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

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 $\chi^2$ test statistics.$^7$–$^{11}$ If there is association between one marker and the trait locus, it is expected that the $\chi^2$ 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.$^6$ 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. 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. 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 $T^2$ tests 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 $X^{(A)}_i$ and $Y^{(U)}_i$ for the affected sibling and normal sibling of the $i$-th sib-pair, respectively, by one of the following two ways:

(i) **Haplotype/allele coding**: For the affected sibling of the $i$-th sib-pair, let $G^{(A)}_{ij}$ be his/her genotype at marker $H_j$.

Define $X^{(A)}_i = \left( z^{(A)}_{ij1}, \ldots, z^{(A)}_{ijn_j} \right)^T$, where $z^{(A)}_{ijl}$ is the number of alleles $H_{jl}$ for the affected sibling of the $i$-th sib-pair — that is,

$$z^{(A)}_{ijl} = \begin{cases} 2 & \text{if } G^{(A)}_{ijl} = H_{jl}H_{jl} \\ 1 & \text{if } G^{(A)}_{ijl} = H_{jl}H_{jl}, l \neq k \\ 0 & \text{else} \end{cases} \quad (i=1, \ldots, J; j=1, \ldots, n_j)$

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

(ii) **Genotype coding**: Note that $G^{(A)}_{ij}$ can be one of $n_j(n_j + 1)/2$ possible choices: $n_j$ homozygous genotypes $H_{jl}H_{jl}$, and $n_j(n_j - 1)/2$ heterozygous genotypes $H_{jl}H_{jl}, k < l$. Depending on the genotype, let us define an indicator vector $X^{(A)}_{ij} = (x^{(A)}_{ij1}, \ldots, x^{(A)}_{ijn_j}, x^{(A)}_{ij12}, \ldots, x^{(A)}_{ijn_j1}, \ldots, x^{(A)}_{ij(n_j-1)n_j})^T$.

Here, $x^{(A)}_{ijl}$ is the indicator variable of genotype $H_{jl}H_{jk}$ defined by

$$x^{(A)}_{ijl} = \begin{cases} 1 & \text{if } G^{(A)}_{ijl} = H_{jl}H_{jk} \\ 0 & \text{else} \end{cases} \quad (i=1, \ldots, J; j=1, \ldots, n_j; k < l)$$

is the indicator variable of genotype $H_{jl}H_{jk}$ defined by $x^{(A)}_{ijk} = \begin{cases} 1 & \text{if } G^{(A)}_{ijk} = H_{jl}H_{jk} \\ 0 & \text{else} \end{cases} \quad (i=1, \ldots, J; j=1, \ldots, n_j; k < l) \quad \text{— that is, the total number } n_j(n_j + 1)/2 \text{ of genotypes of marker } H_j \text{ minus } 1 \text{ to remove the redundancy.}$

Let $X^{(A)}_i = (X^{(A)}_{i1}, \ldots, X^{(A)}_{ijn_j})^T$ be the combined genotype coding of the $J$ markers $H_1, \ldots, H_J$. The dimension of $X^{(A)}_i$ is $\sum_{j=1}^{J} n_j(n_j + 1)/2 - J$.

For the unaffected sibling of the $i$-th sib-pair, let $G^{(U)}_i$ be his/her genotype at marker $H_j$. One may define a vector $Y^{(U)}_i$ 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 $\bar{X}^{(A)} = \sum_{i=1}^{N} X^{(A)}_i/N$ and $\bar{Y}^{(U)} = \sum_{i=1}^{N} Y^{(U)}_i/N$ 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 α = 0.01 using one marker, H₁, or two markers, H₁ and H₂. In model I, one bi-allelic marker H₁ is used, P(H₁) = P(H₂) = 0.50. In model II, two bi-allelic markers H₁ and H₂ are used, P(Hᵢ) = 0.5, i, j = 1, 2, ΔHᵢHⱼ = 0.05. In model III, one quadric-allelic marker H₁ is used, P(H₁) = P(H₂) = 0.35, P(H₂) = P(H₄) = 0.15.

| Model | Test | df | # type I error rates | Mean | Std Dev | Minimum | Maximum |
|-------|------|----|----------------------|------|---------|----------|---------|
| I     | TG   | 1  | 100                  | 0.010808 | 0.0014264 | 0.0066 | 0.0140 |
| N=200 | TG   | 2  | 100                  | 0.011240 | 0.0013923 | 0.0082 | 0.0152 |
| II    | TG   | 2  | 100                  | 0.011286 | 0.0014717 | 0.0070 | 0.0146 |
| N=200 | TG   | 4  | 100                  | 0.012352 | 0.0015899 | 0.0088 | 0.0160 |
| III   | TG   | 3  | 100                  | 0.011660 | 0.0014348 | 0.0074 | 0.0160 |
| N=200 | TG   | 9  | 100                  | 0.014352 | 0.0018710 | 0.0102 | 0.0196 |
| III   | TG   | 3  | 100                  | 0.01186  | 0.0015669 | 0.0074 | 0.0160 |
| N=300 | TG   | 9  | 100                  | 0.013076 | 0.0017027 | 0.0084 | 0.0176 |

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 α = 0.01, using one marker H₁ or two markers H₁ and H₂. Three models are considered. In model I, one marker H₁ is used in analysis: H₁ is a bi-allelic marker with equal allele frequency P(H₁) = P(H₂) = 0.50. In model II, two bi-allelic markers H₁ and H₂ are used in analysis, where P(Hᵢ) = 0.5, i, j = 1, 2,
In model III, one marker $H_1$ is used in analysis, where $H_1$ is a quadri-allelic marker with allele frequencies $P(H_{21}) = 0.35$, $P(H_{22}) = 0.15$. 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 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 $H_2$.

| 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  |

$\Delta_{H_1,H_2} = 0.05$. 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 $H_2$.

| 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                  | 0.014536 | 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 $T_C$. 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 $T_C$ and $T_H$ is 3 and 9, respectively, for model III. Hence, the number of degrees of freedom for test $T_C$ 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. For less optimistic models, Table 5 lists penetrance probabilities taken from Fan and Knapp. 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 $\Delta_{jk} = P(H_{jk}D) - P(H_{jk})P(D)$, $k = 1, \ldots, n_j$. Here, $P(H_{jk}D)$ is the frequency of haplotype $H_{jk}D$.

#### 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   |

and $P(H_{jk})$ is the frequency of allele $H_{jk}$. For two bi-allelic markers $H_1$ and $H_2$, let $\Delta_{H1,H2} = P(H_{11}H_{21}) - P(H_{11})P(H_{21})$ 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 $\Delta_{GD} = P(H_{11}DH_{21}) - P(H_{11})\Delta_{21} - P_D\Delta_{H1,H2} - P(H_{21})\Delta_{11} - P(H_{11})P_D\Delta_{H1,DH21}$ be the measure of the third order LD. Here, $P(H_{11}DH_{21})$ is the frequency of haplotype $H_{11}DH_{21}$.

Figure 1 shows power curves of $T_H$ and $T_C$ against the measure of LD $\Delta_{11}$ at a significance level $\alpha = 0.05$ using two bi-allelic marker $H_1$ and $H_2$, when $P(H_{11}) = P(H_{21}) = 0.50$, $i = 1, 2$, $P_D = 0.15$ and $N = 200$ sib-pairs for the first set of parameters of the four genetic models of Table 4. The power curves of $T_H$ and $T_C$ are calculated based on one marker $H_1$. In the graphs, Delta_11 = Delta_11; the other parameters are given in the legend of the Figure. Figure 2 shows power curves of $T_H$ and $T_C$ against the measure of LD $\Delta_{11}$ at a significance level $\alpha = 0.05$ using two bi-allelic marker $H_1$ and $H_2$, when $P(H_{11}) = P(H_{21}) = 0.50$, $i = 1, 2$, $P_D = 0.15$ 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 $T_H$ and $T_C$ 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 $T_C$. This is consistent with the results of Fan and Knapp for population case-control studies and Fan et al. for nuclear family data. This is most likely due to the large number of degrees of freedom of the test statistic $T_C$. The power of $T_H$ (or $T_C$) based on two markers $H_1$ and $H_2$ is generally higher than that of $T_H$ (or $T_C$), 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     |
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 $\Delta_{11}$ at a significance level $\alpha = 0.05$ using a quadri-allelic marker $H_1$, when $P(H_{i1}) = P(H_{i2}) = 0.35$.

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 $T_H$ and $T_G$ are calculated based on one marker $H_i$. In the graphs, $\Delta_{11} = \Delta_1 = P(H_1D) - P(H_1)P_D$ is a measure of linkage disequilibrium (LD) between marker $H_1$ and disease locus $D$; in addition, the other parameters are given by $\Delta_2 = P(H_2D) - P(H_2)P_D = \Delta_{11}$, $\Delta_{H,H_2} = P(H_1H_2) - P(H_1)P(H_2) = 0.05$, and $\Delta_{1D2} = P(H_1DH_2) - P(H_1)\Delta_1 - P_0\Delta_{H,H_2} - P(H_2)\Delta_1 - P(H_1)P_DP(H_2) = \Delta_{11}/3$.
The other parameters are given in the legend of the Figure. Figure 4 shows power curves of $TH$ and $TG$ at a significance level $\alpha = 0.05$ using a quadri-allelic marker $H_1$, when $P(H_{1i}) = P(H_{12}) = 0.35$, $P(H_{13}) = P(H_{14}) = 0.15$, $P_D = 0.15$ 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, $TH$ generally has a higher power than that of $TG$.

In addition to the power curves of $TH$ and $TG$, which are based on sib-pair data, Figures 3 and 4 show the simulated power curves of $STH$ and $STG$, 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 $STH$ and $STG$: 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 sibships of size 3, two are normal and one is affected.

Figure 2. Power curves of $TH$ and $TG$ at a significance level $\alpha = 0.05$, using two bi-allelic markers $H_1$ and $H_2$, when $P(H_{1i}) = P(H_{2i}) = 0.50$, $i = 1,2$, $P_D = 0.15$ 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 = $\Delta_{11} = P(H_1D) - P(H_1)P_D$ is a measure of linkage disequilibrium (LD) between marker $H_1$ and disease locus $D$; in addition, the other parameters are given by $\Delta_{21} = P(H_2D) - P(H_2)P_D = \Delta_{11}$, $\Delta_{H1H2} = P(H_1H_2D) - P(H_1)P(H_2D) = 0.05$, and $\Delta_{102} = P(H_1DH_2) - P(H_1)\Delta_{21} - P_D\Delta_{H1H2} - P(H_1\Delta_{11} - P(H_1)P_D)P(H_2D) = \Delta_{11}/3$.

$P(H_{1i}) = P(H_{14}) = 0.15$, $P_D = 0.15$ 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 $T_{H1}$ and $T_{G1}$ at a significance level $\alpha = 0.05$ using a quadri-allelic marker $H_1$, when $P(H_{1i}) = P(H_{12}) = 0.35$, $P(H_{13}) = P(H_{14}) = 0.15$, $P_D = 0.15$ 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, $TH$ generally has a higher power than that of $TG$.
$N/4 = 50$ 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 $N/2 = 100$; 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.

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 $\Delta_{11}$ of LD is...
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 statistic $T_{H}$ 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 $\chi^2$-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

---

**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 $P(H_1) = P(H_2) = 0.35$, $P(H_3) = P(H_4) = 0.15$, $P_D = 0.15$ and $N = 600$ sib-pairs for the second set of parameters of the four genetic models of Table 5. $\Delta_{11} = P(H_1 D) - P(H_1)P_D$ 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.
Intuitively, sibships of large size contain more information than that of a sib-pair. The test statistics $T_{H}$ and $T_{C}$ 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 on two coding methods—‘haplotype/allele coding’ and ‘genotype coding’—paired Hotelling’s $T^{2}$ test statistics $T_{H}$ and $T_{C}$ are proposed for high-resolution association studies, using normal sibs as controls for affected sibs. The test statistic 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 $T_{C}$ based on the ‘genotype coding’ method. This is most likely due to the large number of degrees of freedom of $T_{C}$. Moreover, the type I error rates of the test statistic $T_{H}$ are lower than those of test statistic $T_{C}$.

For population case-control association studies, false-positives 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 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 Knapp19 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 $T_{C}$ 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 Clayton2 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 $T_{C}$ 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 $T_{C}$ 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 $X^{(H)}$ and $Y^{(U)}$, 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.
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 $A_1 = \text{(the first sibling is affected)}, U_2 = \text{(the second sibling is unaffected)}$. 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 $f_{DD} = 1 - f_{Dd}, f_{Dd} = 1 - f_{dd}$ and $f_{dd} = 1 - f_{Dd}$. Denote the disease prevalence in population by $A = f_{DD}f_{Dd}^2 + 2f_{DD}f_{Dd}f_{dd} + f_{dd}f_{dd}^2 = 1 - A$. Assume that the affected status of an individual depends only on his/her own genotype at the disease locus. Let us denote the event $(i \text{ IBD}) = \text{the sib-pair share i gene identical by descent (IBD) at the disease locus D}$. Then the joint probability

$$
P(A_1, U_2) = P(A_1, U_2 | 2 \text{ IBD})/4 + P(A_1, U_2 | 1 \text{ IBD})/2 + P(A_1, U_2 | 0 \text{ IBD})/4$$

$$= \frac{1}{4} \left[ \sum_{i \in \{D,d\}} f_{ij}P_{ij} + 2 \sum_{j \in \{D,d\}} f_{ij}P_{ij} + \sum_{i,j \in \{D,d\}} f_{ij}P_{ij} + AA \right],$$

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 $H_kH_k$ in an affected sibling is given by:

$$a_{jk} = P[G_{ij}^{(A)} = H_kH_k | A_1, U_2] = P[G_{ij}^{(A)} = H_kH_k, A_1, U_2 | (2 \text{ IBD}) \cup (1 \text{ IBD}) \cup (0 \text{ IBD})] / P(A_1, U_2)$$

$$= \frac{1}{4} \left[ \sum_{i \in \{D,d\}} f_{ij}P(H_k | H_k)P(H_k | H_k) + \sum_{j \in \{D,d\}} f_{ij}P(H_k | H_k)P(H_k | H_k) + \sum_{i,j \in \{D,d\}} f_{ij}P(H_k | H_k)P(H_k | H_k) + AA \right] / P(A_1, U_2).$$

Similarly, the frequency of homozygous genotype $H_kH_k$ in an unaffected sibling is given by:

$$a_{jk} = P[G_{ij}^{(U)} = H_kH_k | A_1, U_2] = P[G_{ij}^{(U)} = H_kH_k, A_1, U_2 | (2 \text{ IBD}) \cup (1 \text{ IBD}) \cup (0 \text{ IBD})] / P(A_1, U_2)$$

$$= \frac{1}{4} \left[ \sum_{i \in \{D,d\}} f_{ij}P(H_k | H_k)P(H_k | H_k) + \sum_{j \in \{D,d\}} f_{ij}P(H_k | H_k)P(H_k | H_k) + \sum_{i,j \in \{D,d\}} f_{ij}P(H_k | H_k)P(H_k | H_k) + AA \right] / P(A_1, U_2).$$

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

Similarly, the frequency of the heterozygous genotype $H_kH_l, k \neq l$, in an affected sibling can be calculated as follows:

$$a_{jk} = P[G_{ij}^{(A)} = H_kH_l | A_1, U_2] = P[G_{ij}^{(A)} = H_kH_l, A_1, U_2 | (2 \text{ IBD}) \cup (1 \text{ IBD}) \cup (0 \text{ IBD})] / P(A_1, U_2)$$

$$= \frac{1}{4} \left[ \sum_{i \in \{D,d\}} f_{ij}P(H_k | H_k)P(H_l | H_l) + \sum_{j \in \{D,d\}} f_{ij}P(H_k | H_k)P(H_l | H_l) + \sum_{i,j \in \{D,d\}} f_{ij}P(H_k | H_k)P(H_l | H_l) + AA \right] / P(A_1, U_2).$$

Similarly, the frequency of the heterozygous genotype $H_kH_l, k \neq l$, in an unaffected sibling can be calculated as follows:

$$a_{jk} = P[G_{ij}^{(U)} = H_kH_l | A_1, U_2] = P[G_{ij}^{(U)} = H_kH_l, A_1, U_2 | (2 \text{ IBD}) \cup (1 \text{ IBD}) \cup (0 \text{ IBD})] / P(A_1, U_2)$$

$$= \frac{1}{4} \left[ \sum_{i \in \{D,d\}} f_{ij}P(H_k | H_k)P(H_l | H_l) + \sum_{j \in \{D,d\}} f_{ij}P(H_k | H_k)P(H_l | H_l) + \sum_{i,j \in \{D,d\}} f_{ij}P(H_k | H_k)P(H_l | H_l) + AA \right] / P(A_1, U_2).$$
Note that $\hat{a}_{kl}$ can be calculated by the formula for $a_{kl}$ by substituting $\hat{t}_{ik}$ using $T^2$ 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, $a_{iD} = 0$ for all $j$, the haplotype frequencies are equal to the product of the allele frequencies; for example, $P(H_iD) = P(H_i)P_D$, $P(H_0D) = P(H_0)P_D$, and $P(H_0D) = P(H_0)P_D$. From equations (4) and (5), $a_{iD} = \hat{a}_{iD} = 2P(H_iD)P(H_0D)$. Therefore, the expectation $E\{X_2(D) - Y_2(D)|A_1, U_2\} = 0$ for the ‘genotype coding’ method.

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

$$E\{z_{ik}(D)|A_1, U_2\} = 2\hat{a}_{ik} + \sum_{j\neq k} a_{ij}, E\{z_{ik}(U)|A_1, U_2\} = 2\hat{a}_{ik} + \sum_{j\neq k} \hat{a}_{ij}.$$  

(6)

From equation (6), expectation $E\{z_{ik}(D) - z_{ik}(U)|A_1, U_2\} = 2P(H_iD) - 2P(H_0D) = 0$ 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 multilocus 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 disease-genotype 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., McGinnes, 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 Wijmjan, 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 Wijmjan, 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., Ehnh, 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, John 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’, Natue Vol 409, pp. 928–933.

15. Kong, A., Guibjartsson, 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 tagging: 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., Wijmjan, M. et al. (2005), ‘High resolution $T^2$ association tests of complex diseases based on family data’, Am. Hum. Genet. Vol 69, pp. 187–208.

21. Loukola, A., Chalot, M., Penn, S.G. et al. (2004), ‘Comprehensive evaluation of the association between prostate cancer and genotype/haplotypes in CYP17A1, CYP3A4, and SDR5A2’, 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. Wijmjan, 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 design’, Am. J. Hum. Genet. Vol 72, pp. 671–680.

27. Curtis, D. (1997), ‘Use of siblings as controls in case-control association studies’, Am. 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 $A_1 = (\text{the first sibling is affected}), U_2 = (\text{the second sibling is unaffected})$. For ‘haplotype/allele coding’, the coding vector of the affected sibling in the $i$-th sib-pair is $X_{ik}^{(A)} = (z_{i1}^{(A)}, \ldots, z_{i(n_0-1)}^{(A)}, \ldots, z_{in_1}^{(A)}, \ldots, z_{in_0}^{(A)})^T$. Similarly, $Y_{ik}^{(U)} = (z_{i1}^{(U)}, \ldots, z_{i(n_0-1)}^{(U)}, \ldots, z_{i1}^{(U)}, \ldots, z_{in_0}^{(U)})^T$ is the coding vector of the normal sibling. Denote the variance–covariance matrix of $X_{ik}^{(A)} = Y_{ik}^{(U)}$ by $\Sigma_{xy} = Var(X_{ik}^{(A)}) - Y_{ik}^{(U)}|A_1, U_2 = Var(X_{ik}^{(A)}|A_1, U_2) - Cov(X_{ik}^{(A)}, Y_{ik}^{(U)}|A_1, U_2) - Cov(Y_{ik}^{(U)}, X_{ik}^{(A)}|A_1, U_2) + Var(Y_{ik}^{(U)}|A_1, U_2)$. The elements of the above variance–covariance matrices are given in Appendices A, B, and C: $Var(X_{ik}^{(A)}|A_1, U_2)$ and
Var(Y_{ij}^{(U)}|A_1, U_2) in Appendix A, and Cov(X_{ij}^{(U)}, Y_{ij}^{(U)}|A_1, U_2) in Appendices B and C. Using quantities of E(z_{ijk}^{(A)}|A_1, U_2) and E(z_{ijk}^{(A)}|A_1, U_2) in the Appendix to the manuscript, E(\chi_{ij}^{(A)} - Y_{ij}^{(U)}|A_1, U_2) can be calculated. The non-centrality parameter \lambda_{ij} of Hotelling's statistics T_{ij} is given by \lambda_{ij} = NE(\chi_{ij}^{(A)} - Y_{ij}^{(U)}|A_1, U_2)/\Sigma_{h_{ij}}^{-1}E(\chi_{ij}^{(A)} - Y_{ij}^{(U)}|A_1, U_2).

For the 'genotype coding' method, the coding vector of the affected sibling in the i-th sub-pair is X_{ij}^{(A)} = (x_{ij1}, \ldots, x_{ij(n-1)}, x_{ijn+1}, \ldots, x_{ijn})^T \quad j = \ldots, J.

Similarly, Y_{ij}^{(U)} = (y_{ij1}, \ldots, y_{ijn+1}, \ldots, y_{ijn})^T is the coding vector of the normal sibling. Let a_{ijkl} and a_{ijk} be the frequencies of genotype H_{ik}H_{jk} in affected and unaffected siblings given in the Appendix to the manuscript. Then,

E[X_{ij}^{(A)}|A_1, U_2] = (a_{ij11}, \ldots, a_{ij(n-1)(n-1)}, a_{ij22}, \ldots, a_{ijnn})^T, \quad E[Y_{ij}^{(U)}|A_1, U_2] = (a_{ij11}, \ldots, a_{ij(n-1)(n-1)}, a_{ij22}, \ldots, a_{ijnn})^T.

Using E[X_{ij}^{(A)}|A_1, U_2] and E[Y_{ij}^{(U)}|A_1, U_2], one may calculate the expectation E(\chi_{ij}^{(A)} - Y_{ij}^{(U)}|A_1, U_2) = E(X_{ij}^{(A)} - Y_{ij}^{(U)}|A_1, U_2)^T. Let \Sigma_{g_{ij}} = Cov(X_{ij}^{(A)} - Y_{ij}^{(U)}|A_1, U_2) = Var(X_{ij}^{(A)}|A_1, U_2) - Cov(X_{ij}^{(A)}, Y_{ij}^{(U)}|A_1, U_2) - Cov(Y_{ij}^{(U)}, X_{ij}^{(A)}|A_1, U_2) + Var(Y_{ij}^{(U)}|A_1, U_2) be the variance–covariance matrix of \chi_{ij}^{(A)} - Y_{ij}^{(U)}. Then the non-centrality parameter \lambda_{ij} of Hotelling's statistics T_{ij} is given by \lambda_{ij} = NE(\chi_{ij}^{(A)} - Y_{ij}^{(U)}|A_1, U_2)\Sigma_{g_{ij}}^{-1}E(\chi_{ij}^{(A)} - Y_{ij}^{(U)}|A_1, U_2).

For j \neq g, assume that markers H_i and H_k flank disease locus D in the order of H_iDH_k. Let P(H_{ik}H_{jk}H_{k}\text{H}_{g}) be frequencies of haplotype H_{ik}H_{jk}H_{k}\text{H}_{g}. The frequencies of other haplotypes are denoted accordingly. For the i-th sub-pair, let G_{ij}^{(A)} be the disease genotype of the unaffected sibling and G_{ij}^{(U)} be the disease genotype of the affected sibling. To calculate the covariance between z_{ijk}^{(A)} and z_{ijkl}^{(A)}, denote for j \neq g, k \neq l', h \neq h',

\begin{align*}
Var(z_{ijk}^{(A)}|A_1, U_2) &= E[(z_{ijk}^{(A)})^2|A_1, U_2] - [E(z_{ijk}^{(A)}|A_1, U_2)]^2 \\
&= 4a_{ijk} + \sum_{l \neq k} a_{ijkl} - 2a_{jkl} + \sum_{l \neq k} a_{jkl}. \\

Similarly, the covariance between the number of alleles H_{ik} and the number of alleles H_{ik} | \neq k, in the affected and unaffected siblings can be calculated as

Cov(z_{ijkl}^{(A)}, z_{ijkl}^{(A)}|A_1, U_2) &= E(z_{ijkl}^{(A)}z_{ijkl}^{(A)}|A_1, U_2) - E(z_{ijkl}^{(A)}|A_1, U_2)E(z_{ijkl}^{(A)}|A_1, U_2) \\
&= P(G_{ij}^{(A)} = H_{ik}H_{jk}|A_1, U_2) - \sum_{l \neq k} a_{jkl} + \sum_{l \neq k} a_{jkl}. \\

For j \neq g, assume that markers H_i and H_k flank disease locus D in the order of H_iDH_k. Let P(H_{ik}H_{jk}H_{k}\text{H}_{g}) be frequencies of haplotype H_{ik}H_{jk}H_{k}\text{H}_{g}. The frequencies of other haplotypes are denoted accordingly. For the i-th sub-pair, let G_{ij}^{(A)} be the disease genotype of the unaffected sibling and G_{ij}^{(U)} be the disease genotype of the affected sibling. To calculate the covariance between z_{ijk}^{(A)} and z_{ijkl}^{(A)}, denote for j \neq g, k \neq l', h \neq h',

\begin{align*}
Cov(z_{ijkl}^{(A)}, z_{ijkl}^{(A)}|A_1, U_2) &= E(z_{ijkl}^{(A)}z_{ijkl}^{(A)}|A_1, U_2) - E(z_{ijkl}^{(A)}|A_1, U_2)E(z_{ijkl}^{(A)}|A_1, U_2) \\
&= P(G_{ij}^{(A)} = H_{ik}H_{jk}|A_1, U_2) - \sum_{l \neq k} a_{jkl} + \sum_{l \neq k} a_{jkl}. \\

Appendix A

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

Var(X_{ij}^{(A)}|A_1, U_2) = Var(z_{ij11}, \ldots, z_{ijn+1}, \ldots, z_{ij(n-1)(n-1)}, a_{ij22}, \ldots, a_{ijnn})^T|A_1, U_2],

Var(Y_{ij}^{(U)}|A_1, U_2) = Var(z_{ij11}, \ldots, z_{ijn+1}, \ldots, z_{ijn+1})^T|A_1, U_2],

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

Var(z_{ijk}^{(A)}|A_1, U_2) = E[(z_{ijk}^{(A)})^2|A_1, U_2] - [E(z_{ijk}^{(A)}|A_1, U_2)]^2 = 4a_{ijk} + \sum_{l \neq k} a_{ijkl} - 2a_{jkl} + \sum_{l \neq k} a_{jkl}. \\

Similarly, the covariance between the number of alleles H_{ik} and the number of alleles H_{ik} | \neq k, in the affected and unaffected sibling can be calculated as

Cov(z_{ijkl}^{(A)}, z_{ijkl}^{(A)}|A_1, U_2) = E(z_{ijkl}^{(A)}z_{ijkl}^{(A)}|A_1, U_2) - E(z_{ijkl}^{(A)}|A_1, U_2)E(z_{ijkl}^{(A)}|A_1, U_2) = P(G_{ij}^{(A)} = H_{ik}H_{jk}|A_1, U_2) - \sum_{l \neq k} a_{jkl} + \sum_{l \neq k} a_{jkl}. \\

For j \neq g, assume that markers H_i and H_k flank disease locus D in the order of H_iDH_k. Let P(H_{ik}H_{jk}H_{k}\text{H}_{g}) be frequencies of haplotype H_{ik}H_{jk}H_{k}\text{H}_{g}. The frequencies of other haplotypes are denoted accordingly. For the i-th sub-pair, let G_{ij}^{(A)} be the disease genotype of the unaffected sibling and G_{ij}^{(U)} be the disease genotype of the affected sibling. To calculate the covariance between z_{ijk}^{(A)} and z_{ijkl}^{(A)}, denote for j \neq g, k \neq l', h \neq h',
Sibship $T^2$ association tests

\[
\begin{align*}
\hat{G}_{\text{IBD}}^{(A)} &= \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jk}H_{gl}H_{gh}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jk}H_{gl}) P(H_{jk}H_{gh}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jk}H_{gh}) P(H_{jk}H_{gl}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jk}H_{gl}) P(H_{jk}H_{gh}) \\
\end{align*}
\]

\[
\begin{align*}
\hat{G}_{\text{IBD}}^{(A)} &= \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jk}H_{gl}H_{gh}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jk}H_{gl}) P(H_{jk}H_{gh}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jk}H_{gh}) P(H_{jk}H_{gl}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jk}H_{gl}) P(H_{jk}H_{gh}) \\
\end{align*}
\]

\[
\begin{align*}
\hat{G}_{\text{IBD}}^{(A)} &= \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jk}H_{gl}H_{gh}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jak}H_{gh}) P(H_{jak}H_{gl}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jak}H_{gh}) P(H_{jak}H_{gl}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jak}H_{gl}) P(H_{jak}H_{gh}) \\
\end{align*}
\]

\[
\begin{align*}
\hat{G}_{\text{IBD}}^{(A)} &= \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jk}H_{gl}H_{gh}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jak}H_{gh}) P(H_{jak}H_{gl}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jak}H_{gh}) P(H_{jak}H_{gl}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jak}H_{gl}) P(H_{jak}H_{gh}) \\
\end{align*}
\]

\[
\begin{align*}
\hat{G}_{\text{IBD}}^{(A)} &= \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jk}H_{gl}H_{gh}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jak}H_{gh}) P(H_{jak}H_{gl}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jak}H_{gh}) P(H_{jak}H_{gl}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jak}H_{gl}) P(H_{jak}H_{gh}) \\
\end{align*}
\]

\[
\begin{align*}
\hat{G}_{\text{IBD}}^{(A)} &= \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jk}H_{gl}H_{gh}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jak}H_{gh}) P(H_{jak}H_{gl}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jak}H_{gh}) P(H_{jak}H_{gl}) \\
+ \frac{1}{4} \sum_{i,t\in[D,d]} f_{i,t} \hat{G}_{ij}^{(A)} P(H_{jak}H_{gl}) P(H_{jak}H_{gh}) \\
\end{align*}
\]
For $k = 1, \ldots, n_j - 1$ and $h = 1, \ldots, n_g - 1, j \neq g$, the covariance
\[
\text{Cov}(z^{(A)}_{ijk}, z^{(A)}_{ijh} | A_1, U_2) = E[z^{(A)}_{ijk} z^{(A)}_{ijh} | A_1, U_2] \equiv 4g^{2}_{kkhh} + 2g^{(A)}_{kkhh} + 2g^{(A)}_{kkhh} + 2g^{(A)}_{kkhh} \left[ 2a_{ghk} + 2g_{ghh} \right].
\]
Similarly, for $k = 1, \ldots, n_j - 1$ and $h = 1, \ldots, n_g - 1, j \neq g$, the covariance
\[
\text{Cov}(z^{(U)}_{ijk}, z^{(U)}_{ijh} | A_1, U_2) = E[z^{(U)}_{ijk} z^{(U)}_{ijh} | A_1, U_2] \equiv 4g^{2}_{kkhh} + 2g^{(U)}_{kkhh} + 2g^{(U)}_{kkhh} \left[ 2a_{ghk} + 2g_{ghh} \right].
\]
where $g^{(A)}_{kkhh}$, $g^{(U)}_{kkhh}$, $g^{(A)}_{kkhh}$, and $g^{(U)}_{kkhh}$ are the expected genotype frequencies in the normal sibling as follows:
\[
g^{(A)}_{kkhh} = E[1_{(c_y^{(A)} = H_j H_k)} 1_{(c_y^{(A)} = H_j H_k)} | A_1, U_2],
\]
\[
g^{(U)}_{kkhh} = E[1_{(c_y^{(U)} = H_j H_k)} 1_{(c_y^{(U)} = H_j H_k)} | A_1, U_2],
\]
\[
g^{(A)}_{kkhh} = E[1_{(c_y^{(A)} = H_j H_k)} 1_{(c_y^{(A)} = H_j H_k)} | A_1, U_2],
\]
\[
g^{(U)}_{kkhh} = E[1_{(c_y^{(U)} = H_j H_k)} 1_{(c_y^{(U)} = H_j H_k)} | A_1, U_2].
\]

To calculate $g^{(A)}_{kkhh}$, $g^{(U)}_{kkhh}$, $g^{(A)}_{kkhh}$, and $g^{(U)}_{kkhh}$, one may use the formula of $g^{(A)}_{kkhh}$, $g^{(U)}_{kkhh}$, $g^{(A)}_{kkhh}$, and $g^{(U)}_{kkhh}$ by substituting $f_{it}$ using $f_{it}$.

**Appendix B**

The conditional covariance
\[
\text{Cov}(Y^{(U)}_i, X^{(A)}_i | A_1, U_2) = E[Y^{(U)}_i X^{(A)}_i | A_1, U_2] - E[Y^{(U)}_i | A_1, U_2]E[X^{(A)}_i | A_1, U_2] = E[Y^{(U)}_i X^{(A)}_i | A_1, U_2] - E[Y^{(U)}_i | A_1, U_2]E[X^{(A)}_i | A_1, U_2].
\]
For the ‘haplotype/allele coding’ method, the expectations $E[Y^{(U)}_i | A_1, U_2]$ and $E[X^{(A)}_i | A_1, U_2]$ are given by two quantities $E(z^{(A)}_{ijk} | A_1, U_2)$ and $E(z^{(U)}_{ijk} | A_1, U_2)$ (see Appendix to the paper). To get $E[Y^{(U)}_i X^{(A)}_i | A_1, U_2]$, we will calculate $E[z^{(A)}_{ijk} z^{(A)}_{ijh} | A_1, U_2]$ and $E[z^{(U)}_{ijk} z^{(U)}_{ijh} | A_1, U_2]$, $l \neq k$ in this Appendix. In Appendix C, we will calculate the expectation $E[z^{(U)}_{ijk} z^{(U)}_{ijh} | A_1, U_2]$ for $j \neq g$. Note that:
\[
E[z^{(U)}_{ijk} z^{(U)}_{ijh} | A_1, U_2] = E \left( 21_{(c^{(A)}_y = H_j H_k)} \sum_{l \neq k} 1_{(c^{(A)}_y = H_j H_k)} A_1 U_2 \right).
\]

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
\[
E[1_{(c^{(A)}_y = H_j H_k)} 1_{(c^{(A)}_y = H_j H_k)} | A_1, U_2] = P[G^{(A)}_y = H_j H_k, G^{(A)}_y = H_j H_k, A_1, U_2],
\]
(2IBD) $\cup$ (1IBD) $\cup$ (0IBD)]
\[
= \frac{1}{4} \sum_{i,t \in (D,d)} f_{it} f_{it} P[G^{(A)}_y = H_j H_k, G^{(A)}_y = H_j H_k] C^{(D)}_{id} = st, G^{(D)}_{id} = st
\]
\[
+ \frac{1}{2} \sum_{i,t \in (D,d)} f_{it} f_{it} P[G^{(A)}_y = H_j H_k, G^{(A)}_y = H_j H_k] C^{(I)}_{id} = st, G^{(I)}_{id} = st
\]
\[
+ \frac{1}{4} \sum_{i,t \in (D,d)} f_{it} f_{it} P[G^{(A)}_y = H_j H_k, G^{(A)}_y = H_j H_k] C^{(0)}_{id} = st, G^{(0)}_{id} = st
\]
\[
= \frac{1}{4} \sum_{i,t \in (D,d)} f_{it} f_{it} P[H_j H_k] P(H_j H_k)
\]
\[
+ \frac{1}{2} \sum_{i,t \in (D,d)} f_{it} f_{it} P[H_j H_k] P(H_j H_k)
\]
\[
+ \frac{1}{4} \sum_{i,t \in (D,d)} f_{it} f_{it} P[H_j H_k] P(H_j H_k). \tag{4}
\]

For $l \neq k$, one may calculate the expectation
\[
E[1_{(c^{(A)}_y = H_j H_k)} 1_{(c^{(A)}_y = H_j H_k)} | A_1, U_2] = P[G^{(A)}_y = H_j H_k, G^{(A)}_y = H_j H_k, A_1, U_2],
\]
(2IBD) $\cup$ (1IBD) $\cup$ (0IBD)]
\[
= \frac{1}{4} \sum_{i,t \in (D,d)} f_{it} f_{it} P[G^{(A)}_y = H_j H_k, G^{(A)}_y = H_j H_k] C^{(D)}_{id} = st, G^{(D)}_{id} = st
\]
\[
+ \frac{1}{2} \sum_{i,t \in (D,d)} f_{it} f_{it} P[G^{(A)}_y = H_j H_k, G^{(A)}_y = H_j H_k] C^{(I)}_{id} = st, G^{(I)}_{id} = st
\]
\[
+ \frac{1}{4} \sum_{i,t \in (D,d)} f_{it} f_{it} P[G^{(A)}_y = H_j H_k, G^{(A)}_y = H_j H_k] C^{(0)}_{id} = st, G^{(0)}_{id} = st
\]
\[
= st, G^{(D)}_{id} = st, G^{(I)}_{id} = st, G^{(0)}_{id} = st.
\]
Similarly, one has the following expectation

$$E \left[ \sum_{i,j \in \{D,d\}} f_{ij} \cdot 2P(G_{ij})P(H_{ij})P(H_{ij}q) + P(H_{ij})P(H_{ij}r) \right].$$

(5)

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

$$E[1(G_{jk}^{(U)}=H_j H_k) \cdot 1(G_{lk}^{(A)}=H_l H_k) | A_1, U_1]$$

$$= \sum_{i,j \in \{D,d\}} f_{ij} \cdot 2P(G_{ij}^{(C)}) = H_j H_k, G_{ij}^{(A)}$$

$$= \sum_{i,j \in \{D,d\}} f_{ij} \cdot 2P(H_{ij})P(H_{ij}q) + P(H_{ij}r)P(H_{ij}q)$$

(6)

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

$$E[1(G_{lk}^{(U)}=H_l H_k) \cdot 1(G_{lk}^{(A)}=H_l H_k) | A_1, U_1]$$

$$= \sum_{i,j \in \{D,d\}} f_{ij} \cdot 2P(G_{ij}^{(C)}) = H_j H_k, G_{ij}^{(A)}$$

$$= \sum_{i,j \in \{D,d\}} f_{ij} \cdot 2P(H_{ij})P(H_{ij}q) + P(H_{ij}r)P(H_{ij}q)$$

(7)

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

$$E[1(G_{lk}^{(U)}=H_l H_k) \cdot 1(G_{lk}^{(A)}=H_l H_k) | A_1, U_1]$$

$$= \sum_{i,j \in \{D,d\}} f_{ij} \cdot 2P(G_{ij}^{(C)}) = H_j H_k, G_{ij}^{(A)}$$

$$= \sum_{i,j \in \{D,d\}} f_{ij} \cdot 2P(H_{ij})P(H_{ij}q) + P(H_{ij}r)P(H_{ij}q)$$

(8)

By using equations (4), (5), (6), (7) and (8), we may calculate $E[1(G_{jk}^{(A)} \neq H_j H_k) | A_1, U_1]$ in (3). If $k \neq l$, then

$$E[1(G_{jk}^{(A)} \neq H_j H_k) | A_1, U_1]$$

$$= E \left[ \left( 2 \cdot 1(G_{jk}^{(U)}=H_j H_k) + \sum_{m \neq k} 1(G_{mk}^{(U)}=H_m H_k) \right) \right.$$
First, one may calculate the expectation
\[
E[1_{(G^{(U)}_y = H_kH_j)}] (1_{G^{(A)}_y = H_kH_j}) A_1 U_2
\]
\[
= P[G^{(U)}_y = H_kH_j, G^{(A)}_y = H_kH_j, A_1, U_2, (2 \text{ IBD}) \cup (1 \text{ IBD}) \cup (0 \text{ IBD})]
\]
\[
= \frac{1}{4} \sum_{t,q} j_{t,q} P[G^{(U)}_y = H_kH_j, G^{(A)}_y = H_kH_j, G^{(U)}_D = st, G^{(A)}_D = st]
\]
\[
+ \frac{1}{2} \sum_{t,q} j_{t,q} P[G^{(U)}_y = H_kH_j, G^{(A)}_y = H_kH_j, G^{(U)}_D = st, G^{(A)}_D = tq]
\]
\[
+ \frac{1}{4} \sum_{t,q} j_{t,q} P[G^{(U)}_y = H_kH_j, G^{(A)}_y = H_kH_j, G^{(U)}_D = st, G^{(A)}_D = qr]
\]
\[
= \frac{1}{4} \sum_{t,q} j_{t,q} P(H_{jk}) P(H_{jk}) P(H_{jk}) P(H_{jk}) (10)
\]

For \(n \neq k, l\), one may have the following expectation
\[
E[1_{(G^{(U)}_y = H_kH_j)}] (1_{G^{(A)}_y = H_kH_j}) A_1 U_2
\]
\[
= P[G^{(U)}_y = H_kH_j, G^{(A)}_y = H_kH_j, A_1, U_2, (2 \text{ IBD}) \cup (1 \text{ IBD}) \cup (0 \text{ IBD})]
\]
\[
= \frac{1}{4} \sum_{t,q} j_{t,q} P[G^{(U)}_y = H_kH_j, G^{(A)}_y = H_kH_j, G^{(U)}_D = st, G^{(A)}_D = st]
\]
\[
+ \frac{1}{2} \sum_{t,q} j_{t,q} P[G^{(U)}_y = H_kH_j, G^{(A)}_y = H_kH_j, G^{(U)}_D = st, G^{(A)}_D = tq]
\]
\[
+ \frac{1}{4} \sum_{t,q} j_{t,q} P[G^{(U)}_y = H_kH_j, G^{(A)}_y = H_kH_j, G^{(U)}_D = st, G^{(A)}_D = qr]
\]
\[
= \frac{1}{4} \sum_{t,q} j_{t,q} P(H_{jk}) P(H_{jk}) [P(H_{jk}) P(H_{jk})]
\]
\[
+ P(H_{jk}) P(H_{jk})]
\]
\[
= \frac{1}{4} \sum_{t,q} j_{t,q} P[H_{jk}] P(H_{jk}) (11)
\]

For \(m \neq k, l\), one may have the following expectation
\[
E[1_{(G^{(U)}_y = H_kH_m)}] (1_{G^{(A)}_y = H_kH_m}) A_1 U_2
\]
\[
= P[G^{(U)}_y = H_kH_m, G^{(A)}_y = H_kH_m, A_1, U_2, (2 \text{ IBD}) \cup (1 \text{ IBD}) \cup (0 \text{ IBD})]
\]
\[
= \frac{1}{4} \sum_{t,q} j_{t,q} P[G^{(U)}_y = H_kH_m, G^{(A)}_y = H_kH_m, G^{(U)}_D = st, G^{(A)}_D = st]
\]
\[
+ \frac{1}{2} \sum_{t,q} j_{t,q} P[G^{(U)}_y = H_kH_m, G^{(A)}_y = H_kH_m, G^{(U)}_D = st, G^{(A)}_D = tq]
\]
\[
+ \frac{1}{4} \sum_{t,q} j_{t,q} P[G^{(U)}_y = H_kH_m, G^{(A)}_y = H_kH_m, G^{(U)}_D = st, G^{(A)}_D = qr]
\]
\[
= \frac{1}{4} \sum_{t,q} j_{t,q} P(H_{jk}) P(H_{jk}) P(H_{jk}) P(H_{jk}) (12)
\]

For \(m \neq k, l\) and \(n \neq m, k, l\), one may have the following expectation:
\[
E[1_{(G^{(U)}_y = H_kH_m)}] (1_{G^{(A)}_y = H_kH_m}) A_1 U_2
\]
\[
= \sum_{t,q} j_{t,q} P[G^{(U)}_y = H_kH_m, G^{(A)}_y = H_kH_m, A_1, U_2, (2 \text{ IBD}) \cup (1 \text{ IBD}) \cup (0 \text{ IBD})]
\]
\[
= \frac{1}{4} \sum_{t,q} j_{t,q} P[G^{(U)}_y = H_kH_m, G^{(A)}_y = H_kH_m, G^{(U)}_D = st, G^{(A)}_D = st]
\]
\[
+ \frac{1}{2} \sum_{t,q} j_{t,q} P[G^{(U)}_y = H_kH_m, G^{(A)}_y = H_kH_m, G^{(U)}_D = st, G^{(A)}_D = tq]
\]
\[
+ \frac{1}{4} \sum_{t,q} j_{t,q} P[G^{(U)}_y = H_kH_m, G^{(A)}_y = H_kH_m, G^{(U)}_D = st, G^{(A)}_D = qr]
\]
\[
= \frac{1}{4} \sum_{t,q} j_{t,q} P(H_{jk}) P(H_{jk}) (13)
\]

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
\[
E[1_{(G^{(U)}_y = H_kH_m)}] A_1 U_2
\]
\[
= \sum_{t,q} j_{t,q} P(H_{jk}) P(H_{jk}) 1_{A_1} U_2
\]
\[
= \sum_{t,q} j_{t,q} [2 \cdot 1_{(G^{(U)}_y = H_kH_m)} + \sum_{l \neq k} 1_{(G^{(U)}_y = H_kH_m)}] A_1 U_2
\]
\[
= \sum_{t,q} j_{t,q} [2 \cdot 1_{(G^{(U)}_y = H_kH_m)} + \sum_{l \neq k} 1_{(G^{(U)}_y = H_kH_m)}] A_1 U_2 (14)
\]
Suppose that blocks/markers $H_j$ and $H_g$ flank disease locus $D$ in the order $H_fH_g$. The expectation

$$\text{E}[1(c_{ij}^{U}=H_f,H_g)] = P[c_{ij}^{U} = H_f,H_g, c_{kg}^{A} = H_k,H_g, A_1, U_2, (2\text{IBD}) \cup (1\text{IBD}) \cup (0\text{IBD})]$$

$$= \frac{1}{4} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} u P[c_{ij}^{(U)} = H_f,H_g, c_{kg}^{(A)} = H_k,H_g, G_{1D}^{(U)} = st, G_{1D}^{(A)} = st]$$

$$+ \frac{1}{2} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} q P[c_{ij}^{(U)} = H_f,H_g, c_{kg}^{(A)} = H_k,H_g, G_{1D}^{(U)} = st, G_{1D}^{(A)} = tq]$$

$$+ \frac{1}{2} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} q P[c_{ij}^{(U)} = H_f,H_g, c_{kg}^{(A)} = H_k,H_g, G_{1D}^{(U)} = st, G_{1D}^{(A)} = qr]$$

$$= \frac{1}{4} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} u P(H_f,H_g)P(H_f,H_g)$$

$$+ \frac{1}{2} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} q P(H_f,H_g)P(H_f,H_g)$$

$$+ P(H_f,H_g)P(H_f,H_g)$$

$$+ \frac{1}{2} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} q P(H_f,H_g)P(H_f,H_g)$$

$$= \frac{1}{4} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} u P(H_f,H_g)P(H_f,H_g)$$

$$+ \frac{1}{2} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} q P(H_f,H_g)P(H_f,H_g)$$

$$= \frac{1}{4} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} u$$

$$+ \frac{1}{2} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} q$$

$$= \frac{1}{4} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} u$$

If $k \neq k'$, the expectation

$$\text{E}[1(c_{ij}^{U}=H_f,H_g) = 1(c_{ij}^{U}=H_f,H_g)]$$

$$= P[c_{ij}^{U} = H_f,H_g, c_{kg}^{A} = H_k,H_g, A_1, U_2, (2\text{IBD}) \cup (1\text{IBD}) \cup (0\text{IBD})]$$

$$= \frac{1}{4} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} u P[c_{ij}^{(U)} = H_f,H_g, c_{kg}^{(A)} = H_k,H_g, G_{1D}^{(U)} = st, G_{1D}^{(A)} = st]$$

$$+ \frac{1}{2} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} q P[c_{ij}^{(U)} = H_f,H_g, c_{kg}^{(A)} = H_k,H_g, G_{1D}^{(U)} = st, G_{1D}^{(A)} = tq]$$

$$+ \frac{1}{2} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} q P[c_{ij}^{(U)} = H_f,H_g, c_{kg}^{(A)} = H_k,H_g, G_{1D}^{(U)} = st, G_{1D}^{(A)} = qr]$$

$$= \frac{1}{4} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} u P(H_f,H_g)P(H_f,H_g)$$

$$+ \frac{1}{2} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} q P(H_f,H_g)P(H_f,H_g)$$

$$= \frac{1}{4} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} u$$

$$+ \frac{1}{2} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} q$$

$$= \frac{1}{4} \sum_{t,s \in \{D,d\}} \tilde{f}_{ds} u$$

© HENRY STEWART PUBLICATIONS 1479–7364. HUMAN GENOMICS. VOL. 2. NO. 2. 90–112 JUNE 2005
\[
\begin{align*}
&= \frac{1}{t} \sum_{\ell \in \{1, \ldots, d\}} \text{Var}(X_{j\ell}) + \text{Cov}(X_{j\ell}, X_{j\ell}') + \text{Cov}(X_{j\ell}, X_{j\ell}')' + \text{Var}(X_{j\ell}'),
&= \frac{1}{t} \sum_{\ell \in \{1, \ldots, d\}} \text{Var}(X_{j\ell}) + \text{Cov}(X_{j\ell}, X_{j\ell}') + \text{Cov}(X_{j\ell}, X_{j\ell}')' + \text{Var}(X_{j\ell}'),
&= \frac{1}{t} \sum_{\ell \in \{1, \ldots, d\}} \text{Var}(X_{j\ell}) + \text{Cov}(X_{j\ell}, X_{j\ell}') + \text{Cov}(X_{j\ell}, X_{j\ell}')' + \text{Var}(X_{j\ell}'),
&= \frac{1}{t} \sum_{\ell \in \{1, \ldots, d\}} \text{Var}(X_{j\ell}) + \text{Cov}(X_{j\ell}, X_{j\ell}') + \text{Cov}(X_{j\ell}, X_{j\ell}')' + \text{Var}(X_{j\ell}'),
&= \frac{1}{t} \sum_{\ell \in \{1, \ldots, d\}} \text{Var}(X_{j\ell}) + \text{Cov}(X_{j\ell}, X_{j\ell}') + \text{Cov}(X_{j\ell}, X_{j\ell}')' + \text{Var}(X_{j\ell}'),
&= \frac{1}{t} \sum_{\ell \in \{1, \ldots, d\}} \text{Var}(X_{j\ell}) + \text{Cov}(X_{j\ell}, X_{j\ell}') + \text{Cov}(X_{j\ell}, X_{j\ell}')' + \text{Var}(X_{j\ell}'),
&= \frac{1}{t} \sum_{\ell \in \{1, \ldots, d\}} \text{Var}(X_{j\ell}) + \text{Cov}(X_{j\ell}, X_{j\ell}') + \text{Cov}(X_{j\ell}, X_{j\ell}')' + \text{Var}(X_{j\ell}'),
&= \frac{1}{t} \sum_{\ell \in \{1, \ldots, d\}} \text{Var}(X_{j\ell}) + \text{Cov}(X_{j\ell}, X_{j\ell}') + \text{Cov}(X_{j\ell}, X_{j\ell}')' + \text{Var}(X_{j\ell}'),
&= \frac{1}{t} \sum_{\ell \in \{1, \ldots, d\}} \text{Var}(X_{j\ell}) + \text{Cov}(X_{j\ell}, X_{j\ell}') + \text{Cov}(X_{j\ell}, X_{j\ell}')' + \text{Var}(X_{j\ell}'),
&= \frac{1}{t} \sum_{\ell \in \{1, \ldots, d\}} \text{Var}(X_{j\ell}) + \text{Cov}(X_{j\ell}, X_{j\ell}') + \text{Cov}(X_{j\ell}, X_{j\ell}')' + \text{Var}(X_{j\ell}'),
\end{align*}
\]

Using results of equations (19), (20) and (21), one may calculate \( \text{Var}(X_{j\ell}^U|A_1, U_2) \) and \( \text{Var}(Y_{j\ell}^U|A_1, U_2) \) for the ‘genotype coding’ method.

**Appendix E**

In this Appendix, we calculate the following covariance matrix for the ‘genotype coding’ method:

\[
\text{Cov}(Y_{ij}^U, X_{j\ell}^U|A_1, U_2) = E[Y_{ij}^U X_{j\ell}^U|A_1, U_2] - E[Y_{ij}^U|A_1, U_2] E[X_{j\ell}^U|A_1, U_2] = E[Y_{ij}^U X_{j\ell}^U|A_1, U_2] - E[Y_{ij}^U|A_1, U_2] E[X_{j\ell}^U|A_1, U_2].
\]

The probability \( P(A_1, U_2) \) is given in the Appendix to the manuscript, and the components of expectations \( E[X_{j\ell}^U|A_1, U_2] \) and \( E[Y_{ij}^U|A_1, U_2] \) are given in equations (1) and (2). For \( E[Y_{ij}^U X_{j\ell}^U|A_1, U_2] \), we note the following results:

- the expectation \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (4); for \( l \neq k \), the expectation \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (5); for \( l \neq k \), \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (6); for \( l \neq k \), \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (7); for \( l \neq k \), \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (8); for \( l \neq k \), \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (9); for \( l \neq k \), \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (10); for \( l \neq k \), \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (11); for \( l \neq k \), \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (12); for \( l \neq k \), \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (13).

In addition, \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (14); \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (15); \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (16); \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) is given by (17); \( E[1(G_{j\ell}^U = H_{j\ell} H_{j\ell})] \) 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 \( \Delta_{11} \) 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 $\chi^2$-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 $T_H$, $T_G$, $T_{HI}$ and $T_{GI}$ 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.