

A Unified Framework for Detecting Genetic Association with Multiple SNPs in a Candidate Gene or Region: Contrasting Genotype Scores and LD Patterns between Cases and Controls

WEI PAN

*Division of Biostatistics, School of Public Health, University of Minnesota,
Minneapolis, MN 55455*

February 9, 2009; revised March 30, 2009

Running title: A Unified Framework to Test Genetic Association

Correspondence author: Wei Pan

Telephone: (612) 626-2705

Fax: (612) 626-0660

Email: weip@biostat.umn.edu

Address: Division of Biostatistics, MMC 303,
School of Public Health, University of Minnesota,
Minneapolis, Minnesota 55455-0392, U.S.A.

ABSTRACT

It is critical to develop and apply powerful statistical tests for genetic association studies due to typically weak associations with complex human diseases or phenotypes. For population-based case-control studies with unphased multilocus genotype data, most of the existing methods are based on comparing genotype scores, e.g. allele frequencies, between the case and control groups. Another class of approaches are motivated to contrast linkage disequilibrium (LD) patterns between the two groups. It is expected that no single test can be uniformly most powerful across all situations, and different tests may perform better under different scenarios. A recent effort has been devoted to combining the above two classes of approaches, which however has some potential drawbacks. Here we propose a general and simple framework to unify the above two classes of approaches: it is based on the simple idea to incorporate LD measurements, in addition to genotype scores, as covariates in a logistic regression model, from which various tests can be constructed by taking advantage of the nice properties of the score statistics for the logistic model. It also has an advantage in easily accommodating covariates and other study designs. We use simulated data to show that our proposed tests performed well across several scenarios. In particular, in contrast to either of the two classes of the tests that is only powerful in detecting only one, but not both, of the two types of the distributional differences between cases and controls, our proposed tests are sensitive to both.

Key Words: Genome-wide association study; Linkage disequilibrium (LD); LD contrast test; Logistic regression; Multilocus analysis; Score test; SNP; Sum of squared score tests.

1. INTRODUCTION

As genome-wide association studies have become feasible with many being underway, there have been intensive researches in developing powerful statistical tests to detect genetic association with multiple single-nucleotide polymorphisms (SNPs) in a candidate region. There are at least two compelling reasons for such developments. First, for complex human diseases or phenotypes, recent studies have confirmed usually weak genetic associations, often with estimated relative risks only from 1.1 to 1.5, hence powerful statistical tests are needed to discover genetic loci or variants weakly associated with a disease (Altshuler et al 2008). Second, it is known that there is no uniformly most powerful test for multiple parameters (Cox and Hinkley 1974), as for association studies with multiple SNPs. Instead, different tests are expected to perform best under different scenarios. For example, the standard multivariate score test and a sum of squared score test can be each regarded as an *estimated* most powerful tests (Pan 2009) under different situations.

Among the existing statistical tests for population-based case-control studies with unphased multilocus genotype data, there are several classes of tests that aim to detect different aspects of distributional differences of genotypes between cases and controls. The first class includes some most popular approaches, such as Hotelling's T^2 test (Fan and Knapp 2003; Xiong et al 2002) and single-locus-based tests (Roeder et al 2005). These tests are based on comparing genotype scores between the case and control groups. Another class includes a linkage disequilibrium (LD) contrast (LDC) test (Zaykin et al 2006) and a modified LDC (mLDC) test (Wang et al 2007). The main idea of these tests is to capitalize on possible differences of LD patterns between the case and control groups (Hayes et al 2004). Because of their different targets, different classes of the tests have high power for different aspects of distributional differences of genotypes between the case and control groups. As expected, if the distributional difference between the two groups is only or mainly in their mean genotype scores, the

first class of the tests is expected to be more powerful than the second class. Similarly, if the main difference lies only in their varying LD patterns, the second class will be more powerful, as to be confirmed by our simulation studies. For any real data, since the distributional difference could be in either or both aspects, it would be desirable to have a test that are powerful in detecting either and both types of the distributional differences. For this purpose, Wang et al (2009) attempted to combine the ideas of the two classes of approaches and proposed a Normal-based likelihood ratio test (LRT) to compare both the mean vectors and covariance matrices of genotype scores between the two groups. However, there are several potential drawbacks associated with their method. First, a covariance matrix may not be most suitable for an LDC test as pointed out by Zaykin et al (2006). More importantly, it cannot incorporate a background-corrected LD measurement as proposed by Wang et al (2007), which was adopted in the mLDC test shown to perform better than the correlation-based LDC test. Second, their LRT is based on the working assumption that genotype scores have a multivariate Normal distribution, which does not hold due to the discreteness of any genotype score. Although the LRT test is still valid with this incorrect Normality assumption, the power of the test can be negatively influenced, as to be shown. Finally, due to the use of permutations and the need to calculate large covariance matrices and their determinants, the method is computationally slow.

Here we propose a general framework under logistic regression to contrast both genotype scores and LD patterns simultaneously. The basic idea is to incorporate some terms measuring LD patterns, in addition to genotype scores, as covariates into a logistic regression model. As for popular main-effects logistic regression, by taking advantage of some nice properties of the associated score statistics, we can construct a class of tests, not just a single one. Specifically, in addition to the usual multivariate score test, we can also apply a sum of squared score (SSU) test or its weighted version (SSUw), and univariate or single-locus-based minP (UminP) test. The SSU, SSUw

and UminP tests have been shown to perform well across a wide range of scenarios in main-effects logistic regression models to compare genotype scores between the two groups (Chapman and Whittaker 2008; Pan 2009). In this way, even only for the purpose of contrasting LD patterns between the two groups, we have an expanded set of LDC-type tests. For example, rather than taking a simple average of the terms measuring LD differences between the two groups as in the LDC and mLDC tests, we can take a weighted average by putting higher weights to those terms with smaller variances as implemented in the SSUw test, or take the most significant one as in the UminP test. Additionally and importantly, rather than depending on computationally intensive permutations as in the LDC, mLDC and LRT tests, we can use simple and accurate asymptotic distributions to calculate p-values for these new tests. Furthermore, the proposed regression framework can easily accommodate covariates and other study designs.

A potential technical difficulty in incorporating LD measurements into logistic regression is the large number of regression coefficients induced by the added LD terms, leading to possibly a large number of degrees of freedom (DF) and thus possibly reduced power for a test. The SSU and SSUw tests, and to a lesser degree the UminP test, are robust to the large number of parameters to be tested. In fact, the SSU test (and thus SSUw) is closely related to an empirical Bayes test specifically developed to test parameters in a high-dimensional space (Goeman et al 2006; Pan 2009). Hence, the robustness of these tests facilitates their use in a largely expanded logistic regression model including many LD-related terms. As to be shown using simulated data, if information on genetic association is mainly embedded in varying LD patterns between the case and control groups, these tests are among the most powerful. Otherwise, they may still maintain high power, and in particular are more powerful than the LDC test (Zaykin et al 2006), mLDC test (Wang et al 2007) and Normal-based LRT (Wang et al 2009).

2. METHODS

2.1 Data

We consider a population-based case-control study with a disease indicator as the binary trait $Y_i = 0$ or 1 , and some genotyped markers in a candidate gene or region $X_i = (X_{i1}, \dots, X_{ik})'$. We adopt the popular dosage coding for X_{ij} : $X_{ig,j} = 0, 1$ or 2 represents the copy number of one of the two alleles present at locus j for subject i , though other coding schemes can be equally adopted. Without loss of generality, for the given data $\{(Y_i, X_i) : i = 1, \dots, m\}$, we assume that the first n_0 subjects are controls with $Y_i = 0$, and the next n_1 are cases with $Y_i = 1$. The study goal is to test whether there is any association between the trait and any markers.

2.2 Statistical Tests Based on Contrasting Genotype Scores

2.2.1 Logistic regression model

Many tests are based on a *main-effects* logistic regression model:

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + \sum_{j=1}^k X_{ij}\beta_j, \quad (1)$$

for which we would like to test the null hypothesis $H_{0,1}$: $\beta = (\beta_1, \beta_2, \dots, \beta_k)' = 0$ versus $H_{1,1}$: $\beta \neq 0$.

2.2.2 Multivariate score test

To test $H_{0,1}$, a multivariate score test can be constructed with the score statistic

$$U = \sum_{i=1}^m (Y_i - \bar{Y})X_i \quad (2)$$

with its covariance matrix consistently estimated by the expected Fisher information matrix

$$V = \bar{Y}(1 - \bar{Y}) \sum_{i=1}^m (X_i - \bar{X})(X_i - \bar{X})',$$

with $\bar{Y} = \sum_{i=1}^m Y_i/m$ and $\bar{X} = \sum_{i=1}^m X_i/m$. The test statistic is

$$T_{Sco} = U'V^{-1}U, \quad (3)$$

which, under H_0 , has an asymptotic chi-squared distribution χ_r^2 with degrees of freedom $r=\text{rank}(V)$, typically $r = k$. A potential problems with the test is that, with many SNPs, the test can be low-powered because of the cost of the large DF.

The above score test is closely related to the generalized Hotelling's T^2 test (Fan and Knapp 2003; Xiong et al 2002) with test statistic

$$T_H = (\bar{X}^1 - \bar{X}^0)' S^{-1} (\bar{X}^1 - \bar{X}^0),$$

where $\bar{X}^0 = \sum_{i=1}^{n_0} X_i/n_0$ and $\bar{X}^1 = \sum_{i=n_0+1}^m X_i/n_1$ are the mean genotype scores for the control and case groups respectively, and S is the pooled sample covariance matrix. On the other hand, as shown by Clayton et al (2004), we have

$$U = (0 - n_1/m) \sum_{i=1}^{n_0} X_i + (1 - n_1/m) \sum_{i=n_0+1}^m X_i = n_0 n_1 (\bar{X}^1 - \bar{X}^0)/m$$

Hence, T_{Sco} and T_H only differ in how the covariance matrix of the mean difference between the two groups is estimated.

2.2.3 Single-locus-based test

The multivariate score test involves the use of a possibly large covariance matrix, which may cause problems. A single-locus-based test would test individual SNPs one-by-one, e.g., by univariate score tests, then combine the individual tests. The most popular and simplest is the so-called univariate minP (UminP) test that takes the minimum p-value of the individual tests, and then do a multiple testing adjustment. Equivalently, one can combine individual score test statistics by their maximum:

$$T_{UminP} = \max_{1 \leq j \leq k} U_j^2/v_j, \quad (4)$$

where U_j is the j th component of U and $v_j = \text{Var}(U_j)$ is the j th diagonal element of V .

A multiple test adjustment based on either permutation or the Bonferroni method is commonly used. Because the Bonferroni adjustment is known to be conservative, it

is more common to use a permutation method by permuting Y . Here we propose using a faster simulation based approach (Seaman and Muller-Myhsok 2005; Chapman and Whittaker 2008) to estimate the asymptotic null distribution. First, because the asymptotic null distribution of U is known to be $U \sim N(0, V)$, we simulate B iid copies of $U^{(b)}$'s from $N(0, V)$ with $b = 1, 2, \dots, B$. Second, based on each $U^{(b)}$, we calculate $T_{UminP}^{(b)} = \max_{1 \leq j \leq k} U_j^{(b)2} / v_j$. Third, the p-value of the UminP test is simply $\sum_{b=1}^B I[T_{UminP} < T_{UminP}^{(b)}] / B$. We used $B = 1000$ throughout.

In our experience, the above simulation-based method is much faster than the permutation-based method. In addition to permuting the data, the permutation method requires refitting the model or re-calculating the test statistic repeatedly from each permuted dataset, which is typically more time-consuming than drawing a random sample from the null distribution of the test statistic. With the score test here, due to the available closed-form of the score statistic U (2) and its covariance matrix V 's invariance to permutation, the difference in computing time between the two methods is not dramatic, though in a simulation to be discussed later, the simulation method was 50% and 75% faster than the permutation method for sample sizes $m = 1000$ and $m = 2000$ respectively. If the Wald test is used, the difference will be much larger: the permutation method requires repeatedly fitting a regression model, each in an iterative algorithm to estimate β , while one only needs to fit the model once for the simulation method.

2.2.4 Sum of squared score tests: SSU and $SSUw$

Next we describe two tests based on the multivariate score statistic. The key feature of the two tests is their ignoring the full or non-diagonal elements of the covariance matrix of the score statistic. The first one is to simply use the sum of the squared scores as the test statistic

$$T_{SSU} = U'U. \tag{5}$$

This test is equivalent to the permutation-based version of the test of Goeman et al (2006). Although the Goeman test was derived under an empirical Bayes framework to test on a large number of parameters, as arising in microarray gene expression data, Chapman and Whittaker (2008) and Pan (2009) found that the Goeman test worked impressively well across a wide range of scenarios for the main-effects logistic model (1).

A weighted form of the SSU test is the SSUw test with

$$T_{SSUw} = U' \text{Diag}(V)^{-1} U = \sum_{l=1}^k U_j^2 / v_j. \quad (6)$$

Hence, compared to the UminP test that takes the maximum of the individual univariate score test statistics, the SSUw takes their sum (or equivalently, average). The SSUw test can be interpreted as an *estimated* most powerful test (Pan 2009), which also partially explains the good performance of SSU. Often SSU and SSUw perform similarly, but for some situations SSUw can be more or less powerful as to be shown here.

Asymptotically, each of the two test statistics is a quadratic form of Normal variates, $Q = U'W^{-1}U$, with $W = I$ or $W = \text{Diag}(V)$ respectively. It is well known (e.g. Johnson and Kotz, 1970, p.150) that the distribution of Q is a weighted sum of k independent chi-squared variates with DF=1, $\sum_{j=1}^k c_j \chi_1^2$, where c_j 's are the eigen values of VW^{-1} . Furthermore, by the results of Zhang (2005), $\sum_{j=1}^k c_j \chi_1^2$ can be well approximated by $a\chi_d^2 + b$ with

$$a = \frac{\sum_{j=1}^k c_j^3}{\sum_{j=1}^k c_j^2}, \quad b = \sum_{j=1}^k c_j - \frac{\left(\sum_{j=1}^k c_j^2\right)^2}{\sum_{j=1}^k c_j^3}, \quad d = \frac{\left(\sum_{j=1}^k c_j^2\right)^3}{\left(\sum_{j=1}^k c_j^3\right)^2}.$$

To calculate a p-value, for example, for the SSU test, we use

$$Pr(T_{SSU} > s | H_0) \approx Pr(\chi_d^2 > (s - b)/a).$$

2.3 Statistical Tests Based on Contrasting LD Patterns

Genetic associations can be revealed based on contrasting the LD patterns, e.g. measured by composite correlation matrices, between the case and control groups (Abecasis and Cookson 2000; Hayes 2004; Nielsen et al 2004; Schaid 2004; Zaykin et al 2006). For the dosage coding, the composite correlation between two loci can be estimated by their Pearson correlation coefficient. Suppose that the Pearson correlation matrices for the control and case groups are $R_0 = (r_{jl}^0)$ and $R_1 = (r_{jl}^1)$ respectively; for example,

$$r_{jl}^0 = \frac{\sum_{i=1}^{n_0} (X_{ij} - \bar{X}_{.j}^0)(X_{il} - \bar{X}_{.l}^0)}{s_j^0 s_l^0},$$

where $\bar{X}_{.j}^0 = \sum_{i=1}^{n_0} X_{ij}/n_0$ and s_j^0 are the sample mean and sample standard deviation of the genotype scores at locus j for controls, respectively. An LD contrast (LDC) test is constructed with

$$T_{LDC} = \text{Trace}((R_1 - R_0)'(R_1 - R_0)), \quad (7)$$

and its p-value is obtained based on permutation by shuffling the disease labels Y_i 's.

Wang et al (2007) proposed a modified LDC (mLDC) test to better account for background LDs. The method works in the following steps. First, a linear mixed-effects model is fitted for the case and control groups separately. For example, for controls, we have

$$X_{ij} = \gamma_j + b_i + e_{ij}, \quad i = 1, \dots, n_0 \text{ and } j = 1, \dots, k,$$

where γ_j is a fixed effect of locus j , $b_i \sim N(0, \sigma_b)$ is an iid random effect for subject i , and $e_{ij} \sim N(0, \sigma_e)$ is an iid random error term. Second, based on the above fitted model, we obtain the best linear unbiased predictor (BLUP) of X_{ij} , say \hat{X}_{ij} . Then we form the BLUP-corrected residuals, $z_{ij} = X_{ij} - \hat{X}_{ij}$. Note the different use of the sample mean across all the subjects in the control or case group in the LDC test.

Third, we form the cross-products of the residuals:

$$\lambda_{jl}^0 = \sum_{i=1}^{n_0} z_{ij} z_{il} = \sum_{i=1}^{n_0} (X_{ij} - \hat{X}_{ij})(X_{il} - \hat{X}_{il}), \quad \lambda_{jl}^1 = \sum_{i=n_0+1}^m z_{ij} z_{il} = \sum_{i=n_0+1}^m (X_{ij} - \hat{X}_{ij})(X_{il} - \hat{X}_{il}),$$

and their corresponding matrices $\Lambda^0 = (\lambda_{ji}^0)$ and $\Lambda^1 = (\lambda_{ji}^1)$. Finally the test statistic is

$$T_{mLDC} = \text{Trace} \left((\Lambda^1 - \Lambda^0)' (\Lambda^1 - \Lambda^0) \right). \quad (8)$$

Again permutation is used to obtain a p-value.

2.3 Testing on Differences of Means and of Covariance Matrices

It can be shown that the generalized Hotelling's T^2 test assesses the mean difference of the genotype scores between the control and case groups, while the LDC test compares their correlation matrices. As noted by Zaykin et al (2006), rather than the Pearson correlation matrix, the covariance matrix of each group can be also used, though the latter did not work as well as the former. Based on this observation, Wang et al (2009) proposed a Normal-based likelihood ratio test (LRT) to compare the mean and covariance matrix differences simultaneously. Specifically, suppose that the pooled and group-specific sample means are

$$\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i, \quad \bar{X}^0 = \frac{1}{n_0} \sum_{i=1}^{n_0} X_i, \quad \bar{X}^1 = \frac{1}{n_1} \sum_{i=n_0+1}^m X_i,$$

and the sample covariance matrices are

$$S = \frac{1}{m} \sum_{i=1}^m (X_i - \bar{X})'(X_i - \bar{X}),$$

$$S_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} (X_i - \bar{X}^0)'(X_i - \bar{X}^0), \quad S_1 = \frac{1}{n_1} \sum_{i=n_0+1}^m (X_i - \bar{X}^1)'(X_i - \bar{X}^1).$$

Then, the test statistic is

$$T_{LRT} = m \log |S| - n_0 \log |S_0| - n_1 \log |S_1|. \quad (9)$$

A permutation procedure is used to obtain its p-value.

Wang et al (2009) also proposed using principle component analysis (PCA) to reduce the dimension of the genotype vector X_i . Specifically, one applies PCA to the

pooled sample and retains only the first k_1 PCs explaining at least 85% of the total variation. Then the LRT is applied to the k_1 PCs as before. The resulting test is called LRT_{pc}.

As a result of dimensional reduction, the LRT_{pc} test has a smaller DF and thus possibly higher power than the LRT. Furthermore, for some data, the genotype scores of some linked SNPs may be linearly dependent, leading to a singular covariance matrix and consequently non-applicability of the LRT.

2.4 A Unified Framework: Logistic Regression

There are several potential drawbacks with the LRT and LRT_{pc} tests of Wang et al (2009). First, the tests use sample covariance matrices, not Pearson's correlation matrices (to estimate composite correlations) as preferred by Zaykin et al (2006). More importantly, it is not clear how their tests can use BLUP-corrected residuals as in the mLDC test advocated by Wang et al (2007). As shown by Wang et al (2007, 2009), the mLDC test performed better than the LDC test, providing a compelling reason to use BLUP-corrected residuals, not the sample covariance matrices. Second, the LRT and LRT_{pc} tests are based on the Normality assumption on the distribution of the genotype scores. Given each genotype score at any locus has only three possible values, its distribution cannot be Normal. Although the non-Normality does not automatically render the LRT and LRT_{pc} tests invalid, their power can be negatively influenced by this incorrect working assumption. Third, the (asymptotic) null distributions of the two test statistics are unknown, hence computationally demanding permutations have to be used to calculate p-values. Since for each permuted dataset, it is time-consuming to calculate the sample covariance matrices and their determinants, the LRT and LRT_{pc} are much slower than any of the other tests considered in this article. Here we provide an alternative that overcomes the above three shortcomings of the LDC and LDC_{pc} tests. The main idea is to introduce some relevant terms into a logistic regression model, and the corresponding score statistic will auto-

matically contrast the means of these terms, including both genotype scores and LD patterns, between the case and control groups. Accordingly, not just a single test, but a class of tests as for the main-effects logistic model (1) can be constructed. First, we consider a simple and straightforward way by incorporating two-way interactions into a Full logistic regression model:

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + \sum_{j=1}^k X_{ij}\beta_j + \sum_{j \neq l} X_{ij}X_{il}\beta_{j,l}, \quad (10)$$

and test a new null hypothesis $H_{0,2}$: $\beta_1 = \dots = \beta_k = \beta_{1,2} = \dots = \beta_{k-1,k} = 0$ with any of the tests introduced in section 2.2, for all of which we can quickly calculate their p-values either by their asymptotic null distributions or by simulation.

As for a main-effect shown earlier, the score statistic for an interaction, say for $\beta_{1,2}$, is

$$U_{1,2} = n_1 n_0 \left(\sum_{i=n_0+1}^m X_{i1}X_{i2}/n_1 - \sum_{i=1}^{n_0} X_{i1}X_{i2}/n_0 \right) / m,$$

which contrasts the mean cross-products between the two groups in the same spirit of the LDC test. Conceptually, the idea is related to using an LD measure to represent interactions between two unlinked loci (Zhao et al 2006). More generally, rather than using the two-way interactions, we can add various LD measurements, e.g. the cross-products of the BLUP-corrected residuals into an LD+main-effects-logistic regression model:

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + \sum_{j=1}^k X_{ij}\beta_j + \sum_{j \neq l} z_{ij}z_{il}\beta_{j,l}. \quad (11)$$

and test $H_{0,2}$: $\beta_1 = \dots = \beta_k = \beta_{1,2} = \dots = \beta_{k-1,k} = 0$ with any of the tests introduced in section 2.2. It is noted that the score statistic $U = (U'_\beta, U'_{\beta\beta})'$ contains two parts: the first part U_β is for main effects parameters $(\beta_1, \dots, \beta_k)'$, comparing the mean difference of the genotype scores between the two groups, and the second part $U_{\beta\beta}$ for cross-product terms contrasts the LD patterns.

We can also construct tests based on contrasting LD patterns only, for example,

with an LD-only logistic regression model:

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + \sum_{j \neq l} z_{ij} z_{il} \beta_{j,l}. \quad (12)$$

and test $H_{0,3}$: $\beta_{1,2} = \dots = \beta_{k-1,k} = 0$. In particular, the test statistic T_{SSU} for $H_{0,3}$ is the same as T_{mLDC} (up to a constant), but they differ in how to approximate their null distributions: the former is based on its asymptotic distribution while the latter is based on computationally more intensive permutations. Regardless of their different ways in estimating the null distributions, the analogy between our proposed tests based on logistic regression and the LDC/mLDC tests is the same as that between the logistic regression-based score test and the generalized Hotelling's T^2 test, as discussed before.

An advantage of our proposed logistic regression framework is that it provides flexibility in incorporating various ways to combine the information in genotype scores and LD patterns. Not only various tests can be directly applied to test $H_{0,2}$ in the LD+main-effects model (11), but also we can combine several tests, some of them test $H_{0,1}$ in the main-effects model (1) to contrast genotype scores while others test $H_{0,3}$ in the LD-only model (12) to contrast LD patterns. There is also flexibility in how to combine. Here for illustration, we will use the minP method that takes the minimum of the p-values from the two SSU tests applied to models (1) and (12) respectively, then use the Bonferroni correction for multiple testing, as suggested by a reviewer. Specifically, if the p-values from the SSU tests applied to the two models are $P_{SSU,Main}$ and $P_{SSU,LD}$, then the p-value of the minP test is

$$P_{minP} = 2 \min(P_{SSU,Main}, P_{SSU,LD}). \quad (13)$$

Note that it seems possible to apply the simulation method as discussed in section 2.2.3 to estimate the null distribution of the minP method to avoid the conservativeness of the Bonferroni correction, and it is under our current investigation.

A summary of the various methods is offered in Table 1. For our proposed logistic regression, each of the listed tests in A) can be applied to each model listed in B).

3. RESULTS

3.1 Simulation I: Haplotype Effect Model

3.1.1 Simulation Set-ups

We performed a simulation study following the set-ups given in Wang et al (2009) with $k = 10$ marker SNPs and $n_1 = 500$ cases and $n_0 = 500$ controls. First, we generated a latent vector from a multivariate Normal distribution with a compound symmetry (CS) correlation structure with correlation coefficient ρ_0 ; that is, the correlation between the two latent variables at any two loci j and l is $\rho_{jl} = \rho_0$. Second, the latent vector was dichotomized to yield a haplotype with minor allele frequencies (MAFs) at various loci randomly chosen between 0.1 and 0.4. Third, for two independently generated haplotypes $h_1 = s_{1,1}s_{1,2}\dots s_{1,10}$ and $h_2 = s_{2,1}s_{2,2}\dots s_{2,10}$ for any subject, if we coded $s_{t,j} = 1$ or 0 for the minor or major allele at any locus j , the non-null haplotype effects on a continuous liability trait y^* were $g_t = |\sum_{j=1}^{10} s_{t,j} - 5|/5$ for $t = 1$ and 2; specifically, we had $y^* = g_1 + g_2 + e$ for a non-null model, and $y^* = e$ for a null-model, where $e \sim N(0, \sigma)$ is a random error, and σ controlled noise level in the data. Fourth, a disease status Y was generated as the binary indicator $I(y^* > 1.64)$. For each set-up, we simulated 1000 datasets, from which we obtained an empirical size or power for each test as its proportion of incorrectly or correctly rejecting its null hypothesis respectively. With 1000 independent replications, the Monte Carlo standard error of an empirical size or power, say \hat{p} , is $\sqrt{\hat{p}(1 - \hat{p})/1000} \leq 0.016$.

3.1.2 Simulation Results

Type I error rate: For the null models with various values of ρ_0 , all the tests except the multivariate score test based on the LD-only or LD+main-effects-logistic

model maintained their Type I error rates close to the nominal significance level at 0.05 (Tables 2-3), but the above two score tests could have much larger Type I error rates (not shown). For example, for the LD-only model, the test sizes of the score test increased from 0.067 for $\rho_0 = 0$ to 0.107 for $\rho_0 = 0.8$. For their inflated Type I error rates, we will not consider the two score tests in the LD-only and LD + Main-effects models in the following comparisons.

Power: We can draw the following conclusions consistently across all scenarios with three noise levels σ . First, we consider how power changed with correlation coefficient ρ_0 . For the LRT, LRTpc and various LDC-type tests, the power increased with ρ_0 , reflecting increasing LD differences. In contrast, for other tests comparing genotype scores with or without contrasting LD patterns between the two groups, as ρ_0 was increased, their power first increased, then decreased. Second, for a small ρ_0 (e.g. $\rho_0 = 0$), the SSU test (closely followed by SSUw) based on the Full-logistic model was the overall winner; for an intermidate ρ_0 , the SSU (or SSUw) tests based on either the Full model, Main-effects model, or LD+main-effects model performed similarly and were best, but as ρ_0 was increased, the SSU test based on the LD+main-effects model was the winner; for a large ρ_0 , the SSU test based on the LD-only logistic model had the highest power. Third, between the LRT and LRTpc tests, the former won for a large ρ_0 while the latter was the winner for a small ρ_0 . Fourth, as shown by Wang et al (2007), the mLDC test consistently performed better than the LDC test, though the SSU test in the LD-only model was much better than the mLDC test. *Fifth, for a smaller ρ_0 , the SSU test in the LD+main-effects model had an edge over the minP test, but as ρ_0 increased, the power of the latter exceeded that of the former.*

In summary, when LD was weak, the information about the trait-genotype association was mainly contained in genotype score differences, for which the SSU or SSUw test based on a logistic regression model with main-effects (and possibly other

high-order terms) worked best. *On the other hand, as the strength of LD was increased, there was a stronger difference in LD patterns, but not in genotype scores, between the two groups, for which case the SSU or SSUw test based on either LD-only or LD+main-effects logistic regression was the most powerful.*

To gauge the overall performance, we averaged the power of seven representative tests over the values of ρ_0 (Fig 1). For each noise level σ , the minP test was the overall winner, followed closely by the SSU test in the LD+main-effects model, and then by the SSU test in the main-effects model.

3.2 Simulation II: Two-locus Disease Model

3.2.1 Simulation Set-up

To mimic real LD patterns, we used haplotype data of Wang et al (2009), containing two segments on chromosome 21 based on the CEU HapMap samples (Thorisson et al 2005). Wang et al (2009, Table I) listed the haplotypes and their frequencies in the two gene regions, each with 7 SNPs. The first region included 16 different haplotypes, of which three haplotypes had a cumulative frequency of 73%. The second one had only 13 haplotypes with the two most frequent having a cumulative frequency of 69%. We combined two independently sampled haplotypes to obtain a vector of marker genotypes for each subject. As in Wang et al (2009), we used the third SNP in each of the two regions as the causal SNPs, say u_1 and u_2 , with MAF equal to 0.14 and 0.066 respectively, and simulated a binary trait using a dominant genetic model:

$$\text{Logit Pr}(Y_i = 1) = \theta_0 + \theta_1 I_{i,u_1} + \theta_2 I_{i,u_2} + \theta_{12} I_{i,u_1} I_{i,u_2}, \quad (14)$$

where I_{i,u_1} and I_{i,u_2} were binary indicators of the presence of the minor alleles in the two disease-causing SNPs for subject i .

For the null model, we used $\theta_1 = \theta_2 = \theta_{12} = 0$. We considered four types of non-null models (Chatterjee et al 2006): i) a purely epistatic model with $\theta_1 = \theta_2 = 0$ and $\theta_{12} = \theta > 0$; ii) an additive model for odds ratio (OR) with $\theta = (\theta_1, \theta_2)$ and $\theta_{12} = \theta_1 + \theta_2 + \log(e^{\theta_1} + e^{\theta_2} - 1)$; iii) a multiplicative OR model (i.e. a main-effects model

with only non-zero marginal effects and zero interacting effects) with $\theta_1 = \theta_2 = \theta > 0$ and $\theta_{12} = 0$; iv) a cross-over model with $\theta_1 = \log(0.9)$, $\theta_2 = 0$ and $\theta_{12} = \theta > 0$. For all the models, we used $\theta_0 = -\log(9) \approx -2.2$. For each type of the true genetic model, we varied the value of θ .

We excluded the two causal SNPs from the observed data, and combined the remaining 12 marker SNPs as in a candidate region. We used a sample size of $n_1 = 500$ cases and $n_0 = 500$ controls for each simulated dataset; 1000 replicates were used for each simulation set-up, from which we obtained an empirical size or power for each test as its proportion of incorrectly or correctly rejecting its null hypothesis.

3.2.2 Simulation Results

Type I error rate: For the null model, each of the tests seemed to maintain a correct test size: the Type I error rates were always close to the nominal significance level at 0.05 (Tables 4-5),

Power: First, the tests based on only contrasting LD patterns appeared to be always low powered, while those based on (only or partly) comparing mean genotype scores seemed to have higher power, suggesting that the information content for genetic associations was mainly contained in mean genotype scores. Second, for the purely epistatic model and cross-over model, the SSUw test based on the Full model was the overall winner, closely followed by the UminP test based on the Full model. On the other hand, for the other two underlying genetic models, the multivariate score test based on the main-effects model seemed to be the overall winner, followed by the SSUw and possibly SSU tests. Third, although the tests based on LD+main-effects model were not most powerful, they had much higher power than the LDC and mLDC tests, and than the LRT and LRTpc tests. *Fourth, between the SSU test for the LD+main-effects model and the minP test, neither dominated the other: each performed better than the other for two genetic models.*

In overall, averaged over the various parameter values in each genetic model, the

power of each of seven representative tests is shown in Fig 2. It is clear that, due to the presence of interactions, the SSU test for the Full model was the overall winner, closely followed by the SSU for the main-effects model, the minP test or the SSU for the LD+main-effects model.

3.3 Simulation III: HapMap data for gene CHI3L2

3.3.1 Simulation Set-up

We conducted a simulation study based on real LD patterns within the CHI3L2 gene as observed in the HapMap data. We downloaded the SNPs of the CHI3L2 gene for the 90 CEU (Utah residents with ancestry from northern and western Europe) individuals from the HapMap site in June 2008. As in Wang and Elston (2007), first, we excluded SNPs with $MAF \leq 0.2$, leaving 23 SNPs. Second, we did a single imputation for each of the missing genotypes by randomly drawing an observed genotype at the same locus. Third, we deleted redundant SNPs that were perfectly correlated with other SNPs, leading to 17 remaining SNPs. Fourth, we repeatedly sampled (with replacement) subjects from the 90 CEU individuals. As Wang and Elston, we chose the SNP rs2182114 as disease-causing: if its genotype score was X_{i0} for subject i , we generated a disease indicator Y_i from a logistic regression model

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + \log(OR)X_{i0}, \quad (15)$$

where we chose $\beta_0 = -\log 4$ to give a background (i.e. not caused by the SNP) disease probability of 0.2, and the association strength between the disease and causal SNP was reflected by the odds ratio (OR), which ranged from 1 (i.e. no association) to 1.5. Finally, following the case-control design, we sampled $n_1 = 500$ cases (with $Y_i = 1$) and $n_0 = 500$ controls (with $Y_i = 0$). We excluded the disease-causing SNP when applying various statistical tests. For each set-up, we simulated 1000 datasets.

3.3.2 Simulation Results

Type I error rate: For the null model with $OR=1$, each of the tests seemed to

maintain correct test size with the Type I error rate close to the nominal significance level of 0.05 (Tables 6), Due to the linear dependence among the SNPs, no result was available for the Normal-based LRT method.

Power: The Full logistic regression model gave the highest power, closely followed by the Main-effects model and LD+Main-effects model. Although there were no explicit interaction terms in the true model (15), due to the connection between an interaction and LD (Zhao et al 2006), the Full model, as well as the LD+Main-effects model, performed well for capturing distributional differences in both genotype scores and LD patterns between the two groups. The Main-effects model also worked well, presumably because the main distributional difference was in genotype scores and it gained power from its small DF. The LDC-type tests all had low power, though the mLDC test and the LD-only model performed better than the LDC test. Across the various values of OR, the relative performance of the methods was consistent, as shown in Fig 3.

4. DISCUSSION

We have proposed a general framework based on logistic regression to unify two general approaches to detecting genetic association based on unphased genotype data: rather than choosing between comparing genotype scores and contrasting LD patterns between the case and control groups, as implemented in the Hotelling's T^2 test and LDC test respectively, we can construct tests under this unified framework to *simultaneously* assess the two aspects of the distributional difference. The main idea is to incorporate some LD measurements, as those for genotype scores, as covariates in a logistic regression model. In this way, we can take advantage of the score statistic and its associated nice asymptotic properties, to construct a class of suitable tests, such as the SSU, SSUw and UminP tests, in addition to the standard multivariate score test, whose null distributions can be easily estimated based on the asymptotic Normality of the score statistic. Even if one is only interested in contrasting LD patterns between

the case and control groups, logistic regression offers a general framework to construct some alternative LDC-type tests. For example, rather than taking the simple average of the LD differences as in the LDC, mLDC and our proposed SSU tests, we can take a weighted average as in the SSUw test, or take the most significant difference as in the UminP test, to summarize the distributional difference between the two groups. As shown for contrasting genotypes scores, the above alternative tests can yield higher power under some situations (Chapman and Whittaker 2008; Pan 2009). Importantly, due to the nice asymptotic properties of the score vector for logistic regression (or any generalized linear model, GLM), it is easy to derive the asymptotic null distributions and thus calculate the p-values for these alternative tests. The logistic regression (or more generally GLM) framework not only makes it straightforward to accommodate other phenotypes (e.g. quantitative traits) and incorporate other covariates, such as environmental or other risk factors, but also facilitates the extension of the proposed tests to other robust non-likelihood-based estimation methods (Zhao et al 2003) or other study designs, e.g. family-based association studies (e.g. Fan and Xiong 2003; Baksh et al 2005), and to combining population- and family-based association studies (Chen and Lin 2008; Hsu et al 2009).

Our numerical results have confirmed the good performance of the proposed tests. As expected, if the distributional difference between cases and controls lies in both genotype scores and LD patterns (while neither aspect dominates the other), the SSU test based on an LD+main-effects logistic regression model was most powerful. On the other hand, when distributional difference is mainly in only one of the two aspects, the SSU (or SSUw) test based on one of the two models, a main-effects (for genotype scores) or an LD-only (for LD patterns) logistic model, was most powerful, while that based on the other model may be quite low powered. In contrast, under these situations, although the SSU (or SSUw) test based on an LD+main-effects logistic model may not be most powerful, it still maintains high power. In particular, in all

our examples, one of the SSU and SSUw tests based on an LD+main-effects logistic regression model was always more powerful than the LDC and mLDC tests, and than two recently proposed Normal-based LRT and LRTpc tests that compare the mean and covariance differences between the two groups. Furthermore, our proposed tests are computationally much faster than the four permutation-based tests.

Although we have focused on the application of any association test to a candidate gene or region, we may extend the methodology to genome-wide association studies. One strategy is to apply some existing algorithms to detect haplotype blocks as defined as chromosome regions with SNPs in high LD, then apply an association test to each of the haplotype blocks (Cardon and Abecasis 2003). A potential drawback is that haplotype blocks may not be uniquely defined or detected by different methods. An alternative is to apply an association test to several fixed-sized sliding windows along a chromosome (Durrant et al 2004), or to all possible varying-sized windows with a small preset maximum number of SNPs inside (Cheng et al 2005). Guo et al (2009) showed that, though computationally quite demanding, exploring all possible sliding-window sizes could improved statistical power over that based on only haplotype blocks or single SNP loci. In any case, our proposed tests can be coupled with a haplotype-block searching or sliding-window strategy for application to genome-wide association studies.

Acknowledgement

The author is grateful to the reviewers for extremely helpful and constructive comments. The author thanks Dr Tao Wang for generously sharing his R code for the mLDC test, and Dr Xuexia Wang for clarifications on the LRT/LRTpc tests. This research was partially supported by NIH grants GM081535 and HL65462.

REFERENCES

Abecasis GR, Cookson WO (2000) GOLD—graphical overview of linkage disequilibrium.

Bioinformatics 16:182-183.

Altshuler D, Daly M, Lander ES (2008). Genetic mapping in human disease. *Science* 322:881-888.

Baksh MF, Balding DJ, Vyse TJ, Whittaker JC (2005) A likelihood ratio approach to family-based association studies with covariates. *Annals of Human Genetics* 70:131-139.

Cardon LR, Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. *Trends Genet* 19:135-140.

Chapman JM, Whittaker J (2008) Analysis of multiple SNPs in a candidate gene or region. *Genetic Epidemiology*, Published Online: 21 Apr 2008.

Chatterjee N, Kalalioglu Z, Moslehi R, Peters U, Wacholder S (2006). Powerful multi-locus tests of genetic association in the presence of gene-gene and gene-environment interactions. *Am J Hum Genet* 79:1002-1016.

Chen Y-H, Lin H-W (2008) Simple association analysis combining data from trios/sibships and unrelated controls. *Genet Epidemiol* 32:520-527.

Cheng R, Ma JZ, Elston RC et al. (2005) Fine mapping functional sites or regions from case-control data using haplotypes of multiple linked SNPs. *Ann Hum Genet* 69:102-112.

Clayton D, Chapman J, Cooper J (2004) Use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol* 27:415-428.

Durrant C, Zongdervan KT, Cardon LR et al. (2004) Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am J Hum Genet* 75:35-43.

Fan R, Knapp M (2003) Genome association studies of complex diseases by case-control designs. *Am J Hum Genet* 72:850-868.

- Fan R, Xiong M (2003) Linkage and association studies of QTL for nuclear families by mixed models. *Biostatistics* 4:75-95.
- Goeman JJ, van de Geer S, van Houwelingen HC (2006). Testing against a high dimensional alternative. *J R Stat Soc B* 68:477-493.
- Guo Y, Li, J, Bonham A, Wang Y, Deng H (2009). Gain in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: a comparison of association-mapping strategies. To appear *European Journal of Human Genetics*. Published on-line.
- Hayes MG, del Bosque-Plata L, Tsuchiya T, Hanis CL, Cox NJ (2005). Patterns of linkage disequilibrium in the type 2 diabetes gene calpain-10. *Diabetes* 54:3573-3576.
- Hsu L, Starr JR, Zheng Y, Schwartz SM (2009) On combining triads and unrelated subjects data in candidate gene studies: An application to data on testicular cancer. *Hum Hered* 67:88-103.
- Johnson NL, Kotz S (1970) *Distributions in Statistics, Continuous Univariate Distributions*, vol 2. Boston: Houghton-Mifflin.
- Lin, DY (2005) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* 21:781-787.
- Nielsen DM, Ehm MG, Zaykin DV, Weir BS (2004) Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. *Genetics* 168:1029-1040.
- Pan, W. (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. To appear in *Genetic Epidemiology*. Published Online on Jan 23 2009. DOI: 10.1002/gepi.20402.
Available at www.biostat.umn.edu/rrs.php as Research Report 2008-018, Division of Biostatistics, University of Minnesota.

- Plenge RM et al. (2007) TRAF1-C5 as a risk locus for rheumatoid arthritis—a genomewide study. *N Engl J Med* 357:1199-209.
- Roeder K, Bacanu SA, Sonpar V, Zhang X, Devlin B (2005). Analysis of single-locus tests to detect gene/disease associations. *Genet Epidemiol* 28: 207-219.
- Seaman SR, Muller-Myhsok B (2005) Rapid simulation of p values for product methods and multiple testing adjustment in association studies. *Am J Hum Genet* 76:399-408.
- Schaid DJ (2004) Linkage disequilibrium testing when linkage phase is unknown. *Genetics* 166:505-512.
- Wang T, Elston RC (2007) Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am J Hum Genet* 80:353-360.
- Wang T, Zhu X, Elston RC (2007) Improving power in contrasting linkage-disequilibrium patterns between cases and controls. *Am J Hum Genet* 80:911-920.
- Wang X, Zhang S, Sha Q (2009). A new association test to test multiple-marker association. *Genetic Epidemiology* (in press).
- Xiong M, Zhao J, Boerwinkle E (2002) Generalized T^2 test for genome association studies. *Am J Hum Genet* 70:1257-1268.
- Zaykin DV, Meng Z, Ehm MG (2006) Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. *Am J Hum Genet* 78:737-746.
- Zhang J-T (2005) Approximate and asymptotic distributions of Chi-squared-type mixtures with applications. *Journal of the American Statistical Association* 100:273-285.
- Zhao LP, Li S, Khalid N (2003) A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *Am J Hum Genet* 72:1231-1250.

Table 1: A summary of various methods. Column Eq gives the equation numbers in the text, and column p-value indicates whether a p-value is calculated based on the permutation, simulation, Bonferroni adjustment, or asymptotic distribution.

Methods	Acronym	Eq	Reference	Contrast		p-value
				Genotype	LD	
<i>Logistic regression-based:</i>						
A) Tests:			Pan (2009)			
A1. Sum of squared score	SSU	(5)				Asympt
A2. Weighted SSU	SSUw	(6)				Asympt
A3. Single-locus score	UminP	(4)				Simul
A4. Multivariate score	Sco	(3)				Asympt
B) Models:			Here			
B1. Main-effects		(1)		Yes	No	
B2. Full		(10)		Yes	Yes	
B3. LD-only		(12)		No	Yes	
B4. LD+Main-effects		(11)		Yes	Yes	
C) Combining A1 applied to B1 & B3	minP	(13)	Here	Yes	Yes	Bonfer
LD contrast test	LDC	(7)	Zaykin et al (2006)	No	Yes	Permut
Modified LD contrast test	mLDC	(8)	Wang et al (2007)	No	Yes	Permut
Normal likelihood ratio tests:			Wang et al (2009)			
1) with original SNPs	LRT	(9)		Yes	Yes	Permut
2) with principle components	LRTpc			Yes	Yes	Permut

Zhao J, Jin L, Xiong M (2006). Test for interaction between two unlinked loci. *Am J Hum Genet* 79:831-845.

Table 2: Empirical sizes and powers of various tests with nominal significance level 0.05 for simulation I. The latent variables for genotypes had a correlation structure of $CS(\rho_0)$ with #SNP=10, and sample size $n_0 = n_1 = 500$.

Model	ρ_0	LRT	LRTpc	Full model				Main-effects			
				SSU	SSUw	UminP	Sco	SSU	SSUw	UminP	Sco
Null	.0	.066	.064	.057	.055	.042	.059	.065	.060	.055	.060
	.2	.064	.071	.058	.053	.062	.043	.068	.073	.057	.059
	.4	.062	.066	.062	.063	.055	.060	.068	.062	.060	.067
	.6	.032	.051	.061	.057	.049	.039	.067	.067	.076	.052
	.8	.058	.064	.059	.056	.055	.046	.062	.061	.060	.068
$\sigma = 3.0$.0	.150	.171	.423	.405	.218	.113	.240	.238	.140	.228
	.2	.239	.268	.574	.540	.339	.134	.562	.543	.350	.322
	.4	.343	.320	.475	.420	.330	.151	.587	.557	.404	.310
	.6	.435	.405	.305	.248	.315	.169	.507	.463	.402	.296
	.8	.536	.466	.182	.141	.270	.143	.377	.302	.361	.320
$\sigma = 2.5$.0	.203	.260	.600	.583	.315	.159	.373	.363	.220	.344
	.2	.372	.429	.743	.717	.510	.209	.749	.742	.500	.509
	.4	.499	.494	.627	.575	.489	.215	.763	.744	.574	.461
	.6	.630	.582	.434	.377	.465	.235	.683	.640	.575	.424
	.8	.734	.665	.276	.216	.420	.217	.531	.471	.507	.498
$\sigma = 2.0$.0	.377	.474	.831	.814	.505	.281	.656	.647	.346	.612
	.2	.614	.680	.918	.899	.736	.397	.933	.932	.739	.764
	.4	.745	.739	.833	.784	.723	.394	.929	.922	.803	.714
	.6	.854	.827	.650	.591	.703	.393	.863	.843	.793	.683
	.8	.937	.879	.415	.345	.646	.388	.727	.677	.716	.758

Table 3: Empirical sizes and powers of various tests with nominal significance level 0.05 for simulation I. The latent variables for genotypes had a correlation structure of $CS(\rho_0)$ with #SNP=10, and sample size $n_0 = n_1 = 500$.

Model	ρ_0	LDC	mLDC	LD-only			LD + Main-effects			
				SSU	SSUw	UminP	SSU	SSUw	UminP	minP
Null	.0	.060	.040	.050	.053	.065	.063	.063	.060	.052
	.2	.066	.064	.061	.054	.065	.077	.062	.069	.068
	.4	.068	.072	.063	.058	.054	.063	.068	.058	.081
	.6	.042	.044	.042	.030	.055	.063	.055	.060	.053
	.8	.049	.049	.045	.046	.055	.060	.053	.051	.057
$\sigma = 3.0$.0	.054	.058	.067	.055	.057	.181	.122	.083	.182
	.2	.070	.125	.204	.145	.115	.557	.364	.180	.500
	.4	.120	.238	.374	.277	.161	.633	.523	.273	.624
	.6	.272	.405	.554	.422	.256	.577	.569	.349	.600
	.8	.436	.514	.639	.506	.310	.444	.555	.343	.609
$\sigma = 2.5$.0	.049	.061	.079	.067	.058	.294	.172	.103	.295
	.2	.065	.187	.332	.224	.150	.745	.565	.286	.697
	.4	.174	.382	.563	.425	.252	.812	.715	.408	.794
	.6	.372	.577	.751	.604	.397	.772	.764	.501	.853
	.8	.555	.701	.825	.715	.497	.615	.753	.539	.806
$\sigma = 2.0$.0	.056	.091	.125	.089	.094	.560	.340	.178	.562
	.2	.072	.347	.548	.397	.253	.937	.837	.496	.918
	.4	.259	.604	.793	.664	.437	.950	.914	.659	.940
	.6	.551	.814	.932	.836	.621	.919	.917	.744	.946
	.8	.768	.929	.971	.932	.755	.833	.947	.791	.965

Table 4: Empirical sizes and powers of various tests with nominal significance level 0.05 for simulation II. The genotypes were generated from haplotype frequencies of two regions based on the HapMap data with #SNP=12, and sample size $n_0 = n_1 = 500$.

Model	θ	LRT	LRT _{pc}	Full model				Main-effects			
				SSU	SSU _w	UminP	Sco	SSU	SSU _w	UminP	Sco
Null	0	.045	.040	.062	.054	.061	.031	.050	.057	.057	.040
Epist	1.59	.235	.078	.391	.496	.461	.250	.291	.279	.312	.427
	1.92	.398	.123	.637	.751	.712	.457	.490	.479	.493	.679
	2.17	.544	.169	.824	.888	.855	.646	.691	.652	.634	.836
	2.26	.611	.190	.867	.916	.888	.704	.755	.725	.695	.878
Addit	(.62, .80)	.362	.153	.747	.816	.722	.420	.698	.649	.613	.829
	(.62, .87)	.407	.173	.800	.862	.776	.468	.745	.700	.658	.872
	(.89, .80)	.544	.206	.936	.937	.897	.712	.923	.898	.864	.974
	(.89, .87)	.586	.231	.944	.949	.913	.736	.932	.910	.882	.979
Multi	1.6	.178	.086	.407	.445	.392	.187	.375	.326	.339	.516
	1.8	.285	.121	.697	.730	.624	.357	.653	.572	.562	.772
	2.0	.424	.179	.869	.892	.818	.545	.823	.793	.740	.924
	2.2	.583	.235	.947	.956	.917	.721	.929	.909	.872	.978
Cross	1.89	.330	.101	.512	.620	.644	.367	.377	.358	.413	.551
	2.16	.483	.140	.719	.815	.810	.577	.564	.546	.579	.751
	2.25	.546	.155	.786	.873	.858	.646	.641	.609	.630	.810
	2.40	.644	.190	.863	.925	.904	.742	.751	.734	.716	.877

Table 5: Empirical sizes and powers of various tests with nominal significance level 0.05 for simulation II. The genotypes were generated from haplotype frequencies of two regions based on the HapMap data with #SNP=12, and sample size $n_0 = n_1 = 500$.

Model	θ	LDC	mLDC	LD-only			LD + Main-effects			
				SSU	SSUw	UminP	SSU	SSUw	UminP	minP
Null	0	.045	.049	.050	.043	.051	.051	.042	.057	.044
Epist	1.59	.149	.146	.128	.111	.158	.239	.157	.206	.228
	1.92	.242	.246	.239	.204	.268	.450	.302	.386	.439
	2.17	.319	.341	.332	.283	.383	.629	.459	.534	.603
	2.26	.382	.407	.385	.346	.409	.701	.518	.584	.674
Addit	(.62, .80)	.122	.161	.168	.131	.188	.503	.263	.378	.585
	(.62, .87)	.139	.188	.191	.143	.213	.549	.286	.429	.635
	(.89, .80)	.151	.252	.275	.203	.236	.804	.479	.686	.870
	(.89, .87)	.163	.271	.304	.226	.265	.825	.515	.718	.932
Multi	1.6	.073	.102	.098	.076	.102	.245	.130	.183	.265
	1.8	.109	.142	.146	.112	.143	.436	.227	.337	.523
	2.0	.133	.209	.225	.163	.204	.647	.351	.504	.747
	2.2	.176	.265	.305	.220	.269	.823	.515	.676	.885
Cross	1.89	.213	.193	.189	.168	.227	.344	.246	.312	.308
	2.16	.302	.306	.301	.259	.357	.537	.392	.492	.504
	2.25	.344	.355	.345	.290	.406	.601	.459	.550	.567
	2.40	.416	.439	.412	.366	.484	.713	.548	.622	.679

Table 6: Empirical sizes and powers of various tests with nominal significance level 0.05 for simulation III. The genotypes were generated from haplotype frequencies of two regions based on the HapMap data with #SNP=16, and sample size $n_0 = n_1 = 500$.

OR	LRT _{pc}	Full model				Main-effects			
		SSU	SSU _w	UminP	Sco	SSU	SSU _w	UminP	Sco
1.0	.058	.043	.045	.058	.050	.054	.056	.052	.057
1.1	.129	.142	.149	.123	.060	.128	.133	.128	.080
1.2	.316	.395	.410	.349	.103	.370	.370	.333	.160
1.3	.571	.693	.713	.638	.209	.683	.687	.649	.346
1.4	.818	.897	.915	.871	.358	.896	.901	.883	.579
1.5	.944	.973	.981	.965	.561	.975	.975	.973	.794

OR	LDC	mLDC	LD-only			LD + Main-effects			
			SSU	SSU _w	UminP	SSU	SSU _w	UminP	minP
1.0	.051	.052	.048	.048	.059	.047	.052	.059	.046
1.1	.058	.085	.076	.066	.074	.119	.078	.086	.108
1.2	.070	.133	.132	.106	.129	.349	.178	.205	.307
1.3	.122	.219	.229	.185	.198	.634	.370	.433	.593
1.4	.192	.360	.368	.301	.303	.879	.578	.710	.845
1.5	.274	.520	.518	.443	.436	.969	.787	.891	.957

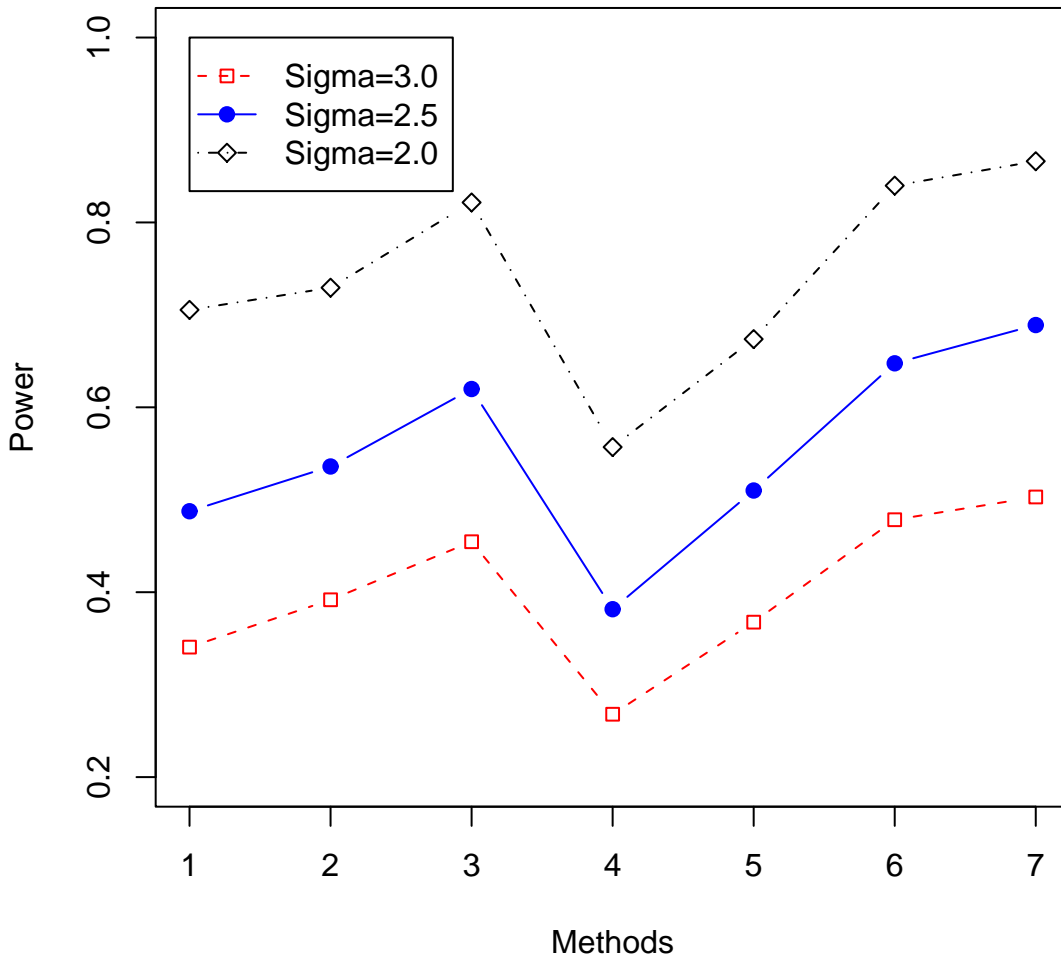


Figure 1: Average power in simulation I for each of methods 1-7, corresponding to the LRT, Full-SSU test, Main-SSU test, mLDC test, LD-only-SSU test, LD+Main-SSU test and minP test.

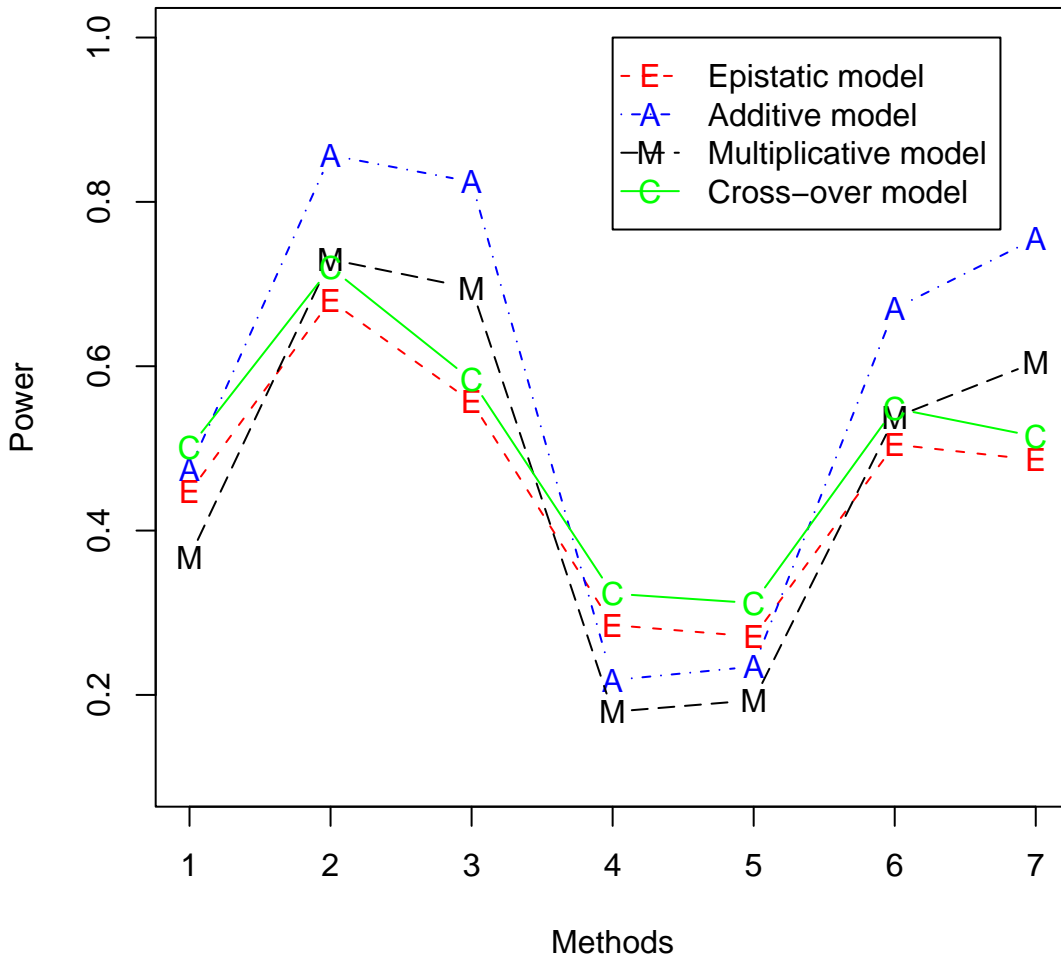


Figure 2: Average power in simulation II for each of methods 1-7, corresponding to the LRT, Full-SSU test, Main-SSU test, mLDC test, LD-only-SSU test, LD+Main-SSU test and minP test.

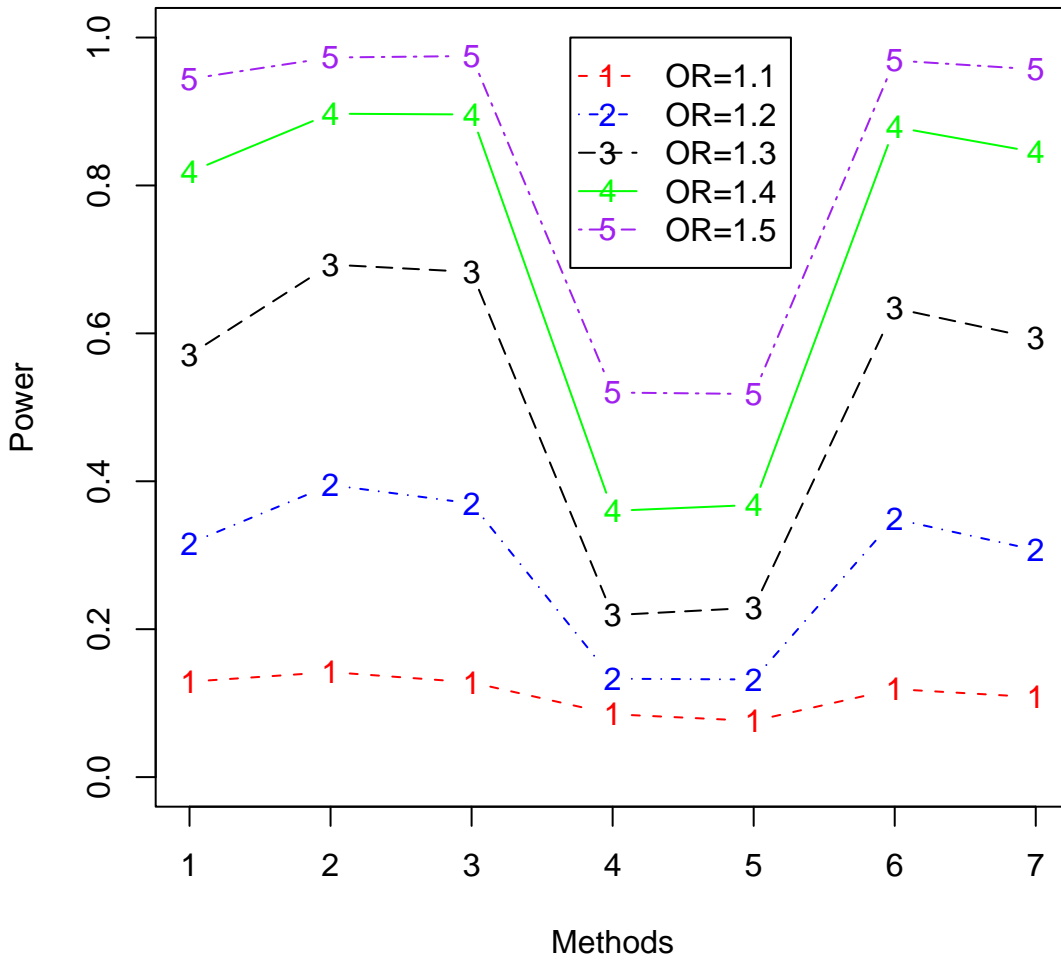


Figure 3: Power in simulation III for each of methods 1-7, corresponding to the LRTpc, Full-SSU test, Main-SSU test, mLDC test, LD-only-SSU test, LD+Main-SSU test and minP test.