Published in final edited form as:
Genet Epidemiol. 2010 November ; 34(7): 680–688. doi:10.1002/gepi.20529.

Powerful Multi-marker Association Tests: Unifying Genomic Distance-Based Regression and Logistic Regression

Fang Han and **Wei Pan**
Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455

Abstract

To detect genetic association with common and complex diseases, many statistical tests have been proposed for candidate gene or genome-wide association studies with the case-control design. Due to linkage disequilibrium (LD), multi-marker association tests can gain power over single-marker tests with a Bonferroni multiple testing adjustment. Among many existing multi-marker association tests, most target to detect only one of many possible aspects in distributional differences between the genotypes of cases and controls, such as allele frequency differences, while a few new ones aim to target two or three aspects, all of which can be implemented in logistic regression. In contrast to logistic regression, a genomic-distance based regression (GDBR) approach aims to detect some high-order genotypic differences between cases and controls. A recent study has confirmed the high power of GDBR tests. At this moment, the popular logistic regression and the emerging GDBR approaches are completely unrelated; for example, one has to choose between the two. In this article, we reformulate GDBR as logistic regression, opening a venue to constructing other powerful tests while overcoming some limitations of GDBR. For example, asymptotic distributions can replace time-consuming permutations for deriving p-values, and covariates, including gene-gene interactions, can be easily incorporated. Importantly, this reformulation facilitates combining GDBR with other existing methods in a unified framework of logistic regression. In particular, we show that Fisher's p-value combining method can boost statistical power by incorporating information from allele frequencies, Hardy-Weinberg disequilibrium (HWD), LD patterns and other higher-order interactions among multi-markers as captured by GDBR.

Keywords

Fisher's method; Genome-wide association study; GWAS; multi-marker analysis; score test; SNP

1. INTRODUCTION

With the completion of the first-wave GWASs, some lessons have been learned (Altshuler et al 2008). First, effect sizes for common variants are typically small to modest: often the estimated odds ratios are only from 1.1 to 1.4 with a mode at only 1.2 (Flint and Mackay 2009). Given a small effect size and the typical sample size of a few thousand individuals, the power of the standard single-marker test in most GWASs to detect weak association becomes low. Second, in spite of some successes of GWASs, for most common diseases the proportion of the overall phenotypic variance explained by discovered disease-susceptibility loci remains very low (Maher 2008). It is likely that only a small fraction of causal loci have been identified. Because most GWASs applied the univariate single-marker analysis with a

Correspondence author: Wei Pan, Telephone: (612) 626-2705, Fax: (612) 626-0660, weip@biostat.umn.edu, Address: Division of Biostatistics, MMC 303, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455–0392, U.S.A.

conservative Bonferroni adjustment for multiple testing, which may have low power, the development and application of more powerful statistical tests will increase the chance of discovering more disease loci. On the other hand, in spite of many existing multi-marker tests, most of which can be implemented in logistic regression, since there is no uniformly most powerful test, it has become difficult for a geneticist to choose a suitable test from many existing ones. At the same time, it has been increasingly recognized that many existing tests may be powerful in some situations but not in others, since they solely target to detect one or few aspects of genotypic distribution differences between cases and controls. It should be more productive to combine information and aim to detect multiple aspects of distributional differences. It is a main goal of this article to show that indeed it is possible to combine multiple types of information in multi-marker genotypes within a unified framework of logistic regression to construct powerful tests.

Although the most general approach to detecting genetic association for population-based case-control studies with unphased multi-marker genotype data is to compare the distributions of the genotypes between the case and control groups, as targeted by the B-statistic (Zhang and Liu 2007), due to the complexity of such multivariate discrete distributions, almost all existing methods, for simplicity and feasibility, aim to detect only one aspect of distributional differences. Below are five classes of approaches to detecting various targeted differences. As discussed by Won and Elston (2008, 2009), there are three types of information demonstrating genotypic differences between cases and controls: in allele frequencies, in parameters for Hardy-Weinberg disequilibrium (HWD), and in parameters of linkage disequilibrium (LD), each of which can be targeted to construct the corresponding class of association tests. The first class, perhaps most popular, aims to compare mean genotype scores, e.g. allele frequencies, between the case and control groups. This class includes some classic and most popular tests, such as Hotelling's T^2 test (Fan and Knapp 2002; Xiong et al 2003) and combining single-marker-based tests (Roeder et al 2005). Some emerging powerful tests, such as the sum of squared score (SSU) test (Pan 2009), also belongs to this class. It is worth noting that these tests can be implemented within the framework of logistic regression. The second class contrasts the HWD trends between affected and unaffected individuals (Feder et al 1996; Nielsen et al 1998; Deng et al 2000; Jiang et al 2003). The third class includes an LD contrast (LDC) test (Zaykin et al 2006) and its modifications (Wang et al 2007). Note that the majority of the tests in the first three classes can be implemented via logistic regression (Kim et al 2009; Pan 2010). The fourth and fifth classes include genomic distance-based regression (GDBR) (Wessel and Schork 2006) and genotype or haplotype similarity-based methods (Tzeng et al 2003ab; Schaid et al 2005; Yuan et al 2006; Sha et al 2007; Wei et al 2008) respectively. A close connection between the last two has been elucidated by Lin and Schaid (2009). Furthermore, since a GDBR method has been shown to empirically perform best among several candidate tests (Lin and Schaid 2009), we will skip the fifth and focus on the fourth class. Note that, although a popular class of haplotype-based tests (Schaid et al 2002; Zhao et al 2003a, 2003b) is not included in the above, it is closely related to the first class with logistic regression (Chapman et al 2003; Clayton et al 2004). Therefore, combining the first four classes of the tests in a unified framework is desirable, and is a major goal of this article.

There have been some recent attempts to combine multiple types of tests. Song and Elston (2006) proposed a test statistic that is a weighted sum of two test statistics from the first two classes respectively, while Wang and Shete (2008) proposed combining the p-values of two tests, one from each class. Chen and Chatterjee (2007) proposed a test with the use of information from both allele frequencies and HWE. As a unification of the first and third classes, Wang et al (2009b) proposed a Normal-based likelihood ratio test (NLRT) to compare both the mean vectors and covariance matrices of genotype scores between the two groups, which however depends on the incorrect normality assumption on *discrete* genotype

scores and is computationally intensive for its use of permutations to derive p-values. As an alternative, Pan (2010) proposed a general framework under logistic regression to contrast both genotype scores and LD patterns simultaneously. Kim et al (2009) considered the use of information in the HWD, in addition to allele frequencies and LD patterns, for only one- and two-marker logistic regression models, but other types of the tests (e.g. SSU) were not discussed. Here we will generalize the approach of Kim et al (2009) by including more than two markers and other tests.

There has been no attempt to combine the GDBR methods with other existing tests. This has become an important issue given that a very recent simulation study (Lin and Schaid 2009) found that a GDBR method (Haplo-match) (Wessel and Schork 2006) worked best among several methods compared, though unfortunately, some competitive tests, the SSU and UminP tests based on a main-effects logistic regression model (Chapman and Whittaker 2008; Pan 2009), were not included. Here we aim to reformulate GDBR into a unified framework of logistic regression. In particular, we show that the F-test in GDBR is either exactly or approximately equivalent to the SSU test for a corresponding logistic regression model. *Significant benefits* of reformulating GDBR as logistic regression include its immediate extensions with the use of other (e.g. score or UminP) tests that may be more powerful in some situations, the availability of the asymptotic distributions (in contrast to the use of computing-intensive permutations), the applicability to test combination and test selection procedures (Pan et al 2010), and ready adjustments for covariates, e.g. for population stratification and gene-gene interactions. Importantly, under this unified framework, it becomes feasible to simultaneously assess various aspects of the distributional differences of genotypes between the case and control groups.

2. METHODS

2.1 Review: Logistic regression and its associated tests

Given n independent observations (Y_i, X_i) with $Y_i = 0$ or 1 as disease status and $X_i = (X_{i1}, \ldots, X_{ik})'$ as genotype scores at k SNPs for subject $i = 1, \ldots, n$, we would like to test for any possible association between the disease and genotypes. We assume that there are n_1 cases and n_0 controls. The k SNPs are possibly in LD, as drawn from a candidate region or an LD block. Unless specified otherwise, we use the dosage coding for X_{ij} under an additive genetic model: $X_{ij} = 0$, 1 or 2, representing the number of the copies of a specific allele present in SNP j of subject i, though other genetic models can be adopted. Many multi-marker association tests are based on fitting a logistic regression model

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

A global test of any possible association between the disease and SNPs can be formulated as jointly testing on the multiple parameters β_j's with the null hypothesis H_0: $\beta = (\beta_1, \ldots, \beta_k)' = 0$, typically by one of the three asymptotically equivalent tests, the likelihood ratio test (LRT), Wald test and score test. Under H_0, any of the three test statistics has an asymptotic chi-squared distribution with degrees of freedom DF=k. The generalized Hotelling's T^2 test (Fan and Knapp 2003; Xiong et al 2002) is closely related to the above score test (Clayton et al 2004). Other tests, such as the SSU and UminP tests may be more powerful and can be also applied (Pan 2009). A potential problem with the above tests is that the logistic model includes only main-effects while ignoring other high-order terms. For example, if there is little marginal effects but substantial epistatic effects, ignoring interaction terms may lead to reduced power.

Since the derivations will be similar regardless of the terms included in a logistic regression model, below we derive the score vector and its covariance matrix based on model (1):

$$U=\sum_{i=1}^{n}(Y_i-)X_i,\quad V=(1-)\sum_{i=1}^{n}(X_i-\bar{\mathrm{X}}1)(X_i-\bar{\mathrm{X}}1)', \tag{2}$$

where $=\sum_{i=1}^{n}Y_i/n$ and $\bar{\mathrm{X}}=\sum_{i=1}^{n}X_i/n$. Three representative tests are

$$T_{\mathrm{Score}}=U'V^{-1}U,\ T_{\mathrm{SSU}}=U'U,\ T_{\mathrm{UminP}}=\max_{j=1}^{k}U_j^2/v_j, \tag{3}$$

where U_j is the jth element of U and v_j is the (j, j)th diagonal element of V. Under H_0, the score statistic has an asymptotic chi-squared distribution with DF=rank(U), the SSU has an approximate chi-squared distribution (Pan 2009), while the distribution of $\mathrm{T_{UminP}}$ can be numerically obtained (Conneely and Boehnke 2007).

The SSU test is equivalent to the permutation-based version of the empirical Bayes score test proposed by Goeman et al (2006) for high-dimensional microarray data (Pan 2009). Ballard et al (2009) commented that Goeman's test gives the same result as a variance component-based score test of Tzeng and Zhang (2007).

2.2 Review: Genomic distance-based regression (GDBR)

A distinguishing feature of the class of GDBR methods is the use of a similarity measure to compare any two subjects in a study. The similarity measure can be for genotypes based on identity-by-state (IBS), or for diplotypes based on counting measure or matching measure (Lin and Schaid 2009). A GDBR method can be summarized as the following:

1. Calculate an $n \times n$ distance matrix for all pairs of subjects by $D = (D_{ij}) = (1 - S_{ij})$ with $0 \le S_{ij} \le 1$ as a similarity measure between subjects i and j;
2. Calculate $A=(-D_{ij}^2/2)$;
3. Center A to get $G = (I - 11'/n)A(I - 11'/n)$;
4. Code the $n \times 1$ outcome vector y with elements $y_i = -1$ or 1;
5. Calculate the projection matrix $H = y(y'y)^{-1}y'$;
6. Calculate the pseudo-F statistic as

$$F=\frac{tr(HGH)}{tr[(I-H)G)I-H)]},$$

where $tr(A)$ is the trace of matrix A.

To obtain a p-value, permutations (by shuffling y) are used. If G is an outer product matrix, e.g. when the distance is Euclidean, the above F-test reduces to the usual F-test in multivariate analysis of variance (MANOVA) (McArdle and Anderson 2001).

Three similarity measures were found to work well by Lin and Schaid (2009). The first is a similarity measure of genotypes, called geno-sim or simply G. Suppose that g_i^l and g_j^l are the genotypes of the lth locus for subjects i and j respectively, then the similarity between the two subjects is the average of IBS for the k loci:

$$S_{ij}^{G}=\frac{\sum_{l=1}^{k} s(g_i^l, g_j^l)}{2k},$$

where $s(g_i^l, g_j^l)$=0, 1 or 2, is the IBS at locus l for subjects i and j. $S^G=(S_{ij}^G)$ is the similarity matrix called geno-sim.

The next two similarity matrices are based on inferred haplotypes from genotypes. Suppose that $h_{iu}=(h_{iu1}/h_{iu2})$ is the uth possible diplotype for subject i, where $u=1, \ldots, n_{h_i}$, and n_{h_i} is the number of possible diplotypes for subject i. Given the unphased genotype data, suppose $P(h_{iu}|g_i)$ is the posterior probability that subject i has the uth diplotype. A similarity measure given by Wessel and Schork (2006) is

$$S_{ij}^{H_1}=\frac{1}{2k}\sum_u\sum_v P(h_{iu}|g_i)P(h_{jv}|g_i)\ \max\{\sum_l s(h_{iu1}^l, h_{jv1}^l)+s(h_{iu2}^l, h_{jv2}^l), \sum_l s(h_{iu1}^l, h_{jv2}^l)+s(h_{iu2}^l, h_{jv1}^l)\},$$

where h_{iuc}^l is the allele at locus l on chromosome c = 1 or 2 for the uth diplotype of subject i. The score $s(h_{iu1}^l, h_{jv1}^l)=1$ if $h_{iu1}^l=h_{jv1}^l$, and $s(h_{iu1}^l, h_{jv1}^l)=0$ otherwise. $S_{ij}^{H_1}$ gives the expected haplotype-similarity over the posterior distribution of the haplotype pairs given the observed genotypes, and max() ensures that the similarity does not depend on the order of the two haplotypes in each haplotype pair. The resulting similarity matrix $S^{H_1}=(S_{ij}^{H_1})$ is called haplo-sim or simply H1.

Another haplotype-based measure is

$$S_{ij}^{H_2}=\frac{1}{2}\sum_u\sum_v P(h_{iu}|g_i)P(h_{jv}|g_i)\ \max\{s(h_{iu1}, h_{jv1})+s(h_{iu2}, h_{jv2}), s(h_{iu1}, h_{jv2})+s(h_{iu2}, h_{jv1})\},$$

where $s(h_{iu1}, h_{jv1})$ = 1 if all alleles in the two haplotypes are equal, and $s(h_{iu1}, h_{jv1})$ = 0 otherwise. Hence, rather than counting the number of equal alleles between two haplotypes as in $S_{ij}^{H_1}$, it counts the number of equal haplotypes in $S_{ij}^{H_2}$; the two similarity measures correspond to the “counting measure” and “matching measure” of Tzeng et al (2003a). The resulting similarity matrix $S^{H_2}=(S_{ij}^{H_2})$ is called haplo-match or simply H2.

Given unphased genotype data, an EM algorithm can be used to infer haplotypes, as implemented in R package `haplo.stats` (Schaid et al 2002).

2.3. New formulation of GDBR as logistic regression

The above F-test was developed as an extension of MANOVA (McArdle and Anderson 2001) with only a distance matrix D available. If G is an outer product matrix, say $G=ZZ'$ with an $n\times p$ matrix Z, the above F-test is simply testing H_0':B=0 in a multivariate linear model

$$Z=yB+\varepsilon, \tag{4}$$

where y is an $n\times 1$ vector of elements 1 and −1 for cases and controls respectively, B is a $1\times p$ vector of unknown regression coefficients, and ε is an $n\times p$ matrix of random errors.

To assess possible association between Z and group memberships y (or equivalently, Y), rather than regressing Z on y as in GDBR, we regress Y on Z via a logistic regression model:

$$\text{Logit Pr}(Y=1)=\beta_0+Z\beta, \tag{5}$$

where the assessment of possible association can be accomplished by testing on the unknown $p \times 1$ vector of unknown regression coefficients in null hypothesis $H_0 : \beta = 0$, for which we can apply any of the score, SSU and UminP tests. Importantly, the null distribution of each test can be easily obtained, making it possible to combine them with other tests.

In general, G may not be an outer product matrix. Nevertheless, we can approximate $G \approx ZZ'$ with an $n \times (n - 1)$ matrix Z found by Gower's (1966) principle coordinates analysis. We simply applied the classical multi-dimensional scaling as implemented in R package `cmdscale`. Practically we should not include all the derived coordinates into a logistic regression model. As in principle components analysis, we could select just the first few coordinates with the largest eigenvalues; in the simulations, we only included the components with the absolute values of eigenvalues larger than 10^{-8}, and excluded any column of Z with too small elements whose absolute values had a mean less than 10^{-5}.

We regard Z as representing some complex high-order interactions among the SNPs. The reformulation of GDBR as logistic regression opens a new venue to constructing other novel powerful tests: for example, to test the same H_0 in the logistic regression model (5), one can use a variety of tests, as discussed before. In particular, we prove in Appendix that, under the conditions that the distance matrix G is an outer product matrix and that we have an equal number of cases and controls ($n_1 = n_0$), the permutation-based F-test in GDBR is equivalent to the SSU test in the corresponding logistic regression; more generally, they are expected to be close if $G \approx ZZ'$ and $n_1 \approx n_0$. Furthermore, formulating GDBR as logistic regression makes it possible to derive asymptotic distributions and to combine GDBR with other tests that incorporate various types of information drawn from the data.

2.4 Unification of GDBR and logistic regression

We aim to incorporate genotype scores, HWD parameters, LD measurements, and dissimilarity-derived scores as covariates into a logistic model, then characterize its Type I error and power properties. In particular, we would like to assess whether such an expanded model and its associated tests can maintain high power by combining multiple types of information in genotypic distributional differences between the case and control groups.

Suppose that X is an $n \times k$ genotype matrix with the dosage coding; that is, $X_{ij} = 0$, 1 or 2 represents the number of the copies of an allele in locus j for subject i. We denote XX as the $n \times (k(k + 1)/2)$ cross-product matrix with the ith row as $(XX)_{i.}=(X_{i1}^2, X_{i1}X_{i2}, \ldots, X_{i1}X_{ik}, X_{i2}^2, X_{i2}X_{i3}, \ldots, X_{ik}^2)$. Suppose Z_G, Z_{H_1} and Z_{H_2} are the matrices derived from similarity matrices S^G, S^{H1} and S^{H2} respectively. We will consider the following logistic regression models:

L1: Logit Pr($Y = 1$) = $\beta_0 + X\beta_1$,

L2: Logit Pr($Y = 1$) = $\beta_0 + X\beta_1 + XX\beta_2$,

L3: Logit Pr($Y = 1$) = $\beta_0 + X\beta_1 + XX\beta_2 + Z_G\beta_3 + Z_{H_1}\beta_4 + Z_{H_2}\beta_5$,

corresponding to three null hypotheses:

$H_{0,1}$: $\beta_1 = 0$,

$H_{0,2}$: $\beta_1 = 0$ and $\beta_2 = 0$,

$H_{0,3}$: $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$, $\beta_4 = 0$, and $\beta_5 = 0$.

Model L1, perhaps the most popular main-effects model in use, aims to detect the mean difference between the genotypes scores of the case and control groups. In addition to the mean difference in genotype scores, model L2 incorporates the possible differences in HWD parameters (through the squared terms of the genotype scores) and in LD patterns (through the pairwise cross-products or interactions of the SNPs) (Kim et al 2009). Model L3 is our proposed new one, aiming to capitalize on all four types of information: mean genotype scores, HWD parameters, pairwise genotype score interactions, and other high-order interactions. Since different dissimilarity matrices may capture different aspects of complex high-order interactions among the SNPs, we use all three types of dissimilarity matrices in model L3, though other simplified models can be also considered (see Supplementary Materials).

To test $H_{0,1}$ in model L1, we can simply apply the SSU, score and UminP tests as discussed in section 2.1. To test the null hypothesis for either L2 and L3, we apply Fisher's (1932) method, though other methods can be also utilized (see Supplementary Materials). For example, to test $H_{0,2}$, we first apply the SSU to test its two components $\beta_1 = 0$ and $\beta_2 = 0$ separately, obtaining twp p-values p_1 and p_2; then we apply Fisher's method to combine the two p-values to obtain a final p-value. Specifically, given $L = 2$ or 5 p-values, $p_1, \ldots, p_L$, obtained from L SSU tests on individual components of $H_{0,2}$ or $H_{0,3}$, Fisher's method combines the p-values as

$$T_F(p_1, \ldots, p_L) = \prod_{j=1}^{L} p_j.$$

To obtain a final p-value, we propose using a simulation based approach. First, we note that each individual test is based on a component of the whole score vector U. Second, because of the asymptotic null distribution of U is known as $U \sim N(0, V)$, we can simulate B iid copies of U^b's from $N(0, V)$ with $b = 1, 2 \ldots, B$. Based on each U^b, we can calculate individual p-values as $p_1^b, \ldots, p_L^b$, and thus $T_F(p_1^b, \ldots, p_L^b)$. Third, the final p-value for $TF(p_1, \ldots, p_L)$ is simply $\sum_{b=1}^{B} I[T_F(p_1, \ldots, p_L) < T_F(p_1^b, \ldots, p_L^b)]/B$. We used $\mathbf{B} = 10^3$ for simulated data and $B = 10^6$ for the ALS data.

2.5 Simulations

We followed the simulation set-ups of Lin and Schaid (2009). To mimic real human LD structures, they used the genotypes of the 60 unrelated CEU samples (i.e. parents of the 30 trios) in the HapMap data (Thorisson et al 2005). They considered 13 regions from 12 chromosomes: eight 4-SNP regions on eight chromesomes, four 8-SNPs regions on four chromesomes, and a region of 25 SNPs on chromosome 17. These regions represented a wide spectrum of LD patterns and allele frequencies: the pairwise r^2 and minor allele frequencies within each region ranged from 0.002 to 1, and from 0.05 to 0.45, respectively; for more details, see Table 1 of Lin and Schaid (2009). In each region, each SNP was sequentially treated as disease-causal. Conditional on the copy number of the minor allele in each causal SNP as 0, 1 and 2, the disease probabilities were assigned as 0.029, 0.076 and 0.214, respectively. These conditional probabilities mimic the penetrances of Alzheimer's disease for the APOE-4 genotype. In each region, the non-causal SNPs were used to detect possible association with disease.

For each scenario, we simulated 1000 datasets; in each dataset, the sample size was $n = 100$ with $n_1 = 50$ cases and $n_0 = 50$ controls. Similar results (not shown) were obtained for a larger sample size with 100 cases and 100 controls.

For each statistical test, its overall power was calculated by averaging its power over all scenarios (i.e. over each causal SNP across the 13 regions). In addition, the power was stratified on a few factors: i) marker informativity: whether the average of MAFs of markers was above 0.215; ii) number of markers being used in the region (3, 7 or 24) after the causal SNP was excluded; iii) causal allele frequency; iv) preponderance of the most common high-risk haplotype: whether the relative frequency of the most common high-risk haplotype over other high-risk haplotypes was above 0.8; v) LD pattern: whether the squared correlation coefficient r^2 between the causal SNP and its adjacent makers was above 0.6 ("high"), between 0.15 and 0.6 ("moderate"), or below 0.15 ("low").

2.6 ALS data

We applied the methods to a data set drawn from a genome-wide association study on amyotrophic lateral sclerosis (ALS) (Schymick et al 2007). ALS is a fatal neurodegenerative disease leading to paralysis and death. Despite persistent efforts in elucidating the genetic components of ALS, little is known. The original study assayed 555352 unique SNPs for each of the $n_1 = 276$ patients with sporadic ALS and $n_0 = 268$ controls. Schymick et al (2007) took a single-marker approach by testing SNP by SNP with either a 2-DF or 1-DF chi-squared test, and identified 34 most significant SNPs, none of which however reached a genome-wide significance level after a Bonferroni adjustment. A 1-DF test is based on the dosage coding of a SNP, while a 2-DF test creates 2 dummy indicators for the three alleles of the SNP. For this dataset, since a 2-DF test seemed to give more significant p-values, we adopted the 2-DF coding of SNPs by default.

We randomly picked up 9 SNPs from the list of the 34 most significant common SNPs of Schymick et al (2007). For each of the 9 SNPs, we extracted 10 neighboring SNPs upstream and another 10 downstream, then applied the default LD blocking algorithm implemented in Haploview (v4.1) (Barrett et al 2005) to each 21-SNP region for the control group. The number of the SNPs inside each LD block surrounding each of the nine SNPs ranged from 2 to 19.

3. RESULTS

3.1 Simulations: Type I errors

By removing each of the four SNPs and using the remaining 3 SNPs as markers in a region of chromosome 6, we simulated a null case, in which disease probability did not depend on any SNP. Figure 1 shows the Type I error rates of various methods. It is clear that all the tests had their Type I error rates well controlled below the nominal level of 0.05. As discussed by Lin and Schaid (2009), the lower Type I error rates could be due to the discreteness of the data resulting from the small pool of the CEU samples.

3.2 Simulations: Comparing GDBR and logistic regression

Figure 1 shows that the permutation-based F-test in GDBR with any similarity matrix (G or H1 or H2) had almost the same overall power as the SSU test in the corresponding logistic regression model (i.e. with the corresponding decomposed G or H1 or H2 matrix as predictors). The slight reduction in power in logistic regression was due to possible information loss in extracting a few major components from a distance matrix.

An advantage of formulating GDBR as logistic regression is that it opens door to applying other tests based on logistic models, such as the multivariate score test or UminP test, in addition to the SSU test. Although on average, the score and UminP tests were not as powerful as SSU (or GDBR) test for these data, in some situations they could be more powerful. For example, Table 1 shows the power of the tests applied to a region on chromosome 6 with 4 SNPs; each SNP was treated as causal and removed from the test data. It can be seen that, for a given similarity matrix, the UminP test sometimes had higher power than the SSU test or GDBR F-test. Similarly, when the causal SNP_0 is the first or second one, among the three distance metrics, although on average the GDBR with H2 (or the corresponding SSU test) was the most powerful, the GDBR (or SSU) with similarity matrix G1 or H1 could be more powerful than that with H2.

3.3 Simulations: Incorporating GDBR into logistic regression

In terms of the overall power, it is clear from Figure 1 that Fisher's method in model L3 was most powerful among all the tests, showing the power gain by combining multiple types of information. Fisher's method in model L2 performed second best. The SSU test in model L1 had power close to that of the F-test in GDBR with distance matrix H2.

We also conducted stratified power analysis and reached the same conclusion; these and other results are available in the Supplementary Materials.

3.4 ALS data

We applied both single-marker tests and various multi-marker tests to the nine SNPs and their LD blocks, respectively. In Table 2 we only show the results for three SNPs while those for other ones are in the Supplementary Materials. For the univariate tests applied to each individual SNP, we included both a 2-DF score test and a 1-DF score test in logistic regression. For the GDBR tests, we used 10^3 permutations for the GDBR tests, and used $B = 10^6$ simulations for combining p-values, while asymptotic distributions were used to obtain p-values for other tests. Since each of the three types of the similarity matrices might have many small eigenvalues, we decided to include only those first few components with the largest positive eigenvalues so that their sum is at least 95% of the total sum of all positive eigenvalues; including those components with negative eigenvalues yielded similar results (not shown).

It is clear that the GDBR tests gave p-values in close agreement with those of the SSU tests in the corresponding logistic regression models, but the latter could give significance levels beyond that of the former based on only 1000 permutations. Importantly, for each logistic regression model, in addition to the SSU test, we could also apply the score test and UminP test, which, for example, for both the similarity matrices G and H1, gave more significant p-values than those of the SSU test (and hence the GDBR F-test) for SNP rs7976059. It is noted that, in most cases, Fisher's method gave most significant p-values, showing possible power gains by incorporating the use of multiple types of information across multiple markers.

4. DISCUSSION

Here we have presented an approach to utilizing distributional differences in allele frequencies, HWD parameters, LD patterns and some complex high-order interactions among multiple markers. We regard that the effectiveness of a GDBR method comes from its ability to capture some complex high-order interactions among SNPs through its genomic distance metric. A key technical advance is to reformulate GDBR as logistic regression, which not only sheds light on the close connection between the class of the GDBR methods

and the majority of existing association tests based on logistic regression, but also opens a new and potentially productive venue to constructing novel and powerful association tests under the unified and general framework of logistic regression. In addition to facilitating combining GDBR with many existing logistic regression-based approaches, our formulation of GDBR as logistic regression offered additional advantages. For example, rather than depending on permutations to derive p-values for GDBR, which may be just too time-consuming to achieve a stringent significance level as in GWAS, we can recourse to well-known asymptotic results. In addition, it is straightforward to incorporate covariates, including principle components or eigenvectors in adjustment for population stratification (Price et al 2006; Li and Yu 2008; Lee et al 2009). Furthermore, we can also detect SNP-disease association in a logistic regression model in the presence of gene-gene or gene-environment interactions (Pan 2010b).

In general, there does not exist a uniformly most powerful test on multiple parameters as in the current situation with multi-marker association testing (Cox and Hinkley 1974). There is always a trade-off between targeting more aspects of possible distributional differences by including more terms in a logistic regression model versus possible loss of power due to the increasing DF in the model. More research is needed to develop adaptive tests that can determine what terms to be included based on the given data, as illustrated by test selection as studied by Pan et al (2010).

Finally, we comment on an implication of our result to high-dimensional data analysis. Since both GDBR (Zapala and Schork 2006) and Goeman's test (2006) have been successfully applied to high-dimensional microarray data and the two approaches appear quite different, it may give an impression that two approaches are distinct competitors to each other. In light of our new result on the connection between the F-test in GDBR and the SSU test in logistic regression (see Appendix) and the known result of the equivalence between Goeman's test and the SSU test (Pan 2009), the two approaches are closely related.

Software in R implementing the proposed new tests will be posted on our web site http://www.biostat.umn.edu/_weip/prog.html.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by NIH grants GM081535 and HL65462. The authors are most grateful to Wan-Yu Lin for sharing his R code and data for GDBR analysis, and to the study team who provided the ALS data in the dbGaP. The authors thank the reviewers for helpful comments.

APPENDIX Relationship between the F-test in GDBR and the SSU test in logistic regression

We first consider the situation that G is an outer product matrix, say $G = ZZ'$. In GDBR, the F-test tests H_0':B=0 in multivariate linear model (4). The least squares estimate of B is

$$\hat{\mathrm{B}}=(y'y)^{-1}y'Z=(n_1\bar{\mathrm{Z}}_1 - n_0\bar{\mathrm{Z}}_0)/n,$$

where n_0 and n_1 are the numbers of the cases and controls, respectively, $n = n_0 + n_1$, and $\bar{Z}_1$ and $\bar{Z}_0$ are the sample means of Z for the case and control groups respectively. With the

corresponding fitted values $\hat{Z} = y\hat{B}$ and residuals $R = Z - \hat{Z} = (I - H)Z$, the total sum of squares and cross-product (SSCP) matrix can be partitioned into: $Z'Z = \hat{Z}'\hat{Z} + R'R$. Then it is easy to verify that

$$F = \frac{tr(HGH)}{tr[(I-H)G(I-H)]} = \frac{tr(')}{tr(R'R)} = \frac{1}{tr(R'R)/tr(')} \propto \frac{1}{[tr(')+tr(R'R)]/tr(')} = \frac{tr(')}{tr(Z'Z)}.$$

Under permutations, $tr(Z'Z)$ is fixed as a constant, hence F-statistic is equivalent to

$$tr(') = n\hat{B}'\hat{B} = (n_1\bar{Z}_1 - n_0\bar{Z}_0)'(n_1\bar{Z}_1 - n_0\bar{Z}_0)/n. \quad (6)$$

On the other hand, to test H_0: β = 0 in logistic regression model (5), the score vector, as shown by Clayton et al (2004), is

$$U = \frac{n_0 n_1}{n}(\bar{Z}_1 - \bar{Z}_0),$$

and thus the SSU test statistic

$$T_{SSU} = U'U = \frac{n_0^2 n_1^2}{n^2}(\bar{Z}_1 - \bar{Z}_0)'(\bar{Z}_1 - \bar{Z}_0). \quad (7)$$

Comparing (6) and (7), we see that the F-statistic and SSU-statistic are equivalent if $n_1 = n_0$. More generally, if $G \approx ZZ'$ and $n_1 \approx n_0$, the two statistics are expected to be close (up to a constant).

REFERENCES

Altshuler D, Daly M, Lander ES. Genetic mapping in human disease. Science. 2008; 322:881–888. [PubMed: 18988837]

Ballard DH, Cho J, Zhao H. Comparisons of multi-marker association methods to detect association between a candidate region and disease. Genetic Epidemiology. 2009; 34:201–212. [PubMed: 19810024]

Chapman JM, Whittaker J. Analysis of multiple SNPs in a candidate gene or region. Genetic Epidemiology. 2008; 32:560–566. [PubMed: 18428428]

Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations 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]

Chen J, Chatterjee N. Exploiting Hardy-Weinberg equilibrium for efficient screening of single SNP associations from case-control studies. Human Heredity. 2007; 63:196–204. [PubMed: 17317968]

Clayton D, Chapman J, Cooper J. Use of unphased multilocus genotype data in indirect association studies. Genetic Epidemiology. 2004; 27:415–428. [PubMed: 15481099]

Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. Am J Hum Genet. 2007; 81:1158–1168. [PubMed: 17966093]

Cox, DR.; Hinkley, DV. Theoretical Statistics. London: Chapman and Hall; 1974.

Deng HW, Chen WM, Recker RR. QTL fine mapping by measuring and testing for Hardy-Weinberg and linkage disequilibrium at a series of linked marker loci in extreme samples of population. American Journal of Human Genetics. 2000; 66:1027–1045. [PubMed: 10712216]

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

Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, et al. A novel MHC class I like gene is mutated in patients with hereditary haemochromatosis. Nature Genetics. 1996; 13:399–408. [PubMed: 8696333]

Fisher, RA. Statistical Methods for Research Workers. 4th edition. London: Oliver & Boyd; 1932.

Flint J, Mackay T. Genetic architecture of quantitative traits in mice, flies and humans. Genome Research. 2009; 19:723–733. [PubMed: 19411597]

Goeman JJ, van de Geer S, van Houwelingen HC. Testing against a high dimensional alternative. J R Stat Soc B. 2006; 68:477–493.

Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966; 53:325–338.

Jiang R, Dong J, Wang D, Sun FZ. Fine-scale mapping using Hardy-Weinberg disequilibrium. Annals of Human Genetics. 2001; 65:207–219. [PubMed: 11427179]

Kim S, Morris NJ, Won S, Elston RC. Single-marker and two-marker association tests for unphased case-control genotype data, with a power comparison. Genetic Epidemiology. 2010; 34:67–77. [PubMed: 19557751]

Lee AB, Luca D, Klei L, Devlin B, Roeder K. Discovering Genetic Ancestry Using Spectral Graph Theory. Genetic Epidemiology. 2010; 34:51–59. [PubMed: 19455578]

Li Q, Yu K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. Genetic Epidemiology. 2008; 32:215–226. [PubMed: 18161052]

Li Q, Wacholder S, Hunter DJ, Hoover RN, Chanock S, Thomas G, Yu K. Genetic background comparison using distance-based regression, with applications in population stratification evaluation and adjustment. Genetic Epidemiology. 2009; 33:432–441. [PubMed: 19140130]

Lin WY, Schaid DJ. Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. Genet Epidemiol. 2009; 33:183–197. [PubMed: 18814307]

Maher B. Personal genomes: the case of the missing heritability. Nature. 2008; 456:18–21. [PubMed: 18987709]

McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology. 2001; 82:290–297.

Nielsen DM, Ehm MG, Weir BS. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. American Journal of Human Genetics. 1998; 63:1531–1540. [PubMed: 9867708]

Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genetic Epidemiology. 2009; 33:497–507. [PubMed: 19170135]

Pan W. 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. Human Heredity. 2010; 69:1–13. [PubMed: 19797904]

Pan W. Statistical Tests of Genetic Association in the Presence of Gene-Gene and Gene-Environment Interactions. Human Heredity. 2010b; 69:131–142. [PubMed: 19996610]

Pan W, Han F, Shen X. Test Selection with Application to Detecting Disease Association with Multiple SNPs. Human Heredity. 2010; 69:120–130. [PubMed: 19996609]

Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 2006; 38:904–909. [PubMed: 16862161]

Roeder K, Bacanu SA, Sonpar V, Zhang X, Devlin B. Analysis of single-locus tests to detect gene/disease associations. Genet Epidemiol. 2005; 28:207–219. [PubMed: 15637715]

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

Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN. Non-parametric tests of association of multiple genes with human disease. Am J Hum Genet. 2005; 76:780–793. [PubMed: 15786018]

Sha Q, Chen H-S, Zhang S. A new association test using haplotype similarity. Genet Epidemiol. 2007; 31:577–593. [PubMed: 17443704]

Song KA, Elston RC. A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. Stat Med. 2006; 25:105–126. [PubMed: 16220513]

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

Tzeng J-Y, Devlin B, Wasserman L, Roeder K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet. 2003a; 72:891–902. [PubMed: 12610778]

Tzeng J-Y, Byerley W, Devlin B, Roeder K, Wasserman L. Outlier detection and false discovery rates for whole-genome DNA matching. J Am Stat Assoc. 2003b; 98:236–246.

Tzeng J-Y, Zhang D. Haplotype-based association analysis via variance-components score test. Am J Hum Genet. 2007; 81:927–938. [PubMed: 17924336]

Wang J, Shete S. A test for genetic association that incorporates information about deviation from Hardy-Weinberg proportions in cases. Am J Hum Genet. 2008; 83:5363.

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]

Wang X, Zhang S, Sha Q. A new association test to test multiple-marker association. Genetic Epidemiology. 2009; 33:164–171. [PubMed: 18720476]

Wei Z, Li M, Rebbeck T, Li H. U-statistics-based tests for multiple genes in genetic association studies. Annals of Human Genetics. 2008; 72:821–833. [PubMed: 18691161]

Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. Am J Hum Genet. 2006; 79:792–806. [PubMed: 17033957]

Won S, Elston RC. The power of independent types of genetic information to detect association in a case-control study design. Genetic Epidemiology. 2008; 32:731–756. [PubMed: 18481783]

Won S, Kim S, Elston RC. Phase uncertainty in case-control association studies. Genetic Epidemiology. 2009; 33:463–478. [PubMed: 19194981]

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

Yuan A, Yue Q, Apprey V, Bonney G. Detecting disease gene in DNA haplotype sequences by nonparametric dissimilarity test. Hum Genet. 2006; 120:253–261. [PubMed: 16807758]

Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proc Natl Acad Sci USA. 2006; 103:19430–19435. [PubMed: 17146048]

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combing p-values. Genetic Epidemiology. 2002; 22:170–185. [PubMed: 11788962]

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]

Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. Nature Genetics. 2007; 39:1167–1173. [PubMed: 17721534]

Zhao H, Pfiffer R, Gail MH. Haplotype analysis in population genetics and association studies. Pharmacogenomics. 2003a; 4:171–178. [PubMed: 12605551]

Zhao LP, Li S, Khalid N. Assessing haplotype-based association with multiple SNPs in case-control studies. American Journal of Human Genetics. 2003b; 72:1231–1250. [PubMed: 12704570]

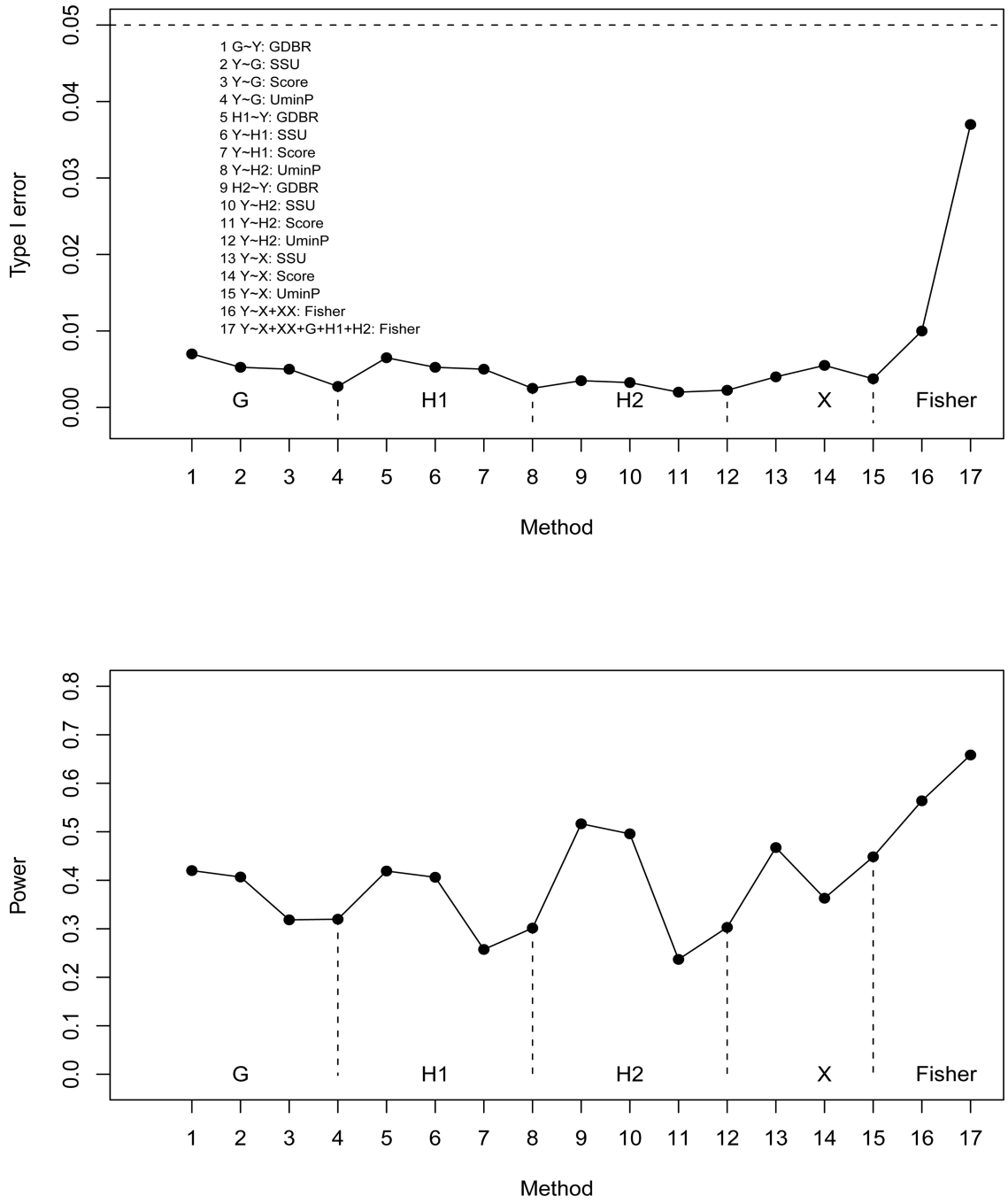

Figure 1.
Empirical Type I error and average power of various tests at the nominal level of 0.05 for simulated data.

Table 1

Empirical power from 1000 simulations for a region with 4 SNPs on chromosome 6. Each SNP_0 was treated as causal and removed from the simulated data.

				Logistic								
	GDBR			G			H1			H2		
SNP0	*G*	*H*1	*H*2	Score	SSU	UminP	Score	SSU	UminP	Score	SSU	UminP
1	0.252	0.249	0.126	0.148	0.242	0.117	0.140	0.242	0.113	0.151	0.116	0.135
2	0.262	0.265	0.120	0.161	0.248	0.135	0.145	0.250	0.128	0.141	0.115	0.149
3	0.013	0.006	0.381	0.145	0.017	0.152	0.132	0.009	0.148	0.202	0.363	0.228
4	0.005	0.007	0.038	0.006	0.006	0.006	0.006	0.006	0.006	0.018	0.038	0.025

Table 2

P-values of the single-marker and multi-marker tests for the ALS data.

					Logistic regression								
		GDBR			G			H1			H2		
SNP	**LD blk #SNPs**	***G***	***H*1**	***H*2**	**Score**	**SSU**	**UminP**	**Score**	**SSU**	**UminP**	**Score**	**SSU**	**UminP**
rs4363506	3	< .001	< .001	< .001	1.64×10^{-5}	3.21×01^{-6}	2.44×10^{-5}	7.50×10^{-5}	3.82×10^{-6}	3.72×10^{-5}	0.0196	1.30×10^{-5}	0.0001
rs7976059	4	0.018	0.022	0.007	0.0037	0.0161	0.0061	0.0051	0.0192	0.0079	0.0057	0.0060	0.0531
rs10773543	2	< .001	< .001	< .001	0.0004	2.46×10^{-5}	7.62×10^{-5}	0.0002	1.53×10^{-5}	5.14×10^{-5}	0.0001	3.13×10^{-5}	0.0001

			Logistic regression					
	Single-marker		L1			L2	L3	
SNP	1-DF	2-DF	Score	SSU	UminP	Fisher	Fisher	
rs4363506	3.64×10^{-6}	1.52×10^{-6}	3.90×10^{-5}	8.95×10^{-6}	1.57×10^{-5}	$< 1.00\times10^{-6}$	$< 1.00\times10^{-6}$	
rs7976059	0.0008	6.56×10^{-5}	0.0012	0.0303	0.0190	0.0416	8.80×10^{-5}	
rs10773543	0.0003	3.55×10^{-5}	5.98×10^{-5}	2.98×10^{-6}	2.23×10^{-5}	$< 1.00\times10^{-6}$	$< 1.00\times10^{-6}$	