

# Genome Association Studies of Complex Diseases by Case-Control Designs

Ruzong Fan<sup>1,2</sup> and Michael Knapp<sup>2</sup>

<sup>1</sup>Department of Statistics, Texas A&M University, College Station, TX; and <sup>2</sup>Institute of Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn

One way to perform linkage-disequilibrium (LD) mapping of genetic traits is to use single markers. Since dense marker maps—such as single-nucleotide polymorphism and high-resolution microsatellite maps—are available, it is natural and practical to generalize single-marker LD mapping to high-resolution haplotype or multiple-marker LD mapping. This article investigates high-resolution LD-mapping methods, for complex diseases, based on haplotype maps or microsatellite marker maps. The objective is to explore test statistics that combine information from haplotype blocks or multiple markers. Based on two coding methods, genotype coding and haplotype coding, Hotelling's  $T^2$  statistics  $T_G$  and  $T_H$  are proposed to test the association between a disease locus and two haplotype blocks or two markers. The validity of the two  $T^2$  statistics is proved by theoretical calculations. A statistic  $T_C$ , an extension of the traditional  $\chi^2$  method of comparing haplotype frequencies, is introduced by simply adding the  $\chi^2$  test statistics of the two haplotype blocks together. The merit of the three methods is explored by calculation and comparison of power and of type I errors. In the presence of LD between the two blocks, the type I error of  $T_C$  is higher than that of  $T_H$  and  $T_G$ , since  $T_C$  ignores the correlation between the two blocks. For each of the three statistics, the power of using two haplotype blocks is higher than that of using only one haplotype block. By power comparison, we notice that  $T_C$  has higher power than that of  $T_H$ , and  $T_H$  has higher power than that of  $T_G$ . In the absence of LD between the two blocks, the power of  $T_C$  is similar to that of  $T_H$  and higher than that of  $T_G$ . Hence, we advocate use of  $T_H$  in the data analysis. In the presence of LD between the two blocks,  $T_H$  takes into account the correlation between the two haplotype blocks and has a lower type I error and higher power than  $T_G$ . Besides, the feasibility of the methods is shown by sample-size calculation.

## Introduction

With the development of the Human Genome Project and of high-resolution microsatellite and early chromosomewide haplotype maps of the human genome, enormous amounts of genetic data on human chromosomes are becoming available. The opportunities for genomewide scans to map complex-disease genes are tremendous. However, it is not yet clear how to extract the most useful information for mapping complex-disease genes. To fully utilize the massive amount of genetic data for mapping complex-disease genes, novel mathematical and statistical methods are crucial. One urgent need is to explore statistical approaches of high-resolution haplotype or multiple-marker linkage disequilibrium (LD) mapping of complex diseases. One way to perform LD mapping of genetic traits is to use single markers. Since dense marker maps—such as single-nucleotide polymorphism (SNP) and high-resolution microsatellite

maps—are available (Broman et al. 1998; The International SNP Map Work Group 2001; Kong et al. 2002), it is natural and practical to generalize single-marker LD mapping to high-resolution haplotype or multiple-marker LD mapping. With the recent discovery of haplotype block structures in the human genome and with the development of early chromosomewide haplotype maps, it is important to develop better statistical methods for analysis of data on SNPs, haplotype patterns, and related patterns of LD. The chromosomewide haplotype maps are expected to be key resources for mapping complex-disease genes. For example, a systematic case-control analysis of common haplotype variants in the human genome would reveal major causative genetic contributions to a disease.

For a case-control study, one can use a  $\chi^2$  statistic to test the null hypothesis that the marker allele or haplotype frequencies are equal in the cases and controls on the basis of a multiple-allele marker (Olson and Wijsman 1984; Chapman and Wijsman 1998; Nielsen et al. 1998; Kaplan and Morris 2001). The method, however, can not be directly used for multiple markers or haplotype blocks, since the phase of a double heterozygote may be unknown (Ott 1999, p. 7). For multiple biallelic markers, such as SNPs, Xiong et al. (2002) proposed a Hotelling's  $T^2$  statistic for LD mapping of qualitative

Received September 20, 2002; accepted for publication January 3, 2003; electronically published March 19, 2003.

Address for correspondence and reprints: Dr. Ruzong Fan, Texas A&M University, 447 Blocker Building, College Station, TX 77843-3143. E-mail: rfan@stat.tamu.edu

© 2003 by The American Society of Human Genetics. All rights reserved. 0002-9297/2003/7204-0008\$15.00

traits for case-control studies, which can not be used for haplotype block data with multiple haplotypes. Hence, it is necessary to develop methods of genomic LD mapping of qualitative trait loci based on haplotype block data or multiple-marker data for case-control studies.

This paper investigates methods of high-resolution LD mapping for complex diseases based on haplotype maps or microsatellite marker maps. The objective is to explore test statistics that combine information from haplotype blocks or multiple markers. For uniformity of notation, we use ‘‘haplotype blocks’’ or simply ‘‘blocks’’ in our analysis, which can be changed to ‘‘multiallelic markers.’’ We are interested in developing statistical methods for efficient use of genomic patterns of LD to identify genetic variants that contribute to qualitative complex diseases on the basis of multiallelic markers or haplotype block data in a case-control study. Based on two coding methods, genotype coding and haplotype coding, we propose Hotelling’s  $T^2$  statistics to test the association between a disease locus and two haplotype blocks. The statistical property of the above  $T^2$  statistics will be investigated. An extension of traditional  $\chi^2$  method of comparing haplotype frequencies is proposed by simply adding two  $\chi^2$  test statistics of the two haplotype blocks together. The merit of the three methods will be explored by calculation and comparison of power and type I errors. Also, the feasibility of the methods is shown by calculation of sample sizes.

**Methods**

*Test Statistics*

Suppose that a disease locus  $D$  is flanked by two haplotype blocks  $H_1$  and  $H_2$ , where  $H_1$  is a haplotype block on the left-hand side of  $D$  and  $H_2$  is a haplotype block on the right-hand side. Let us denote the haplotypes of block  $H_1$  by  $H_{11}, \dots, H_{1l}$  and the haplotypes of block  $H_2$  by  $H_{21}, \dots, H_{2r}$ , where  $l$  and  $r$  denote the number of observed haplotypes of blocks  $H_1$  and  $H_2$ , respectively. Consider a case-control design with  $N$  cases from an affected population and  $M$  controls from a unaffected population. Let us define a coding vector for each case or control by one of the following two ways (Schaid 1996, p. 430).

*Genotype coding.*—For the  $i$ th case, let  $\text{Hap}_{1i}$  be his/her two haplotypes at block  $H_1$ , and let  $\text{Hap}_{2i}$  be his/her two haplotypes at block  $H_2$ . Depending on the haplotypes  $\text{Hap}_{1i}$  (or  $\text{Hap}_{2i}$ ), let us define an indicator vector  $X_{1i}$  (or  $X_{2i}$ ) that contains exactly one component with value 1 and other components with value 0. That is,

$$X_{1i} = [x_{1i1}, \dots, x_{1i(l-1)}, x_{1i12}, \dots, x_{1i1l}, \dots, x_{1i(l-1)l}]^T, \quad X_{2i} = [x_{2i1}, \dots, x_{2i(r-1)}, x_{2i12}, \dots, x_{2i1r}, \dots, x_{2i(r-1)r}]^T, \text{ and}$$

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix},$$

where the indicator variables  $x_{1ij}, x_{1ijk}, j < k, x_{2is}$ , and  $x_{2ist}, s < t$  are defined by

$$\begin{aligned} x_{1ij} &= \begin{cases} 1 & \text{if Hap}_{1i} = H_{1j}H_{1j} \\ 0 & \text{else} \end{cases}, \\ x_{2is} &= \begin{cases} 1 & \text{if Hap}_{2i} = H_{2s}H_{2s} \\ 0 & \text{else} \end{cases}, \\ x_{1ijk} &= \begin{cases} 1 & \text{if Hap}_{1i} = H_{1j}H_{1k} \\ 0 & \text{else} \end{cases}, \\ x_{2ist} &= \begin{cases} 1 & \text{if Hap}_{2i} = H_{2s}H_{2t} \\ 0 & \text{else} \end{cases}. \end{aligned} \tag{1}$$

The dimension of  $X_{1i}$  (or  $X_{2i}$ ) is  $l(l+1)/2 - 1$  (or  $r(r+1)/2 - 1$ )—that is, the total number  $l(l+1)/2$  (or  $r(r+1)/2$ ) of genotypes of haplotype block  $H_1$  (or  $H_2$ ) minus 1 to remove the redundancy.

*Haplotype coding.*—Define  $X_i = [z_{1i1}, \dots, z_{1i(l-1)}, z_{2i1}, \dots, z_{2i(r-1)}]^T$ , where  $z_{uij}$  is the number of haplotypes  $H_{uj}$  for the  $i$ th case, i.e.,

$$\begin{aligned} z_{1ij} &= \begin{cases} 2 & \text{if Hap}_{1i} = H_{1j}H_{1j} \\ 1 & \text{if Hap}_{1i} = H_{1j}H_{1k}, k \neq j \\ 0 & \text{else} \end{cases}, \\ z_{2is} &= \begin{cases} 2 & \text{if Hap}_{2i} = H_{2s}H_{2s} \\ 1 & \text{if Hap}_{2i} = H_{2s}H_{2t}, t \neq s \\ 0 & \text{else} \end{cases}. \end{aligned} \tag{2}$$

For the  $i$ th control, one may define a vector  $Y_i$  in the same way. To illustrate the above coding methods, table 1 gives an example of genotype and haplotype codings for a block  $H_1$  with three haplotypes. The coding method for block  $H_2$  is similar.

Let  $\bar{X} = \sum_{i=1}^N X_i/N$  and  $\bar{Y} = \sum_{i=1}^M Y_i/M$  be the mean

**Table 1**  
**Genotype and Haplotype Codings for a Block  $H_1$  with Three Haplotypes**

Hap <sub>1i</sub>	CODING FOR METHOD	
	Genotype Coding	Haplotype Coding
$H_{11}H_{11}$	10000	20
$H_{12}H_{12}$	01000	02
$H_{13}H_{13}$	00000	00
$H_{11}H_{12}$	00100	11
$H_{11}H_{13}$	00010	10
$H_{12}H_{13}$	00001	01

vectors. Define a pooled-sample variance-covariance matrix by

$$S = \frac{1}{N + M - 2} \times \left[ \sum_{i=1}^N (X_i - \bar{X})(X_i - \bar{X})^T + \sum_{i=1}^M (Y_i - \bar{Y})(Y_i - \bar{Y})^T \right].$$

A Hotelling's  $T^2$  statistic can be defined as

$$T^2 = \frac{NM}{N + M} (\bar{X} - \bar{Y})^T S^{-1} (\bar{X} - \bar{Y})$$

(Hotelling 1931; Anderson 1984). Hereafter, we will denote the Hotelling's  $T^2$  for haplotype coding as  $T_H$  and the Hotelling's  $T^2$  for genotype coding as  $T_G$ . Assume that the sample sizes  $N$  and  $M$  are 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  $\chi^2$  with  $l + r - 2$  (or  $[l(l + 1)/2 - 1] + [r(r + 1)/2 - 1]$ ) df. Under the alternative hypothesis of association,  $T_H$  (or  $T_G$ ) is asymptotically distributed as noncentral  $\chi^2$ . If only one haplotype block  $H_1$  is used in the analysis, the Hotelling's  $T^2$  for haplotype coding will be denoted as  $T_{H1}$ , and the Hotelling's  $T^2$  for genotype coding will be denoted as  $T_{G1}$ . Under the null hypothesis of no association, the statistic  $T_{H1}$  (or  $T_{G1}$ ) is asymptotically distributed as central  $\chi^2$  with  $l - 1$  (or  $l(l + 1)/2 - 1$ ) df. Under the alternative hypothesis of association,  $T_{H1}$  (or  $T_{G1}$ ) is asymptotically distributed as noncentral  $\chi^2$ . Similarly, one may introduce test statistics  $T_{H2}$  (or  $T_{G2}$ ) if only one haplotype block  $H_2$  is used in the analysis.

If each of haplotype block  $H_u$  has only two haplotypes  $H_{u1}, H_{u2}, u = 1, 2$ , then the Hotelling's  $T^2$  by haplotype coding described above coincides with the test statistic introduced by Xiong et al. (2002). To see this, notice that  $z_{ui1} = 1 + (z_{ui1} - 1)$ , where  $z_{ui1} - 1$  is equal to the indicator variable  $X_{ij}$  (defined in Xiong et al. [2002], p. 1257). Hence, our method generalizes the method of using two biallelic markers in Xiong et al. (2002) to two haplotype blocks with multiple haplotypes.

In the definition above, we consider only two haplotype blocks  $H_1$  and  $H_2$ . In practice, the test statistics  $T_H$  and  $T_G$  can be easily generalized to multiple haplotype blocks. To make the notation as simple as possible, we will focus on two haplotype blocks throughout the present article. In appendices A, B, and C, we will justify the use of the Hotelling's  $T^2$  as an appropriate statistic to test association between the disease locus and the haplotype blocks by either the genotype coding method or the haplotype coding method. The basic idea is to show that the expectation of difference  $\bar{X} - \bar{Y}$  is equal to 0 if there is no association between the disease locus

and the haplotype blocks. Then one may construct a test statistic based on the difference vector  $\bar{X} - \bar{Y}$ , which leads to the Hotelling's  $T^2$ .

Noncentrality Parameters

Let  $\Sigma_{A1} = \text{Cov}\{[z_{1i1}, \dots, z_{1i(l-1)}]|\text{Aff}\}$  and  $\Sigma_{A2} = \text{Cov}\{[z_{2i1}, \dots, z_{2i(r-1)}]|\text{Aff}\}$  be variance-covariance matrices of vectors  $[z_{1i1}, \dots, z_{1i(l-1)}]^T$  and  $[z_{2i1}, \dots, z_{2i(r-1)}]^T$ , respectively, in affected individuals. Similarly, let  $\Sigma_{\bar{A}1} = \text{Cov}\{[z_{1i1}, \dots, z_{1i(l-1)}]|\text{Unaff}\}$  and  $\Sigma_{\bar{A}2} = \text{Cov}\{[z_{2i1}, \dots, z_{2i(r-1)}]|\text{Unaff}\}$  be variance-covariance matrices of column vectors  $[z_{1i1}, \dots, z_{1i(l-1)}]^T$  and  $[z_{2i1}, \dots, z_{2i(r-1)}]^T$  in controls. Let  $\Delta_1 = [\Delta_{11}, \dots, \Delta_{1(l-1)}]^T$  (or  $\Delta_2 = [\Delta_{21}, \dots, \Delta_{2(r-1)}]^T$ ) be the column vector of measures of LD between haplotype block  $H_1$  (or  $H_2$ ) and the disease locus  $D$ . Let  $\alpha_D$  be the average effect of gene substitution and  $\bar{\alpha}_D = -\alpha_D$ , and let  $A$  be the disease prevalence in population and  $\bar{A} = 1 - A$  (appendix A). Based on  $E(z_{uij}|\text{Aff})$  and  $E(z_{uij}|\text{Unaff})$ , given in equation (A5) of appendix A, the noncentrality parameter  $\lambda_{H_u}$  of Hotelling's test statistic  $T_{H_u}$  is given by

$$\lambda_{H_u} = \frac{4(\alpha_D/A - \bar{\alpha}_D/\bar{A})^2 NM}{N + M} \times \vec{\Delta}_u^T \left[ \frac{(N - 1)\Sigma_{Au} + (M - 1)\Sigma_{\bar{A}u}}{N + M - 2} \right]^{-1} \vec{\Delta}_u, u = 1, 2.$$

The elements of variance-covariance matrices  $\Sigma_{Au}$  and  $\Sigma_{\bar{A}u}$  are calculated in appendix D. If the haplotype  $H_u$  has only two haplotypes— $H_{u1}, H_{u2}$  and  $N = M$ —then  $\lambda_{H_u} = 4N\Delta_{u1}^2(\alpha_D/A - \bar{\alpha}_D/\bar{A})^2[\text{Var}(z_{ui1}|\text{Aff}) + \text{Var}(z_{ui1}|\text{Unaff})]^{-1}$ , where  $\text{Var}(z_{ui1}|\text{Aff})$  and  $\text{Var}(z_{ui1}|\text{Unaff})$  are given in equations (D1) and (D2) in appendix D.

Let  $\Sigma_A = \text{Cov}\{[z_{1i1}, \dots, z_{1i(l-1)}, z_{2i1}, \dots, z_{2i(r-1)}]|\text{Aff}\}$  be a variance-covariance matrix of column vector  $[z_{1i1}, \dots, z_{1i(l-1)}, z_{2i1}, \dots, z_{2i(r-1)}]^T$  in affected individuals. Similarly, let  $\Sigma_{\bar{A}} = \text{Cov}\{[z_{1i1}, \dots, z_{1i(l-1)}, z_{2i1}, \dots, z_{2i(r-1)}]|\text{Unaff}\}$  be a variance-covariance matrix of column vector  $[z_{1i1}, \dots, z_{1i(l-1)}, z_{2i1}, \dots, z_{2i(r-1)}]^T$  in controls. Let us denote

$$\vec{\Delta} = \begin{pmatrix} \vec{\Delta}_1 \\ \vec{\Delta}_2 \end{pmatrix}.$$

Then the noncentrality parameter  $\lambda_H$  of Hotelling's test statistic  $T_H$  is given by

$$\lambda_H = \frac{4(\alpha_D/A - \bar{\alpha}_D/\bar{A})^2 NM}{N + M} \times \vec{\Delta}^T \left[ \frac{(N - 1)\Sigma_A + (M - 1)\Sigma_{\bar{A}}}{N + M - 2} \right]^{-1} \vec{\Delta}.$$

The elements of variance-covariance matrices  $\Sigma_A$  and  $\Sigma_{\bar{A}}$  are calculated in appendices D and E. The noncentrality parameter  $\lambda_G$  (or  $\lambda_{G1}$  or  $\lambda_{G2}$ ) of  $T_G$  (or  $T_{G1}$  or  $T_{G2}$ ) is given in appendix F.

For a case-control study using only one haplotype block  $H_1$ , one may use a  $\chi^2$  statistic to test the null hypothesis that the haplotype frequencies are equal in the cases and controls (Olson and Wijsman 1994; Chapman and Wijsman 1998; Kaplan and Morris 2001). Assume that  $N$  cases and  $N$  controls are sampled. Then the test statistic is given by  $T_{C1} = 2N \sum_{j=1}^l (\hat{p}_{1j} - \hat{q}_{1j})^2 / (\hat{p}_{1j} + \hat{q}_{1j})$ , where  $\hat{p}_{1j}$  is the frequency of haplotype  $H_{1j}$  in the cases, and  $\hat{q}_{1j}$  is the frequency of haplotype  $H_{1j}$  in the controls. Using haplotype block  $H_2$ , one may construct a similar test statistic  $T_{C2} = 2N \sum_{j=1}^r (\hat{p}_{2j} - \hat{q}_{2j})^2 / (\hat{p}_{2j} + \hat{q}_{2j})$ , where  $\hat{p}_{2j}$  is the frequency of haplotype  $H_{2j}$  in the cases, and  $\hat{q}_{2j}$  is the frequency of haplotype  $H_{2j}$  in the controls. Using both haplotype blocks  $H_1$  and  $H_2$ , one may construct a test statistic  $T_C = T_{C1} + T_{C2}$  by summing  $T_{C1}$  and  $T_{C2}$  together. If the two statistics  $T_{C1}$  and  $T_{C2}$  are independent,  $T_C$  is asymptotically distributed as central  $\chi^2_{l+r-2}$ , with  $l+r-2$  df under the null hypothesis of no association. Under the alternative hypothesis, it is asymptotically distributed as noncentral  $\chi^2_{l+r-2}(\lambda_C)$ , where  $\lambda_C = \lambda_{C1} + \lambda_{C2}$ ,

$$\lambda_{C1} = 2N \sum_{j=1}^l \frac{[P(H_{1j}|Aff) - P(H_{1j}|Unaff)]^2}{P(H_{1j}|Aff) + P(H_{1j}|Unaff)}$$

$$= 2N \left[ \frac{\alpha_D}{A} - \frac{\bar{\alpha}_D}{\bar{A}} \right]^2 \sum_{j=1}^l \frac{\Delta_{1j}^2}{\Delta_{1j}(\alpha_D/A + \bar{\alpha}_D/\bar{A}) + 2P(H_{1j})},$$

and  $\lambda_{C2} = 2N[\alpha_D/A - \bar{\alpha}_D/\bar{A}]^2 \sum_{j=1}^r \Delta_{2j}^2 / [\Delta_{2j}(\alpha_D/A + \bar{\alpha}_D/\bar{A}) + 2P(H_{2j})]$ . To calculate  $\lambda_{C1}$ , one needs to notice that  $P(H_{1j}|Aff) = E(z_{1j}|Aff)/2$ , and so the conditional expected frequencies  $P(H_{1j}|Aff) = \Delta_{1j}\alpha_D/A + P(H_{1j})$  and  $P(H_{1j}|Unaff) = \Delta_{1j}\bar{\alpha}_D/\bar{A} + P(H_{1j})$  (appendices A and C). However, the independence of  $T_{C1}$  and  $T_{C2}$  can be true only in the case that there is linkage equilibrium between the two blocks. Hence,  $T_C$  may not be a valid test statistic unless one has strong evidence that the two blocks are in linkage equilibrium.

**Results**

*Type I Errors*

To explore the performance of the test statistics, we calculate type I errors for statistics  $T_C$ ,  $T_H$ , and  $T_G$  for the four scenarios in table 2. We simulate 10,000 samples under an assumption of penetrance probabilities  $(f_{DD}, f_{Dd}, f_{dd}) = (0.05, 0.05, 0.05)$ , which implies that the disease is not associated with the two haplotype blocks. Every sample contains 100 cases and 100 controls ( $N = M = 100$ ). For each sample, we calculate the em-

**Table 2**

**Type I Errors at Significance Level  $\alpha = 0.01$  using Two Haplotype Blocks  $H_1, l = 2, P(H_{11}) = P(H_{12}) = 0.50$  and  $H_2$  with  $N = M = 100$**

MODEL AND TEST	SIZE <sup>a</sup>	TYPE I ERROR			
		Mean	SD	Minimum	Maximum
<b>I:<sup>b</sup></b>					
$T_G$	101	.012	.001	.009	.014
$T_H$	101	.011	.001	.008	.014
$T_C$	101	.026	.001	.022	.030
<b>II:<sup>c</sup></b>					
$T_G$	101	.015	.001	.012	.018
$T_H$	101	.012	.001	.010	.015
$T_C$	101	.017	.001	.014	.021
<b>III:<sup>d</sup></b>					
$T_G$	101	.020	.001	.018	.023
$T_H$	101	.013	.001	.011	.015
$T_C$	101	.016	.001	.012	.018
<b>IV:<sup>e</sup></b>					
$T_G$	101	.020	.001	.016	.023
$T_H$	101	.013	.001	.011	.016
$T_C$	101	.010	.001	.007	.012

<sup>a</sup> Size is the total number of type I errors calculated for each statistic under a specific model.

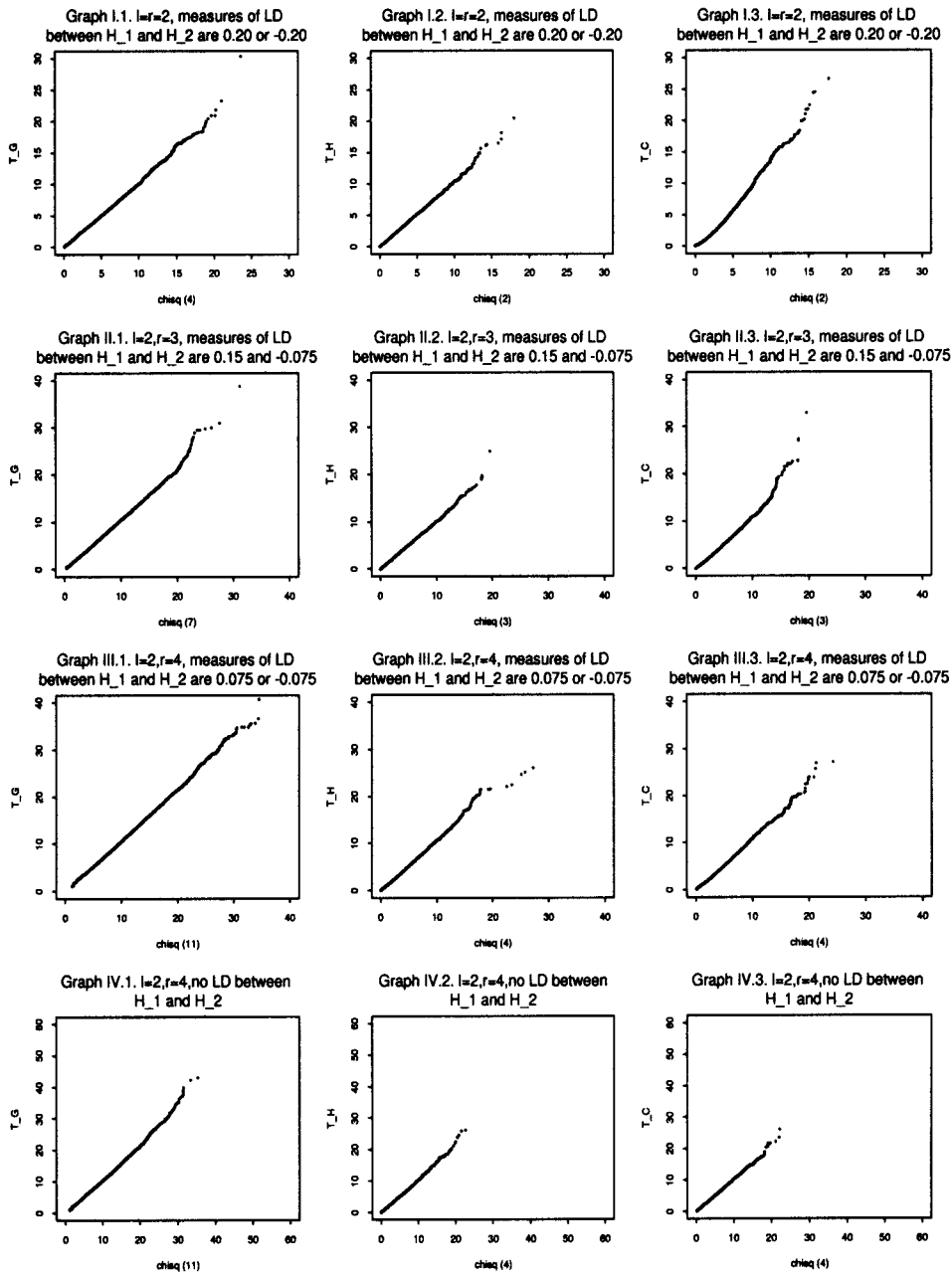
<sup>b</sup> In model I,  $r = 2, P(H_{21}) = P(H_{22}) = 0.50$ , and  $\Delta_{H_{11}H_{21}} = -\Delta_{H_{11}H_{22}} = 0.20$ .

<sup>c</sup> In model II,  $r = 3, P(H_{21}) = 0.4, P(H_{22}) = P(H_{23}) = 0.30, \Delta_{H_{11}H_{21}} = 0.15$ , and  $\Delta_{H_{11}H_{22}} = \Delta_{H_{11}H_{23}} = -0.075$ .

<sup>d</sup> In model III,  $r = 4, P(H_{21}) = P(H_{22}) = P(H_{23}) = P(H_{24}) = 0.25, \Delta_{H_{11}H_{21}} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.075, \Delta_{H_{11}H_{22}} = P(H_{11}H_{22}) - P(H_{11})P(H_{22}) = 0.075, \Delta_{H_{11}H_{23}} = P(H_{11}H_{23}) - P(H_{11})P(H_{23}) = -0.075$ , and  $\Delta_{H_{11}H_{24}} = P(H_{11}H_{24}) - P(H_{11})P(H_{24}) = -0.075$ .

<sup>e</sup> In model IV, all parameters are the same as those of model III, except that  $\Delta_{H_{11}H_{21}} = \Delta_{H_{11}H_{22}} = \Delta_{H_{11}H_{23}} = \Delta_{H_{11}H_{24}} = 0$ .

pirical test statistics  $T_C, T_H$ , and  $T_G$ . The type I error is calculated by dividing the count of those empirical test statistics, which are greater than or equal to the cut-off point at the significance level  $\alpha = 0.01$ , by 10,000. We repeat the above process a total of 100 times to get 101 type I errors for each of the test statistics  $T_C, T_H$ , and  $T_G$  for the four models in table 2. On the basis of the 101 type I errors of each statistic, we calculate their mean, standard deviation (SD), minimum, and maximum, which are presented in table 2. For model I in table 2, a strong LD between the two blocks  $H_1, l = 2$ , and  $H_2, r = 2$ , is assumed ( $\Delta_{H_{11}H_{21}} = 0.20$ ); in this case, the type I error of  $T_C$  (mean 0.026) is much greater than those of  $T_H$  (mean 0.011) and  $T_G$  (0.012). In model II in table 2, we assume that block  $H_1$  has two haplotypes and the block  $H_2$  has three haplotypes, and the measures of LD are  $\Delta_{H_{11}H_{21}} = 0.15$  and  $\Delta_{H_{11}H_{22}} = -0.075$ ; in this case, the type I error of  $T_C$  (mean 0.017) is the highest, and  $T_G$  (mean 0.015) has higher type I error than  $T_H$  (mean 0.012). In models III and IV of table 2, the block  $H_2$  has four haplotypes; in model III, the measures



**Figure 1** QQ plot at significance level  $\alpha = 0.01$  using two haplotype blocks  $H_1, l = 2$ , and  $H_2$ . In graphs I.1, I.2, and I.3, all parameters are the same as those of model I in table 2. In graphs II.1, II.2, and II.3, all parameters are the same as those of model II in table 2. In graphs III.1, III.2, and III.3, all parameters are the same as those of model III in table 2. In graphs IV.1, IV.2, and IV.3, all parameters are the same as those of model IV in table 2.

of LD are  $\Delta_{H_{11}H_{21}} = \Delta_{H_{11}H_{22}} = 0.075$  and  $\Delta_{H_{11}H_{23}} = \Delta_{H_{11}H_{24}} = -0.075$ , and in model IV, the two blocks are in linkage equilibrium; in these two cases, the type I errors of  $T_G$  (mean 0.20) are the highest, which may be due to the large degree of freedom of  $T_G$ . With LD (model III),  $T_C$  (mean 0.16) has a slightly higher type I error than  $T_H$  (mean 0.13); without LD (model VI),  $T_H$  (mean

0.013) has a slightly higher type I error than  $T_C$  (mean 0.010). Figure 1 shows the QQ plot for each statistic of  $T_C, T_H$ , and  $T_G$  for the four models in table 2. Each of the QQ plots in figure 1 is drawn by comparing 10,000 sample statistic values with 10,000 related  $\chi^2$ -distribution values ( $X$ -axis). These QQ plots are consistent with the results of table 2. Moreover, it is evident that the

**Table 3**

**First Set of Parameters of Simulated Genetic Models**

Model Type	$f_{DD}$	$f_{Dd}$	$f_{dd}$
Heterogeneous recessive	1.00	.05	.05
Heterogeneous dominant	1.00	.95	.05
Additive	1.00	.50	.0
Multiplicative	.81	.045	.0025

type I error level of statistic  $T_H$  is reasonable for  $N = M = 100$ .

*Power Calculation and Comparison*

To calculate the noncentrality parameters, we assume a deterministic population genetic model. Assume that a single disease mutation was introduced into the population  $T$  generations ago, with a frequency  $P_D$ . First, we consider only one haplotype block  $H_u, u = 1, 2$ . At the initial generation of the occurrence of the mutation, the haplotype frequencies  $P(H_{u1}D)(0) = P_D$  and  $P(H_{uj}D)(0) = 0, j = 2, \dots, l$ , if  $u = 1$ , or  $j = 2, \dots, r$ , if  $u = 2$ . Moreover,  $P(H_{u1}d)(0) = P(H_{u1}) - P_D$  and  $P(H_{uj}d)(0) = P(H_{uj}), j = 2, \dots, l$  if  $u = 1$ , or  $j = 2, \dots, r$  if  $u = 2$ . Let  $\theta_u$  be the recombination fraction between haplotype block  $H_u$  and disease locus  $D, u = 1, 2$ . Given a map distance  $\lambda_u$  between haplotype block  $H_u$  and disease locus  $D$ , the recombination fraction  $\theta_u$  can be calculated by Haldane's map function  $\theta_u = [1 - \exp(-2\lambda_u)]/2$ , under the assumption of no interference. At generation  $T$ , the haplotype frequencies can be approximately calculated by  $P(H_{uj}D)(T) = P(H_{uj}D)(0)e^{-T\theta_u} + P_D P(H_{uj})(1 - e^{-T\theta_u})$  and  $P(H_{uj}d)(T) = P(H_{uj}d)(0)e^{-T\theta_u} + P_d P(H_{uj})(1 - e^{-T\theta_u}), j = 1, \dots, l$ , if  $u = 1$ , or  $j = 2, \dots, r$ , if  $u = 2$ . Second, we consider both haplotype blocks  $H_1$  and  $H_2$ . At the initial generation of the occurrence of mutation, the haplotype frequencies  $P(H_{11}DH_{21})(0) = P_D$  and  $P(H_{1j}DH_{2s})(0) = 0, j = 1, \dots, l, s = 1, \dots, r$ , and  $(j, s) \neq (1, 1)$ . That is, the disease-susceptibility allele  $D$  was carried by haplotype  $H_{11}H_{21}$  at the initial generation of mutation. The other initial haplotype frequencies are  $P(H_{11}dH_{21})(0) = P(H_{11}H_{21}) - P_D$  and  $P(H_{1j}dH_{2s})(0) = P(H_{1j}H_{2s}), j = 1, \dots, l, s = 1, \dots, r$  and  $(j, s) \neq (1, 1)$ .

At generation  $T$ , the haplotype frequencies can be approximately calculated by  $P(H_{1j}DH_{2s})(T) = \Delta_{jDs}(0)e^{-T(\theta_1+\theta_2)} + P(H_{1j})\Delta_{2s}(0)e^{-T\theta_2} + P(H_{2s})\Delta_{1j}(0)e^{-T\theta_1} + P(H_{1j})P_D P(H_{2s})$  and  $P(H_{1j}dH_{2s})(T) = P(H_{1j}H_{2s}) - P(H_{1j}DH_{2s})(T), j = 1, \dots, l, s = 1, \dots, r$ , where  $\Delta_{jDs}(0) = P(H_{1j}DH_{2s})(0) - P(H_{1j})\Delta_{2s}(0) - P(H_{2s})\Delta_{1j}(0) - P(H_{1j})P_D P(H_{2s})$  is the measure of initial LD at the three loci for haplotypes  $H_{1j}$  and  $H_{2s}$ ,  $\Delta_{1j}(0) = P(H_{1j}D)(0) - P(H_{1j})P_D$  is the measure of initial LD between haplotype  $H_{1j}$  and disease locus  $D$ , and  $\Delta_{2s}(0) = P(DH_{2s})(0) -$

$P_D P(H_{2s})$  is the measure of initial LD between haplotype  $H_{2s}$  and disease locus  $D$  (Akey et al. 2001).

To make a power comparison, we consider four genetic models: heterogeneous recessive, heterogeneous dominant, additive, and multiplicative. First, we consider optimistic penetrance probabilities and genotype relative risks given in table 3 (Nielson et al. 1998). For less optimistic models, with lower penetrance probabilities and genotype relative risks, we consider the four models in table 4. For each model in table 4, the population disease prevalence is  $\sim 0.05$  and the sib recurrence risk is  $\sim 0.06$  (Iles 2002). We assume that the distance between the two haplotype blocks is 4 cM. The block  $H_1$  is located at position 0 cM, and the block  $H_2$  is located at position 4 cM. Since the disease locus  $D$  is usually unknown, we assume that it is located in the interval between  $H_1$  and  $H_2$ . Given the location of disease locus  $D$ , the map distance  $\lambda_u$  between  $H_u$  and  $D$  can be used to calculate the recombination fraction  $\theta_u$  by Haldane's map function,  $u = 1, 2, \lambda_1 + \lambda_2 = 4$  cM. To calculate the power, we first partition the interval of 4 cM between block  $H_1$  and  $H_2$  to be 100 subintervals with 101 end-points. Given that the disease locus  $D$  is located at an end-point, we may perform power calculation at this locus. We assume that the haplotype  $H_1$  has two haplotypes  $H_{11}$  and  $H_{12}$  with equal frequencies,  $P_D = 0.10, N = M = 100$ , and  $T = 50$  for the four models in table 3. For the four models in table 4,  $P_D = 0.30, N = M = 500$ . For each genetic model in table 4, figures 2, 3, and 4 show power curves of  $T_C, T_{H1}, T_G, T_{C2}, T_{H2}$ , and  $T_{G2}$  for  $r = 2, 3, 4$  haplotypes of block  $H_2$ , respectively. The related parameters, such as measures of LD between block  $H_1$  and block  $H_2$ , are given in the legend of each figure. First, it is clear from these three figures that the power of using two haplotype blocks is generally higher than that of using one block. When the disease locus  $D$  is far from block  $H_2$ , the power of using two haplotype blocks is significantly higher. When the disease locus  $D$  is close to block  $H_2$ , the power of using two haplotype blocks is similar to that of using only one block  $H_2$ . Second, the power of  $T_C$  is generally higher than or similar to that of  $T_{H1}$ , and the power of

**Table 4**

**Second Set of Parameters of Simulated Genetic Models**

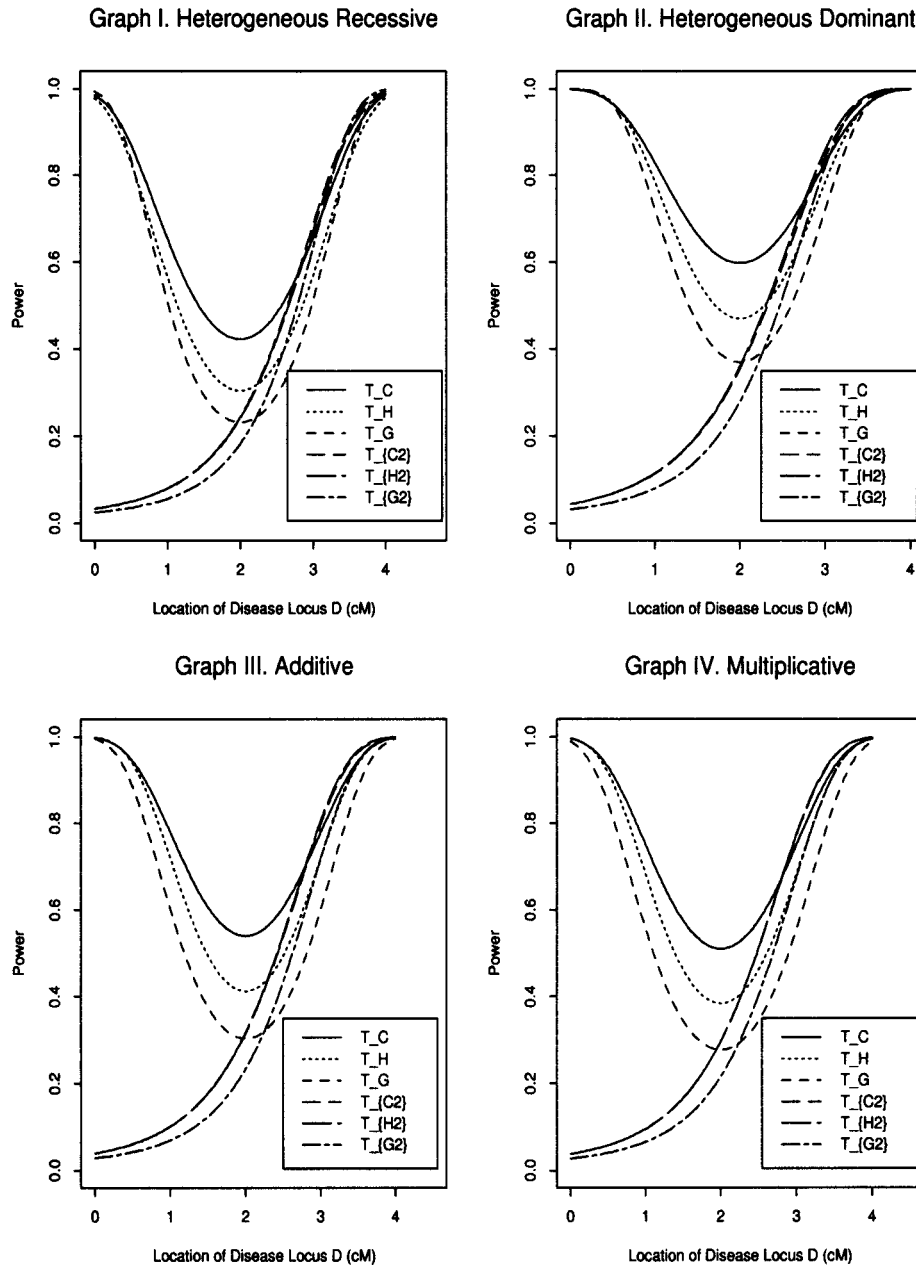
Model Type	$f_{DD}$	$f_{Dd}$	$f_{dd}$
Heterogeneous recessive <sup>a</sup>	.16	.04	.04
Heterogeneous dominant <sup>b</sup>	.08	.08	.02
Additive <sup>c</sup>	.108	.0675	.028
Multiplicative <sup>d</sup>	.12	.06	.03

<sup>a</sup>  $f_{DD} = 4f_{Dd} = 4f_{dd}$ .

<sup>b</sup>  $f_{DD} = f_{Dd} = 4f_{dd}$ .

<sup>c</sup>  $f_{DD} = 4f_{dd}, f_{Dd} = (f_{DD} + f_{dd})/2$ .

<sup>d</sup>  $f_{DD} = 4f_{dd}, f_{Dd} = 2f_{dd}$ .

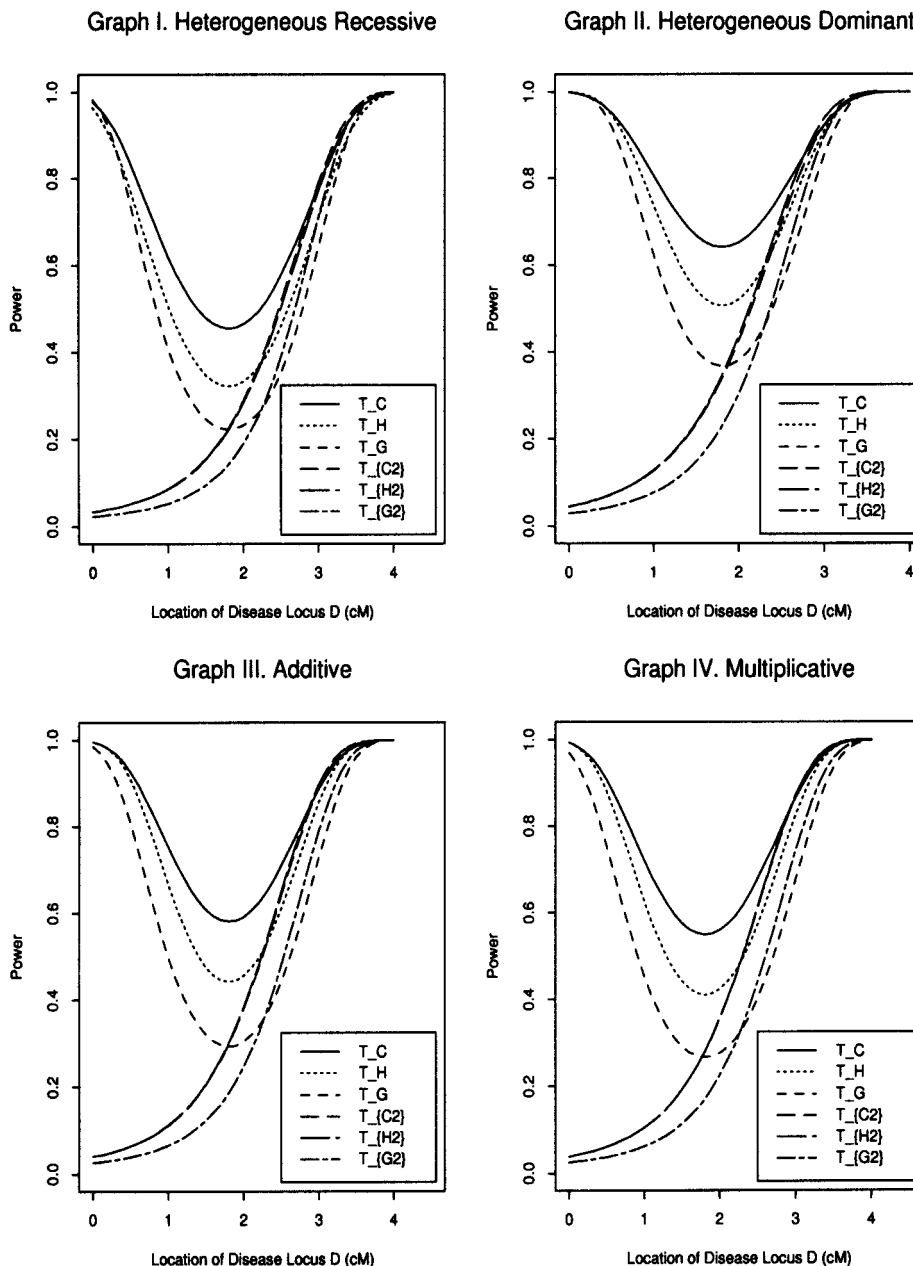


**Figure 2** Power curves of  $T_C$ ,  $T_H$ ,  $T_G$ ,  $T_{C2}$ ,  $T_{H2}$ , and  $T_{G2}$  at significance level  $\alpha = 0.01$ , using two haplotype blocks  $H_1$ ,  $l = 2$ , and  $H_2$ ,  $r = 2$ , when  $P(H_{11}) = P(H_{12}) = P(H_{21}) = P(H_{22}) = 0.50$ ,  $\Delta_{H_{11}H_{21}} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.075$ ,  $\Delta_{H_{11}H_{22}} = P(H_{11}H_{22}) - P(H_{11})P(H_{22}) = -0.075$ ,  $P_D = 0.30$ ,  $N = M = 500$ ,  $T = 50$ , for the four genetic models in table 4.

$T_H$  is higher than or similar to that of  $T_C$ . This may be due to the lack of consideration of correlation between the two blocks by  $T_C$  (see the type I error comparison in table 2). Third, the power of  $T_{C2}$  is similar to that of  $T_{H2}$  and higher than that of  $T_{G2}$ .

To explore the effect of the degree of LD on the test statistics, figure 5 plots power curves under an assumption of linkage equilibrium between the two blocks  $H_1$  and  $H_2$  for four models in table 4. From the four graphs

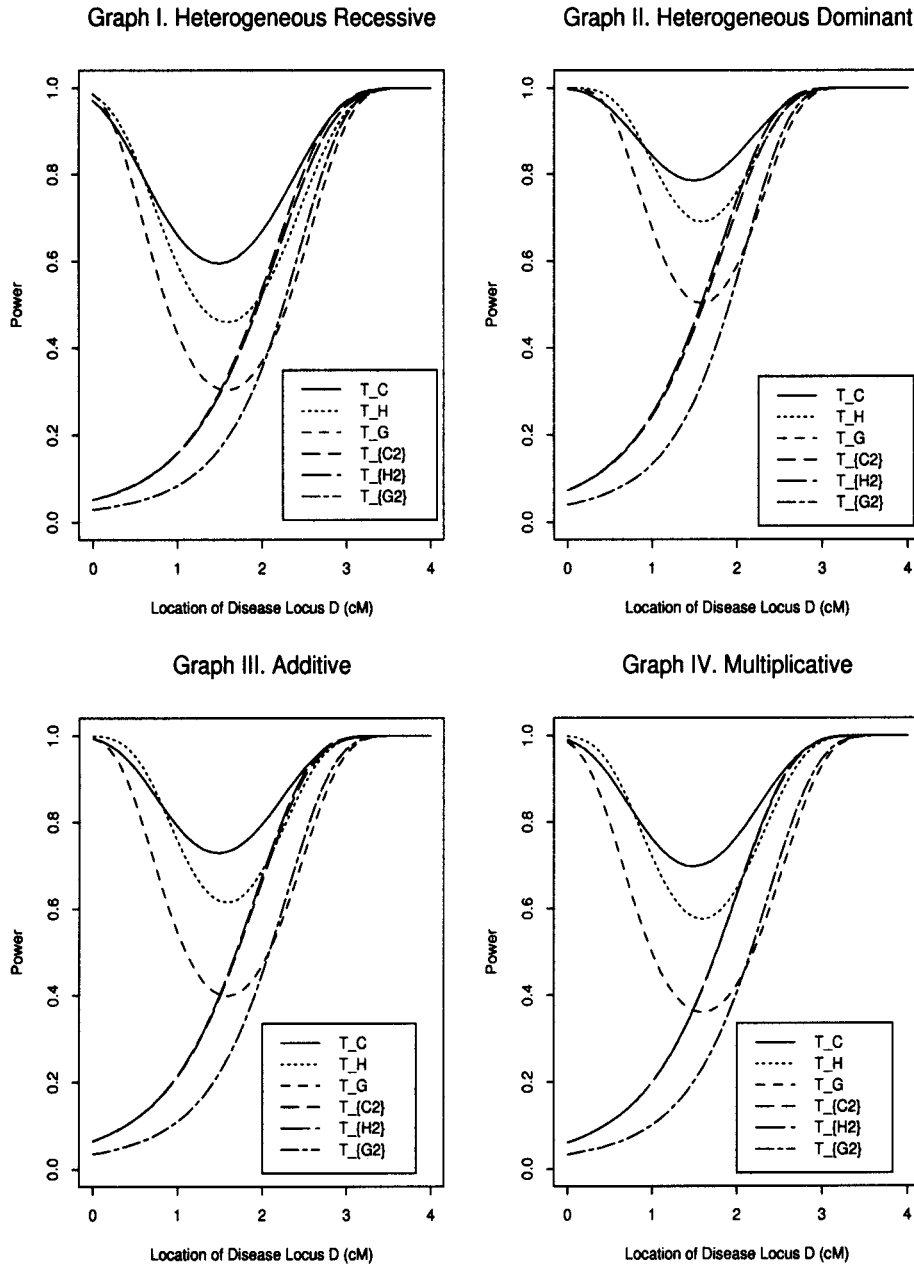
of figure 5, the power of  $T_H$  is similar to or slightly higher than that of  $T_C$ , except for heterogeneous recessive and multiplicative models, in which the power of  $T_H$  is slightly lower than that of  $T_C$ . In all graphs of figure 5, the power of  $T_C$  and  $T_H$  is higher than that of  $T_G$ . Figure 6 plots power curves for different mutation ages of the disease allele  $D$  for four models in table 4. For the four models in table 4, the power is very high for a disease mutation of  $T = 30$ , high for  $T = 40$ , and relatively



**Figure 3** Power curves of  $T_C$ ,  $T_H$ ,  $T_G$ ,  $T_{C2}$ ,  $T_{H2}$ , and  $T_{G2}$  at significance level  $\alpha = 0.01$ , using two haplotype blocks  $H_1$ ,  $l = 2$ , and  $H_2$ ,  $r = 3$ , when  $P(H_{11}) = P(H_{12}) = 0.5$ ,  $P(H_{21}) = 0.4$ ,  $P(H_{22}) = P(H_{23}) = 0.30$ ,  $\Delta_{H_{11}H_{21}} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.075$ ,  $\Delta_{H_{11}H_{22}} = P(H_{11}H_{22}) - P(H_{11})P(H_{22}) = -0.0375$ ,  $\Delta_{H_{11}H_{23}} = P(H_{11}H_{23}) - P(H_{11})P(H_{23}) = -0.0375$ ,  $P_D = 0.30$ ,  $N = M = 500$ ,  $T = 50$ , for the four genetic models in table 4.

high for  $T = 50$  generations old. Figure 7 plots power curves of  $T_H$  for different disease frequencies  $P_D$  for the four models in table 4. For recessive disease model in table 4, a disease with frequency  $P_D \geq 0.30$  would have high power if the haplotype block is close to the disease locus. For the other three models in table 4, a disease with frequency  $P_D \geq 0.20$  would have high power if the haplotype block is close to the disease locus (fig. 7).

Corresponding to the six figures for the less optimistic models in table 4, we provide six figures for the optimistic models in table 3 on our Web site. The power of the heterogeneous recessive model in table 3 is low (figs. 1, 2, and 3 on our Web site). In contrast, the power of the heterogeneous recessive model in table 4 is reasonably high (figs. 2, 3, and 4). In the absence of LD, the power of  $T_H$  is similar to or slightly higher than that of



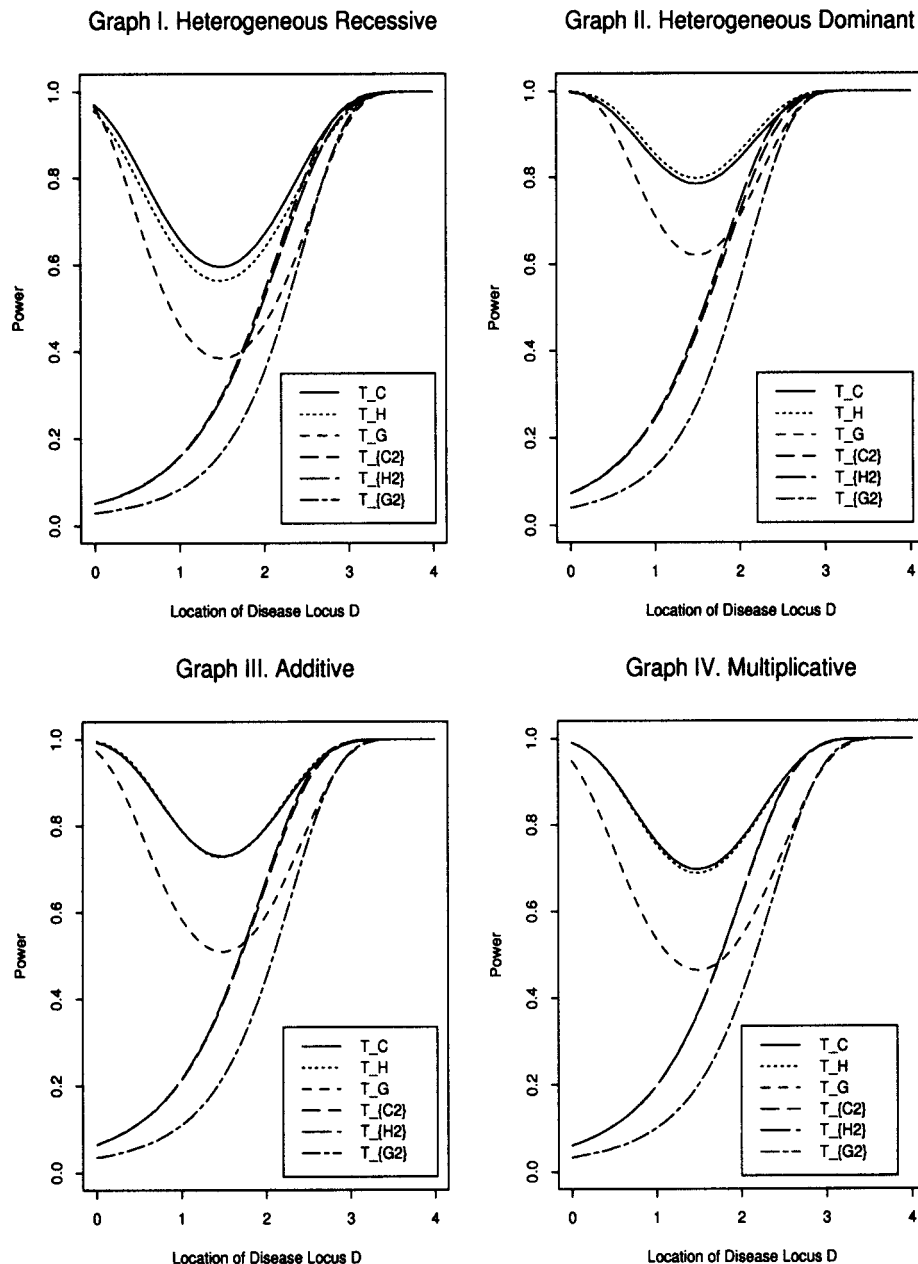
**Figure 4** Power curves of  $T_C$ ,  $T_H$ ,  $T_G$ ,  $T_{C2}$ ,  $T_{H2}$ , and  $T_{G2}$  at significance level  $\alpha = 0.01$  using two haplotype blocks  $H_1$ ,  $l = 2$ , and  $H_2$ ,  $r = 4$ , when  $P(H_{11}) = P(H_{12}) = 0.5, P(H_{21}) = P(H_{22}) = P(H_{23}) = P(H_{24}) = 0.25$ ,  $\Delta_{H_{11}H_{21}} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.075$ ,  $\Delta_{H_{11}H_{22}} = P(H_{11}H_{22}) - P(H_{11})P(H_{22}) = 0.075$ ,  $\Delta_{H_{11}H_{23}} = P(H_{11}H_{23}) - P(H_{11})P(H_{23}) = -0.075$ ,  $\Delta_{H_{11}H_{24}} = P(H_{11}H_{24}) - P(H_{11})P(H_{24}) = -0.075$ ,  $P_D = 0.30$ ,  $N = M = 500$ ,  $T = 50$ , for the four genetic models in table 4.

$T_C$  for the four models in table 3 (fig. 4 on our Web site). For recessive disease model in table 3, the power is low even for very young disease mutation ( $T = 10$ ) (fig. 5 on our Web site). For the recessive disease model in table 3, a disease with frequency  $P_D \geq 0.15$  would have high power if the haplotype block is close to the disease locus. For the other three models in table 3, a

disease with frequency  $P_D \geq 0.10$  would have high power (fig. 6 on our Web site).

*Sample Size*

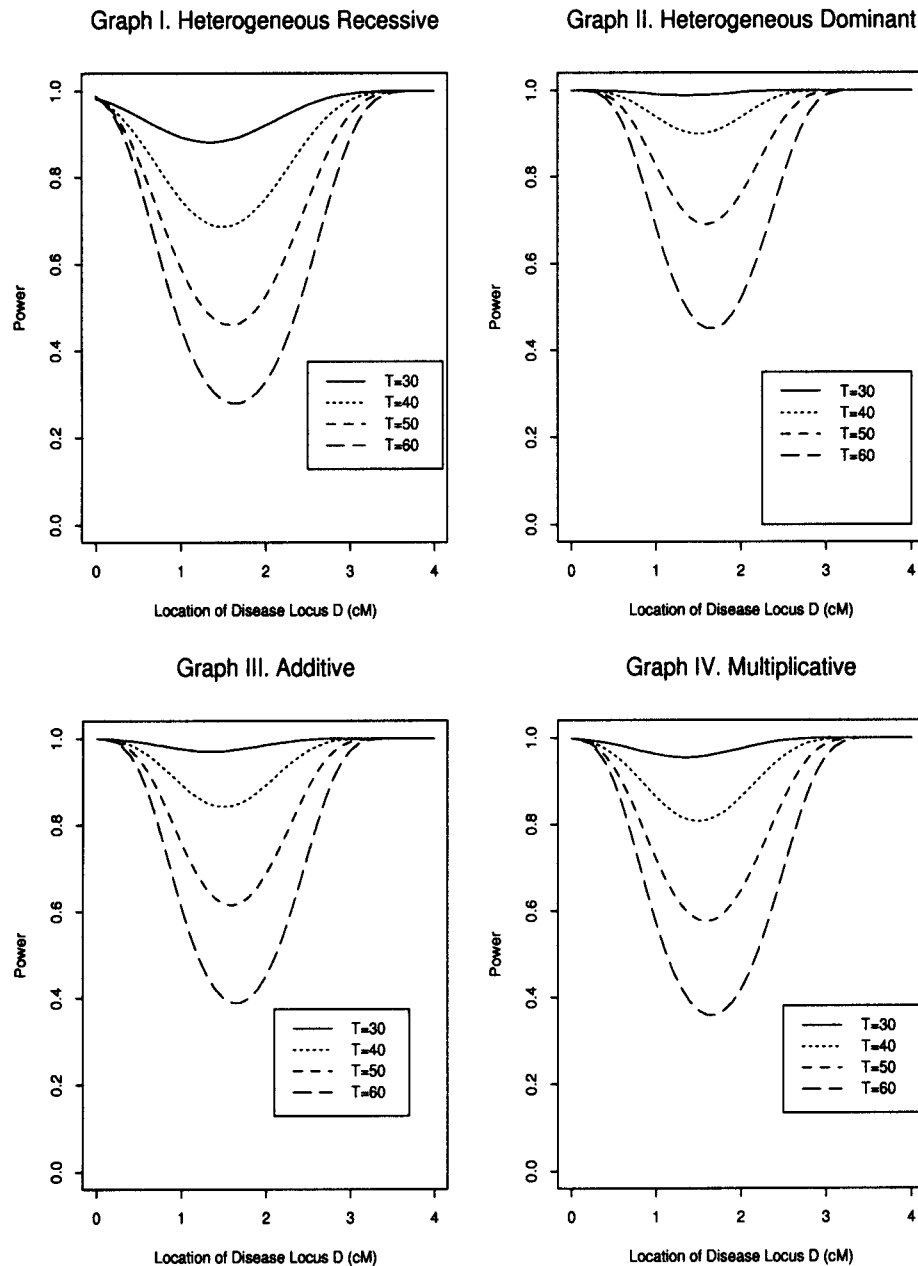
Table 5 gives sample size required for the four genetic models in table 3 at significance level .01 and 80%



**Figure 5** Power curves of  $T_C$ ,  $T_H$ , and  $T_G$  at significance level  $\alpha = 0.01$  using two haplotype blocks  $H_1$ ,  $l = 2$ , and  $H_2$ ,  $r = 4$ , when  $P(H_{11}) = P(H_{12}) = 0.5, P(H_{21}) = P(H_{22}) = P(H_{23}) = P(H_{24}) = 0.25$ ,  $\Delta_{H_{11}H_{21}} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.0$ ,  $\Delta_{H_{11}H_{22}} = P(H_{11}H_{22}) - P(H_{11})P(H_{22}) = 0.0$ ,  $\Delta_{H_{11}H_{23}} = P(H_{11}H_{23}) - P(H_{11})P(H_{23}) = 0.0$ ,  $\Delta_{H_{11}H_{24}} = P(H_{11}H_{24}) - P(H_{11})P(H_{24}) = 0.0$ ,  $P_D = 0.30$ ,  $N = M = 500$ ,  $T = 50$ , for the four genetic models in table 4.

power using two haplotype blocks  $H_1$ ,  $l = 2$ , and  $H_2$ ,  $r = 4$ . Except for heterozygous recessive disease with low disease-allele frequency  $P_D = 0.05$ , the sample sizes required are  $<400$  and are feasible in practice. For most cases, the sample sizes required are  $<100$ . Table 6 gives the sample sizes required for the four genetic models in table 4 at significance level 0.01 and 80% power, using

two haplotype blocks  $H_1$ ,  $l = 2$ , and  $H_2$ ,  $r = 4$ . Compared with the sample sizes in table 5 for the four models in table 3, the sample sizes in table 6 for the four models in table 4 are much greater. For the recessive disease model in table 4, the sample sizes required for low frequency ( $P_D \leq 0.10$ ) are  $>5,000$ , and so it may not be realistic to recruit enough patients for such disease stud-



**Figure 6** Power curves of  $T_H$  for different mutation ages at significance level  $\alpha = 0.01$ , using two haplotype blocks  $H_1$ ,  $l = 2$ , and  $H_2$ ,  $r = 4$ , when  $P(H_{11}) = P(H_{12}) = 0.5$ ,  $P(H_{21}) = P(H_{22}) = P(H_{23}) = P(H_{24}) = 0.25$ ,  $\Delta_{H_{11}H_{21}} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.075$ ,  $\Delta_{H_{11}H_{22}} = P(H_{11}H_{22}) - P(H_{11})P(H_{22}) = 0.075$ ,  $\Delta_{H_{11}H_{23}} = P(H_{11}H_{23}) - P(H_{11})P(H_{23}) = -0.075$ ,  $\Delta_{H_{11}H_{24}} = P(H_{11}H_{24}) - P(H_{11})P(H_{24}) = -0.075$ ,  $P_D = 0.30$ ,  $N = M = 500$ , for the four genetic models in table 4.

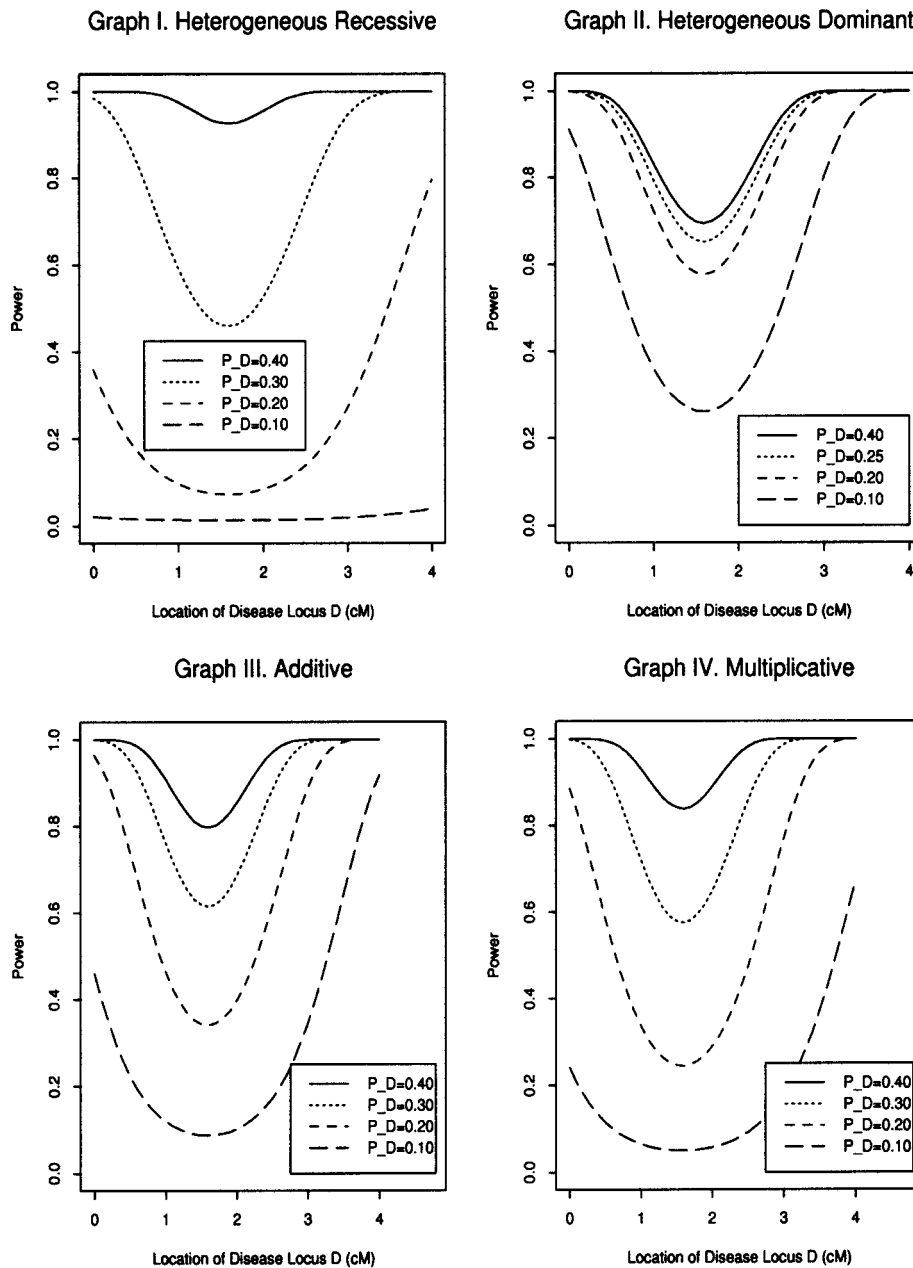
ies. For all dominant disease models and recessive disease models with high disease frequency ( $P_D = 0.20$  or  $0.30$ ), the sample sizes required are  $<1,000$  and are feasible in practice. For the additive and multiplicative disease models in table 4, the sample sizes required are  $<1,000$ , except for low-disease-frequency cases ( $P_D = 0.05$ ) or old disease mutations ( $T \geq 50$ ).

For the sample sizes given in tables 5 and 6, we per-

form an empirical power calculation by 10,000 replicates. The results for  $T_H$  are pretty consistent with the theoretical value of 0.80.

## Discussion

The objective of this paper is to explore methods for high-resolution haplotype or multiple-marker genome-



**Figure 7** Power curves of  $T_H$  for different disease frequency at significance level  $\alpha = 0.01$ , using two haplotype blocks  $H_{13}$ ,  $l = 2$ , and  $H_{23}$ ,  $r = 4$ , when  $P(H_{11}) = P(H_{12}) = 0.5$ ,  $P(H_{21}) = P(H_{22}) = P(H_{23}) = P(H_{24}) = 0.25$ ,  $\Delta_{H_{11}H_{21}} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.075$ ,  $\Delta_{H_{11}H_{22}} = P(H_{11}H_{22}) - P(H_{11})P(H_{22}) = 0.075$ ,  $\Delta_{H_{11}H_{23}} = P(H_{11}H_{23}) - P(H_{11})P(H_{23}) = -0.075$ ,  $\Delta_{H_{11}H_{24}} = P(H_{11}H_{24}) - P(H_{11})P(H_{24}) = -0.075$ ,  $T = 50$ ,  $N = M = 500$ , for the four genetic models in table 4.

association studies of complex diseases by case-control designs. We investigated test statistics that combine information from haplotype blocks or multiple markers. We introduced two Hotelling's  $T^2$  statistics  $T_G$  and  $T_H$  to test association between a disease locus and two haplotype blocks on the basis of two coding methods, genotype coding and haplotype coding. By theoretical analysis, we showed that they are valid test statistics.

Ignoring the correlation between the two blocks, one may use an extension sum statistic,  $T_C$ , of two traditional  $\chi^2$  test statistics,  $T_{C1}$  and  $T_{C2}$ , for comparing haplotype frequencies in cases and controls. For each of the three statistics, the power of using two haplotype blocks is higher than that of using only one haplotype block. By power comparison, we notice that  $T_C$  has higher power than  $T_H$ , and  $T_H$  has higher power than  $T_G$ .

**Table 5**

**Sample Sizes Required for the Four Genetic Models in Table 3, at Significance Level 0.01 and 80% Power, Using Two Haplotype Blocks  $H_1$ ,  $l = 2$ , and  $H_2$ ,  $r = 4$**

MODEL TYPE AND $P_D$	REQUIRED SAMPLE SIZE														
	$T = 20$			$T = 30$			$T = 40$			$T = 50$			$T = 60$		
	$T_C$	$T_H$	$T_G$	$T_C$	$T_H$	$T_G$	$T_C$	$T_H$	$T_G$	$T_C$	$T_H$	$T_G$	$T_C$	$T_H$	$T_G$
Heterogeneous recessive:															
.05	2,296	3,005	2,145	2,534	3,297	2,455	2,798	3,619	2,806	3,088	3,976	3,202	3,409	4,369	3,650
.10	193	279	241	213	303	271	235	329	303	258	358	339	285	390	379
.15	57	85	82	62	92	92	68	99	102	75	107	113	83	116	125
Heterogeneous dominant:															
.05	45	47	60	50	53	67	55	59	75	60	65	84	66	73	94
.10	26	24	29	29	27	33	32	30	38	35	34	43	39	38	48
.15	19	15	17	21	17	21	23	20	24	26	23	28	28	26	32
Additive:															
.05	22	18	21	25	21	24	27	24	28	30	27	32	33	31	37
.10	20	16	18	22	18	21	24	21	24	27	24	28	29	27	32
.15	18	13	15	20	16	18	22	18	21	24	20	24	26	23	28
Multiplicative:															
.05	29	34	47	32	38	51	35	41	56	38	46	62	42	50	68
.10	15	17	23	17	19	25	19	21	28	21	23	31	23	25	34
.15	12	11	15	13	13	17	14	14	19	15	16	22	17	18	24

NOTE.—Data shown are for  $P(H_{11}) = P(H_{12}) = 0.5$ ,  $P(H_{21}) = P(H_{22}) = P(H_{23}) = P(H_{24}) = 0.25$ ,  $\Delta_{H_{11}H_{21}} = P(H_{11}H_{21}) - P(H_{11})P(H_{21}) = 0.075$ ,  $\Delta_{H_{11}H_{22}} = P(H_{11}H_{22}) - P(H_{11})P(H_{22}) = 0.075$ ,  $\Delta_{H_{11}H_{23}} = P(H_{11}H_{23}) - P(H_{11})P(H_{23}) = -0.075$ ,  $\Delta_{H_{11}H_{24}} = P(H_{11}H_{24}) - P(H_{11})P(H_{24}) = -0.075$ , and  $\theta_1 = \theta_2 = 0.005$ .

In the absence of LD between the two blocks, the power of  $T_C$  is similar to that of  $T_H$  and is higher than that of  $T_G$ . In the presence of LD between the two blocks, the type I error of  $T_C$  is higher than those of  $T_H$  and  $T_G$ . Hence, we advocate to use  $T_H$  in the data analysis. In the presence of LD between the two blocks,  $T_H$  takes into account of the correlation between the two haplotype blocks and has the lowest type I error and a higher power than  $T_G$ . On the one hand,  $T_G$  has the lowest power, although it takes into account the correlation between the two haplotype blocks. On the other hand, the type I error of  $T_G$  gets bigger as the number of haplotypes increases, which may be due to the large degree of freedom. Therefore,  $T_G$  is less favorable than  $T_H$ .

Several empirical studies showed that the haplotypes have block structures in human genome, and each haplotype block has limited diversity (Daly et al. 2001; Goldstein 2001; Patil et al. 2001; Reich et al. 2001; Rioux et al. 2001; Stephens et al. 2001; Gabriel et al. 2002). The haplotype blocks are punctuated by apparent sites of recombination or hot-spot areas. Within a haplotype block, there are only a few (2–4) haplotypes, and LD decays only gradually with distance. Within the hot-spot areas, however, there may have been several recombination events, and thus LD decays rapidly with distance. The recombination events are clustered to be hot spots. These patterns of LD are very relevant to

genomewide association studies for mapping complex-disease genes. However, the general properties of haplotype structure in human genome are not fully understood. It is necessary to characterize patterns of LD in the human genome, and to investigate approaches of high resolution LD mapping of complex traits based on haplotype block data.

The test statistics, such as  $T_H$  and  $T_G$ , that are based on multiallelic markers or haplotype blocks can usually lead to a large number of df. However, when haplotype block data are used, the df would not be very large if one took into account the recent discovery of haplotype structure in human genome. Although a haplotype block may enclose many SNPs, it takes only a few SNPs to uniquely identify each of the haplotypes in the block. This implies that the number of df when haplotype block data are used may be even less than that when multiple SNP markers are used in an analysis. Moreover, haplotype block data already take into account the haplotype structure and potentially are more powerful.

In our analysis, only two haplotype blocks are discussed. One could generalize the method to use multiple haplotype blocks in the analysis. One interesting topic is to study the merit of a generalized  $T_H$  that uses multiple haplotype blocks, instead of the current version of  $T_H$ , which uses only two haplotype blocks. Moreover, the methods can be generalized to analyze pedigree data, including sib pairs (Cordell and Clayton 2002). Other

**Table 6**

**Sample Sizes Required for the Four Genetic Models in Table 4 at Significance Level 0.01 and 80% Power Using Two Haplotype Blocks  $H_1$ ,  $l = 2$ , and  $H_2$ ,  $r = 4$**

MODEL TYPE AND $P_D$	REQUIRED SAMPLE SIZE											
	$T = 30$			$T = 40$			$T = 50$			$T = 60$		
	$T_C$	$T_H$	$T_G$	$T_C$	$T_H$	$T_G$	$T_C$	$T_H$	$T_G$	$T_C$	$T_H$	$T_G$
Heterogeneous recessive:												
.05	94,373	117,340	78,996	104,275	129,519	91,520	115,217	142,977	105,797	127,309	157,850	122,031
.10	6233	7965	5707	6883	8763	6555	7601	9644	7517	8394	10618	8607
.20	477	657	550	525	716	620	579	782	698	639	854	785
.30	125	183	173	138	198	192	152	215	214	167	233	238
Heterogeneous dominant:												
.05	469	576	771	517	635	851	570	699	939	629	770	1,036
.10	196	233	302	216	257	335	238	284	370	262	313	410
.20	109	121	145	120	134	162	132	149	181	145	165	202
.30	93	96	104	102	107	117	112	119	133	123	133	150
Additive:												
.05	1,296	1,607	2,188	1,430	1,771	2,412	1,578	1,952	2,658	1,741	2,151	2,931
.10	421	520	703	464	572	774	512	630	853	564	694	940
.20	162	195	260	178	215	287	196	237	317	216	261	350
.30	102	119	157	112	132	174	123	145	192	136	160	212
Multiplicative:												
.05	2,397	2,979	4,070	2,646	3,284	4,487	2,921	3,620	4,946	3,224	3,992	5,454
.10	669	836	1,141	737	919	1,256	813	1,012	1,382	897	1,114	1,522
.20	203	254	347	224	279	382	246	307	419	271	337	460
.30	107	133	182	118	146	200	130	160	219	143	176	241

NOTE.—All other parameters except the penetrance probabilities are the same as those in table 5.

issues, such as population-stratification effects and methods of combining population and pedigree data, are exciting research topics (Ardlie et al. 2000; Rannala and Reeve 2001). If the data contain individuals with missing genotypes within the haplotype blocks or with genotyping errors, some potential problems can arise in actual data analysis. The effect of uncertainty in the haplotype block's start and stop positions is unclear. More investigations will be necessary to cope with these challenges.

### Acknowledgments

We thank two reviewers for very detailed and thoughtful critiques, which made the paper more clear. R.F. was supported partially by a research fellowship from the Alexander von Humboldt Foundation, Germany, and an International Research Travel Assistance Grant from Texas A&M University. M.K. was supported by grant KN 370/1-1 (Project D1 of FOR 423) from the Deutsche Forschungsgemeinschaft.

### Appendix A

Suppose that the disease locus has two alleles  $D$  and  $d$ ,  $D$  being the allele for disease susceptibility and  $d$  being normal. Assume that the disease-susceptibility allele  $D$  has population frequency  $P_D$ , and normal allele  $d$  has population frequency  $P_d$ . 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  $f_{DD} \geq f_{Dd} \geq f_{dd}$ . 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 the population by  $A = f_{DD}P_D^2 + 2f_{Dd}P_DP_d + f_{dd}P_d^2$ , and  $\bar{A} = \bar{f}_{DD}P_D^2 + 2\bar{f}_{Dd}P_DP_d + \bar{f}_{dd}P_d^2 = 1 - A$ . As in quantitative genetics, let us introduce some notation. Let  $a = f_{DD} - (f_{DD} + f_{dd})/2$ ,  $d = f_{Dd} - (f_{DD} + f_{dd})/2$ ,  $\delta_D = 2d$ , and  $\alpha_D = a + (P_d - P_D)d$ . In terms of quantitative genetics,  $\alpha_D$  is the average effect of gene substitution, and  $\delta_D$  is the dominant deviation (Falconer and Mackay 1996). Similarly, denote  $\bar{a} = \bar{f}_{DD} - (\bar{f}_{DD} + \bar{f}_{dd})/2 = -a$ ,  $\bar{d} = \bar{f}_{Dd} - (\bar{f}_{DD} + \bar{f}_{dd})/2 = -d$ ,  $\bar{\delta}_D = 2\bar{d} = -2d$ , and  $\bar{\alpha}_D = \bar{a} + (P_d - P_D)\bar{d} = -\alpha_D$ . Denote the measures of LD between haplotype  $H_{1j}$  of the first haplotype block  $H_1$  and the disease locus  $D$  by  $\Delta_{1j} = P(H_{1j}D) - P(H_{1j})P_D$ ,  $j = 1, \dots, l$ , and the measures of LD between haplotype  $H_{2s}$  of the second haplotype block  $H_2$  and the disease locus  $D$  by  $\Delta_{2s} = P(H_{2s}D) - P(H_{2s})P_D$ ,  $s = 1, \dots, r$ . For  $u = 1, 2$ , the

frequencies of heterozygous genotype  $H_{uj}H_{uk}, j \neq k$ , in affected and unaffected individuals are calculated in appendix B as

$$a_{ujk} = P(H_{uj}H_{uk}|\text{Aff}) = 2\left\{-\Delta_{uj}\Delta_{uk}\delta_D/A + [\Delta_{uj}P(H_{uk}) + \Delta_{uk}P(H_{uj})]\alpha_D/A + P(H_{uk})P(H_{uj})\right\} \quad (\text{A1})$$

$$\bar{a}_{ujk} = P(H_{uj}H_{uk}|\text{Unaff}) = 2\left\{-\Delta_{uj}\Delta_{uk}\bar{\delta}_D/\bar{A} + [\Delta_{uj}P(H_{uk}) + \Delta_{uk}P(H_{uj})]\bar{\alpha}_D/\bar{A} + P(H_{uk})P(H_{uj})\right\}. \quad (\text{A2})$$

The frequencies of homozygous genotype  $H_{uj}H_{uj}$  in affected and unaffected individuals are calculated in appendix B as

$$a_{ujj} = P(H_{uj}H_{uj}|\text{Aff}) = (-\delta_D)\Delta_{uj}^2/A + 2\alpha_DP(H_{uj})\Delta_{uj}/A + P(H_{uj})^2 \quad (\text{A3})$$

$$\bar{a}_{ujj} = P(H_{uj}H_{uj}|\text{Unaff}) = (-\bar{\delta}_D)\Delta_{uj}^2/\bar{A} + 2\bar{\alpha}_DP(H_{uj})\Delta_{uj}/\bar{A} + P(H_{uj})^2. \quad (\text{A4})$$

Under the null hypothesis of no association between the haplotype blocks  $H_u, u = 1, 2$  and the disease locus  $D$ —that is,  $\Delta_{uj} = 0$  for all  $j$ , equations (A1), (A2), (A3) and (A4), imply the expectation  $E(\bar{X} - \bar{Y}) = 0$  for genotype coding method. In appendix C, we show

$$E(z_{ujj}|\text{Aff}) = 2\left[\Delta_{uj}\alpha_D/A + P(H_{uj})\right], \quad E(z_{ujj}|\text{Unaff}) = 2\left[\Delta_{uj}\bar{\alpha}_D/\bar{A} + P(H_{uj})\right]. \quad (\text{A5})$$

Hence, we have  $E(z_{ujj}|\text{Aff}) - E(z_{ujj}|\text{Unaff}) = 2\Delta_{uj}[\alpha_D/A - \bar{\alpha}_D/\bar{A}]$ , which implies the expectation  $E(\bar{X} - \bar{Y}) = 0$  for the haplotype coding method, under the null hypothesis of no association between the haplotype blocks  $H_u, u = 1, 2$  and the disease locus  $D$ .

## Appendix B

Notice that  $P(H_{uj}D) = \Delta_{uj} + P(H_{uj})P_D, P(H_{uj}d) = -\Delta_{uj} + P(H_{uj})P_d, P(H_{uk}D) = \Delta_{uk} + P(H_{uk})P_D,$  and  $P(H_{uk}d) = -\Delta_{uk} + P(H_{uk})P_d$  for  $u = 1, 2$ . Using the expression  $\alpha_D = (f_{DD} - f_{dd})/2 + (P_d - P_D)[f_{Dd} - (f_{DD} + f_{dd})/2] = P_D f_{DD} + P_d f_{Dd} - P_D f_{Dd} - P_d f_{dd}$ , the frequency of genotype  $H_{uj}H_{uk}, j \neq k$ , in affected can be calculated as

$$\begin{aligned} P(H_{uj}H_{uk}|\text{Aff}) &= 2\left\{P(H_{uj}D)P(H_{uk}D)f_{DD} + [P(H_{uj}D)P(H_{uk}d) + P(H_{uj}d)P(H_{uk}D)]f_{Dd} + P(H_{uj}d)P(H_{uk}d)f_{dd}\right\}/A \\ &= 2\Delta_{uj}\Delta_{uk}\frac{f_{DD} - 2f_{Dd} + f_{dd}}{A} + 2P(H_{uk})P(H_{uj}) + 2[\Delta_{uj}P(H_{uk}) + \Delta_{uk}P(H_{uj})]\frac{P_D f_{DD} + P_d f_{Dd} - P_D f_{Dd} - P_d f_{dd}}{A} \\ &= 2\Delta_{uj}\Delta_{uk}\frac{-\delta_D}{A} + 2[\Delta_{uj}P(H_{uk}) + \Delta_{uk}P(H_{uj})]\frac{\alpha_D}{A} + 2P(H_{uk})P(H_{uj}). \end{aligned}$$

Similarly, the frequency of genotype  $H_{uj}H_{uj}$  in affected can be calculated as

$$\begin{aligned} P(H_{uj}H_{uj}|\text{Aff}) &= [P(H_{uj}D)^2 f_{DD} + 2P(H_{uj}D)P(H_{uj}d)f_{Dd} + P(H_{uj}d)^2 f_{dd}]/A \\ &= (-\delta_D)\Delta_{uj}^2/A + 2\alpha_DP(H_{uj})\Delta_{uj}/A + P(H_{uj})^2. \end{aligned}$$

Similarly, we may prove equations (A2) and (A4).

### Appendix C

Notice that  $\sum_k \Delta_{uk} = 0$  and  $\sum_k P(H_{uk}) = 1$ . From equations (A1) and (A3), the expectation of numbers of haplotypes  $H_{uj}$  in affected is equal to

$$\begin{aligned} E(z_{uj}|\text{Aff}) &= 2P(H_{uj}H_{uj}|\text{Aff}) + \sum_{k \neq j} P(H_{uj}H_{uk}|\text{Aff}) \\ &= 2\left[(-\delta_D)\Delta_{uj}^2/A + 2\alpha_D P(H_{uj})\Delta_{uj}/A + P(H_{uj})^2\right] \\ &\quad + 2\sum_{k \neq j} \left[\Delta_{uj}\Delta_{uk} \frac{-\delta_D}{A} + [\Delta_{uj}P(H_{uk}) + \Delta_{uk}P(H_{uj})] \frac{\alpha_D}{A} + P(H_{uk})P(H_{uj})\right], \\ &= 2\left[\Delta_{uj}\alpha_D/A + P(H_{uj})\right]. \end{aligned}$$

Similarly, one may show that the expectation of numbers of haplotypes  $H_{uj}$  in unaffected is equal to  $E(z_{uj}|\text{Unaff}) = 2[\Delta_{uj}\bar{\alpha}_D/\bar{A} + P(H_{uj})]$ .

### Appendix D

Using the notations of  $a_{ujk}, \bar{a}_{ujk}, j \neq k, a_{uij}, \bar{a}_{uij}$  in equations (A1), (A2), (A3), and (A4), we calculate the variance-covariance matrices  $\Sigma_{A1}$  and  $\Sigma_{\bar{A}1}$ . First, we calculate the variance of the number of haplotypes  $H_{uj}$  in affected by equations (A3) and (A5)

$$\begin{aligned} \text{Var}(z_{uj}|\text{Aff}) &= E(z_{uj}^2|\text{Aff}) - [E(z_{uj}|\text{Aff})]^2 \\ &= 2a_{uij} + (2a_{uij} + \sum_{k \neq j} a_{ujk}) - (2a_{uij} + \sum_{k \neq j} a_{ujk})^2 \\ &= 2\left[-\delta_D \Delta_{uj}^2/A + 2\alpha_D P(H_{uj})\Delta_{uj}/A + P(H_{uj})^2\right] \\ &\quad + 2\left[\Delta_{uj}\alpha_D/A + P(H_{uj})\right] - 4\left[\Delta_{uj}\alpha_D/A + P(H_{uj})\right]^2 \\ &= 2\Delta_{uj}^2[-\delta_D/A - 2\alpha_D^2/A^2] \\ &\quad + 2\Delta_{uj}[1 - 2P(H_{uj})]\alpha_D/A + 2P(H_{uj})[1 - P(H_{uj})]. \end{aligned} \tag{D1}$$

Similarly, the variance of the number of haplotypes  $H_{uj}$  in controls is

$$\text{Var}(z_{uj}|\text{Unaff}) = 2\Delta_{uj}^2[-\bar{\delta}_D/\bar{A} - 2\bar{\alpha}_D^2/\bar{A}^2] + 2\Delta_{uj}[1 - 2P(H_{uj})]\bar{\alpha}_D/\bar{A} + 2P(H_{uj})[1 - P(H_{uj})]. \tag{D2}$$

By use of equations (A1) and (A5), the covariance between the number of haplotypes  $H_{uj}$  and the number of haplotypes  $H_{uk}, j \neq k$  in affected individuals is

$$\begin{aligned} \text{Cov}(z_{uij}, z_{uik}|\text{Aff}) &= E(z_{uij}z_{uik}|\text{Aff}) - E(z_{uij}|\text{Aff})E(z_{uik}|\text{Aff}) \\ &= P(H_{uj}H_{uk}|\text{Aff}) - E(z_{uij}|\text{Aff})E(z_{uik}|\text{Aff}) \\ &= 2\left[-\Delta_{uj}\Delta_{uk} \frac{\delta_D}{A} + [P(H_{uj})\Delta_{uk} + P(H_{uk})\Delta_{uj}] \frac{\alpha_D}{A} + P(H_{uj})P(H_{uk})\right] \\ &\quad - 4\left[\Delta_{uj}\alpha_D/A + P(H_{uj})\right]\left[\Delta_{uk}\alpha_D/A + P(H_{uk})\right] \\ &= 2\Delta_{uj}\Delta_{uk}[-\delta_D/A - 2\alpha_D^2/A^2] - 2[P(H_{uj})\Delta_{uk} + P(H_{uk})\Delta_{uj}]\alpha_D/A \\ &\quad - 2P(H_{uj})P(H_{uk}). \end{aligned}$$

Similarly, the covariance between the number of haplotypes  $H_{uj}$  and the number of haplotypes  $H_{uk}$ ,  $j \neq k$  in controls is

$$\text{Cov}(z_{uj}, z_{uk} | \text{Unaff}) = 2\Delta_{uj}\Delta_{uk}[-\bar{\delta}_D/\bar{A} - 2\bar{\alpha}_D^2/\bar{A}^2] - 2[P(H_{uj})\Delta_{uk} + P(H_{uk})\Delta_{uj}]\bar{\alpha}_D/\bar{A} - 2P(H_{uj})P(H_{uk}) .$$

## Appendix E

To calculate the covariance between  $z_{1ij}, z_{2is}$ , denote for  $j \neq k, s \neq t$

$$\begin{aligned} g_{jiss} &= E[1_{(H_{1j}H_{1j})}1_{(H_{2s}H_{2s})} | \text{Aff}] \\ &= \left[ P(H_{1j}DH_{2s})^2 f_{DD} + 2P(H_{1j}DH_{2s})P(H_{1j}dH_{2s})f_{Dd} + P(H_{1j}dH_{2s})^2 f_{dd} \right] / A , \\ g_{jist} &= E[1_{(H_{1j}H_{1j})}1_{(H_{2s}H_{2t})} | \text{Aff}] \\ &= 2 \left[ P(H_{1j}DH_{2s})P(H_{1j}DH_{2t})f_{DD} + [P(H_{1j}DH_{2s})P(H_{1j}dH_{2t}) \right. \\ &\quad \left. + P(H_{1j}dH_{2s})P(H_{1j}DH_{2t})]f_{Dd} + P(H_{1j}dH_{2s})P(H_{1j}dH_{2t})f_{dd} \right] / A , \\ g_{jkss} &= E[1_{(H_{1j}H_{1k})}1_{(H_{2s}H_{2s})} | \text{Aff}] \\ &= 2 \left[ P(H_{1j}DH_{2s})P(H_{1k}DH_{2s})f_{DD} + [P(H_{1j}DH_{2s})P(H_{1k}dH_{2s}) \right. \\ &\quad \left. + P(H_{1j}dH_{2s})P(H_{1k}DH_{2s})]f_{Dd} + P(H_{1j}dH_{2s})P(H_{1k}dH_{2s})f_{dd} \right] / A , \text{ and} \\ g_{jkst} &= E[1_{(H_{1j}H_{1k})}1_{(H_{2s}H_{2t})} | \text{Aff}] \\ &= 2 \left[ [P(H_{1j}DH_{2s})P(H_{1k}DH_{2t}) + P(H_{1j}DH_{2t})P(H_{1k}DH_{2s})]f_{DD} \right. \\ &\quad \left. + [P(H_{1j}DH_{2s})P(H_{1k}dH_{2t}) + P(H_{1j}dH_{2s})P(H_{1k}DH_{2t})]f_{Dd} \right. \\ &\quad \left. + [P(H_{1j}DH_{2t})P(H_{1k}dH_{2s}) + P(H_{1j}dH_{2t})P(H_{1k}DH_{2s})]f_{Dd} \right. \\ &\quad \left. + [P(H_{1j}dH_{2s})P(H_{1k}dH_{2t}) + P(H_{1j}dH_{2t})P(H_{1k}dH_{2s})]f_{dd} \right] / A . \end{aligned} \tag{E1}$$

For  $j = 1, \dots, l-1$  and  $s = 1, \dots, r-1$ , the covariance

$$\begin{aligned} \text{Cov}(z_{1ij}, z_{2is} | \text{Aff}) &= E(z_{1ij}z_{2is} | \text{Aff}) - E(z_{1ij} | \text{Aff})E(z_{2is} | \text{Aff}) \\ &= 4g_{jiss} + 2 \sum_{t \neq s} g_{jist} + 2 \sum_{k \neq j} g_{jkss} + \sum_{k \neq j} \sum_{t \neq s} g_{jkst} \\ &\quad - 4[\Delta_{1j}\alpha_D/A + P(H_{1j})][\Delta_{2s}\alpha_D/A + P(H_{2s})] . \end{aligned}$$

Similarly, for  $j = 1, \dots, l-1$  and  $s = 1, \dots, r-1$ , the covariance

$$\begin{aligned} \text{Cov}(z_{1ij}, z_{2is} | \text{Unaff}) &= E(z_{1ij}z_{2is} | \text{Unaff}) - E(z_{1ij} | \text{Unaff})E(z_{2is} | \text{Unaff}) \\ &= 4\bar{g}_{jiss} + 2 \sum_{t \neq s} \bar{g}_{jist} + 2 \sum_{k \neq j} \bar{g}_{jkss} + \sum_{k \neq j} \sum_{t \neq s} \bar{g}_{jkst} - 4[\Delta_{1j}\bar{\alpha}_D/\bar{A} + P(H_{1j})][\Delta_{2s}\bar{\alpha}_D/\bar{A} + P(H_{2s})] , \end{aligned}$$

where  $\bar{g}_{jiss}, \bar{g}_{jist}, \bar{g}_{jkss}$  and  $\bar{g}_{jkst}$  are expected genotype frequencies in controls like those defined in equation (E1) for cases.

## Appendix F

To calculate the noncentrality parameter  $\lambda_G$ , we notice first that the expectation

$$E(\bar{X} - \bar{Y}) = \begin{pmatrix} E\bar{X}_1 - E\bar{Y}_1 \\ E\bar{X}_2 - E\bar{Y}_2 \end{pmatrix},$$

where  $E(\bar{X}_1 - \bar{Y}_1)$  is equal to  $[a_{111} - \bar{a}_{111}, \dots, a_{1(l-1)(l-1)} - \bar{a}_{1(l-1)(l-1)}, a_{112} - \bar{a}_{112}, \dots, a_{11l} - \bar{a}_{11l}, \dots, a_{1(l-1)l} - \bar{a}_{1(l-1)l}]^T$ , and  $E(\bar{X}_2 - \bar{Y}_2)$  is equal to  $(a_{211} - \bar{a}_{211}, \dots, a_{2(r-1)(r-1)} - \bar{a}_{2(r-1)(r-1)}, a_{212} - \bar{a}_{212}, \dots, a_{21r} - \bar{a}_{21r}, \dots, a_{2(r-1)r} - \bar{a}_{2(r-1)r})^T$ .

Let  $\Sigma_G$  be the variance-covariance matrix of genotype coding  $X_i$ . Then its elements can be calculated by  $\text{Var}(x_{uij}|\text{Aff}) = a_{uij} - a_{uij}^2$ , where  $j = 1, \dots, l - 1$  if  $u = 1$  and  $j = 1, \dots, r - 1$  if  $u = 2$ ,  $\text{Var}(x_{uijk}|\text{Aff}) = a_{uijk} - a_{uijk}^2$ ,  $j \neq k$ ,  $\text{Cov}(x_{uij}, x_{ui(j+k)}|\text{Aff}) = -a_{uij}a_{ui(j+k)}$  if  $k \geq 1$ ,  $\text{Cov}(x_{uij}, x_{uimk}|\text{Aff}) = -a_{uij}a_{uimk}$  for  $m \neq k$ ,  $\text{Cov}(x_{uijk}, x_{uist}|\text{Aff}) = -a_{uijk}a_{uist}$  for  $j \neq k$  and  $s \neq t$ .

Using the notation in equations (A1), (A3), and (E1), the covariances between  $x_{1ij}, x_{1ijk}$  and  $x_{2is}, x_{2ist}$  are given by

$$\begin{aligned} \text{Cov}(x_{1ij}, x_{2is}|\text{Aff}) &= g_{jiss} - a_{1ij}a_{2ss}, \\ \text{Cov}(x_{1ij}, x_{2ist}|\text{Aff}) &= g_{jist} - a_{1ij}a_{2st}, \\ \text{Cov}(x_{1ijk}, x_{2is}|\text{Aff}) &= g_{jkss} - a_{1jk}a_{2ss}, \text{ and} \\ \text{Cov}(x_{1ijk}, x_{2ist}|\text{Aff}) &= g_{jkst} - a_{1jk}a_{2st}. \end{aligned}$$

Similarly, we may calculate the variance-covariance matrix  $\Sigma_{\bar{G}}$  for the controls. Then the noncentrality parameter  $\lambda_G$  of  $T_G$  is given by

$$\lambda_G = \frac{NM}{N + M} (E\bar{X} - E\bar{Y})^T \left[ \frac{(N - 1)\Sigma_G + (M - 1)\Sigma_{\bar{G}}}{N + M - 2} \right]^{-1} (E\bar{X} - E\bar{Y}).$$

Using the variance-covariance matrices  $\Sigma_{G1}$  and  $\Sigma_{\bar{G}1}$  of the genotype coding vector  $X_{1i}$  in affected and unaffected individuals, one may calculate the noncentrality parameter  $\lambda_{G1}$  similarly.

### Electronic-Database Information

Accession numbers and URLs for data presented herein are as follows:

R.F.'s Web site, [http://stat.tamu.edu/~rfan/paper.html/case\\_control\\_Figs\\_supplement.pdf](http://stat.tamu.edu/~rfan/paper.html/case_control_Figs_supplement.pdf) and [http://stat.tamu.edu/~rfan/paper.html/case\\_control\\_powsim.pdf](http://stat.tamu.edu/~rfan/paper.html/case_control_powsim.pdf) (for supplementary information)

### References

Akey J, Jin L, Xiong MM (2001) Haplotype vs. single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet* 9:291-300  
 Anderson TW (1984) An introduction to multivariate statistical analysis, 2nd edition. John Wiley and Sons, New York  
 Ardlie KG, Lunetta KL, Seielstad M (2002) Testing for population subdivision and association in four case-control studies. *Am J Hum Genet* 71:304-311  
 Broman KW, Murray JC, Sheffed VC, White RL, Weber JL (1998) Comprehensive human genetic map: individual and sex-specific variation in recombination. *Am J Hum Genet* 63:861-869

Chapman NH, Wijsman EM (1998) Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. *Am J Hum Genet* 63:1872-1885  
 Cordell HJ, Clayton DG (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* 70:124-141  
 Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. *Nat Genet* 29:229-232  
 Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edition. Longman, London  
 Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. *Science* 296:2225-2229  
 Goldstein GB (2001) Islands of LD. *Nat Genet* 29:109-111  
 Hotelling H (1931) The generalization of student's ratio. *Ann Math Stat* 2:360-378  
 Iles MM (2002) Effect of mode of inheritance when calculating

- the power of a transmission/disequilibrium test study. *Hum Hered* 53:153–157
- The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933
- Kaplan N, Morris R (2001) Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. *Genet Epidemiol* 20:432–457
- Kong A, Gudbjartsson DF, Sainz J, Jonsson GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high resolution recombination map of the human genome. *Nat Genet* 31:241–247
- Nielsen DM, Ehm MG, Weir BS (1998) Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am J Hum Genet* 63:1531–1540
- Olson JM, Wijsman EM (1994) Design and sample size considerations in the detection of linkage disequilibrium with a marker locus. *Am J Hum Genet* 55:574–580
- Ott J (1999) *Analysis of human genetic linkage*, 3rd edition. Johns Hopkins University Press, Baltimore and London
- Patil NP, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BTN, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SPA, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294:1719–1723
- Rannala B, Reeve JP (2001) High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *Am J Hum Genet* 69:159–178 (erratum 69:172)
- Reich DE, Cargill M, Bolk S, Ireland J, Sabet RC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. *Nature* 411:199–204
- Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, et al (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet* 29:223–228
- Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* 13:423–449
- Schaid DJ, Rowland C (1998) Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. *Am J Hum Genet* 63:1492–1506
- Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, et al (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 293:489–493
- Xiong MM, Zhao J, Boerwinkle E (2002) Generalized  $T^2$  test for genome association studies. *Am J Hum Genet* 70:1257–1268