ $j = 1,\ldots,J$. Similarly, $Y_{ij}^{(U)} = (x_{ij1}^{(U)}, \ldots, x_{ij(n_j-1)}^{(U)}, x_{ij12}^{(U)}, \ldots, x_{ij1n_j}^{(U)}, \ldots, x_{ij(n_j-1)n_j}^{(U)})^\tau$ is the coding vector of the normal sibling. Let a_{jkl} and $\bar{a}_{jkl}$ be the frequencies of genotype $H_{jk}H_{jl}$ in affected and unaffected siblings given in the Appendix to the manuscript. Then,

$$\mathrm{E}[X_{ij}^{(A)}|A_1, U_2] = (a_{j11}, \ldots, a_{j(n_j-1)(n_j-1)}, a_{j12}, \ldots, a_{j1n_j}, \ldots, a_{j(n_j-1)n_j})^\tau, \tag{1}$$

$$\mathrm{E}[Y_{ij}^{(U)}|A_1, U_2] = (\bar{a}_{j11}, \ldots, \bar{a}_{j(n_j-1)(n_j-1)}, \bar{a}_{j12}, \ldots, \bar{a}_{j1n_j}, \ldots, \bar{a}_{j(n_j-1)n_j})^\tau. \tag{2}$$

Using $\mathrm{E}[X_{ij}^{(A)}|A_1, U_2]$ and $\mathrm{E}[Y_{ij}^{(U)}|A_1, U_2]$, one may calculate the expectation $\mathrm{E}(\bar{X}^{(A)} - \bar{Y}^{(U)}|A_1, U_2) = (\mathrm{E}[X_{i1}^{(A)} - Y_{i1}^{(U)}|A_1, U_2]^\tau, \ldots, \mathrm{E}[X_{iJ}^{(A)} - Y_{iJ}^{(U)}|A_1, U_2]^\tau)^\tau$. Let $\Sigma_{geno} = \mathrm{Cov}(X_i^{(A)} - Y_i^{(U)}|A_1, U_2) = \mathrm{Var}(X_i^{(A)}|A_1, U_2) - \mathrm{Cov}(X_i^{(A)}, Y_i^{(U)}|A_1, U_2) - \mathrm{Cov}(Y_i^{(U)}, X_i^{(A)}|A_1, U_2) + \mathrm{Var}(Y_i^{(U)}|A_1, U_2)$ be the variance–covariance matrix of $X_i^{(A)} - Y_i^{(U)}$. Then the non-centrality parameter λ_G of Hotelling's statistics T_G is given by $\lambda_G = N\mathrm{E}[\bar{X}^{(A)} - \bar{Y}^{(U)}|A_1, U_2]^\tau[\Sigma_{geno}]^{-1}\mathrm{E}[\bar{X}^{(A)} - \bar{Y}^{(U)}|A_1, U_2]$. The elements of the above variance–covariance matrices are given in Appendices D and E: $\mathrm{Var}(X_i^{(A)}|A_1, U_2)$ and $\mathrm{Var}(Y_i^{(U)}|A_1, U_2)$ in Appendix D, and $\mathrm{Cov}(X_i^{(A)}, Y_i^{(U)}|A_1, U_2)$ in Appendix E.

Appendix A

Consider the 'haplotype/allele coding' method. The variance–covariance matrices are

$$\begin{aligned}\mathrm{Var}(X_i^{(A)}|A_1, U_2) &= \mathrm{Var}[(z_{i11}^{(A)}, \ldots, z_{i1(n_1-1)}^{(A)}, \ldots, z_{iJ1}^{(A)}, \ldots, z_{iJ(n_J-1)}^{(A)})^\tau|A_1, U_2],\\ \mathrm{Var}[Y_i^{(U)}|A_1, U_2] &= \mathrm{Var}[(z_{i11}^{(U)}, \ldots, z_{i1(n_1-1)}^{(U)}, \ldots, z_{iJ1}^{(U)}, \ldots, z_{iJ(n_J-1)}^{(U)})^\tau|A_1, U_2].\end{aligned}$$

The variance of the number of the alleles H_{jk} in the affected sibling and unaffected sibling can be calculated as

$$\begin{aligned}\mathrm{Var}(z_{ijk}^{(A)}|A_1, U_2) &= \mathrm{E}[(z_{ijk}^{(A)})^2|A_1, U_2] - [\mathrm{E}(z_{ijk}^{(A)}|A_1, U_2)]^2\\ &= 4a_{jkk} + \sum_{l\neq k} a_{jkl} - \left[2a_{jkk} + \sum_{l\neq k} a_{jkl}\right]^2,\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(z_{ijk}^{(U)}|A_1, U_2) &= \mathrm{E}[(z_{ijk}^{(U)})^2|A_1, U_2] - [\mathrm{E}(z_{ijk}^{(U)}|A_1, U_2)]^2\\ &= 4\bar{a}_{jkk} + \sum_{l\neq k} \bar{a}_{jkl} - \left[2\bar{a}_{jkk} + \sum_{l\neq k} \bar{a}_{jkl}\right]^2.\end{aligned}$$

Similarly, the covariance between the number of alleles H_{jk} and the number of alleles H_{jl}, $l \neq k$, in the affected sibling and the unaffected sibling can be calculated as

$$\begin{aligned}\mathrm{Cov}(z_{ijk}^{(A)}, z_{ijl}^{(A)}|A_1, U_2) &= \mathrm{E}(z_{ijk}^{(A)} z_{ijl}^{(A)}|A_1, U_2)\\ &\quad - \mathrm{E}(z_{ijk}^{(A)}|A_1, U_2)\mathrm{E}(z_{ijl}^{(A)}|A_1, U_2)\\ &= P(G_{ij}^{(A)} = H_{jk}H_{jl}|A_1, U_2)\\ &\quad - \left[2a_{jkk} + \sum_{k'\neq k} a_{jkk'}\right]\left[2a_{jll} + \sum_{l'\neq l} a_{jll'}\right]\\ &= a_{jkl} - \left[2a_{jkk} + \sum_{k'\neq k} a_{jkk'}\right]\left[2a_{jll} + \sum_{l'\neq l} a_{jll'}\right],\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(z_{ijk}^{(U)}, z_{ijl}^{(U)}|A_1, U_2) &= \mathrm{E}(z_{ijk}^{(U)} z_{ijl}^{(U)}|A_1, U_2)\\ &\quad - \mathrm{E}(z_{ijk}^{(U)}|A_1, U_2)\mathrm{E}(z_{ijl}^{(U)}|A_1, U_2)\\ &= \bar{a}_{jkl} - \left[2\bar{a}_{jkk} + \sum_{k'\neq k} \bar{a}_{jkk'}\right]\left[2\bar{a}_{jll} + \sum_{l'\neq l} \bar{a}_{jll'}\right].\end{aligned}$$

For $j \neq g$, assume that markers H_j and H_g flank disease locus D in the order of H_jDH_g. Let $P(H_{jk}DH_{gh})$ be frequencies of haplotype $H_{jk}DH_{gh}$. The frequencies of other haplotypes are denoted accordingly. For the *i*-th sib-pair, let $G_{iD}^{(U)}$ be the disease genotype of the unaffected sibling and $G_{iD}^{(A)}$ be the disease genotype of the affected sibling. To calculate the covariance between $z_{ijk}^{(A)}, z_{igh}^{(A)}$, denote for $j \neq g$, $k \neq k'$, $h \neq h'$,

$$\begin{aligned}g_{kkhh}^{(A,jg)} &= \mathrm{E}[1_{(G_{ij}^{(A)}=H_{jk}H_{jk})}1_{(G_{ig}^{(A)}=H_{gh}H_{gh})}|A_1, U_2]\\ &= P[G_{ij}^{(A)} = H_{jk}H_{jk}, G_{ig}^{(A)} = H_{gh}H_{gh}, A_1, U_2,\\ &\qquad (2\,\mathrm{IBD}) \cup (1\,\mathrm{IBD}) \cup (0\,\mathrm{IBD})]/P(A_1, U_2)\\ &= \Bigg[\frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}\bar{f}_{st} P[G_{ij}^{(A)} = H_{jk}H_{jk},\\ &\qquad G_{ig}^{(A)} = H_{gh}H_{gh}, G_{iD}^{(A)} = st, G_{iD}^{(U)} = st]\\ &\quad + \frac{1}{2}\sum_{s,t,q\in\{D,d\}} f_{st}\bar{f}_{tq} P[G_{ij}^{(A)} = H_{jk}H_{jk},\\ &\qquad G_{ig}^{(A)} = H_{gh}H_{gh}, G_{iD}^{(A)} = st, G_{iD}^{(U)} = tq]\\ &\quad + \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} f_{st}\bar{f}_{qr} P[G_{ij}^{(A)} = H_{jk}H_{jk},\\ &\qquad G_{ig}^{(A)} = H_{gh}H_{gh}, G_{iD}^{(A)} = st, G_{iD}^{(U)} = qr]\Bigg]/P(A_1, U_2)\end{aligned}$$

$$
\begin{aligned}
&= \Bigg[\frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}\bar{f}_{st}P(H_{jk}sH_{gh})P(H_{jk}tH_{gh})\\
&\quad +\frac{1}{2}\sum_{s,t,q\in\{D,d\}} f_{st}\bar{f}_{tq}P(H_{jk}tH_{gh})P(H_{jk}sH_{gh})P_q\\
&\quad +\frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}P(H_{jk}sH_{gh})P(H_{jk}tH_{gh})\bar{A}\Bigg]/P(A_1,U_2)
\end{aligned}
$$

$$
\begin{aligned}
g^{(A,jg)}_{kkhh'} &= \mathrm{E}[1_{(G^{(A)}_{ij}=H_{jk}H_{jk})}1_{(G^{(A)}_{ig}=H_{gh}H_{gh'})}|A_1,U_2]\\
&= P[G^{(A)}_{ij}=H_{jk}H_{jk},\, G^{(A)}_{ig}=H_{gh}H_{gh'},A_1,U_2,\\
&\qquad (2\,\mathrm{IBD})\cup(1\,\mathrm{IBD})\cup(0\,\mathrm{IBD})]/P(A_1,U_2)\\
&= \Bigg[\frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}\bar{f}_{st}P[G^{(A)}_{ij}=H_{jk}H_{jk},\, G^{(A)}_{ig}=H_{gh}H_{gh'},\\
&\qquad\qquad G^{(A)}_{iD}=st,\, G^{(U)}_{iD}=st]\\
&\quad +\frac{1}{2}\sum_{s,t,q\in\{D,d\}} f_{st}\bar{f}_{tq}P[G^{(A)}_{ij}=H_{jk}H_{jk},\, G^{(A)}_{ig}=H_{gh}H_{gh'},\\
&\qquad\qquad G^{(A)}_{iD}=st,\, G^{(U)}_{iD}=tq]\\
&\quad +\frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} f_{st}\bar{f}_{qr}P[G^{(A)}_{ij}=H_{jk}H_{jk},\, G^{(A)}_{ig}=H_{gh}H_{gh'},\\
&\qquad\qquad G^{(A)}_{iD}=st,\, G^{(U)}_{iD}=qr\Bigg]/P(A_1,U_2)\\
&= \Bigg[\frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}\bar{f}_{st}(P(H_{jk}sH_{gh})P(H_{jk}tH_{gh'})\\
&\quad + P(H_{jk}tH_{gh})P(H_{jk}sH_{gh'}))\\
&\quad +\frac{1}{2}\sum_{s,t,q\in\{D,d\}} f_{st}\bar{f}_{tq}(P(H_{jk}sH_{gh})P(H_{jk}tH_{gh'}).\\
&\qquad +P(H_{jk}tH_{gh})P(H_{jk}sH_{gh'}))P_q\\
&\quad +\frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}(P(H_{jk}sH_{gh})P(H_{jk}tH_{gh'})\\
&\qquad +P(H_{jk}tH_{gh})P(H_{jk}sH_{gh'}))\bar{A}\Bigg]/P(A_1,U_2)
\end{aligned}
$$

$$
\begin{aligned}
g^{(A,jg)}_{kk'hh} &= \mathrm{E}[1_{(G^{(A)}_{ij}=H_{jk}H_{jk'})}1_{(G^{(A)}_{ig}=H_{gh}H_{gh})}|A_1,U_2]\\
&= P[G^{(A)}_{ij}=H_{jk}H_{jk'},G^{(A)}_{ig}=H_{gh}H_{gh},A_1,U_2,\\
&\qquad (2\mathrm{IBD})\cup(1\mathrm{IBD})\cup(0\mathrm{IBD})]/P(A_1,U_2)\\
&= \Bigg[\frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}\bar{f}_{st}P[G^{(A)}_{ij}=H_{jk}H_{jk'},G^{(A)}_{ig}=H_{gh}H_{gh},\\
&\qquad\qquad G^{(A)}_{iD}=st,G^{(U)}_{iD}=st]\\
&\quad +\frac{1}{2}\sum_{s,t,q\in\{D,d\}} f_{st}\bar{f}_{tq}P[G^{(A)}_{ij}=H_{jk}H_{jk'},G^{(A)}_{ig}=H_{gh}H_{gh},\\
&\qquad\qquad G^{(A)}_{iD}=st,G^{(U)}_{iD}=tq]\\
&\quad +\frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} f_{st}\bar{f}_{qr}P[G^{(A)}_{ij}=H_{jk}H_{jk'},G^{(A)}_{ig}=H_{gh}H_{gh},\\
&\qquad\qquad G^{(A)}_{iD}=st,G^{(U)}_{iD}=qr\Bigg]/P(A_1,U_2)\\
&= \Bigg[\frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}\bar{f}_{st}(P(H_{jk}sH_{gh})P(H_{jk'}tH_{gh})\\
&\quad +P(H_{jk}tH_{gh})P(H_{jk'}sH_{gh}))\\
&\quad +\frac{1}{2}\sum_{s,t,q\in\{D,d\}} f_{st}\bar{f}_{tq}(P(H_{jk}sH_{gh})P(H_{jk'}tH_{gh})\\
&\quad +P(H_{jk}tH_{gh})P(H_{jk'}sH_{gh}))P_q\\
&\quad +\frac{1}{4}\sum_{s,t\in\{Dd\}} f_{st}(P(H_{jk}sH_{gh})P(H_{jk'}tH_{gh})\\
&\quad +P(H_{jk}tH_{gh})P(H_{jk'}sH_{gh}))\bar{A}\Bigg]/P(A_1,U_2)
\end{aligned}
$$

$$
\begin{aligned}
g^{(A,jg)}_{kk'hh'} &= \mathrm{E}[1_{(G^{(A)}_{ij}=H_{jk}H_{jk'})}1_{(G^{(A)}_{ig}=H_{gh}H_{gh'})}|A_1,U_2]\\
&= P[G^{(A)}_{ij}=H_{jk}H_{jk'},G^{(A)}_{ig}=H_{gh}H_{gh'},A_1,U_2,\\
&\quad (2\mathrm{IBD})\cup(1\mathrm{IBD})\cup(0\mathrm{IBD})]/P(A_1,U_2)\\
&= \Bigg[\frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}\bar{f}_{st}P[G^{(A)}_{ij}=H_{jk}H_{jk'},\\
&\qquad G^{(A)}_{ig}=H_{gh}H_{gh'},G^{(A)}_{iD}=st,G^{(U)}_{iD}=st]\\
&\quad +\frac{1}{2}\sum_{s,t,q\in\{D,d\}} f_{st}\bar{f}_{tq}P[G^{(A)}_{ij}=H_{jk}H_{jk'},\\
&\quad G^{(A)}_{ig}=H_{gh}H_{gh'},G^{(A)}_{iD}=st,G^{(U)}_{iD}=tq]\\
&\quad +\frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} f_{st}\bar{f}_{qr}P[G^{(A)}_{ij}=H_{jk}H_{jk'},\\
&\qquad G^{(A)}_{ig}=H_{gh}H_{gh'},G^{(A)}_{iD}=st,G^{(U)}_{iD}=qr]\Bigg]/P(A_1,U_2)\\
&= \Bigg[\frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}\bar{f}_{st}(P(H_{jk}sH_{gh})P(H_{jk'}tH_{gh'})\\
&\quad +P(H_{jk}tH_{gh})P(H_{jk'}sH_{gh'})\\
&\quad +P(H_{jk}sH_{gh'})P(H_{jk'}tH_{gh})+P(H_{jk}tH_{gh'})P(H_{jk'}sH_{gh}))\\
&\quad +\frac{1}{2}\sum_{s,t,q\in\{D,d\}} f_{st}\bar{f}_{tq}(P(H_{jk}sH_{gh})P(H_{jk'}tH_{gh'})\\
&\quad +P(H_{jk}tH_{gh})P(H_{jk'}sH_{gh'})\\
&\quad +P(H_{jk}sH_{gh'})P(H_{jk'}tH_{gh})+P(H_{jk}tH_{gh'})P(H_{jk'}sH_{gh}))P_q\\
&\quad +\frac{1}{4}\sum_{s,t\in\{D,d\}} f_{st}(P(H_{jk}sH_{gh})P(H_{jk'}tH_{gh'})\\
&\quad +P(H_{jk}tH_{gh})P(H_{jk'}sH_{gh'})+P(H_{jk}sH_{gh'})P(H_{jk'}tH_{gh})\\
&\quad +P(H_{jk}tH_{gh'})P(H_{jk'}sH_{gh}))\bar{A}\Bigg]/P(A_1,U_2).
\end{aligned}
$$

For $k = 1,\ldots,n_j - 1$ and $h = 1,\ldots,n_g - 1$, $j \neq g$, the covariance

$$\begin{aligned}
&\mathrm{Cov}(z_{ijk}^{(A)}, z_{igh}^{(A)} | A_1, U_2) \\
&= \mathrm{E}[z_{ijk}^{(A)} z_{igh}^{(A)} | A_1, U_2] - \mathrm{E}[z_{ijk}^{(A)} | A_1, U_2]\mathrm{E}[z_{igh}^{(A)} | A_1, U_2] \\
&= 4g_{kkhh}^{(A,jg)} + 2\sum_{h' \neq h} g_{kkhh'}^{(A,jg)} + 2\sum_{k' \neq k} g_{kk'hh}^{(A,jg)} \\
&\quad + \sum_{k' \neq k}\sum_{h' \neq h} g_{kk'hh'}^{(A,jg)} - \left[2a_{jkk} + \sum_{k' \neq k} a_{jkk'}\right]\left[2a_{ghh} + \sum_{h' \neq h} a_{ghh'}\right].
\end{aligned}$$

Similarly, for $k = 1,\ldots,n_j - 1$ and $h = 1,\ldots,n_g - 1$, $j \neq g$, the covariance

$$\begin{aligned}
&\mathrm{Cov}(z_{ijk}^{(U)}, z_{igh}^{(U)} | A_1, U_2) \\
&= \mathrm{E}[z_{ijk}^{(U)} z_{igh}^{(U)} | A_1, U_2] - \mathrm{E}[z_{ijk}^{(U)} | A_1, U_2]\mathrm{E}[z_{igh}^{(U)} | A_1, U_2] \\
&= 4\bar{g}_{kkhh}^{(U,jg)} + 2\sum_{h' \neq h} \bar{g}_{kkhh'}^{(U,jg)} + 2\sum_{k' \neq k} \bar{g}_{kk'hh}^{(U,jg)} \\
&\quad + \sum_{k' \neq k}\sum_{h' \neq h} \bar{g}_{kk'hh'}^{(U,jg)} - \left[2\bar{a}_{jkk} + \sum_{k' \neq k} \bar{a}_{jkk'}\right]\left[2\bar{a}_{ghh} + \sum_{h' \neq h} \bar{a}_{ghh'}\right].
\end{aligned}$$

where $\bar{g}_{kkhh}^{(U,jg)}$, $\bar{g}_{kkhh'}^{(U,jg)}$, $\bar{g}_{kk'hh}^{(U,jg)}$ and $\bar{g}_{kk'hh'}^{(U,jg)}$ are the expected genotype frequencies in the normal sibling as follows:

$$\begin{aligned}
\bar{g}_{kkhh}^{(U,jg)} &= \mathrm{E}[1_{(G_{ij}^{(U)} = H_{jk}H_{jk})} 1_{(G_{ig}^{(U)} = H_{gh}H_{gh})} | A_1, U_2], \\
g_{kkhh'}^{(U,jg)} &= \mathrm{E}[1_{(G_{ij}^{(U)} = H_{jk}H_{jk})} 1_{(G_{ig}^{(U)} = H_{gh}H_{gh'})} | A_1, U_2], \\
g_{kk'hh}^{(U,jg)} &= \mathrm{E}[1_{(G_{ij}^{(U)} = H_{jk}H_{jk'})} 1_{(G_{ig}^{(U)} = H_{gh}H_{gh})} | A_1, U_2], \\
g_{kk'hh'}^{(U,jg)} &= \mathrm{E}[1_{(G_{ij}^{(U)} = H_{jk}H_{jk'})} 1_{(G_{ig}^{(U)} = H_{gh}H_{gh'})} | A_1, U_2].
\end{aligned}$$

To calculate $\bar{g}_{kkhh}^{(U,jg)}$, $\bar{g}_{kkhh'}^{(U,jg)}$, $\bar{g}_{kk'hh}^{(U,jg)}$ and $\bar{g}_{kk'hh'}^{(U,jg)}$, one may use the formulae of $g_{kkhh}^{(A,jg)}$, $g_{kkhh'}^{(A,jg)}$, $g_{kk'hh}^{(A,jg)}$ and $g_{kk'hh'}^{(A,jg)}$ by substituting f_{st} using $\bar{f}_{st}$.

Appendix B

The conditional covariance

$$\begin{aligned}
\mathrm{Cov}(Y_i^{(U)}, X_i^{(A)} | A_1, U_2) &= \mathrm{E}[Y_i^{(U)} X_i^{(A)^\tau} | A_1, U_2] \\
&\quad - \mathrm{E}[Y_i^{(U)} | A_1, U_2]\mathrm{E}[X_i^{(A)^\tau} | A_1, U_2] \\
&= \frac{\mathrm{E}[Y_i^{(U)} X_i^{(A)^\tau} 1_{A_1} 1_{U_2}]}{P(A_1, U_2)} \\
&\quad - \mathrm{E}[Y_i^{(U)} | A_1, U_2]\mathrm{E}[X_i^{(A)^\tau} | A_1, U_2].
\end{aligned}$$

For the 'haplotype/allele coding' method, the expectations $\mathrm{E}[Y_i^{(U)} | A_1, U_2]$ and $\mathrm{E}[X_i^{(A)^\tau} | A_1, U_2]$ are given by two quantities $\mathrm{E}(z_{ijk}^{(A)} | A_1, U_2)$ and $\mathrm{E}(z_{ijk}^{(U)} | A_1, U_2)$ (see Appendix to the paper). To get $\mathrm{E}[Y_i^{(U)} X_i^{(A)^\tau} 1_{A_1} 1_{U_2}]$, we will calculate $\mathrm{E}[z_{ijk}^{(U)} z_{ijk}^{(A)} 1_{A_1} 1_{U_2}]$ and $\mathrm{E}[z_{ijk}^{(U)} z_{ijl}^{(A)} 1_{A_1} 1_{U_2}]$, $l \neq k$ in this Appendix. In Appendix C, we will calculate the expectation $\mathrm{E}[z_{ijk}^{(U)} z_{igh}^{(A)} 1_{A_1} 1_{U_2}]$ for $j \neq g$. Note that:

$$\begin{aligned}
&\mathrm{E}[z_{ijk}^{(U)} z_{ijk}^{(A)} 1_{A_1} 1_{U_2}] \\
&\quad = \mathrm{E}\Bigg[\left(2 \cdot 1_{(G_{ij}^{(U)} = H_{jk}H_{jk})} + \sum_{l \neq k} 1_{(G_{ij}^{(U)} = H_{jk}H_{jl})}\right) \\
&\quad \left(2 \cdot 1_{(G_{ij}^{(A)} = H_{jk}H_{jk})} + \sum_{l \neq k} 1_{(G_{ij}^{(A)} = H_{jk}H_{jl})}\right) 1_{A_1} 1_{U_2}\Bigg].
\end{aligned} \tag{3}$$

Since the siblings can share 2, 1 and 0 genes identical by descent (IBD) at the disease locus D with probabilities 1/4, 1/2 and 1/4, respectively, the expectation

$$\begin{aligned}
&\mathrm{E}[1_{(G_{ij}^{(U)} = H_{jk}H_{jk})} 1_{(G_{ij}^{(A)} = H_{jk}H_{jk})} 1_{A_1} 1_{U_2}] \\
&\quad = P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} = H_{jk}H_{jk}, A_1, U_2, \\
&\qquad (2\,\mathrm{IBD}) \cup (1\,\mathrm{IBD}) \cup (0\,\mathrm{IBD})] \\
&\quad = \frac{1}{4} \sum_{s,t \in \{D,d\}} \bar{f}_{st} f_{st} P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} = H_{jk}H_{jk}, \\
&\qquad\qquad G_{iD}^{(U)} = st, G_{iD}^{(A)} = st] \\
&\qquad + \frac{1}{2} \sum_{s,t,q \in \{D,d\}} \bar{f}_{st} f_{tq} P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} = H_{jk}H_{jk}, \\
&\qquad\qquad G_{iD}^{(U)} = st, G_{iD}^{(A)} = tq] \\
&\qquad + \frac{1}{4} \sum_{s,t,q,r \in \{D,d\}} \bar{f}_{st} f_{qr} P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} = H_{jk}H_{jk}, \\
&\qquad\qquad G_{iD}^{(U)} = st, G_{iD}^{(A)} = qr] \\
&\quad = \frac{1}{4} \sum_{s,t \in \{D,d\}} \bar{f}_{st} f_{st} P(H_{jk}s) P(H_{jk}t) \\
&\qquad + \frac{1}{2} \sum_{s,t,q \in \{D,d\}} \bar{f}_{st} f_{tq} P(H_{jk}t) P(H_{jk}s) P(H_{jk}q) \\
&\qquad + \frac{1}{4} \sum_{s,t,q,r \in \{D,d\}} \bar{f}_{st} f_{qr} P(H_{jk}s) P(H_{jk}t) P(H_{jk}q) P(H_{jk}r).
\end{aligned} \tag{4}$$

For $l \neq k$, one may calculate the expectation

$$\begin{aligned}
&\mathrm{E}[1_{(G_{ij}^{(U)} = H_{jk}H_{jk})} 1_{(G_{ij}^{(A)} = H_{jk}H_{jl})} 1_{A_1} 1_{U_2}] \\
&\quad = P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} \\
&\quad = H_{jk}H_{jl}, A_1, U_2, (2\,\mathrm{IBD}) \cup (1\,\mathrm{IBD}) \cup (0\,\mathrm{IBD})] \\
&\quad = \frac{1}{4} \sum_{s,t \in \{D,d\}} \bar{f}_{st} f_{st} P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} = H_{jk}H_{jl}, G_{iD}^{(U)} \\
&\qquad = st, G_{iD}^{(A)} = st] + \frac{1}{2} \sum_{s,t,q \in \{D,d\}} \bar{f}_{st} f_{tq} P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} \\
&\qquad = H_{jk}H_{jl}, G_{iD}^{(U)} = st, G_{iD}^{(A)} = tq]
\end{aligned}$$

$$+\frac{1}{4}\sum_{s,t,q,r\in\{D,d\}}\bar{f}_{st}f_{qr}P[G_{ij}^{(U)}=H_{jk}H_{jk},\, G_{ij}^{(A)}$$

$$=H_{jk}H_{jl},\, G_{iD}^{(U)}=st,\, G_{iD}^{(A)}=qr]$$

$$=\frac{1}{2}\sum_{s,t,q\in\{D,d\}}\bar{f}_{st}f_{tq}\cdot 2P(H_{jk}t)P(H_{jk}s)P(H_{jl}q)$$

$$+\frac{1}{4}\sum_{s,t,q,r\in\{D,d\}}\bar{f}_{st}f_{qr}P(H_{jk}s)P(H_{jk}t)\Big[P(H_{jk}q)P(H_{jl}r)$$

$$+P(H_{jk}r)P(H_{jl}q)\Big]. \tag{5}$$

Similarly, one has the following expectation

$$\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jl})}1_{(G_{ij}^{(A)}=H_{jk}H_{jk})}1_{A_1}1_{U_2}]$$

$$=\frac{1}{2}\sum_{s,t,q\in\{D,d\}}\bar{f}_{st}f_{tq}\cdot 2P(H_{jl}s)P(H_{jk}t)P(H_{jk}q)$$

$$+\frac{1}{4}\sum_{s,t,q,r\in\{D,d\}}\bar{f}_{st}f_{qr}\Big[P(H_{jk}s)P(H_{jl}t)$$

$$+P(H_{jk}t)P(H_{jl}s)\Big]P(H_{jk}q)P(H_{jk}r). \tag{6}$$

For $l \neq k$, one may calculate the expectation

$$\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jk})}1_{(G_{ij}^{(A)}=H_{jk}H_{jl})}1_{A_1}1_{U_2}]$$

$$=P[G_{ij}^{(U)}=H_{jk}H_{jl},\, G_{ij}^{(A)}$$

$$=H_{jk}H_{jl}, A_1, U_2, (2\,\mathrm{IBD})\cup(1\,\mathrm{IBD})\cup(0\,\mathrm{IBD})]$$

$$=\frac{1}{4}\sum_{s,t\in\{D,d\}}\bar{f}_{st}f_{st}P[G_{ij}^{(U)}=H_{jk}H_{jl},\, G_{ij}^{(A)}=H_{jk}H_{jl},\, G_{iD}^{(U)}$$

$$=st,\, G_{iD}^{(A)}=st]+\frac{1}{2}\sum_{s,t,q\in\{D,d\}}\bar{f}_{st}f_{tq}P[G_{ij}^{(U)}=H_{jk}H_{jl},\, G_{ij}^{(A)}$$

$$=H_{jk}H_{jl},\, G_{iD}^{(U)}=st,\, G_{iD}^{(A)}=tq]+\frac{1}{4}\sum_{s,t,q,r\in\{D,d\}}\bar{f}_{st}f_{qr}P[G_{ij}^{(U)}$$

$$=H_{jk}H_{jl},\, G_{ij}^{(A)}=H_{jk}H_{jl},\, G_{iD}^{(U)}=st,\, G_{iD}^{(A)}=qr]$$

$$=\frac{1}{4}\sum_{s,t\in\{D,d\}}\bar{f}_{st}f_{st}[P(H_{jk}s)P(H_{jl}t)+P(H_{jk}t)P(H_{jl}s)]$$

$$+\frac{1}{2}\sum_{s,t,q\in\{D,d\}}\bar{f}_{st}f_{tq}[P(H_{jk}t)[P(H_{jl}s)P(H_{jl}q)]+P(H_{jl}t)$$

$$\times[P(H_{jk}s)P(H_{jk}q)]]+\frac{1}{4}\sum_{s,t,q,r\in\{D,d\}}\bar{f}_{st}f_{qr}[P(H_{jk}s)P(H_{jl}t)$$

$$+P(H_{jk}t)P(H_{jl}s)][P(H_{jk}q)P(H_{jl}r)$$

$$+P(H_{jk}r)P(H_{jl}q)]. \tag{7}$$

For $l_1 \neq l_2$, $l_1 \neq k$ and $l_2 \neq k$, one may calculate the expectation

$$\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jl_1})}1_{(G_{ij}^{(A)}=H_{jk}H_{jl_2})}1_{A_1}1_{U_2}]$$

$$=P[G_{ij}^{(U)}=H_{jk}H_{jl_1},\, G_{ij}^{(A)}$$

$$=H_{jk}H_{jl_2}, A_1, U_2, (2\,\mathrm{IBD})\cup(1\,\mathrm{IBD})\cup(0\,\mathrm{IBD})]$$

$$=\frac{1}{4}\sum_{s,t\in\{D,d\}}\bar{f}_{st}f_{st}P[G_{ij}^{(U)}=H_{jk}H_{jl_1},\, G_{ij}^{(A)}=H_{jk}H_{jl_2},$$

$$G_{iD}^{(U)}=st,\; G_{iD}^{(A)}=st]$$

$$+\frac{1}{2}\sum_{s,t,q\in\{D,d\}}\bar{f}_{st}f_{tq}P[G_{ij}^{(U)}=H_{jk}H_{jl_1},\, G_{ij}^{(A)}=H_{jk}H_{jl_2},$$

$$G_{iD}^{(U)}=st,\; G_{iD}^{(A)}=tq]$$

$$+\frac{1}{4}\sum_{s,t,q,r\in\{D,d\}}\bar{f}_{st}f_{qr}P[G_{ij}^{(U)}=H_{jk}H_{jl_1},\, G_{ij}^{(A)}=H_{jk}H_{jl_2},$$

$$G_{iD}^{(U)}=st,\; G_{iD}^{(A)}=qr]$$

$$=\frac{1}{2}\sum_{s,t,q\in\{D,d\}}\bar{f}_{st}f_{tq}\cdot 2P(H_{jk}t)P(H_{jl_1}s)P(H_{jl_2}q)$$

$$+\frac{1}{4}\sum_{s,t,q,r\in\{D,d\}}\bar{f}_{st}f_{qr}[P(H_{jk}s)P(H_{jl_1}t)+P(H_{jk}t)P(H_{jl_1}s)]$$

$$\times[P(H_{jk}q)P(H_{jl_2}r)+P(H_{jk}r)P(H_{jl_2}q)]. \tag{8}$$

By using equations (4), (5), (6), (7) and (8), we may calculate $\mathrm{E}[z_{ijk}^{(U)}z_{ijk}^{(A)}1_{A_1}1_{U_2}]$ in (3). If $k \neq 1$, then

$$\mathrm{E}[z_{ijk}^{(U)}z_{ijl}^{(A)}1_{A_1}1_{U_2}]$$

$$=\mathrm{E}\left[\left(2\cdot 1_{(G_{ij}^{(U)}=H_{jk}H_{jk})}+\sum_{m\neq k}1_{(G_{ij}^{(U)}=H_{jk}H_{jm})}\right)\right.$$

$$\left.\times\left(2\cdot 1_{(G_{ij}^{(A)}=H_{jl}H_{jl})}+\sum_{n\neq l}1_{(G_{ij}^{(A)}=H_{jl}H_{jn})}\right)1_{A_1}1_{U_2}\right]$$

$$=4\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jk})}1_{(G_{ij}^{(A)}=H_{jl}H_{jl})}1_{A_1}1_{U_2}]$$

$$+2\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jk})}1_{(G_{ij}^{(A)}=H_{jl}H_{jk})}1_{A_1}1_{U_2}]$$

$$+2\sum_{n\neq k,l}\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jk})}1_{(G_{ij}^{(A)}=H_{jl}H_{jn})}1_{A_1}1_{U_2}]$$

$$+2\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jl})}1_{(G_{ij}^{(A)}=H_{jl}H_{jl})}1_{A_1}1_{U_2}]$$

$$+2\sum_{m\neq k,l}\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jm})}1_{(G_{ij}^{(A)}=H_{jl}H_{jl})}1_{A_1}1_{U_2}]$$

$$+\sum_{m\neq k,l}\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jm})}1_{(G_{ij}^{(A)}=H_{jl}H_{jm})}1_{A_1}1_{U_2}]$$

$$+\sum_{m\neq k,l}\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jm})}1_{(G_{ij}^{(A)}=H_{jl}H_{jk})}1_{A_1}1_{U_2}]$$

$$+\sum_{m\neq k,l}\;\sum_{n\neq m,k,l}\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jm})}1_{(G_{ij}^{(A)}=H_{jl}H_{jn})}1_{A_1}1_{U_2}]$$

$$+\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jl})}1_{(G_{ij}^{(A)}=H_{jl}H_{jk})}1_{A_1}1_{U_2}]$$

$$+\sum_{n\neq k,l}E[1_{(G_{ij}^{(U)}=H_{jk}H_{jl})}1_{(G_{ij}^{(A)}=H_{jl}H_{jn})}1_{A_1}1_{U_2}]. \tag{9}$$

First, one may calculate the expectation

$$\begin{aligned}
&\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jk})}1_{(G_{ij}^{(A)}=H_{jl}H_{jl})}1_{A_1}1_{U_2}]\\
&\quad= P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} = H_{jl}H_{jl},\\
&\qquad\qquad A_1, U_2, (2\,\mathrm{IBD}) \cup (1\,\mathrm{IBD}) \cup (0\,\mathrm{IBD})]\\
&\quad= \frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st} f_{st} P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} = H_{jl}H_{jl},\\
&\qquad\qquad G_{iD}^{(U)} = st,\ G_{iD}^{(A)} = st]\\
&\qquad+ \frac{1}{2}\sum_{s,t,q\in\{D,d\}} \bar{f}_{st} f_{tq} P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} = H_{jl}H_{jl},\\
&\qquad\qquad G_{iD}^{(U)} = st,\ G_{iD}^{(A)} = tq]\\
&\qquad+ \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} = H_{jl}H_{jl},\\
&\qquad\qquad G_{iD}^{(U)} = st,\ G_{iD}^{(A)} = qr]\\
&\quad= \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} P(H_{jk}s)P(H_{jk}t)P(H_{jl}q)P(H_{jl}r)
\end{aligned}\tag{10}$$

For $n \neq k$, l, one may have the following expectation

$$\begin{aligned}
&\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jk})}1_{(G_{ij}^{(A)}=H_{jl}H_{jn})}1_{A_1}1_{U_2}]\\
&\quad= P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} = H_{jl}H_{jn},\\
&\qquad\qquad A_1, U_2, (2\,\mathrm{IBD}) \cup (1\,\mathrm{IBD}) \cup (0\,\mathrm{IBD})]\\
&\quad= \frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st} f_{st} P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} = H_{jl}H_{jn},\\
&\qquad\qquad G_{iD}^{(U)} = st,\ G_{iD}^{(A)} = st]\\
&\qquad+ \frac{1}{2}\sum_{s,t,q\in\{D,d\}} \bar{f}_{st} f_{tq} P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} = H_{jl}H_{jn},\\
&\qquad\qquad G_{iD}^{(U)} = st,\ G_{iD}^{(A)} = tq]\\
&\qquad+ \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} P[G_{ij}^{(U)} = H_{jk}H_{jk}, G_{ij}^{(A)} = H_{jl}H_{jn},\\
&\qquad\qquad G_{iD}^{(U)} = st,\ G_{iD}^{(A)} = qr]\\
&\quad= \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} P(H_{jk}s)P(H_{jk}t)[P(H_{jl}q)P(H_{jn}r)\\
&\qquad+ P(H_{jl}r)P(H_{jn}q)].
\end{aligned}\tag{11}$$

For $m \neq k$, l, one may have the following expectation

$$\begin{aligned}
&\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jm})}1_{(G_{ij}^{(A)}=H_{jl}H_{jl})}1_{A_1}1_{U_2}]\\
&\quad= P[G_{ij}^{(U)} = H_{jk}H_{jm}, G_{ij}^{(A)} = H_{jl}H_{jl},\\
&\qquad\qquad A_1, U_2, (2\,\mathrm{IBD}) \cup (1\,\mathrm{IBD}) \cup (0\,\mathrm{IBD})]\\
&\quad= \frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st} f_{st} P[G_{ij}^{(U)} = H_{jk}H_{jm}, G_{ij}^{(A)} = H_{jl}H_{jl},\\
&\qquad\qquad G_{iD}^{(U)} = st,\ G_{iD}^{(A)} = st]\\
&\qquad+ \frac{1}{2}\sum_{s,t,q\in\{D,d\}} \bar{f}_{st} f_{tq} P[G_{ij}^{(U)} = H_{jk}H_{jm}, G_{ij}^{(A)} = H_{jl}H_{jl},\\
&\qquad\qquad G_{iD}^{(U)} = st,\ G_{iD}^{(A)} = tq]\\
&\qquad+ \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} P[G_{ij}^{(U)} = H_{jk}H_{jm}, G_{ij}^{(A)} = H_{jl}H_{jl},\\
&\qquad\qquad G_{iD}^{(U)} = st,\ G_{iD}^{(A)} = qr]\\
&\quad= \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} [P(H_{jk}s)P(H_{jm}t)\\
&\qquad+ P(H_{jk}t)P(H_{jm}s)]P(H_{jl}q)P(H_{jl}r).
\end{aligned}\tag{12}$$

For $m \neq k,l,n \neq m,k,l$, one way have the following expectation:

$$\begin{aligned}
&\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jm})}1_{(G_{ij}^{(A)}=H_{jl}H_{jn})}1_{A_1}1_{U_2}]\\
&\quad= P[G_{ij}^{(U)} = H_{jk}H_{jm}, G_{ij}^{(A)} = H_{jl}H_{jn},\\
&\qquad\qquad A_1, U_2, (2\,\mathrm{IBD}) \cup (1\,\mathrm{IBD}) \cup (0\,\mathrm{IBD})]\\
&\quad= \frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st} f_{st} P[G_{ij}^{(U)} = H_{jk}H_{jm}, G_{ij}^{(A)} = H_{jl}H_{jn},\\
&\qquad\qquad G_{iD}^{(U)} = st,\ G_{iD}^{(A)} = st]\\
&\qquad+ \frac{1}{2}\sum_{s,t,q\in\{D,d\}} \bar{f}_{st} f_{tq} P[G_{ij}^{(U)} = H_{jk}H_{jm}, G_{ij}^{(A)} = H_{jl}H_{jn},\\
&\qquad\qquad G_{iD}^{(U)} = st,\ G_{iD}^{(A)} = tq]\\
&\qquad+ \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} P[G_{ij}^{(U)} = H_{jk}H_{jm}, G_{ij}^{(A)} = H_{jl}H_{jn},\\
&\qquad\qquad G_{iD}^{(U)} = st,\ G_{iD}^{(A)} = qr]\\
&\qquad\times \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} [P(H_{jk}s)P(H_{jm}t)\\
&\qquad+ P(H_{jk}t)P(H_{jm}s)][P(H_{jl}q)P(H_{jn}r)\\
&\qquad+ P(H_{jl}r)P(H_{jn}q)].
\end{aligned}\tag{13}$$

Using equations (5) (6), (7), (8), (9), (10), (11) and (13), we may calculate terms of equation (7).

Appendix C

For $j \neq g$, the expectation

$$\begin{aligned}
&\mathrm{E}[z_{ijk}^{(U)} z_{igh}^{(A)} 1_{A_1} 1_{U_2}]\\
&\quad= \mathrm{E}\left[\left(2\cdot 1_{(G_{ij}^{(U)}=H_{jk}H_{jk})} + \sum_{k'\neq k} 1_{(G_{ij}^{(U)}=H_{jk}H_{jk'})}\right)\right.\\
&\qquad\left.\times\left(2\cdot 1_{(G_{ig}^{(A)}=H_{gh}H_{gh})} + \sum_{h'\neq h} 1_{(G_{ig}^{(A)}=H_{gh}H_{gh'})}\right) 1_{A_1} 1_{U_2}\right].
\end{aligned}\tag{14}$$

Suppose that blocks/markers H_j and H_g flank disease locus D in the order H_jDH_g. The expectation

$$\begin{aligned}
&\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jk})}1_{(G_{ig}^{(A)}=H_{gh}H_{gh})}1_{A_1}1_{U_2}]\\
&= P[G_{ij}^{(U)} = H_{jk}H_{jk},\, G_{ig}^{(A)} = H_{gh}H_{gh},\\
&\qquad A_1, U_2, (2\,\mathrm{IBD}) \cup (1\,\mathrm{IBD}) \cup (0\,\mathrm{IBD})]\\
&= \frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st} f_{st} P[G_{ij}^{(U)} = H_{jk}H_{jk},\, G_{ig}^{(A)} = H_{gh}H_{gh},\\
&\qquad G_{iD}^{(U)} = st,\; G_{iD}^{(A)} = st]\\
&+ \frac{1}{2}\sum_{s,t,q\in\{D,d\}} \bar{f}_{st} f_{tq} P[G_{ij}^{(U)} = H_{jk}H_{jk},\, G_{ig}^{(A)} = H_{gh}H_{gh},\\
&\qquad G_{iD}^{(U)} = st,\; G_{iD}^{(A)} = tq]\\
&+ \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} P[G_{ij}^{(U)} = H_{jk}H_{jk},\, G_{ig}^{(A)} = H_{gh}H_{gh},\\
&\qquad G_{iD}^{(U)} = st,\; G_{iD}^{(A)} = qr]\\
&= \frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st} f_{st} P(H_{jk}sH_{gh})P(H_{jk}tH_{gh})\\
&\quad + \frac{1}{2}\sum_{s,t,q\in\{D,d\}} \bar{f}_{st} f_{tq} P(H_{jk}tH_{gh})P(H_{jk}s)P(qH_{gh})\\
&+ \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} P(H_{jk}s)P(H_{jk}t)P(qH_{gh})P(rH_{gh}). \qquad (15)
\end{aligned}$$

If $h' \neq h$, the expectation

$$\begin{aligned}
&\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jk})}1_{(G_{ig}^{(A)}=H_{gh}H_{gh'})}1_{A_1}1_{U_2}]\\
&= P[G_{ij}^{(U)} = H_{jk}H_{jk},\, G_{ig}^{(A)} = H_{gh}H_{gh'},\\
&\qquad A_1, U_2, (2\,\mathrm{IBD}) \cup (1\,\mathrm{IBD}) \cup (0\,\mathrm{IBD})]\\
&= \frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st} f_{st} P[G_{ij}^{(U)} = H_{jk}H_{jk},\, G_{ig}^{(A)} = H_{gh}H_{gh'},\\
&\qquad G_{iD}^{(U)} = st,\; G_{iD}^{(A)} = st]\\
&+ \frac{1}{2}\sum_{s,t,q\in\{D,d\}} \bar{f}_{st} f_{tq} P[G_{ij}^{(U)} = H_{jk}H_{jk},\, G_{ig}^{(A)} = H_{gh}H_{gh'},\\
&\qquad G_{iD}^{(U)} = st,\; G_{iD}^{(A)} = tq]\\
&+ \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} P[G_{ij}^{(U)} = H_{jk}H_{jk},\, G_{ig}^{(A)} = H_{gh}H_{gh'},\\
&\qquad G_{iD}^{(U)} = st,\; G_{iD}^{(A)} = qr]\\
&= \frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st} f_{st} [P(H_{jk}sH_{gh})P(H_{jk}tH_{gh'})\\
&\quad + P(H_{jk}tH_{gh})P(H_{jk}sH_{gh'})]\\
&\quad + \frac{1}{2}\sum_{s,t,q\in\{D,d\}} \bar{f}_{st} f_{tq} [P(H_{jk}tH_{gh})P(H_{jk}s)P(qH_{gh'})\\
&\quad + P(H_{jk}tH_{gh'})P(H_{jk}s)P(qH_{gh})]\\
&\quad + \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} P(H_{jk}s)P(H_{jk}t)[P(qH_{gh})P(rH_{gh'})\\
&\quad + P(rH_{gh})P(qH_{gh'})]. \qquad (16)
\end{aligned}$$

If $k \neq k'$, the expectation

$$\begin{aligned}
&\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jk})}1_{(G_{ig}^{(A)}=H_{gh}H_{gh})}1_{A_1}1_{U_2}]\\
&= P[G_{ij}^{(U)} = H_{jk}H_{jk'},\, G_{ig}^{(A)}\\
&= H_{gh}H_{gh'}, A_1, U_2, (2\,\mathrm{IBD}) \cup (1\,\mathrm{IBD}) \cup (0\,\mathrm{IBD})]\\
&= \frac{1}{4}\sum_{s,t,\in\{D,d\}} \bar{f}_{st} f_{st} P[G_{ij}^{(U)} = H_{jk}H_{jk'},\, G_{ig}^{(A)} = H_{gh}H_{gh'},\\
&\qquad G_{iD}^{(U)} = st,\; G_{iD}^{(A)} = st]\\
&+ \frac{1}{2}\sum_{s,t,q\in\{D,d\}} \bar{f}_{st} f_{tq} P[G_{ij}^{(U)} = H_{jk}H_{jk'},\, G_{ig}^{(A)} = H_{gh}H_{gh},\\
&\qquad G_{iD}^{(U)} = st,\; G_{iD}^{(A)} = tq]\\
&+ \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} P[G_{ij}^{(U)} = H_{jk}H_{jk'},\, G_{ig}^{(A)} = H_{gh}H_{gh},\\
&\qquad G_{iD}^{(U)} = st,\; G_{iD}^{(A)} = qr]\\
&= \frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st} f_{st} [P(H_{jk}sH_{gh})P(H_{jk'}tH_{gh})\\
&+ P(H_{jk}tH_{gh})P(H_{jk'}sH_{gh})]\\
&+ \frac{1}{2}\sum_{s,t,q\in\{D,d\}} \bar{f}_{st} f_{tq} [P(H_{jk}tH_{gh})P(H_{jk'}s)P(qH_{gh})\\
&+ P(H_{jk'}tH_{gh})P(H_{jk}s)P(qH_{gh})]\\
&+ \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} [P(H_{jk}s)P(H_{jk'}t) + P(H_{jk}t)P(H_{jk'}s)]\\
&\quad \times P(qH_{gh})P(rH_{gh}). \qquad (17)
\end{aligned}$$

If $k \neq k'$, $h \neq h'$ the expectation

$$\begin{aligned}
&\mathrm{E}[1_{(G_{ij}^{(U)}=H_{jk}H_{jk'})}1_{(G_{ig}^{(A)}=H_{gh}H_{gh'})}1_{A_1}1_{U_2}]\\
&= P[G_{ij}^{(U)} = H_{jk}H_{jk'}, G_{ig}^{(A)} = H_{gh}H_{gh'},\\
&\qquad A_1, U_2, (2\,\mathrm{IBD}) \cup (1\,\mathrm{IBD}) \cup (0\,\mathrm{IBD})]\\
&= \frac{1}{4}\sum_{s,t\in\{D,d\}} \bar{f}_{st} f_{st} P[G_{ij}^{(U)} = H_{jk}H_{jk'}, G_{ig}^{(A)} = H_{gh}H_{gh'},\\
&\qquad G_{iD}^{(U)} = st, G_{iD}^{(A)} = st]\\
&+ \frac{1}{2}\sum_{s,t,q\in\{D,d\}} \bar{f}_{st} f_{tq} P[G_{ij}^{(U)} = H_{jk}H_{jk'}, G_{ig}^{(A)} = H_{gh}H_{gh'},\\
&\qquad G_{iD}^{(U)} = st, G_{iD}^{(A)} = tq]\\
&+ \frac{1}{4}\sum_{s,t,q,r\in\{D,d\}} \bar{f}_{st} f_{qr} P[G_{ij}^{(U)} = H_{jk}H_{jk'}, G_{ig}^{(A)} = H_{gh}H_{gh'},\\
&\qquad G_{iD}^{(U)} = st, G_{iD}^{(A)} = qr]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{4} \sum_{s,t \in \{D,d\}} \bar{f}_{st} f_{st} [P(H_{jk} s H_{gh}) P(H_{jk'} t H_{gh'}) \\
&\quad + P(H_{jk} t H_{gh}) P(H_{jk'} s H_{gh'}) + P(H_{jk} s H_{gh'}) P(H_{jk'} t H_{gh}) \\
&\quad + P(H_{jk} t H_{gh'}) P(H_{jk'} s H_{gh})] \\
&+ \frac{1}{2} \sum_{s,t,q \in \{D,d\}} \bar{f}_{st} f_{tq} [P(H_{jk} t H_{gh}) P(H_{jk'} s) P(q H_{gh'}) \\
&\quad + P(H_{jk'} t H_{gh}) P(H_{jk} s) P(q H_{gh'}) + P(H_{jk} t H_{gh'}) P(H_{jk'} s) P(q H_{gh}) \\
&\quad + P(H_{jk'} t H_{gh'}) P(H_{jk} s) P(q H_{gh})] \\
&+ \frac{1}{4} \sum_{s,t,q,r \in \{D,d\}} \bar{f}_{st} f_{qr} [P(H_{jk} s) P(H_{jk'} t) \\
&\quad + P(H_{jk} t) P(H_{jk'} s)][P(q H_{gh}) P(r H_{gh'}) + P(r H_{gh}) P(q H_{gh'})].
\end{aligned} \tag{18}
$$

Appendix D

For the 'genotype coding' method, the coding vector of the affected sibling in the i-th sib-pair is $X_{ij}^{(A)} = (x_{ij1}^{(A)}, \ldots, x_{ij(n_j-1)}^{(A)}, x_{ij12}^{(A)}, \ldots, x_{ij1n_j}^{(A)}, \ldots, x_{ij(n_j-1)n_j}^{(A)})^\tau$, $j = 1, \ldots, J$. Similarly, $Y_{ij}^{(U)} = (x_{ij1}^{(U)}, \ldots, x_{ij(n_j-1)}^{(U)}, x_{ij12}^{(U)}, \ldots, x_{ij1n_j}^{(U)}, \ldots, x_{ij(n_j-1)n_j}^{(U)})^\tau$ $j = 1, \ldots, J$ is the coding vector of the normal sibling in the i-th sib-pair. Using the expectations $\mathrm{E}[X_{ij}^{(A)} | A_1, U_2]$ and $\mathrm{E}[Y_{ij}^{(U)} | A_1, U_2]$ given in equations (1) and (2), one may calculate the following variance–covariance matrices:

$$
\begin{aligned}
&\mathrm{Var}(X_{ij}^{(A)} | A_1, U_2) \\
&\quad = diag(a_{j11}, \ldots, a_{j(n_j-1)(n_j-1)}, a_{j12}, \ldots, a_{j1n_j}, \ldots, a_{j(n_j-1)n_j}) \\
&\qquad - [X_{ij}^{(A)} | A_1, U_2] \mathrm{E}[X_{ij}^{(A)} | A_1, U_2]^\tau, \\
&\mathrm{Var}(Y_{ij}^{(U)} | A_1, U_2) \\
&\quad = diag(\bar{a}_{j11}, \ldots, \bar{a}_{j(n_j-1)(n_j-1)}, \bar{a}_{j12}, \ldots, \bar{a}_{j1n_j}, \ldots, \bar{a}_{j(n_j-1)n_j}) \\
&\qquad - \mathrm{E}[Y_{ij}^{(U)} | A_1, U_2] \mathrm{E}[Y_{ij}^{(U)} | A_1, U_2]^\tau.
\end{aligned} \tag{19}
$$

The covariances between x_{ijk}, $x_{ijkk'}$ and x_{igh}, $x_{ighh'}$ are given by

$$
\begin{aligned}
\mathrm{Cov}(x_{ijk}^{(A)}, x_{igh}^{(A)} | A_1, U_2) &= g_{kkhh}^{(A,jg)} - a_{jkk} a_{ghh}, \\
\mathrm{Cov}(x_{ijk}^{(A)}, x_{ighh'}^{(A)} | A_1, U_2) &= g_{kkhh'}^{(A,jg)} - a_{jkk} a_{ghh'}, \\
\mathrm{Cov}(x_{ijkk'}^{(A)}, x_{igh}^{(A)} | A_1, U_2) &= g_{kk'hh}^{(A,jg)} - a_{jkk'} a_{ghh}, \\
\mathrm{Cov}(x_{ijkk'}^{(A)}, x_{ighh'}^{(A)} | A_1, U_2) &= g_{kk'hh'}^{(A,jg)} - a_{jkk'} a_{ghh'}.
\end{aligned} \tag{20}
$$

Similarly,

$$
\begin{aligned}
\mathrm{Cov}(x_{ijk}^{(U)}, x_{igh}^{(U)} | A_1, U_2) &= \bar{g}_{kkhh}^{(U,jg)} - \bar{a}_{jkk} \bar{a}_{ghh}, \\
\mathrm{Cov}(x_{ijk}^{(U)}, x_{ighh'}^{(U)} | A_1, U_2) &= \bar{g}_{kkhh'}^{(U,jg)} - \bar{a}_{jkk} \bar{a}_{ghh'}, \\
\mathrm{Cov}(x_{ijkk'}^{(U)}, x_{igh}^{(U)} | A_1, U_2) &= \bar{g}_{kk'hh}^{(U,jg)} - \bar{a}_{jkk'} \bar{a}_{ghh}, \\
\mathrm{Cov}(x_{ijkk'}^{(U)}, x_{ighh'}^{(U)} | A_1, U_2) &= \bar{g}_{kk'hh'}^{(U,jg)} - \bar{a}_{jkk'} \bar{a}_{ghh'}.
\end{aligned} \tag{21}
$$

Using results of equations (19), (20) and (21), one may calculate $\mathrm{Var}(X_i^{(A)} | A_1, U_2)$ and $\mathrm{Var}(Y_i^{(U)} | A_1, U_2)$ for the 'genotype coding' method.

Appendix E

In this Appendix, we calculate the following covariance matrix for the 'genotype coding' method

$$
\begin{aligned}
\mathrm{Cov}(Y_i^{(U)}, X_i^{(A)} | A_1, U_2) &= \mathrm{E}\,[Y_i^{(U)} X_i^{(A)^\tau} | A_1, U_2] \\
&\quad - \mathrm{E}\,[Y_i^{(U)} | A_1, U_2]\, \mathrm{E}\,[X_i^{(A)^\tau} | A_1, U_2] \\
&= \frac{\mathrm{E}\,[Y_i^{(U)} X_i^{(A)^\tau} 1_{A_1}, 1_{U_2}]}{P(A_1, U_2)} \\
&\quad - \mathrm{E}\,[Y_i^{(U)} | A_1, U_2]\, \mathrm{E}\,[X_i^{(A)^\tau} | A_1, U_2].
\end{aligned}
$$

The probability $P(A_1, U_2)$ is given in the Appendix to the manuscript, and the components of expectations $\mathrm{E}\,[X_i^{(A)} | A_1, U_2]$ and $\mathrm{E}\,[Y_i^{(U)} | A_1, U_2]$ are given in equations (1) and (2). For $\mathrm{E}\,[Y_i^{(U)} X_i^{(A)^\tau} 1_{A_1} 1_{U_2}]$, we note the following results:

the expectation $\mathrm{E}\,[1_{(G_{ij}^{(U)} = H_{jk} H_{jk})} 1_{(G_{ij}^{(A)} = H_{jk} H_{jk})} 1_{A_1} 1_{U_2}]$ is given by (4); For $l \neq k$, the expectation $\mathrm{E}\,[1_{(G_{ij}^{(U)} = H_{jk} H_{jk})} 1_{(G_{ij}^{(A)} = H_{jk} H_{jl})} 1_{A_1} 1_{U_2}]$ is given by (5); For $l \neq k$, $\mathrm{E}[1_{(G_{ij}^{(U)} = H_{jk} H_{jl})} 1_{(G_{ij}^{(A)} = H_{jk} H_{jk})} 1_{A_1} 1_{U_2}]$ is given by (6); For $l \neq k$, $\mathrm{E}\,[1_{(G_{ij}^{(U)} = H_{jk} H_{jl})} 1_{(G_{ij}^{(A)} = H_{jk} H_{jl})} 1_{A_1} 1_{U_2}]$ is given by (7); For $l_1 \neq l_2$, $l_1 \neq k$, $l_2 \neq k$, $\mathrm{E}\,[1_{(G_{ij}^{(U)} = H_{jk} H_{jl_1})} 1_{(G_{ij}^{(A)} = H_{jk} H_{jl_2})} 1_{A_1} 1_{U_2}]$ is given by (8); For $l \neq k$, $\mathrm{E}\,[1_{(G_{ij}^{(U)} = H_{jk} H_{jk})} 1_{(G_{ij}^{(A)} = H_{jl} H_{jl})} 1_{A_1} 1_{U_2}]$ is given by (10); For $l \neq k$, $n \neq k, l$, $\mathrm{E}\,[1_{(G_{ij}^{(U)} = H_{jk} H_{jk})} 1_{(G_{ij}^{(A)} = H_{jl} H_{jn})} 1_{A_1} 1_{U_2}]$ is given by (11); For $l \neq k$, $m \neq k, l$, $\mathrm{E}\,[1_{(G_{ij}^{(U)} = H_{jk} H_{jm})} 1_{(G_{ij}^{(A)} = H_{jl} H_{jl})} 1_{A_1} 1_{U_2}]$ is given by (12); For $l \neq k$, $m \neq k, l$, $n \neq m, k, l$, $\mathrm{E}\,[1_{(G_{ij}^{(U)} = H_{jk} H_{jm})} 1_{(G_{ij}^{(A)} = H_{jl} H_{jn})} 1_{A_1} 1_{U_2}]$ is given by (13). In addition, $\mathrm{E}\,[1_{(G_{ij}^{(U)} = H_{jk} H_{jk})} 1_{(G_{ij}^{(A)} = H_{gh} H_{gh})} 1_{A_1} 1_{U_2}]$ is given by (15); $\mathrm{E}\,[1_{(G_{ij}^{(U)} = H_{jk} H_{jk})} 1_{(G_{ij}^{(A)} = H_{gh} H_{gh'})} 1_{A_1} 1_{U_2}]$ is given by (16); $\mathrm{E}\,[1_{(G_{ij}^{(U)} = H_{jk} H_{jk'})} 1_{(G_{ij}^{(A)} = H_{gh} H_{gh})} 1_{A_1} 1_{U_2}]$ is given by (17); Finally, $\mathrm{E}\,[1_{(G_{ij}^{(U)} = H_{jk} H_{jk'})} 1_{(G_{ij}^{(A)} = H_{gh} H_{gh'})} 1_{A_1} 1_{U_2}]$ is given by (18).

Supplementary information: Simulation study

In order to evaluate the accuracy of the non-centrality parameter approximations, we performed simulations for power curves in Figures 1, 2, 3 and 4 of the paper. To do this, we divided the interval (0, 0.065) (or (0, 0.045)) of the LD measure Δ_{11} of LD uniformly into 20 subintervals for Figures 1 and 2 (or Figures 3 and 4). Correspondingly, the 20 subintervals lead to 21 endpoints. For each endpoint, there is a set of parameters for each power curve. Using the set of parameters, 2,500 datasets are simulated for each endpoint. For each dataset, the empirical statistics T_H, T_G, T_{H1} and T_{G1} were calculated. The simulated power is the proportion of the 2,500 simulated datasets for which the empirical statistic is larger than the cut-off point of the corresponding χ^2-distribution at a 0.05 significance level.

From Figures 1, 2, 3 and 4, it can be seen that the theoretical power curves of T_H, T_G, T_{H1} and T_{G1} are perfectly close to the simulated power curves. Thus, the non-centrality parameter approximations are very accurate.

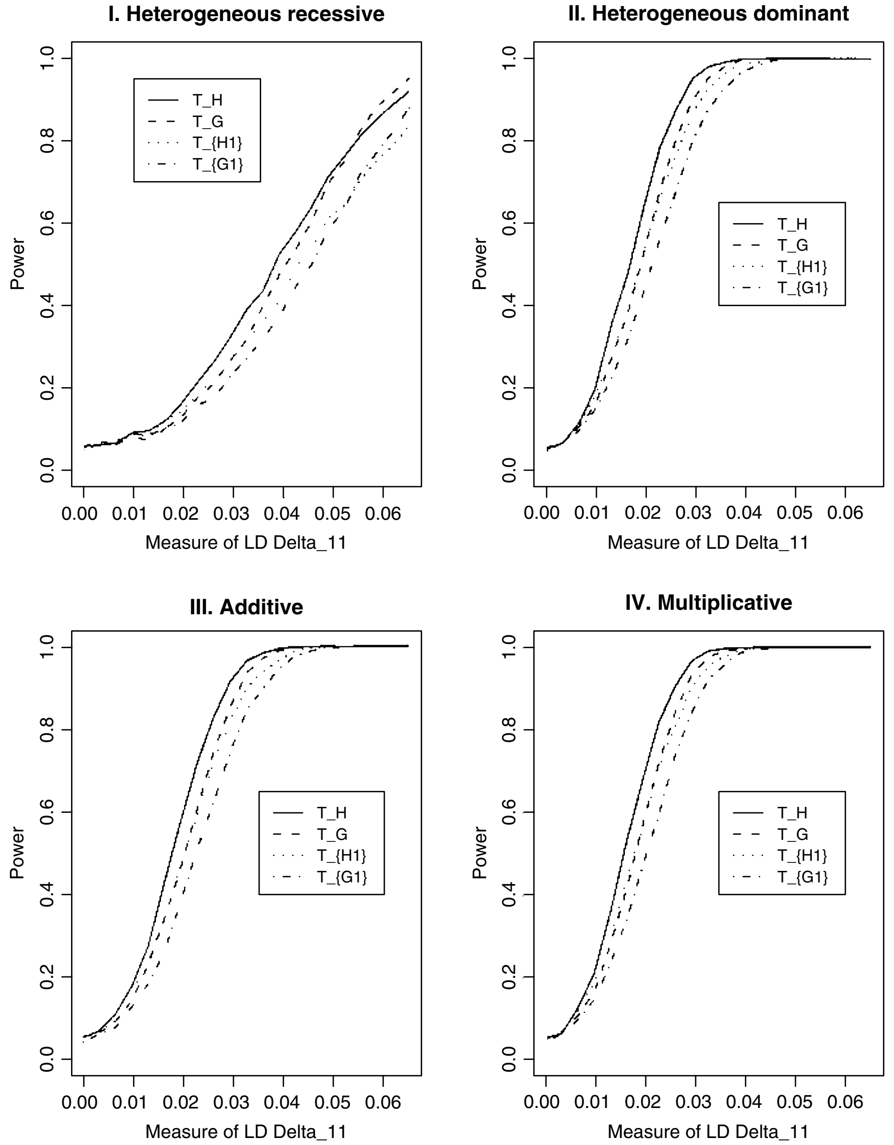

Figure 1. The simulated power curves T_H, T_G, T_{H1} and T_{G1} are plotted. The corresponding parameters are the same as those in Figure 1 of the paper. Abbreviation: LD = linkage disequilibrium.

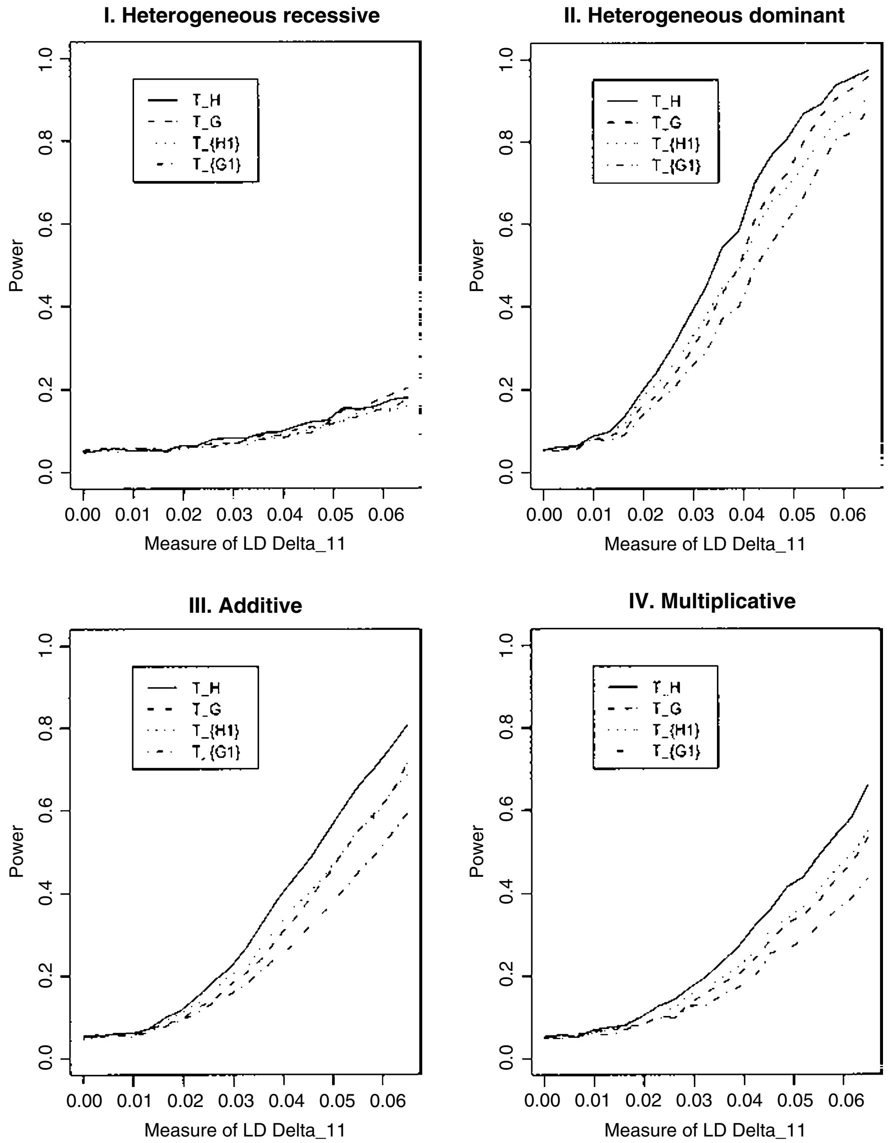

Figure 2. The simulated power curves T_H, T_G, T_{H1} and T_{G1} are plotted. The corresponding parameters are the same as those in Figure 2 of the paper. Abbreviation: LD = linkage disequilibrium.

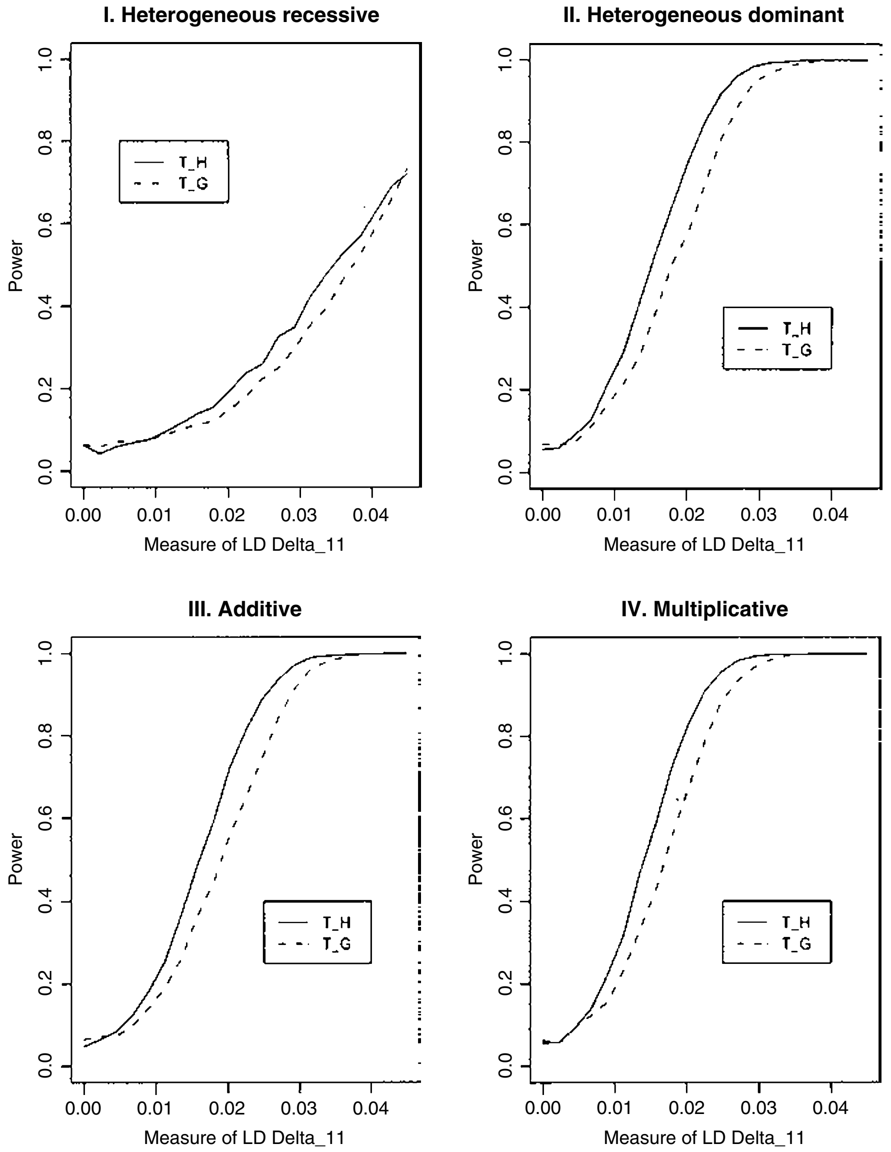

Figure 3. The simulated power curves T_H and T_G are plotted. The corresponding parameters are the same as those of Figure 3 in the paper. Abbreviation: LD = linkage disequilibrium.

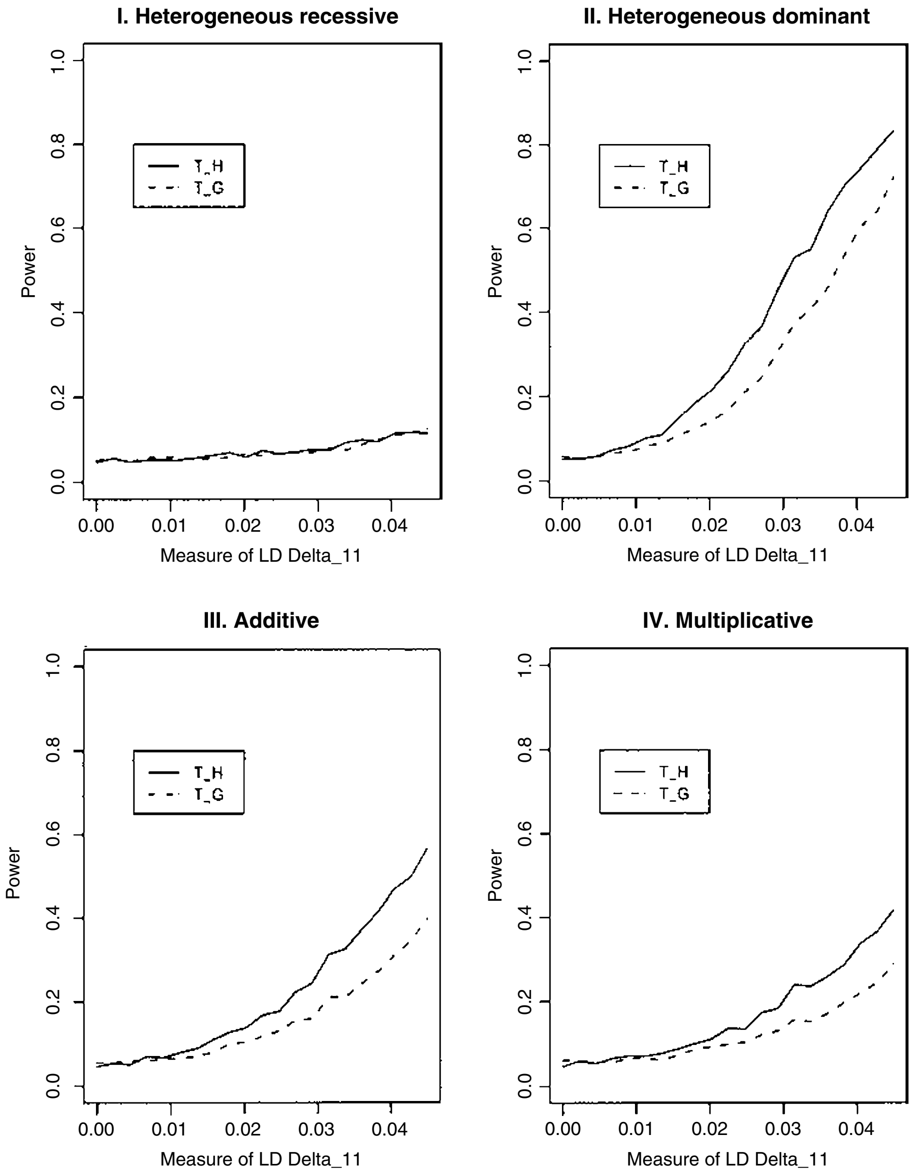

Figure 4. The simulated power curves T_H and T_G are plotted. The corresponding parameters are the same as those of Figure 4 of the paper. Abbreviation: LD = linkage disequilibrium.