A New Association Test to Test Multiple-Marker Association

Xuexia Wang[1], **Shuanglin Zhang**[1,2], and **Qiuying Sha**[1,*]

[1]Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan

[2]Department of Mathematics, Heilongjiang University, Harbin, China

Abstract

As a result of the availability of a very large numbers of single nucleotide polymorphisms, there has been increasing interest in genetic associations involving several closely linked loci. Methods for detection of association between traits and multiple genetic polymorphisms are being rapidly developed, which include the Hotelling's T^2 test and the LD contrast (LDC) tests. The Hotelling's T^2 test can be considered as a test to compare the means of the genotypic score in cases and controls; while the (LDC) tests can be considered as a test to compare the variance-covariance matrices of the genotypic score in cases and controls. In this article, we propose a likelihood ratio test which simultaneously compares the means and the variance-covariance matrices of the genotypic score in cases and controls. We use simulation studies to evaluate the type I error rate of the proposed test, and compare the power of the test with the Hotelling's T^2 test and the LDC tests. The simulation results show that when marginal effects of the disease loci are strong, the proposed test is more powerful than the LDC tests and similar with or slightly less powerful than the Hotelling's T^2 test. If there are interaction effects and weak or no marginal effects, the proposed method is more powerful than the Hotelling's T^2 test and slightly more powerful than the LDC tests.

Keywords

likelihood ratio test; principle component analysis; association study; complex disease

INTRODUCTION

Association tests have been used for detecting the association between the genetic variants and the disease phenotypes. Recently, as the result of the availability of a very large numbers of single nucleotide polymorphisms (SNPs), there has been increasing interest in genetic associations involving several closely linked loci. There is strong evidence that several mutations within a single gene can interact to create a super allele that has a large effect on the observed phenotype [Schaid et al., 2002; Hollox et al., 2001; Clark et al., 1998; Tavtigian et al., 2001; Drysdale et al., 2000], which emphasizes the importance of the analysis of multiple SNPs that jointly represent variation within common transcripts and other functional regions, such as promoters. In recent years, many multi-marker association tests have been proposed which include haplotype-based methods that compare the distribution of haplotypes in cases with those of haplotypes in controls [Sham, 1998; Zhao et al., 2000; Schaid et al., 2002; Zaykin et al., 2002; Sha et al., 2005, 2007]; the Hotelling's T^2 test [Xiong et al., 2002; Fan and Knapp, 2003; Chapman et al., 2003; Wallace et al., 2006];

and the linkage disequilibrium (LD) contrast tests [Hayes et al., 2004; Nielsen et al., 2004; Zaykin et al., 2006; Wang et al., 2007]. Chapman et al. [2003] showed that in most cases the Hotelling's T^2 test is more powerful than haplotype-based methods. However, when marginal effects are weak and interaction effects are strong, the Hotelling's T^2 test which depends on the linear combination of the marginal effects will lose power dramatically. It has been noted that the extent of LD can be different between cases and controls in a region of genetic association, and the case-control LD comparison can aid the analysis in a region of putative association [Hayes et al., 2004]. In the context of association mapping, Nielsen et al. [2004] presented a direct LD comparison approach involving two biallelic loci and noted that, in certain situations, a test that directly compares the LD extent between cases and controls can be a powerful alternative to either haplotype-based or single-marker approaches. More recently, Zaykin et al. [2006] and Wang et al. [2007] suggested new LD contrast (LDC) tests to compare the matrices of pair-wise LD in cases and controls and demonstrated that the new LDC tests may be more powerful than Hotelling's T^2 test in the presence of gene-gene interaction.

The LDC tests can be considered as a test to compare the variance of the genotypic score in cases and controls; while the Hotelling's T^2 test can be considered as a test to compare the mean of the genotypic score in cases and controls. In this article, we propose a test to simultaneously compare the mean and variance of the genotypic score between cases and controls. We used simulation studies to compare the power of the proposed method with the Hotelling's T^2 test and the LDC tests. The simulation results show that when marginal effects of the disease loci are strong, the proposed test is more powerful than the LDC tests and similar with or slightly less powerful than the Hotelling's T^2 test. If there are interaction effects and weak or no marginal effects, the proposed method is more powerful than the Hotelling's T^2 test and slightly more powerful than the LDC tests.

METHOD

Consider a sample of n cases and m controls. Suppose that there are k biallelic markers that have been genotyped for each of the sampled individuals. The jth marker has alleles b_j and B_j. Define a numerical code of the genotype of the jth marker for the ith case:

$$x_{ij} = \begin{cases} 1, & B_j B_j \\ 0, & B_j b_j \\ -1, & b_j b_j \end{cases}.$$

Similarly, we define a numerical code y_{ij} of the genotype of the jth marker for the ith control. Let

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^{\mathrm{T}}, Y_i = (y_{i1}, \ldots, y_{ik})^{\mathrm{T}},$$

$$\bar{\mathrm{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i, \ \bar{\mathrm{Y}} = \frac{1}{m}\sum_{i=1}^{m} Y_i,$$

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

The Hotelling's T^2 [Xiong et al., 2002] test statistic is given by

$$T^2 = \frac{nm}{n+m}(\bar{X} - \bar{Y})^T S^{-1} (\bar{X} - \bar{Y}).$$

When $k = 1$, the Hotelling's T^2 test statistic is the square of the standard student t test statistic which is given by

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{m+n}{mn} S}}.$$

The student t test is a standard test to compare the means of two populations. Thus, we can consider the Hotelling's T^2 test described above as a test to compare the means of genotypic codes in cases and controls.

Let Δ_x and Δ_y denote two $k \times k$ matrices with elements δ_{ij}^x and δ_{ij}^y, respectively. The test statistics of the LDC tests proposed by Zaykin et al. [2006] and Wang et al. [2007] have the form

$$T = \text{trace}[(\Delta_x - \Delta_y)^T (\Delta_x - \Delta_y)] = \sum_{i=1}^{k} \sum_{j=1}^{k} (\delta_{ij}^x - \delta_{ij}^y)^2.$$

In Zaykin et al. [2006], Δ_x and Δ_y are the correlation matrices of the genotype codes in cases and controls, where

$$\delta_{ij}^x = \frac{\sum_{l=1}^{n}(x_{li} - \bar{x}_i)(x_{lj} - \bar{x}_j)}{\sqrt{\sum_{l=1}^{n}(x_{li} - \bar{x}_i)^2 \sum_{l=1}^{n}(x_{lj} - \bar{x}_j)^2}}$$

and

$$\delta_{ij}^y = \frac{\sum_{l=1}^{m}(y_{li} - \bar{y}_i)(y_{lj} - \bar{y}_j)}{\sqrt{\sum_{l=1}^{m}(y_{li} - \bar{y}_i)^2 \sum_{l=1}^{m}(y_{lj} - \bar{y}_j)^2}}$$

In Wang et al. [2007], Δ_x and Δ_y are the modified variance-covariance matrices of the genotype codes in cases and controls. From the statistics, we can see that the LDC tests proposed by Zaykin et al. [2006] and Wang et al. [2007] are in fact used to compare the variance-covariance of the genotypic codes in cases and controls.

In this article we proposed a test statistic to compare the means and the variance-covariance matrices of the genotypic scores in cases and controls simultaneously. Suppose $X_1, X_2 \cdots X_n$ are independent and identically distributed with a mean of $\mu_{\text{case}} = (\mu_1, \ldots, \mu_k)$ and a variance-covariance matrix of Σ_{case}, and $Y_1, Y_2, \ldots, Y_m$ are independent and identically distributed with a mean of $\mu_{\text{control}} = (\mu_1, \ldots, \mu_k)$ and a variance-covariance matrix of Σ_{control}. We consider the null hypothesis

$$H_0 : \mu_{\text{case}} = \mu_{\text{control}} \text{ and } \Sigma_{\text{case}} = \Sigma_{\text{control}}.$$

In the following discussion, we use multivariate normal distribution to deduce the log-likelihood ratio test statistic, and then we use this log-likelihood ratio as our test statistic. Although the multivariate normal assumption maybe violated, the log-likelihood ratio may still be a good test.

If X_i and Y_i both follow a multivariate normal distribution, the log-likelihood function is

$$\begin{aligned}
&\log L(\mu_{\text{case}}, \mu_{\text{control}}, \Sigma_{\text{case}}, \Sigma_{\text{control}}) \\
&= -\frac{n}{2} \log|\Sigma_{\text{case}}| \\
&\quad - \frac{1}{2} \sum_{i=1}^{n} (X_i - \mu_{\text{case}})^{\text{T}} \Sigma_{\text{case}}^{-1} (X_i - \mu_{\text{case}}) \\
&\quad - \frac{m}{2} \log|\Sigma_{\text{control}}| \\
&\quad - \frac{1}{2} \sum_{i=1}^{m} (Y_i - \mu_{\text{control}})^{\text{T}} \Sigma_{\text{control}}^{-1} (Y_i - \mu_{\text{control}}).
\end{aligned}$$

Under the null hypothesis $H_0 : \mu_{\text{case}} = \mu_{\text{control}}$ and $\Sigma_{\text{case}} = \Sigma_{\text{control}}$, the maximum likelihood estimates of the mean and the variance-covariance matrix are $\hat{\mu}$ and $\hat{\Sigma}$, where

$$\hat{\mu} = \frac{1}{(m+n)} \left(\sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i \right)$$

and

$$\hat{\Sigma} = \frac{1}{m+n} \left[\sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^{\text{T}} + \sum_{i=1}^{m} (Y_i - \hat{\mu})(Y_i - \hat{\mu})^{\text{T}} \right].$$

Under the alternative hypothesis, the maximum likelihood estimates of μ_{case}, μ_{control}, Σ_{case} and Σ_{control} are $\hat{\mu}_{\text{case}}$, $\hat{\mu}_{\text{control}}$, $\hat{\Sigma}_{\text{case}}$ and $\hat{\Sigma}_{\text{control}}$, respectively, where

$$\hat{\mu}_{\text{case}} = \frac{1}{n} \sum_{i=1}^{n} X_i, \hat{\mu}_{\text{control}} = \frac{1}{m} \sum_{i=1}^{m} Y_i,$$

$$\hat{\Sigma}_{\text{case}} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu}_{\text{case}})(X_i - \hat{\mu}_{\text{case}})^{\text{T}}$$

and

$$\hat{\Sigma}_{\text{control}} = \frac{1}{m} \sum_{i=1}^{m} (Y_i - \hat{\mu}_{\text{control}})(Y_i - \hat{\mu}_{\text{control}})^{\text{T}}.$$

So, the log-likelihood ratio test statistic is

$$\text{LRT} = 2(\log L_{H_a} - \log L_{H_0}) = -n \log|\hat{\Sigma}_{\text{case}}| - m \log|\hat{\Sigma}_{\text{control}}| + (m+n)\log|\Sigma|. \quad (2)$$

We propose to use a permutation procedure to evaluate the *P*-value of the test instead of using χ^2 distribution. To evaluate the *P*-value of the test, let LRT_0 denote the value of the test statistic based on the original data set. For each permutation, we randomly shuffle the case and control states among the sampled individuals and denote the value of the test statistic based on the permuted data set by LRT^{per}. We perform the permutation procedure many times. Then, the *P*-value of the test is the proportion of the number of permutations with $\text{LRT}^{\text{per}} \quad \text{LRT}_0$.

If there is a very small determinant of $\hat{\Sigma}_{\text{case}}$, $\hat{\Sigma}_{\text{control}}$ or $\hat{\Sigma}$, the $\log |\hat{\Sigma}_{\text{case}}|$, $\log |\hat{\Sigma}_{\text{control}}|$ or $\log |\Sigma|$ will be very sensitive to a small change of eigenvalues. Hence, we propose to use the principle component (PC) analysis to reduce the sensitivity of the determinants. The PC analysis can also increase the power due to reducing the degrees of the freedom of the test [Sha et al., 2005]. To carry out the PC analysis, let q_j be the eigenvector corresponding to the *j*th largest eigenvalue λ_j of the variance-covariance matrix $\hat{\Sigma}_{\text{control}}$ in controls. Then, the *j*th PC of the *i*th individual in controls is $z_{ij} = q_j^{\text{T}} Y_i$ and the total variance in controls explained by the *j*th PC is $\lambda_j/(\lambda_1 + \cdots + \lambda_k)$. Let $z_{i1}, \ldots, z_{il}$ denote the first *l* PCs of Y_i, that can explain the majority of the total variability, where *l* is a pre-specified number and will be discussed below. We use a *l*-dimensional vector $(z_{i1}, \ldots, z_{il})$ instead of a *k*-dimensional vector Y_i, where $l < k$, as the new numerical codes for the multi-marker genotype of individual *i* in controls. Similarly, we calculate the *j*th PC of the *i*th individual in cases as $w_{ij} = q_j^{\text{T}} X_i$ and the *j*th PC of the *i*th individual in pooled data as $v_{ij} = q_j^{\text{T}} G_i$; we use a *l*-dimensional vector $(w_{i1}, \ldots, w_{il})$ as the new numerical codes for the multi-marker genotype of individual *i* in cases. Let $Z_i = (z_{i1}, z_{i2}, \ldots, z_{il})^{\text{T}}$ and $W_i = (w_{i1}, \ldots, w_{il})^{\text{T}}$. Using the new coding scheme, the log-likelihood ratio test statistic will be

$$\text{LRT\_PC} = 2(\log L_{H_a} - \log L_{H_0}) = -n \log|\hat{\Sigma}_{\text{case-pc}}| - m \log|\hat{\Sigma}_{\text{control-pc}}| + (m+n)\log|\hat{\Sigma}_{\text{pc}}|,$$

where

$$\hat{\mu}_{\text{case-pc}} = \frac{1}{n} \sum_{i=1}^{n} W_i, \hat{\mu}_{\text{control-pc}} = \frac{1}{m} \sum_{i=1}^{m} Z_i,$$

$$\hat{\mu}_{\text{pc}} = \frac{1}{(m+n)} \left(\sum_{i=1}^{n} W_i + \sum_{i=1}^{m} Z_i \right),$$

$$\hat{\Sigma}_{\text{case-pc}} = \frac{1}{n} \sum_{i=1}^{n} (W_i - \hat{\mu}_{\text{case-pc}})(W_i - \hat{\mu}_{\text{case-pc}}),$$

$$\hat{\Sigma}_{\text{control-pc}} = \frac{1}{m} \sum_{i=1}^{m} (Z_i - \hat{\mu}_{\text{control-pc}})(Z_i - \hat{\mu}_{\text{control-pc}})^{\text{T}}$$

and

$$\hat{\Sigma}_{\text{pc}} = \frac{1}{m+n} \left[\sum_{i=1}^{n} (W_i - \hat{\mu}_{\text{pc}})(W_i - \hat{\mu}_{\text{pc}})^{\text{T}} + \sum_{i=1}^{m} (Z_i - \hat{\mu}_{\text{pc}})(Z_i - \hat{\mu}_{\text{pc}})^{\text{T}} \right].$$

One question that needs to be answered in the proposed testing procedure is how to choose the value of l. We propose to choose l such that $(\lambda_1 + \lambda_2 + \cdots + \lambda_l)/(\lambda_1 + \lambda_2 + \cdots + \lambda_k)$ is greater than a pre-specified value δ to make sure that the majority of the total variation in the data is explained. We call this δ the cutoff value of PC. In the simulation studies, we choose δ equal to 85% and we will give a further discussion about the choice of δ in the discussion section.

SIMULATIONS

We use simulation studies to evaluate the performance of the proposed test LRT_PC and to compare the power of the LRT_PC with four other tests: the LDC test proposed by Zaykin et al. [2006], the modified LD contrast (MLDC) test proposed by Wang et al. [2007], the Hotelling's T^2 test developed by Xiong et al. [2002], and the single-marker test. For the k biallelic markers, we calculate the 1 degree freedom χ^2 test statistics $t_1, t_2, \ldots, t_k$ and use SMT $= \max\{t_1, t_2, \ldots, t_k\}$ as the test statistic of our single-marker test. For each simulation scenario, we simulate 1,000 replicated data sets and use 1,000 permutations to evaluate the P-values of all the tests.

DATA SETS FOR ASSESSING THE TYPE I ERROR

We generate a data set to evaluate the type I error rate of the proposed method using the simulation setting similar to that of Wang et al. [2007]. Briefly, we simulate a set of markers with four SNPs in a local region or a candidate gene. The haplotypes of the four correlated SNPs are simulated on the basis of a multivariate normal distribution with a pair-wise correlation coefficient ρ. Each allele of a haplotype is generated by dichotomizing the marginal normal distribution, and the cutoff is determined by the minor allele frequency. For each individual, we randomly assign disease status. We consider different minor allele frequencies, different values of ρ, and different sample sizes.

DATA SETS FOR POWER COMPARISON

To assess the power of the proposed tests, we consider two sets of simulations which are generated based on a two-locus disease model and a haplotype effect model.

Haplotype effect model—In this set of simulation we use the same method as given in "*Data sets for assessing the type I error*" section to generate genotypes, except that the number of marker is 10. For a dichotomous trait, we assume that the trait is due to an underlying continuous liability y. The trait y follows a linear model $y = g_1 + g_2 + e$, where

g_1 and g_2 are trait-locus effects of the two haplotypes, and e is a random environment effect. For a haplotype, the trait-locus effect is set in the following way: for a haplotype across the 10 markers, let s_j represent the code of the allele at the jth marker ($s_j = 1$ for minor allele; $s_j = 0$ for major allele) and $S = \sum_{j=1}^{10} s_j$. The trait-locus effect of this haplotype is defined by $g = |S–5|/5$. Disease status is defined by a threshold Z, such that all individuals with $y > Z$ are classified as cases. In our simulations, $Z = 1.65$. We consider a sampling of 100 cases and 100 controls.

Two-locus disease model—In this set of simulation, we assume that the true risk model involves two potentially interacting causal SNPs, S_1 and S_2, residing on two separate candidate genes, G_1 and G_2, respectively. For each gene, there are seven SNPs and the third SNP is the causal SNP. After we generate genotypes, we delete the genotypes at the third SNP (the causal SNP). In our data analysis, we assume that, for each gene, genotype data are available on six marker SNPs. To simulate a realistic LD pattern among the markers within each gene (two genes assume to be in linkage equilibrium), we use haplotype data extracted from two segments on chromosome 21 in the CEU HapMap sample [Thorisson et al., 2005]. The haplotypes and their frequencies are given in Table I.

To generate disease status, we follow the models used by Chatterjee et al. [2006]: purely epistatic model, additively model, and crossover model as given in Table II. Define marginal relative risks (MRR) as

$$\text{MRR}_1 = \frac{p(D|s_1 \geq 1)}{p(D|s_1 = 0)},$$

and

$$\text{MRR}_2 = \frac{p(D|s_2 \geq 1)}{p(D|s_2 = 0)},$$

where D denotes the disease, s_1 and s_2 refer to the number of copies of allele 1 in the causal loci of G_1 and G_2, respectively. The values of the parameters θ_1, θ_2, and θ_{12} are decided by the values of the prevalence (assume to be 0.1 in our simulation), MRR_1, and MRR_2. For each model, we vary the value of MRR_1 in the set $\{1.2, 1.5, 1.75, 2\}$. For the epistatic mode, MRR_2 is determined by MRR_1. For the additive model, we fix the MRR_2 to be 2.0. For the crossover model, we assume $\theta_1 = 0.9$ and MRR_2 is determined by MRR_1. For each model, we consider a sample of 500 cases and 500 controls.

RESULTS

TYPE I ERROR RATES

The estimated type I error rates for different allele frequencies, different correlation coefficients, different sample sizes, and different value of δ (the cutoff value of the PC) are given in Table III. The standard deviation for the estimated type I error is $\sqrt{0.05 \times 0.95/1000} \approx 0.0069$ for the nominal level of 0.05 and 1,000 replicated samples, so the 95% confidence interval is (0.0362, 0.0638). From Table III we can see that the estimated type I error rates of the proposed test are not statistically significantly different from the nominal levels in all the cases.

POWER COMPARISONS

First, we evaluate the effects of δ (the cutoff value of the PC) on the power. We use the haplotype effect models given in the above section to generate data sets. The results for different simulation scenarios are summarized in Figure 1. From Figure 1 we can see that the trends of the power are very similar for different values of correlation coefficients and allele frequencies. The power is increasing as the cutoff value δ increases and it reaches its maximum value at some point of δ, then the power is decreasing as the cutoff value δ increases. Although there is no universal optimal value of δ, there is a wide range of δ in which the power is higher than that of $\delta = 100\%$ that is equivalent to the case of not using PC analysis. In the following discussion, we use the cutoff value $\delta = 85\%$ when we compare the power of the proposed method with the other three methods.

To compare the power of the proposed test with the other four tests, we consider two sets of simulations. One is based on two-locus disease models and the other is based on a haplotype effect model. Under the haplotype effect model, we consider three different scenarios of the high-risk allele frequencies of 0.1, 0.2 and 0.3 and vary the value of correlation coefficient from 0 to 0.8. The results of the power comparisons are summarized in Figure 2. The proposed method LRT_PC is consistently more powerful than the single-marker test and the two LDC tests, LDC and MLDC, in all the cases. Except for the cases of correlation coefficient being 0, the LRT_PC is also more powerful than that of the T^2 test. The powers of the LRT_PC, LDC and MLDC have similar patterns—power is increasing as the value of the correlation coefficient ρ increases. The single-marker test and the Hotelling's T^2 have similar patterns—power is increasing first then decreasing as the value of ρ increases. The pattern of the power of the single-marker test indicates that the marginal effects will be increasing first then decreasing as the value of ρ increases. This may partly explain the power pattern of the other four tests. For large values of ρ such as $\rho = 0.8$, the LRT_PC is much more powerful than the T^2 test and the single-marker test.

The results of the power comparisons under the two-locus disease models are summarized in Figure 3. The MRR given in Table IV as well as the power of the single-marker test in Figure 3 indicate that all of the three two-locus models have strong marginal effects. Figure 3 again shows that the LRT_PC is consistently more powerful than the single-marker test and the two LDC tests, LDC and MLDC. Comparing the power of LRT_PC with that of the T^2 test, we can see that the two tests have similar power under the epistatic and crossover models and the T^2 test is slightly more powerful than the LRT_PC under the additive model. The additive model is favorable to the T^2 test because the T^2 test is derived under the assumption of the additive effect between markers. Under the additive model which assumes no interaction, the LDC and MLDC have almost no power.

Summarizing the results from the two sets of simulations, we can see that the two LDC tests (LDC and MLDC) which are designed to test interaction effects will lose power when there exist marginal effects; the T^2 test will lose power when the marginal effects are weak. The proposed LRT_PC test has reasonable power in all the cases. The LRT_PC is consistently more powerful than the single-marker test and the LDC tests (LDC and MLDC). Except for the additive model, the LRT_PC test is not less powerful than the T^2 test.

DISCUSSION

In this article we proposed a new multi-marker association test for case-control studies. The proposed test statistic compares the difference of the mean and the variance of the genotypic scores in cases and controls simultaneously. At the same time, we propose to use PC analysis to reduce the degrees of the freedom of the test to improve the power of the association test. We compared the power of the proposed test with that of the existing

association tests which include the single-marker test, the Hotelling's T^2 test and two recently developed LDC tests. The simulation results show that the proposed test is consistently more powerful than the single-marker test and the two LDC tests. When there are interaction effects and weak or no marginal effects, our proposed method is more powerful than the Hotelling's T^2 test, when there are strong marginal effects, our proposed method has similar power with or slightly less powerful than the Hotelling's T^2 test (see the three two-locus models).

All the three two-locus models in our simulation studies have strong marginal effects. We also did a simulation study under a two-locus model with weak marginal effects: considering two markers A and B with alleles a, A and b, B, the two-locus high-risk genotypes are {*AAbb, AaBb, aaBB*}. This model has weak marginal effects and strong joint effects. Our simulation results showed that under this model the proposed method is the most powerful one while the single-marker test and the Hotelling's T^2 test are almost no power at all.

In the PC analysis used in this article, we first find PC directions (eigenvectors of the variance-covariance matrix) using the control sample only. Then, we project the original numerical codes of multi-marker genotypes in cases and controls to the PC directions, and get the PC codes for the multi-marker genotypes. Based on the PC codes, we calculate the test statistic LRT_PC. Using this PC analysis, we need to recalculate PC directions and PC codes in each permutation when we use permutation procedure to evacuate the *P*-value of the test. Thus, this PC analysis makes the permutation procedure computationally intensive. An alternative way to do the PC analysis is that we find PC directions using pooled sample (cases and controls together) only. In this way, we do not need to recalculate PC directions and PC codes in each permutation and the permutation procedure will be much faster. We denote the corresponding LRT_PC by using this PC analysis as LRT_PC-Pool. Our simulation studies showed that the power of the LRT_PC-Pool test is only slightly less powerful than the LRT_PC test (results are not shown). Thus, we suggest using this PC analysis whenever computational time becomes a concern.

One remaining question in the proposed LRT_PC is choosing the cutoff value δ in the PC analysis. Our simulation studies show that there is no universal optimal value for δ. However, we feel that the values around 90% are good choices. Although our simulation results show that the optimal value of δ may be less then 50%, our experience shows that we are more likely to miss rare disease associated alleles when we use small value of δ. In general, we need further investigation on choosing the optimal value of δ.

Our method cannot be applied directly to genome-wide association studies. However, we can apply the proposed method to genome-wide association studies by using a sliding window approach. We have done a simulation study to compare the power of the proposed test with the two LDC tests, the Hotelling's T^2 test and the single-marker test by using a sliding window approach. Our simulation study showed that the pattern of the power comparison by using a sliding window approach is similar to that of the other simulation results. However, the power of the five tests is affected by the window sizes or the number of markers in each window. The optimal choice of the window size needs further investigation.

Acknowledgments

# REFERENCES

Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease association due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered. 2003; 56:18–31. [PubMed: 14614235]

Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet. 2006; 79:1002–1016. [PubMed: 17186459]

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet. 1998; 63:595–612. [PubMed: 9683608]

Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K, Arnold K, Ruano G, Liggett SB. Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proc Natl Acad Sci USA. 2000; 97:10483–10488. [PubMed: 10984540]

Fan R, Knapp M. Genome association studies of complex diseases by case-control designs. Am J Hum Genet. 2003; 72:850–868. [PubMed: 12647259]

Hayes MG, Roe CA, Ng M, Bosque-Plata L, Tsuchiya T, Wu X, Ambrose NG, Yairi E, Cook EH, Cox NJ. Case-control differences in linkage disequilibrium as a tool for gene mapping in complex diseases. Am J Hum Genet Suppl. 2004; 54:A146.

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, Swallow DM. Lactase haplotype diversity in the old world. Am J Hum Genet. 2001; 68:160–172. [PubMed: 11095994]

Nielsen DM, Ehm MG, Zaykin DV, Weir BS. Effect of two and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. Genetics. 2004; 168:1029–1040. [PubMed: 15514073]

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score test for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet. 2002; 70:425–434. [PubMed: 11791212]

Sha Q, Dong J, Jiang R, Zhang S. Tests of association between quantitative traits and haplotypes in a reduced-dimensional space. Ann Hum Genet. 2005; 69:715–732. [PubMed: 16266410]

Sha Q, Chen HS, Zhang S. New association tests based on haplotype similarity. Genetic Epidemiol. 2007; 31:577–593.

Sham, P. Statistics in Human Genetics. New York: Oxford University Press, Inc; 1998.

Tavtigian SV, Simard J, Teng DH, Abtin V, Baumgard M, Beck A, Camp NJ, Carillo AR, Chen Y, Dayananth P, Desrochers M, Dumont M, Farnham JM, Frank D, Frye C, Ghaffari S, Gupte JS, Hu R, Iliev D, Janecki T, Kort EM, Laity KE, Leavitt A, Leblanc G, McArthur-Morrison J, Pederson A, Penn B, Peterson KT, Reid JE, Richards S, Schroeder M, Smith R, Snyder SC, Swedlund B, Swensen J, Thomas A, TranchantM,Woodland A, Labrie F, Skolnick MH, Neuhausen S, Rommens J, Cannon-Albright LA. A candidate prostate cancer susceptibility gene at chromosome 17p. Nat Genet. 2001; 27:172–180. [PubMed: 11175785]

Thorisson GA, Smith AV, Krishnan L, Stein LD. The international hapMap project web site. Genome Res. 2005; 15:1591–1593.

Wang T, Zhu X, Elston RC. Improving power in contrasting linkage-disequilibrium patterns between cases and controls. Am J Hum Genet. 2007; 80:911–920. [PubMed: 17436245]

Wallace C, Chapman JM, Clayton DJ. Score test for selective genotyping. Am J Hum Genet. 2006; 78:498–504. [PubMed: 16465623]

Xiong M, Zhao J, Berwinkle E. Generalized t^2 test for genome association studies. Am J Hum Genet. 2002; 70:1257–1268. [PubMed: 11923914]

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered. 2002; 53:79–91. [PubMed: 12037407]

Zaykin DV, Meng Z, Ehm MG. Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. Am J Hum Genet. 2006; 78:737–746. [PubMed: 16642430]

Zhao H, Zhang S, Merikangas KR, et al. Transmission/disequilibrium tests using multiple tightly linked markers. Am J Hum Genet. 2000; 67:936–946. [PubMed: 10968775]
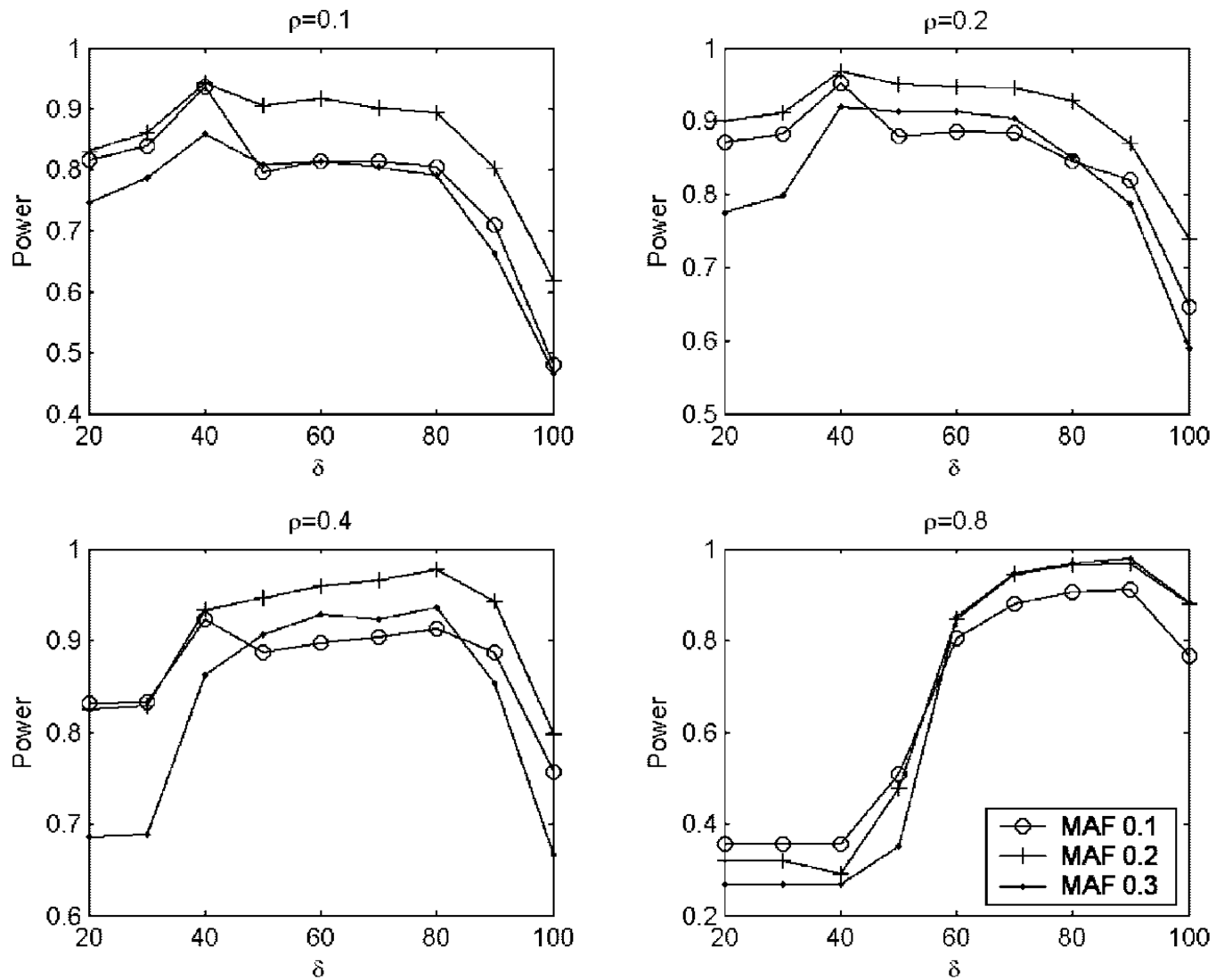
Fig. 1.
Power comparisons of the proposed test for different cutoff values of the PC analysis under the haplotype effect model. δ is the cutoff value of PC. ρ is the correlation coefficient. MAF is the minor allele frequency. The sample consists of 100 cases and 100 controls.
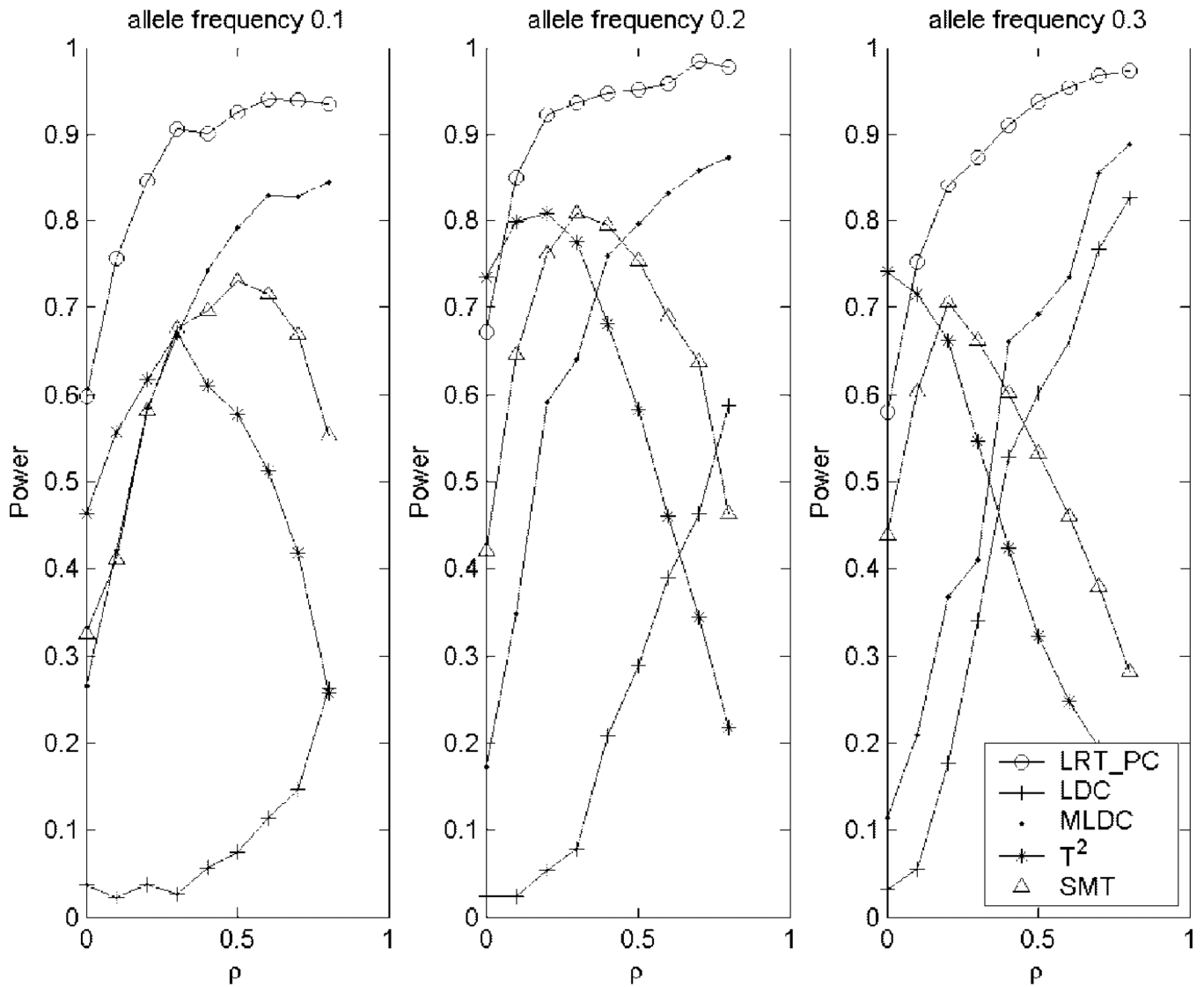
Fig. 2.
Power comparisons of the five tests under the haplotype effect model. The sample consists of 100 cases and 100 controls. LRT_PC: the test proposed in this article; LDC: the LD contrast test proposed by Zaykin et al. [2006]; MLDC: the modified LD contrast test proposed by Wang et al. [2007]; T^2: the Hotelling's T^2 test developed by Xiong et al. [2002]; SMT: the single-marker test; ρ: correlation coefficient.
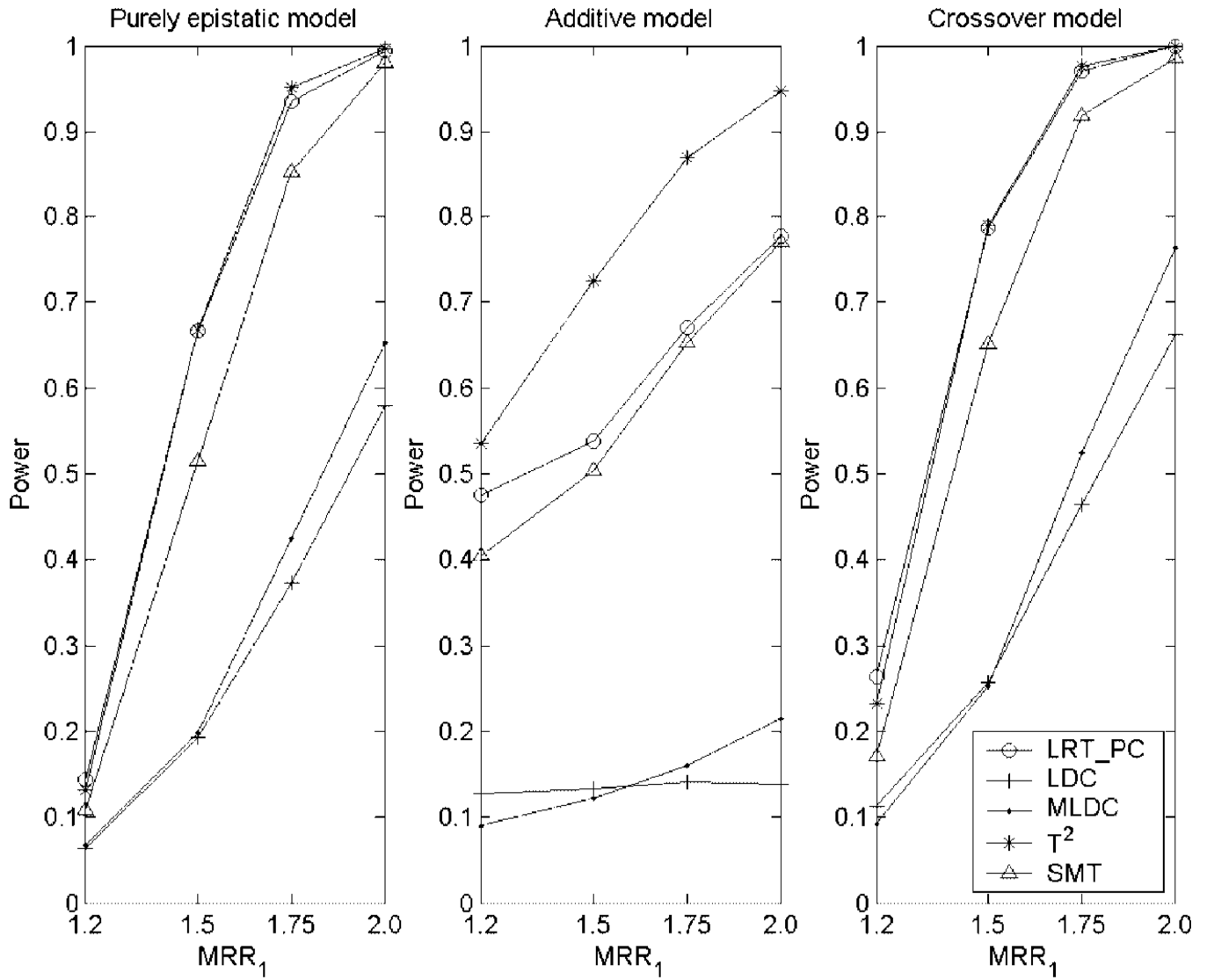
Fig. 3.
Power comparisons of the five tests under three two-locus models. The sample consists of 500 cases and 500 controls. LRT_PC: the test proposed in this article; LDC: the LD contrast test proposed by Zaykin et al. [2006]; MLDC: the modified LD contrast test proposed by Wang et al. [2007]; T^2: the Hotelling's T^2 test developed by Xiong et al. [2002]; SMT: the single-marker test; MRR_1: marginal relative risk.

TABLE I

Haplotypes and their frequencies used in simulating genotype data of the two candidate genes

Gene I (G_1)		Gene II (G_2)	
Haplotype	**Frequency**	**Haplotype**	**Frequency**
0100111	0.4000	1100011	0.4833
1000011	0.2083	1111111	0.0566
1110101	0.1217	1000101	0.0166
1100100	0.0416	0101011	0.2033
1000110	0.0333	1000100	0.0500
1101000	0.0166	1000000	0.0750
1100111	0.0316	1000011	0.0166
0101111	0.0083	1100001	0.0283
1101111	0.0316	1100111	0.0250
1101101	0.0333	1101011	0.0250
1100110	0.0166	1101111	0.0100
1100000	0.0083	0111011	0.0050
1001011	0.0083	1110001	0.0050
1100101	0.0199		
1110111	0.0099		
1111111	0.0099		

TABLE II

Relative risk in three disease models

Model	$s_1 = 0,$ $s_2 = 0$	s_1 1, $s_2 = 0$	$s_1 = 0,$ s_2 1	s_1 1, s_2 1
Purely epistatic	1	1	1	$\theta(>1)$
Additive	1	θ_1	θ_2	$\theta_1 + \theta_2 - 1$
Crossover	1	$\theta_1(<1)$	1	$\theta_{12}(> 1)$

s_1 and s_2 refer to the number of copies of allele 1 in the causal loci of G_1 and G_2, respectively

TABLE III

Type I error rates of the LRT_PC at a significance level of 5%

Sample size	Allele frequency	Correlation coefficient ρ	δ = 65%	δ = 75%	δ = 85%	δ = 95%
200	0.1	0.1	0.042	0.047	0.041	0.040
		0.2	0.052	0.051	0.052	0.052
		0.4	0.050	0.047	0.055	0.053
		0.8	0.058	0.061	0.054	0.050
	0.2	0.1	0.059	0.057	0.068	0.068
		0.2	0.062	0.064	0.057	0.057
		0.4	0.063	0.063	0.063	0.062
		0.8	0.054	0.056	0.05	0.050
	0.3	0.1	0.052	0.051	0.063	0.063
		0.2	0.052	0.046	0.054	0.054
		0.4	0.052	0.064	0.06	0.058
		0.8	0.061	0.054	0.057	0.053
800	0.1	0.1	0.046	0.046	0.043	0.043
		0.2	0.055	0.055	0.046	0.046
		0.4	0.036	0.036	0.045	0.045
		0.8	0.057	0.055	0.052	0.065
	0.2	0.1	0.040	0.040	0.046	0.046
		0.2	0.044	0.044	0.054	0.054
		0.4	0.051	0.051	0.049	0.049
		0.8	0.058	0.053	0.054	0.054
	0.3	0.1	0.049	0.049	0.049	0.049
		0.2	0.040	0.040	0.045	0.045
		0.4	0.042	0.052	0.051	0.051
		0.8	0.054	0.056	0.064	0.041

Note: δ is the cutoff value of PC.

TABLE IV

The values of the parameters under the three two-locus models when MRR_1 varies from 1.2 to 2.0

Model	$MRR_1 = 1.2$	$MRR_1 = 1.5$	$MRR_1 = 1.75$	$MRR_1 = 2.0$	Average MRR
Purely epistatic	$\theta = 2.55$	$\theta = 4.88$	$\theta = 6.82$	$\theta = 8.76$	1.855
	$MRR_2 = 0.8$	$MRR_2 = 2.02$	$MRR_2 = 2.53$	$MRR_2 = 3.04$	
Additive	$\theta_1 = 1.31$	$\theta_1 = 1.86$	$\theta_1 = 2.44$	$\theta_1 = 3.17$	1.806
	$\theta_2 = 2.08$	$\theta_2 = 2.22$	$\theta_2 = 2.38$	$\theta_2 = 2.56$	
	$MRR_2 = 2$	$MRR_2 = 2$	$MRR_2 = 2$	$MRR_2 = 2$	
Crossover	$\theta_1 = 0.9$	$\theta_1 = 0.9$	$\theta_1 = 0.9$	$\theta_1 = 0.9$	1.927
	$\theta_{12} = 2.33$	$\theta_{12} = 4.66$	$\theta_{12} = 6.60$	$\theta_{12} = 8.54$	
	$MRR_2 = 1.38$	$MRR_2 = 2.01$	$MRR_2 = 2.53$	$MRR_2 = 3.05$	