Allelic Based Gene-Gene Interaction in Case-Control Studies

Jeesun Jung^{a, b} Yiqiang Zhao^{a, b}

^aDepartment of Medical and Molecular Genetics and ^bCenter for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, Ind., USA

Key Words

Allelic test • Interaction effect • Score test • Cochran-Armitage method • Epistasis

Abstract

In case-control studies identifying disease susceptibility loci, it has been shown that the interaction caused by multiple single nucleotide polymorphisms (SNPs) within a gene as well as by SNPs at unlinked genes plays an important role in influencing risk of a disease. A novel statistical approach is proposed to detect gene-gene interactions at the allelic level contributing to a disease trait. With a new allelic score inferred from the observed genotypes at two or more unlinked SNPs, we derive a score test from logistic regression and test for association of the allelic scores with a disease trait. Furthermore, F and likelihood ratio tests are derived from Cochran-Armitage regression. By testing for the association, the interaction can be assessed both in cases where the SNP association can be detected and cannot be detected as a main effect in single SNP approach. The analytical power and type I error rates over 6 two-way interaction models are investigated based on the non-centrality parameter approximation of the score test. Simulation studies demonstrate that (1) the power of the score test is asymptotically equivalent to that of the test statistics by the Cochran-Armitage method and (2) the allelic based method provides higher power than two genotypic based methods.

Introduction

With the advance of high-throughput sequencing technologies, genetic association studies to identify susceptibility genes of a common, complex, human disease has been promoted. It has been known that multiple genes along with the environment play interactive roles in contribution to the development of complex diseases. Most ongoing genetic studies have focused on identifying the effect of a single gene; these studies often fail to identify causal association with a disease due to the fact that a disease is the consequence of a complicated network of multiple susceptibility loci, each of which is likely to have only a small effect when considered alone.

Recently many statistical methods have been proposed to detect gene-gene interactions in a case-control study. As a machine learning and data mining approach, multifactor-dimensionality reduction (MDR) was proposed to pool multilocus genotypes into high-risk and low-risk groups and analyze the datasets through cross-validation to calculate prediction accuracy of the testing SNPs [1]. Classical decision tree approaches such as classification and regression trees [2] and support vector machines [3] classify subjects into case and control groups by an optimal hyperplane predicted by SNPs combination. Similar to other statistical model-based methods, the logit regression method [4] is a generalized regression model

Copyright © 2009 S. Karger AG, Basel

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2009 S. Karger AG, Basel
0001-5652/10/0691-0014\$26.00/0

Accessible online at:
www.karger.com/hhe

Jeesun Jung, PhD
Department of Medical & Molecular Genetics, School of Medicine
Indiana University
Indianapolis, IN (USA)
Tel. +1 317 274 3688, Fax +1 317 278 9217, E-Mail jeejung@iupui.edu

that uses a ranking criterion to identify the best model. The most commonly used method is a multiplicative method in a standard logistic model framework which consists of main effects and multiplicative interaction effects of SNPs.

Most currently available methods of either machine learning approaches or model based approaches focus on the association of genotypic combinations of multiple SNPs with a disease trait. Jung et al. [5] proposed a novel approach to search for the interaction at the allelic level for quantitative trait loci (QTL). As stated in Jung et al. [5], there are several differences in comparison between genotype-based methods and allelic-based method. Genotypic-based approaches can increase the degrees of freedom of a test statistic, resulting in a concomitant loss of power when multiple unlinked markers (i.e. more than 3) are tested for interaction. In contrast, the allelic-based method can reduce the number of degrees of freedom by accounting for the allelic levels. Furthermore, the power of the genotypic-based method may be compromised by sparse or empty combination caused by multiple markers with small sample sizes or a lower minor allele frequency (<0.1). However, the allelic-based method is relatively robust even with the lower minor allele frequency and small sample sizes.

In this paper, we extend our previous work detecting interaction of quantitative trait loci (QTL) to a case-control study design comparing those with and those without the disease. By a new definition of the allelic-based gene-gene interaction, which is the nonrandom association of alleles occurring when a particular allele in one gene and a particular allele in another unlinked gene contribute to the risk of a disease trait through their interaction [6], the proposed method tests for association of these allelic scores assigned to each subject with a disease trait. We derive a score test based on the standard logistic regression model and both F test and likelihood ratio test (LRT) statistics based on the Cochran-Armitage (CA) regression model. In addition, the non-centrality parameters approximation of these statistics were derived to evaluate the analytical properties of the test statistics. We perform simulation studies to demonstrate the analytical properties of the score test and CA-based tests and to compare the power of the allelic-based method with that of the genotypic-based methods. The 7 candidate genes of major depression disorder (MDD) case-control data provided in dbGaP [7] were applied to show the feasibility of the proposed method.

Table 1. Allelic scores given genotypes of individual

Genotype	Allelic score			
	AB	Ab	aB	ab
(AA,BB)	1	0	0	0
(AA,Bb)	1/2	1/2	0	0
(AA,bb)	0	1	0	0
(Aa,BB)	1/2	0	1/2	0
(Aa,Bb)	1/4	1/4	1/4	1/4
(Aa,bb)	0	1/2	0	1/2
(aa,BB)	0	0	1	0
(aa,Bb)	0	0	1/2	1/2
(aa,bb)	0	0	0	1

Methods

The allelic-based gene-gene approach proposes to test for non-random association of the allelic combination with a disease trait. The allelic combination of each subject is scored by a conditional probability of having the particular allelic combination given the observed genotypes at the putative unlinked loci of each subject. Table 1 shows the allelic score of subjects whose genotypes of two markers are given.

Statistical Models and Properties

Let D_1^* , D_2^* be two disease loci underlying a disease trait, which are both in Hardy-Weinberg equilibrium (HWE) and are unlinked but associated with a disease through their interaction. Let D_1 and d_1 be two alleles at the first disease locus D_1^* , with frequencies p_{D_1} and p_{d_1} respectively. Let D_2 and d_2 be two alleles at the second disease locus D_2^* , with frequencies p_{D_2} and p_{d_2} . We consider two observed marker loci, each of which is in linkage disequilibrium (LD) with either of the two interacting disease loci; Assume that marker M_1 is in LD with the susceptibility locus D_1^* and marker M_2 is in LD with the other susceptibility locus D_2^* and both are unlinked. An observed marker M_1 has two alleles A and a with frequencies p_A and p_a respectively, and M_2 has two alleles B and b with frequencies p_B and p_b respectively. Let $D_{AD_1} = P(AD_1) - P_AP_{D_1}$ be the measure of LD between D_1^* and M_1 , and $D_{BD_2} = P(BD_2) - P_BP_{D_2}$ be LD measure between D_2^* and M_2 . Denote the penetrance of disease genotype of D_1^* by $f_{kl} = p(Y = 1 | D_1^* \in (D_1D_1, D_1d_1, d_1d_1))$, $k, l \in (D_1, d_1)$, and that of D_2^* by $f_{mn} = p(Y = 1 | D_2^* \in (D_2D_2, D_2d_2, d_2d_2))$, $m, n \in (D_2, d_2)$ respectively. Let $f_{klmn} = P(Y = 1 | D_1^*, D_2^*)$ be the joint penetrance for the two disease loci D_1^* and D_2^* , $k, l \in (D_1, d_1)$, $m, n \in (D_2, d_2)$. As shown in figure 1, the penetrance of a particular allelic combination is calculated using the joint penetrance of genotypic combinations consisting of the particular alleles. Therefore, we can derive the penetrance of two D_1D_2 allelic combination by $f_{D_1D_2} = P(Y = 1 | D_1D_2) = f_{D_1D_1D_2D_2}p_{D_1}p_{D_2} + f_{D_1D_1D_2d_2}p_{D_1}p_{d_2} + f_{D_1d_1D_2D_2}p_{d_1}p_{D_2} + f_{D_1d_1D_2d_2}p_{d_1}p_{d_2}$. In the same way $f_{D_1d_2}, f_{d_1D_2}, f_{d_1d_2}$ can also be calculated. For a control group, the penetrance f is replaced by $1 - f$ such as $1 - f_{kl} = p(Y = 0 | D_1^* \in (D_1D_1, D_1d_1, d_1d_1))$ and $1 - f_{mn} = p(Y = 0 | D_2^* \in (D_2D_2, D_2d_2, d_2d_2))$, $k, l \in (D_1, d_1)$, $m, n \in (D_2, d_2)$ for the two genes.

Fig. 1. Cell combinations of unlinked two disease loci.

Logistic Regression Model and Non-Centrality Parameter

Denote $y_i = 1$ if the i -th subject is affected by a disease (case) and $y_i = 0$ if the i -th subject is not affected by the disease (control). In the non-parametric maximum likelihood solution that allows an arbitrary covariate distribution, Anderson [8], Prentice and Pyke [9] and Scott and Wild [10] showed that fitting a standard prospective logistic regression in a case-control sampling design is equivalent to fitting a retrospective logistic regression except that an intercept in the case-control sampling needs the information of sampling fraction of cases and controls. Therefore, the standard logistic regression model is used due to the equivalence in the parameter estimates of interaction effect. The likelihood function of the standard logistic regression is

$$L(\alpha, \beta_{AB}, \beta_{Ab}, \beta_{aB}) = \prod_{i=1}^N \left\{ \left[P(y_i = 1 | X_{i,AB}, X_{i,aB}, X_{i,Ab}, w_i) \right]^{y_i} \times \left[1 - P(y_i = 1 | X_{i,AB}, X_{i,aB}, X_{i,Ab}, w_i) \right]^{1-y_i} \right\}, \quad (1)$$

where

$$P(y_i = 1 | X_{i,AB}, X_{i,aB}, X_{i,Ab}, w_i) = \exp(\alpha + w_i \gamma + \sum_{k \in \{AB, aB, Ab\}} X_{i,k} \beta_k) / [1 + \exp(\alpha + w_i \gamma + \sum_{k \in \{AB, aB, Ab\}} X_{i,k} \beta_k)].$$

$X_{i,AB} = P(AB | M_1, M_2)$ is an allelic score of AB allelic combination from M_1 and M_2 genotypes of the i -th subject and $\beta_k \in \{AB, Ab, aB\}$ is the interaction effects of the k -th $\in \{AB, Ab, aB\}$ allelic combination. w_i is a covariate such as age or gender and γ is a coefficient of the covariate. The overall proportion of y is R/N , where R is the number of case subjects and N is the number of total subjects. As in Jung et al. [5], table 1 presents the allelic scores given the observed genotypes of two markers. The number of allelic combinations at two markers is 2^2 but since the probability of a particular allele combination is a function of the rest of the probability of three probabilities, so the number of independent variables is $2^2 - 1$. The null hypothesis ($H_0: \beta_{AB} = \beta_{Ab} = \beta_{aB} = 0$) is that there is no association caused by the interaction; therefore, disease risk of the four allelic combinations are not significantly different. Rejecting the null hypothesis would indicate there is an allelic interaction between two markers, especially particular allelic combinations generated by the different genetic characteristics between cases and controls.

Under the assumption of no covariates from (1), an intercept is a nuisance parameter. Let $U^\tau = (U_{AB}, U_{Ab}, U_{aB})^\tau$ be a score function, which is the derivative of the log-likelihood function with respect to $\beta = (\beta_{AB}, \beta_{Ab}, \beta_{aB})$ respectively. Under the null hypothesis $H_0: \beta_{AB} = \beta_{Ab} = \beta_{aB} = 0$ of no interaction effect, the maximum likelihood estimate (MLE) of the intercept is $\hat{\alpha} = \log(\hat{y}/1 - \hat{y})$. The efficient score test statistic under the null hypothesis is

$$S = U^\tau V^{-1} U, \quad (2)$$

where

$$U^\tau = U\beta(\beta_{H_0}, \hat{\alpha}) = (\sum x_{i,AB}(y_i - \bar{y}), \sum x_{i,Ab}(y_i - \bar{y}), \sum x_{i,aB}(y_i - \bar{y}))^\tau$$

and V^{-1} is the submatrix of $I^{-1}(\alpha, \beta_{AB}, \beta_{Ab}, \beta_{aB})$ corresponding to $U^\tau = (U_{AB}, U_{Ab}, U_{aB})^\tau$. Appendix A shows the detailed derivation of U^τ and $I(\alpha, \beta_{AB}, \beta_{Ab}, \beta_{aB})$. The test statistic S is a central chi-square distribution on 3 degrees of freedom.

To investigate the analytical power of the score test, we need to approximate a distribution of S under the alternative hypothesis [11]. The approximation to U under the alternative hypothesis is

$$U^* = U_\beta(\beta_{H_0}, \alpha^*) - I_{\beta\alpha}^* I_{\alpha\alpha}^{*-1} U_\alpha(\beta_{H_0}, \alpha^*) = \sum_i (y_i - \bar{y}) \begin{bmatrix} x_{i,AB} \\ x_{i,Ab} \\ x_{i,aB} \end{bmatrix} - \begin{bmatrix} \bar{x}_{AB} \\ \bar{x}_{Ab} \\ \bar{x}_{aB} \end{bmatrix}, \quad (3)$$

where $U_\alpha(\beta_{H_0}, \alpha^*) = \partial \log L / \partial \alpha$, $I_{\beta\alpha}^*$ are a submatrix of the Fisher information matrix, I , respectively (see Appendix A) and α^* is a solution of the equation $0 = \lim n^{-1} E[U_\alpha(\beta_{H_0}, \alpha)]$. Let Ω and Σ denote the mean and covariance matrix of U^* [11]. The distribution of $S^* = U^{*\tau} \Sigma^{-1} U^*$ is approximately a chi-square distribution on 3 degrees of freedom with non-centrality parameter,

$$\lambda_{score} = \Omega^\tau \Sigma^{-1} \Omega,$$

where

$$\Omega = E(U^*) = \begin{pmatrix} (f_{D_1 D_2} - f_{D_1 d_2} - f_{d_1 D_2} + f_{d_1 d_2}) \\ (f_{D_1} - f_{d_1}) \\ (f_{D_2} - f_{d_2}) \end{pmatrix} \begin{pmatrix} D_{AD_1} & D_{BD_2} \\ D_{AD_1} & D_{bD_2} \\ D_{aD_1} & D_{BD_2} \end{pmatrix} + \begin{pmatrix} (f_{D_1} - f_{d_1}) \\ (f_{D_2} - f_{d_2}) \end{pmatrix} \begin{pmatrix} D_{AD_1} & P_B \\ D_{AD_1} & P_b \\ D_{aD_1} & P_B \end{pmatrix} + \begin{pmatrix} (f_{D_2} - f_{d_2}) \\ (f_{D_1} - f_{d_1}) \end{pmatrix} \begin{pmatrix} D_{BD_2} & P_A \\ D_{bD_2} & P_A \\ D_{BD_2} & P_a \end{pmatrix},$$

$$\Sigma = V(U^*) =$$

$$N\Phi(1-\Phi) \begin{pmatrix} P_A P_B (1-0.5P_a)(1-0.5P_b) & P_A P_B P_b (1-0.5P_a) & P_A P_B P_a (1-0.5P_b) \\ P_A P_B P_b (1-0.5P_a) & P_A P_b (1-0.5P_a)(1-0.5P_B) & P_A P_B P_a P_b \\ P_A P_B P_a (1-0.5P_b) & P_A P_B P_a P_b & P_a P_B (1-0.5P_A)(1-0.5P_b) \end{pmatrix} - \Phi(1-\Phi) P_X^t P_X,$$

where $\Phi = R/N$. Furthermore, $(f_{D_1} - f_{d_1})$, $(f_{D_2} - f_{d_2})$ can be explained by the average effect of the gene substitution at each D_1^* , D_2^* , [11]. $(f_{D_1 D_2} - f_{D_1 d_2} - f_{d_1 D_2} + f_{d_1 d_2})$ is the magnitude of interaction effect. The non-centrality parameter is also a function of LD between a disease locus and a marker. Appendix B shows the detailed derivation of Ω and Σ . We can further investigate which allelic combinations are interacting to influence a disease trait by testing the allelic specific interaction effects under the null hypothesis of $H_0: \beta_C = 0$, in the logistic model with the allelic combinations of interest, $\text{logit}[P(y_i = 1 | w_i, X_{i,C})] = \alpha + w_i \gamma + X_{i,C} \beta_C$. In the extension to multiple markers, the assignment of allelic scores to each subject and modeling logistic regression are straightforward.

Cochran-Armitage Trend Regression Model and Non-Centrality Parameter

With the same allelic scores at two markers in table 1, we can model a linear trend of proportions of cases to total (sum of case and control) at each allelic combination, $p_{k,j} = r_{k,j} / n_{k,j}$, where $n_{k,j} = r_{k,j} + s_{k,j}$ for $k \in (AB, Ab, aB, ab)$, $j \in (0, 1/4, 1/2, 1)$. $r_{k,j}$, $s_{k,j}$ are the number of affected subjects and unaffected subjects having j score at k allelic combination respectively. Table 2 shows the distribution of the allelic combination AB, $Z_{AB,j}$ with the order of $(Z_{AB,0} < Z_{AB,1/4} < \dots < Z_{AB,1})$ in case-control data and the remaining combinations $Z_{Ab,j}$, $Z_{aB,j}$ can be expressed in the same way. It has been shown that regressing $p_{k,j}$ on $Z_{AB,j}$, $Z_{Ab,j}$, $Z_{aB,j}$ is equivalent to regressing y_i on $Z_{AB,j}$, $Z_{Ab,j}$, $Z_{aB,j}$ [13, 14].

As an extension of CA method, the interaction effect of two markers can be modeled as

$$y_i = \alpha + w_i \gamma + Z_{AB,i} \beta_{AB} + Z_{Ab,i} \beta_{Ab} + Z_{aB,i} \beta_{aB} + \varepsilon, \quad (4)$$

where w_i is a covariate such as age or gender and γ is the coefficient of covariate. Under the assumption of no covariates for simplicity, the theoretical regression coefficients can be derived by functions of the genetic effects and of linkage disequilibrium (LD) between a marker and a disease susceptibility locus as follows.

$$\beta = \begin{pmatrix} \alpha \\ \beta_{AB} \\ \beta_{Ab} \\ \beta_{aB} \end{pmatrix} = K^{-1} \begin{pmatrix} R \\ \frac{P_A P_B}{P_A P_b} \\ \frac{P_A P_b}{P_a P_B} \\ \frac{P_a P_B}{P_a P_b} \end{pmatrix} + \begin{pmatrix} (f_{D_1 D_2} - f_{D_1 d_2} - f_{d_1 D_2} + f_{d_1 d_2}) \\ (f_{D_1} - f_{d_1}) \\ (f_{D_2} - f_{d_2}) \end{pmatrix} \begin{pmatrix} 0 \\ D_{AD_1} D_{BD_2} \\ D_{AD_1} D_{bD_2} \\ D_{aD_1} D_{BD_2} \end{pmatrix} + \begin{pmatrix} (f_{D_1} - f_{d_1}) \\ (f_{D_2} - f_{d_2}) \end{pmatrix} \begin{pmatrix} 0 \\ D_{AD_1} P_B \\ D_{AD_1} P_b \\ D_{aD_1} P_B \end{pmatrix} + \begin{pmatrix} (f_{D_2} - f_{d_2}) \\ (f_{D_1} - f_{d_1}) \end{pmatrix} \begin{pmatrix} 0 \\ D_{BD_2} P_A \\ D_{bD_2} P_A \\ D_{BD_2} P_a \end{pmatrix},$$

where

$$K = \begin{pmatrix} 1 & P_A P_B & P_A P_b & P_a P_B \\ P_A P_B & P_A P_B (1-0.5P_a)(1-0.5P_b) & 0.5P_A P_B P_b (1-0.5P_a) & 0.5P_A P_B P_a (1-0.5P_b) \\ P_A P_b & 0.5P_A P_B P_b (1-0.5P_a) & P_A P_b (1-0.5P_a)(1-0.5P_B) & 0.25P_A P_B P_a P_b \\ P_a P_B & 0.5P_A P_B P_a (1-0.5P_b) & 0.25P_A P_B P_a P_b & P_a P_B (1-0.5P_A)(1-0.5P_b) \end{pmatrix}.$$

Likewise, $(f_{D_1} - f_{d_1}), (f_{D_2} - f_{d_2})$ are the average effect of the gene substitution at each D_1^*, D_2^* and $(f_{D_1D_2} - f_{d_1D_2} - f_{d_1D_2} + f_{d_1d_2})$ is the magnitude of interaction effect. Appendix C shows the detailed derivation of the theoretical regression coefficients of the model (4). The global test statistic for the interaction effect is

$$F_{CA} = \frac{(H\hat{\beta})^t [H(X^tX)^{-1}H^t]^{-1} (H\hat{\beta})}{Y^t [I_N - X(X^tX)^{-1}X^t] Y} \frac{N-4}{3},$$

where I_N is the $N \times N$ identity matrix and a test matrix is

$$H = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The F_{CA} follows $F(3, N-4)$ with non-centrality parameter $\lambda_{CA} = 0$ under the null hypothesis $H_0: \beta_{AB} = \beta_{Ab} = \beta_{aB} = 0$. Under the alternative hypothesis, F_{CA} is noncentral to $F(3, N-4)$ with its non-centrality parameter,

$$\lambda_{CA} \approx \frac{R+S}{\sigma^2} (\beta_{AB}, \beta_{Ab}, \beta_{aB}) [HE(X^tX)^{-1}H^t]^{-1} (\beta_{AB}, \beta_{Ab}, \beta_{aB})^t,$$

where σ^2 is the total residual variance and $N = R + S$. Furthermore, an LRT can be derived for the global interaction test of $H_0: \beta_{AB} = \beta_{Ab} = \beta_{aB} = 0$, which is $2(L_{HA} - L_{H0})$ following $\chi^2_{df=3}$. L_{HA} is a likelihood function of maximum likelihood estimates (MLE) $\hat{\alpha}, \hat{\beta}_{AB}, \hat{\beta}_{Ab}, \hat{\beta}_{aB}$ under the alternative hypothesis and L_{H0} is a likelihood function of MLE of $\hat{\alpha}$ under the null hypothesis. Likewise, a specific allelic interaction effect under the null hypothesis of $H_0: \beta_C = 0$ for $y_i = \alpha + w_i\gamma + X_{i,C}\beta_C$ can be tested in the same framework.

Model Comparisons

Due to the possibility that the significant allelic interaction results from the additive effect of one or both SNPs, the allelic based model needs to be compared to a main effects model in order to distinguish the real allelic interaction from the main effects. First, we compare the two-way interaction model with a main effect model consisting only of main effects of two putative SNPs. When a logistic regression model is utilized, the criteria for the model comparison are Akaike Information Criterion (AIC) and the smallest Bayesian Information Criterion (BIC) because of the non-nested terms of an interaction model and a main effect model. When CA regression is utilized, an artificial nesting approach called a J test [15] can be employed as follows:

$$y_i = (1 - \alpha) \cdot f(X, \beta) + \alpha \cdot g(Z, \hat{\delta}),$$

where $f(X, \beta) = \mu + X_{i,AB}\beta_{AB} + X_{i,Ab}\beta_{Ab} + X_{i,aB}\beta_{aB}$ is a two-way interaction model and $g(X, \delta) = \mu + \delta_A Z_{M1} + \delta_B Z_{M2}$ is a main effect model. Since f is linear in β , the comparison requires that one estimates δ and then fits a linear regression and test for $\alpha = 0$ using the ordinary t -statistic [15]. In addition to the test of α , AIC and BIC are also the criteria to select the best models. The same criteria are applied to the three-way interaction model comparison.

Table 2. Distribution of allelic combination, AB for case-control data

Data	$Z_{AB} = P(AB M_1, M_2)$				Total
	0	1/4	1/2	1	
Case ($y_i = 1$)	$r_{AB,0}$	$r_{AB,1/4}$	$r_{AB,1/2}$	$r_{AB,1}$	R
Control ($y_i = 0$)	$s_{AB,0}$	$s_{AB,1/4}$	$s_{AB,1/2}$	$s_{AB,1}$	S
Total	$n_{AB,0}$	$n_{AB,1/4}$	$n_{AB,1/2}$	$n_{AB,1}$	N

Results

Type I Error Rates

To evaluate the statistical robustness of the score test, the F test and the LRT for the allelic based method, simulation studies were carried out to estimate type I error rates at the 1%, 5% significance level. Six models of interaction between two unlinked disease loci were considered (see table 3). Most of the models were designed based on combinations of dominant and recessive inheritance at the genotypic level at each marker. These models are (1) Dominant or Recessive (Dom $\cup$ Rec); (2) Recessive or Recessive (Rec $\cup$ Rec); (3) Dominant and Dominant (Dom $\cap$ Dom); (4) Dominant and Recessive (Dom $\cap$ Rec); (5) Threshold whose disease risk is increased when two or more high risk alleles from either locus are present, and (6) Modified model whose homozygosity at either locus confers disease risk. For each model, we simulated 5,000 datasets using SNaP [16]. Each dataset has 200 cases and 200 controls with two unlinked genes under no LD between markers and disease loci ($D'_{AD_1} = D_{AD_1}/D_{\max} = 0$; $D'_{BD_2} = D_{BD_2}/D_{\max} = 0$; $R^2_{AD_1} = R^2_{BD_2} = 0$), disease risk allele frequencies at two loci of 0.2 ($P_{D_1} = P_{D_2} = 0.2$), and minor allele frequencies at M_1 and M_2 of 0.3 ($P_A = P_B = 0.3$). Table 4 shows the results of the empirical type I error rates of the score test and the F and LRT statistic at 1 and 5% significance level. It illustrates that all type I error rates of the three statistics at each model were close to the nominal values 1% and 5% which suggest that the proposed method is statistically robust. For different parameters such as 0.5 disease-associated allele frequencies ($P_{D_1} = P_{D_2} = 0.5$) and equal allelic frequencies of the two markers ($P_A = P_B = 0.5$), all models were also close to nominal values of 1 and 5% as in table 4 (results not shown).

Analytical Power and Sample Size Calculation

The analytical power and the required sample size of the proposed approach were calculated based on the non-

Table 3. The interaction models: *f* stands for penetrance of genotypic combination

Models	First locus	Second locus			Models	First locus	Second locus		
		D ₂ D ₂	D ₂ d ₂	d ₂ d ₂			D ₂ D ₂	D ₂ d ₂	d ₂ d ₂
Dom ∪ Rec	D ₁ D ₁	<i>f</i>	<i>f</i>	<i>f</i>	Rec ∪ Rec	D ₁ D ₁	<i>f</i>	<i>f</i>	<i>f</i>
	D ₁ d ₁	<i>f</i>	<i>f</i>	<i>f</i>		D ₁ d ₁	<i>f</i>	0	0
	d ₁ d ₁	<i>f</i>	0	0		d ₁ d ₁	<i>f</i>	0	0
Modified	D ₁ D ₁	<i>f</i>	0	<i>f</i>	Threshold	D ₁ D ₁	<i>f</i>	<i>f</i>	0
	D ₁ d ₁	<i>f</i>	0	<i>f</i>		D ₁ d ₁	<i>f</i>	0	0
	d ₁ d ₁	0	0	0		d ₁ d ₁	0	0	0
Dom ∩ Dom	D ₁ D ₁	<i>f</i>	<i>f</i>	0	Dom ∩ Rec	D ₁ D ₁	<i>f</i>	0	0
	D ₁ d ₁	<i>f</i>	<i>f</i>	0		D ₁ d ₁	<i>f</i>	0	0
	d ₁ d ₁	0	0	0		d ₁ d ₁	0	0	0

Table 4. Result of type I error rate (%) at 1% and 5% significance level from 5,000 datasets, each has 200 cases and 200 controls

Model	Logistic regression score test		CA regression F-test		CA regression LRT-test	
	<i>a</i> = 0.01	<i>a</i> = 0.05	<i>a</i> = 0.01	<i>a</i> = 0.05	<i>a</i> = 0.01	<i>a</i> = 0.05
Dom ∪ Rec	1.10	5.50	1.10	5.52	1.12	5.60
Modified	0.78	5.16	0.84	5.16	0.88	5.36
Dom ∩ Dom	0.80	5.36	0.84	5.38	0.86	5.56
Rec ∪ Rec	1.22	5.46	1.26	5.46	1.30	5.64
Threshold	1.0	5.26	1.04	5.26	1.18	5.48
Dom ∩ Rec	0.92	5.60	0.94	5.60	1.00	5.86

centrality parameter $\lambda_{score} = \Omega^T \Sigma^{-1} \Omega$ of the score test. The power to detect the interaction effect was influenced by various parameters such as LD between the putative disease loci and a marker (D_{AD_1} , D_{BD_2}), disease allele frequencies (p_{D_1} , p_{D_2}), observed marker allele frequencies (p_A , p_B), and penetrance (f) which is a proportion of individuals carrying a genotypic combination of two putative disease loci that also develop a disease. For the six two-loci interaction models, the power is calculated against each parameter respectively while the other remaining parameters are fixed at a number that provides the greatest variation across the interaction models.

Figure 2 shows the power curve of the score test S plotted against penetrance f (see table 3) at the 1% significance level. Power is calculated when sample size is 1,000 cases and 1,000 controls with a fixed value $D_{AD_1} = D_{BD_2} = 0.05$, $P_A = P_a = 0.5$; $P_B = P_b = 0.5$. Models ‘Dom ∩ Dom’ and ‘Threshold’ equivalently achieved the highest power and ‘Modified’ obtained the lowest power across the full range of penetrance f . The power of ‘Rec ∪ Rec’ is higher

than that of ‘Dom ∩ Rec’ which in turn is higher than that of ‘Dom ∪ Rec’ model. Figure 3 presents the power against the A allele frequency (p_A) of a marker M_1 , when $p_B = 0.5$ at a marker M_2 . All parameters used in figure 3 are the same as those used in figure 2 except $D_{AD_1} = \{\min(P_{D_1}, P_A) - P_{D_1}P_A\}/2$ and the penetrance $f = 0.5$ at each model. The pattern of the power in figure 3 is very similar to that of the power in figure 2 across the six models except ‘Dom ∩ Rec’ which achieves the second highest power when $p_A > 0.2$, whereas ‘Rec ∪ Rec’ is the second highest power across the full range of penetrance in figure 2.

Figure 4 illustrates the power against a linkage disequilibrium coefficient D_{AD_1} with a fixed value of $D_{BD_2} = 0.05$. The power is calculated under $f = 0.5$, $P_A = P_a = 0.5$, $P_B = P_b = 0.5$ and sample size 1,000 cases and 1,000 controls. The powers of ‘Dom ∪ Dom’ and ‘Threshold’ were identical and the power curves over the 6 models are very similar to those in figure 2 and 3. Figure 5 shows the power against the disease allelic frequency p_{D_1}

Fig. 2. Analytical power against penetrance f at 1% significance level. $P_{D_1} = 0.5$; $P_{D_2} = 0.5$; $P_A = 0.5$; $P_B = 0.5$; $D_{AD_2} = D_{BD_2} = 0.05$; the sample size of cases $R = 1,000$ and controls $S = 1,000$. Dominant or Recessive model is indicated by 'Dom $\cup$ Rec, Dominant and Dominant model by Dom $\cap$ Dom, Recessive or Recessive model by Rec $\cup$ Rec, Dominant and Recessive model by Dom $\cap$ Rec and Recessive and Recessive model by Rec $\cap$ Rec.

Color version available online

Fig. 3. Analytical power against frequency of allele A at marker M_1 when $P_a = 0.5$ at 1% significance level. $P_{D_1} = 0.5$; $P_{D_2} = 0.5$; $P_B = 0.5$; $D_{AD_1} = \{\min(P_{D_1}, P_A) - P_{D_1}, P_A\}/2$; $D_{BD_2} = 0.05$; penetrance $f = 0.5$; the sample size of cases $R = 1,000$ and controls $S = 1,000$; the notations of models are the same as figure 2.

Color version available online

Fig. 4. Analytical power against D_{AD_1} when $D_{BD_2} = 0.05$ at 1% significance level. $P_{D_1} = 0.5$; $P_{D_2} = 0.5$; $P_A = 0.5$; penetrance $f = 0.5$; the sample size of cases $R = 1,000$ and controls $S = 1,000$; the notations of models are the same as figure 2.

Color version available online

Fig. 5. Analytical power against frequency of disease allele D_1 when $P_{D_2} = 0.5$ at 1% significance level. $D_{AD_1} = \{\min(P_{D_1}, P_A) - P_{D_1}, P_A\}/2$, $D_{AD_1} = D_{BD_2} = 0.05$; penetrance $f = 0.5$; the sample size of cases $R = 1,000$ and controls $S = 1,000$; the notations of models are the same as figure 2.

Fig. 6. The required sample size of cases against penetrance f for 80% power achievement at 1% significance level. $P_{D_1} = 0.5$; $P_{D_2} = 0.5$; $P_A = 0.5$; $P_B = 0.5$; $D_{AD_1} = D_{BD_2} = 0.05$; ratio of case to control is 1. Model notations are the same as figure 2.

Table 5. Power (%) at 1% and 5% significance level at $D'_{AQ_1} = D'_{BQ_2} = 0.6$; $R^2_{AQ_1} = R^2_{BQ_2} = 0.1$ ($P_A = P_B = 0.3$) from 2,000 datasets, each has 200 cases and 200 controls

Model	Allelic-based method						Genotypic-based method			
	Logistic regression score test		CA regression F-test		CA regression LR test		CA regression score test		MDR using empirical threshold	
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
Dom $\cup$ Rec	16.40	34.80	16.65	34.85	17.30	35.55	9.85	24.90	6.30	21.45
Modified	22.90	46.75	23.45	46.85	23.95	47.40	13.75	32.00	9.30	27.50
Dom $\cap$ Dom	45.75	68.50	46.15	68.50	46.75	69.00	29.30	52.80	24.50	45.40
Rec $\cup$ Rec	57.35	78.85	57.70	78.85	58.20	79.40	38.10	64.40	31.90	56.70
Threshold	91.65	98.15	91.75	98.15	92.00	98.15	80.45	92.70	73.45	86.65
Dom $\cap$ Rec	96.15	99.30	96.30	99.30	96.45	99.30	88.95	96.70	82.60	93.15

Table 6. Power (%) at 1% and 5% significance level at $D'_{AQ_1} = D'_{BQ_2} = 0.6$; $R^2_{AQ_1} = R^2_{BQ_2} = 0.03$ ($P_A = P_B = 0.05$) from 2,000 datasets, each has 200 cases and 200 controls

Model	Allelic-based method						Genotypic-based method			
	Logistic regression score test		CA regression F-test		CA regression LR test		CA regression score test		MDR using empirical threshold	
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
Dom $\cup$ Rec	1.05	5.70	1.05	5.70	1.05	5.95	0.05	0.30	0	0.25
Modified	1.20	5.58	1.20	5.90	1.20	5.95	0.05	0.30	0	0.20
Dom $\cap$ Dom	1.95	10.05	1.95	10.05	2.15	10.35	0.20	0.95	0.2	1.20
Rec $\cup$ Rec	3.90	13.60	3.90	13.65	4.0	13.90	0.30	1.90	0.35	1.70
Threshold	8.15	22.65	8.30	22.75	8.85	23.15	0.75	4.40	1.05	4.60
Dom $\cap$ Rec	12.2	30.75	12.3	30.80	12.80	31.55	1.71	5.90	1.80	7.45

at D_1^* when p_{D_2} is fixed $p_{D_2} = 0.5$, which have the same parameters as figure 4 except $D_{AD_1} = \{\min(P_{D_1}, P_A) - P_{D_1}P_A\}/2$. Likewise, the power of 'Dom $\cap$ Dom' is the highest and 'Dom $\cup$ Rec' outperforms 'Threshold'. Surprisingly, the power of the 'Modified' model was the third highest power at $p_{D_1} < 0.32$, but it leveled off at a power of approximately 0.6.

Figure 6 shows the required sample size for cases and controls across the full range of penetrance f to achieve 80% power at 1% significance level when we assume the ratio of cases to controls is 1 (equal size). The sample size is calculated under $P_A = P_a = 0.5$; $P_B = P_b = 0.5$ and $D_{AD_1} = D_{BD_2} = 0.05$. As shown in figure 6, at $f > 0.5$, all models need less than 1,500 case (1,500 control) subjects except the 'Modified', 'Dom $\cap$ Dom', and 'Threshold' model which need less than 1,000 case subjects. The 'Modified' model needs the largest sample size over the full range of f , followed by 'Dom $\cup$ Rec' and 'Rec $\cup$ Rec' models.

Power Comparisons

Logistic Regression versus Cochran-Armitage (CA) Regression

In order to compare the score test derived from the logistic regression with the F test and LRT from CA regression, we performed simulation studies using SNaP with two scenarios. The first scenario was defined by common variants of SNPs with MAFs at the two loci of 0.3 ($P_A = P_B = 0.3$), disease-associated allele frequencies of 0.2 ($P_{D_1} = P_{D_2} = 0.2$) for a more realistic situation. The LD between a disease locus and a marker is $D'_{AD_1} = D'_{BD_2} = 0.6$; $R^2_{AD_1} = R^2_{BD_2} = 0.1$. The second scenario was defined by rare variants of SNPs with a minor allelic frequency at the two loci of 0.05 ($P_A = P_B = 0.05$) and disease-associated allele frequencies of 0.5 ($P_{D_1} = P_{D_2} = 0.5$) that are increased due to the floor effect to have the similar LD, $D'_{AD_1} = D'_{BD_2} = 0.6$; $R^2_{AD_1} = R^2_{BD_2} = 0.3$. Both scenarios have a penetrance $f = 0.2$. For each model 2,000 datasets

were simulated. Each dataset has 200 cases and 200 controls. Table 5 shows the empirical power of the score test, F and LRT at the 1 and 5% significance level in the common variants; table 6 shows the results of the rare variants of SNPs. The power of the three tests in both scenarios is asymptotically equivalent across the 6 interaction models.

Allelic-Based Method versus Genotypic-Based Methods

We have compared the allelic-based approach with two genotypic-based approaches. One is a score test of a genotypic-based model using a genotypic score in a logistic regression framework as follows.

$$\begin{aligned} & \text{logit}\left[p(y_i = 1 | w_i \gamma, Z_{i,k=(AA,BB)}, \dots, Z_{i,k=(aa,bb)})\right] \\ & = \alpha + w_i \gamma + \sum_{k=1}^8 Z_{i,k=(jl,mn)} \beta_k, \end{aligned}$$

where $j, l = (A, \text{ or } a), m, n = (B, \text{ or } b)$ and

$$Z_{i,k=(jl,mn)} = \begin{cases} 1 & \text{if } G_k = (jl,mn) \\ 0 & \text{otherwise} \end{cases}.$$

Under the null hypothesis of no interaction between two unlinked loci, $H_0: \beta_{AABB} = \dots = \beta_{aaBb} = 0$. The score test can be derived from the logistic regression, which is approximately equivalent with the F and LRT from the CA regression. The other genotypic-based method we have compared is multifactor dimensionality reduction (MDR). Due to the use of cross-validation to calculate a prediction accuracy, it is not directly comparable [17]. Therefore, we used a chi-square test of homogeneity for high risk and low risk groups between cases and controls and estimated empirical type I error rates from the 5,000 null datasets and the empirical power of each model from the 2,000 datasets after controlling for the type I error rates at each model in both scenarios [17]. Tables 5 and 6 presented the power of the score test for the allelic-based method and the genotypic-based methods. The tables illustrate that the allelic-based method outperformed both a genotypic-based method using a genotypic score and the MDR.

Application to Major Depression Disorder

We applied our method to Major Depression Disorder (MDD) case and control data provided by dbGaP [6]. The 1,754 cases with the disease endpoint of diagnosis and 1,763 controls without the MDD were genotyped on the Perlegen 600K platform. We selected 7 candidate genes related with a serotonergic system and a brain-derived neurotrophic factor (BDNF) associated with a pathway of

MDD. There were 78 SNPs available from the 7 candidate genes (*SLC6A4*, *HTR2A*, *TPH1*, *TPH2*, *ITGB3*, *COMT* and *BDNF*). As quality control procedures, SNPs that have significant deviation from Hardy-Weinberg equilibrium ($p \text{ value} < 6.0 \times 10^{-4}$) and a minor allele frequency less than 0.01 were removed. 64 tag SNPs using Haploview were selected to reduce the number of tests due to the issue of multiple comparison. The procedure for the proposed method to search for genes associated with the MDD through their interaction consists of multiple step. First, two genes among the 7 genes were selected, then we performed a two-way interaction analysis with an SNP from each gene. All SNPs in the two unlinked genes were tested in a pair-wise fashion. 1,470 SNP pair combinations from 21 two-gene combinations were tested. At each two-way interaction, we performed a score test, an F test and an LRT. Only significant SNP pairs that have p values less than 3.4×10^{-5} after adjustment for Bonferroni correction were compared with a model consisting only of a main effect of each SNP to distinguish real allelic interactions. The criteria to select the best model between an interaction model and a main effect model is (1) the p value of the global test is less than 3.4×10^{-5} and (2) the p value of the model comparison between the interaction model and the main effect model is less than 0.01, and (3) the smaller Akaike Information Criterion (AIC) from the following models.

$$\text{Main effect: } H_{A1}: \text{logit}[p(y_i = 1 | w_i, X_{i,A}, X_{i,B})] = \mu + w_i \gamma + X_{i,A} \beta_A + X_{i,B} \beta_B$$

$$\text{Two-way: } H_{A2}: \text{logit}[p(y_i = 1 | w_i, X_{i,AB}, X_{i,Ab}, X_{i,aB})] = \mu + w_i \gamma + X_{i,AB} \beta_{AB} + X_{i,Ab} \beta_{Ab} + X_{i,aB} \beta_{aB}$$

Second, SNP pairs detected by an interaction at the first step were considered for three-way interaction analysis by adding an SNP from one of the 5 genes that were not in the two-way interaction model. This procedure is called a 'forward selection' procedure. Similarly to the two-way interaction analysis, we carried out a comparison test of the significant three-way interaction model with a main effects model. The best three-way interaction model was selected by the same three criteria. Finally, these steps were continued until no further high dimensional interaction models are detected.

Based on the 7 selected genes with 64 tagging SNPs, we identified two interacting SNPs: rs11030104 in *BDNF* gene and rs9526240 in *HTR2A* gene ($p \text{ value} = 2.3 \times 10^{-5}$ for a score and LRT test; $p \text{ value} = 8.11 \times 10^{-6}$ for a comparison test between H_{A2} of a two-way interaction model and H_{A1} of a main effects model). Two pairs of SNPs, rs6265 and rs11030104 in *BDNF*, and rs9526240

and rs6561335 in *HTR2A* are in LD with $R^2 = 0.89$ and $R^2 = 0.8$, respectively. These 4 pair-wise SNPs are also significantly interacting (p values are between 2.4×10^{-5} and 2.6×10^{-5}). Based on the interacting SNPs, we performed a three-way interaction analysis and detected a three-way interaction of rs11030104 in *BDNF*, rs9526240 in *HTR2A* and rs652458 in *TPH1* genes (p value for the interaction model is 5.0×10^{-5} ; p value for the comparison with a main effect is 0.002). The interaction of *BDNF* and *HTR2A* was identified in schizophrenia [18] and in unipolar depression [19]. The association of *SLC6A4*, *HTR2A* was detected in the MDD [20].

Discussion

In this study, a novel statistical approach to identify gene-gene interaction contributing to a disease trait in a case-control sampling design was proposed. In contrast with most of the currently available methods detecting interaction at the genotypic level, the proposed method searches for interaction at the allelic level with a new definition of interaction. Interaction occurs when the contribution to disease of a particular allele inherited in one gene depends on a particular allele inherited in other unlinked genes. Due to the new definition, the interpretation of result may be more straightforward at the allelic rather than at the genotypic level because the interacting alleles can be explained by combinations of disease-associated alleles. The score test derived from a logistic regression model is computationally efficient because one does not need maximum likelihood estimations. As shown in the simulation studies, the power of the score test is asymptotically equivalent to that of the F test and LRT derived from the CA regression. Furthermore, our approach allows an adjustment for non-genetic covariates as nuisance parameters, which may be necessary in complex trait analysis.

The terms 'gene-gene interaction' or 'epistasis' are used with different meanings within different areas of genetics [6]. Recently, Phillips reviewed the conceptual definition and the differences, and also pointed out that there are conceptual barriers to generating a more unified definition [6]. Epistasis was originally defined as a mutation (or an allele) that masks or stimulates the effects of the other mutation on a phenotype. Gene-gene interaction is a broad concept that encompasses a variety of genetic phenomena. In this paper, we use gene-gene interaction in a general way to involve the epistasis. Detection of the genetic interaction at the allelic level is conceptually

similar to epistasis for which the biological interpretation is straightforward.

For the high-order interaction of multiple markers, the issue of multiple comparisons still remains either in independent SNPs or in dependent SNPs due to multiple procedures to distinguish the real allelic interaction from the genetic effects. As illustrated in the application with MDD, each procedure involves testing two or three hypotheses and many combinations of SNPs involve testing many hypothesis testings. Bonferroni correction for the proposed interaction models was used to find the significant sets of SNPs at each step depending on the number of SNPs being analyzed. For each comparison test, a threshold of 0.01 or less was used for the significance level. It has been known that the Bonferroni correction often produces conservative results when LD is ignored. Permutation analysis is an alternative approach to generate an empirical null distribution of each step; however, it may not be computationally feasible for real data applications because of the multiple hierarchical steps for high dimensional interactions. Some machine learning approaches use cross-validation to predict an accuracy rate and permutation analysis to estimate p value, both of which are computationally expensive.

Searching for gene-gene interaction is a necessary procedure in a genome-wide association study because a disease is a complicated network of multiple genes. Effort to develop computationally efficient methods to detect the interacting SNPs at the whole genome level is underway. Future work should extend this approach to the genome-wide association study. Alternatively, an incorporation of biological knowledge of gene pathway to prioritize biologically important genes may reduce the search space and increase power to detect interacting genes associated with a disease.

Acknowledgments

We would like to thank two anonymous reviewers and Peter H. Baenziger for thoughtful critiques that have resulted in a much improved manuscript. This work was supported by NIH/NHLBI R01 HL095086-01. Funding support for Major Depression: Stage 1 Genomewide Association in Population-Based Samples was provided by NIH and the genotyping of samples was provided through the Genetic Association Information Network (GAIN). The dataset(s) used for the analyses described in this manuscript were obtained from the GAIN Database found at <http://view.ncbi.nlm.nih.gov/dbgap-controlled-through/dbGaP> accession number phs000020.v1.p1. Samples and associated phenotype data for Major Depression: Stage 1 Genomewide Association in Population-Based Samples were provided by P. Sullivan.

Appendix A

Let $U^\tau = (U_{AB}, U_{Ab}, U_{aB})^\tau$ be a score function, which is the derivative of the log-likelihood function with respect to $\beta = (\beta_{AB}, \beta_{Ab}, \beta_{aB})$ respectively

$$U^\tau = U_{\beta}(\beta_{H_0}, \hat{\alpha}) = (U_{AB}, U_{Ab}, U_{aB})^\tau = \left(\frac{\partial \log L}{\partial \beta_{AB}}, \frac{\partial \log L}{\partial \beta_{Ab}}, \frac{\partial \log L}{\partial \beta_{aB}} \right)^\tau = \left(\sum x_{i,AB} (y_i - \bar{y}), \sum x_{i,Ab} (y_i - \bar{y}), \sum x_{i,aB} (y_i - \bar{y}) \right)^\tau$$

Let $I(\alpha, \beta_{AB}, \beta_{Ab}, \beta_{aB})$ be the observed Fisher information matrix, which is the second derivative of the log-likelihood function with respect to $(\alpha, \beta_{AB}, \beta_{Ab}, \beta_{aB})$ respectively

$$I(\alpha, \beta_{AB}, \beta_{Ab}, \beta_{aB}) = \bar{y}(1 - \bar{y}) \begin{pmatrix} n & \sum x_{i,AB} & \sum x_{i,Ab} & \sum x_{i,aB} \\ \sum x_{i,AB} & \sum x_{i,AB}^2 & \sum x_{i,AB} x_{i,Ab} & \sum x_{i,AB} x_{i,aB} \\ \sum x_{i,Ab} & \sum x_{i,AB} x_{i,Ab} & \sum x_{i,Ab}^2 & \sum x_{i,Ab} x_{i,aB} \\ \sum x_{i,aB} & \sum x_{i,AB} x_{i,aB} & \sum x_{i,Ab} x_{i,aB} & \sum x_{i,aB}^2 \end{pmatrix} = \begin{pmatrix} I_{\alpha\alpha} & I_{\alpha\beta} \\ I_{\beta\alpha} & I_{\beta\beta} \end{pmatrix},$$

where $I_{\alpha\alpha} = \partial^2 \log L / \partial \alpha^2$, $I_{\alpha\beta} = \partial^2 \log L / \partial \alpha \partial \beta$, $I_{\beta\beta} = \partial^2 \log L / \partial \beta^2$.

Appendix B

Under the assumption of no covariates for simplification, we calculate non-centrality parameter $\lambda_{score} = \Omega^\tau \Sigma^{-1} \Omega$ where Ω and Σ are the mean and covariance matrix of U^* respectively. Let $X = (X_{AB}, X_{Ab}, X_{aB}, X_{ab})$, where $X_{AB} = (X_{1,AB}, X_{2,AB}, \dots, X_{n,AB})^\tau$. $y_i = 1$ if a subject is affected by a disease and $y_i = 0$ if not affected. Let $f_{klmn} = P(y = 1 \mid D_1^*, D_2^*)$ be the joint penetrance for the two disease loci D_1^* and D_2^* , $k, l = (D_1, d_1)$, $m, n = (D_2, d_2)$.

$$\begin{aligned} E(X_{AB} y) &= \sum_{klmn} X_{AB} \times y \times P(y = 1, D_1^* = D_k D_l, D_2^* = D_m D_n, M_1, M_2) \\ &= \sum_{klmn} x_{AB} \times P(y = 1 \mid D_1^*, D_2^*, M_1, M_2) \times P(D_1^*, D_2^*, M_1, M_2) \\ &= 1 \times \left\{ f_{1111} \times P(D_1^* = D_1 D_1, D_2^* = D_2 D_2, M_1 = AA, M_2 = BB) + \dots + f_{2222} \times P(d_1 d_1, d_2 d_2, AA, BB) \right\} \\ &\quad + \frac{1}{2} \times \left\{ f_{1111} \times P(D_1 D_1, D_2 D_2, AA, Bb) + \dots + f_{2222} \times P(d_1 d_1, d_2 d_2, AA, Bb) \right\} \\ &\quad + \frac{1}{2} \times \left\{ f_{1111} \times P(D_1 D_1, D_2 D_2, Aa, BB) + \dots + f_{2222} \times P(d_1 d_1, d_2 d_2, Aa, BB) \right\} \\ &\quad + \frac{1}{4} \times \left\{ f_{1111} \times P(D_1 D_1, D_2 D_2, Aa, Bb) + \dots + f_{2222} \times P(d_1 d_1, d_2 d_2, Aa, Bb) \right\} \\ &= f_{1111} \left\{ p(AD_1)^2 P(BD_2)^2 + \frac{1}{2} \times 2p(AD_1)^2 P(BD_2) P(bD_2) + \frac{1}{2} \times 2p(AD_1) p(aD_1) P(BD_2)^2 + \frac{1}{4} \times 4p(AD_1) p(aD_1) P(BD_2) P(bD_2) \right\} \\ &\quad + \dots + f_{2222} \left\{ p(Ad_1)^2 P(Bd_2)^2 + \frac{1}{2} \times 2p(Ad_1)^2 P(Bd_2) P(bd_2) + \frac{1}{2} \times 2p(Ad_1) p(ad_1) P(Bd_2)^2 + \frac{1}{4} \times 4p(Ad_1) p(ad_1) P(Bd_2) P(bd_2) \right\} \\ E(X_{AB} y) &= P(AD_1) P(BD_2) \{ f_{1111} P(D_1) P(D_2) + f_{1112} P(D_1) P(d_2) + f_{1211} P(d_1) P(D_2) + f_{1212} P(d_1) P(d_2) \} \\ &\quad + P(AD_1) P(Bd_2) \{ f_{1112} P(D_1) P(D_2) + f_{1122} P(D_1) P(d_2) + f_{1212} P(d_1) P(D_2) + f_{1222} P(d_1) P(d_2) \} \\ &\quad + P(Ad_1) P(BD_2) \{ f_{1211} P(D_1) P(D_2) + f_{1212} P(D_1) P(d_2) + f_{2211} P(d_1) P(D_2) + f_{2212} P(d_1) P(d_2) \} \\ &\quad + P(Ad_1) P(Bd_2) \{ f_{1212} P(D_1) P(D_2) + f_{1222} P(D_1) P(d_2) + f_{2212} P(d_1) P(D_2) + f_{2222} P(d_1) P(d_2) \} \end{aligned}$$

Utilizing $P(AD_1) = D_{AD_1} + P_A P_{D_1}$; $P(Ad_1) = D_{Ad_1} + P_A P_{d_1}$; $P(BD_2) = D_{BD_2} + P_B P_{D_2}$; $P(bd_2) = D_{bd_2} + P_B P_{d_2}$ and $f_{D_1 D_2} = f_{1111}P(D_1)P(D_2) + f_{1112}P(D_1)P(d_2) + f_{1211}P(d_1)P(D_2) + f_{1212}P(d_1)P(d_2)$, we derived

$$E(X_{AB}y) = (D_{AD_1} + P_A P_{D_1}) \times (D_{BD_2} + P_B P_{D_2}) f_{D_1 D_2} + \dots + (D_{Ad_1} + P_A P_{d_1}) \times (D_{bd_2} + P_B P_{d_2}) f_{d_1 d_2} \\ = f P_A P_B + (f_{D_1 D_2} - f_{D_1 d_2} - f_{d_1 D_2} + f_{d_1 d_2}) D_{AD_1} D_{BD_2} + (f_{D_1} - f_{d_1}) D_{AD_1} P_B + (f_{D_2} - f_{d_2}) D_{BD_1} P_A.$$

$E(X_{AB}y)$, $E(X_{aB}y)$ can be calculated in the same way.

Using $E(X_{AB}) = P_A P_B$, $E(X_{Ab}) = P_A P_b$, $E(X_{aB}) = P_a P_B$, $E(X_{ab}) = P_a P_b$, $E(X_{AB}^2) = P_A P_B (1 - 0.5P_a)(1 - 0.5P_b)$,

$E(X_{AB}X_{aB}) = 0.5P_A P_B P_b (1 - 0.5P_a)$, ..., $E(X_{aB}X_{ab}) = 1/4 P_a P_B P_b (1 - 2P_a)$,

$$E \begin{pmatrix} \sum x_{i,AB}^2 & \sum x_{i,AB} x_{i,Ab} & \sum x_{i,AB} x_{i,aB} \\ \sum x_{i,AB} x_{i,Ab} & \sum x_{i,Ab}^2 & \sum x_{i,Ab} x_{i,aB} \\ \sum x_{i,AB} x_{i,aB} & \sum x_{i,Ab} x_{i,aB} & \sum x_{i,aB}^2 \end{pmatrix} = N \frac{R}{N} \left(1 - \frac{R}{N} \right) \begin{pmatrix} P_A P_B (1 - 0.5P_a)(1 - 0.5P_b) & 0.5P_A P_B P_b (1 - 0.5P_a) & 0.5P_A P_B P_a (1 - 0.5P_b) \\ 0.5P_A P_B P_b (1 - 0.5P_a) & P_A P_b (1 - 0.5P_a)(1 - 0.5P_b) & 0.25P_A P_B P_a P_b \\ 0.5P_A P_B P_a (1 - 0.5P_b) & 0.25P_A P_B P_a P_b & P_a P_B (1 - 0.5P_a)(1 - 0.5P_b) \end{pmatrix}$$

Therefore, Ω and Σ can be expressed as follows.

$$\Omega = E(U^*) = E \left(\begin{pmatrix} \sum_i y_i x_{i,AB} \\ \sum_i y_i x_{i,Ab} \\ \sum_i y_i x_{i,aB} \end{pmatrix} - \bar{y} \begin{pmatrix} \sum_i x_{i,AB} \\ \sum_i x_{i,Ab} \\ \sum_i x_{i,aB} \end{pmatrix} \right)^T \\ = \left[\frac{R}{N} \begin{pmatrix} P_A P_B \\ P_A P_b \\ P_a P_B \end{pmatrix} + (f_{D_1 D_2} - f_{D_1 d_2} - f_{d_1 D_2} + f_{d_1 d_2}) \begin{pmatrix} D_{AD_1} D_{BD_2} \\ D_{AD_1} D_{bD_2} \\ D_{aD_1} D_{BD_2} \end{pmatrix} + (f_{D_1} - f_{d_1}) \begin{pmatrix} D_{AD_1} P_B \\ D_{AD_1} P_b \\ D_{aD_1} P_B \end{pmatrix} + (f_{D_2} - f_{d_2}) \begin{pmatrix} D_{BD_2} P_A \\ D_{bD_2} P_A \\ D_{BD_2} P_a \end{pmatrix} - \frac{R}{N} \begin{pmatrix} P_A P_B \\ P_A P_b \\ P_a P_B \end{pmatrix} \right] \\ = \left[(f_{D_1 D_2} - f_{D_1 d_2} - f_{d_1 D_2} + f_{d_1 d_2}) \begin{pmatrix} D_{AD_1} D_{BD_2} \\ D_{AD_1} D_{bD_2} \\ D_{aD_1} D_{BD_2} \end{pmatrix} + (f_{D_1} - f_{d_1}) \begin{pmatrix} D_{AD_1} P_B \\ D_{AD_1} P_b \\ D_{aD_1} P_B \end{pmatrix} + (f_{D_2} - f_{d_2}) \begin{pmatrix} D_{BD_2} P_A \\ D_{bD_2} P_A \\ D_{BD_2} P_a \end{pmatrix} \right]$$

$$\Sigma = V(U^*) = I_{\beta\beta} - I_{\beta\alpha} I_{\alpha\alpha}^{-1} I_{\alpha\beta}$$

$$= \bar{y}(1 - \bar{y}) \begin{pmatrix} \sum x_{i,AB}^2 & \sum x_{i,AB} x_{i,Ab} & \sum x_{i,AB} x_{i,aB} \\ \sum x_{i,AB} x_{i,Ab} & \sum x_{i,Ab}^2 & \sum x_{i,Ab} x_{i,aB} \\ \sum x_{i,AB} x_{i,aB} & \sum x_{i,Ab} x_{i,aB} & \sum x_{i,aB}^2 \end{pmatrix} - \frac{1}{N} \bar{y}^2 (1 - \bar{y})^2 \begin{pmatrix} \sum x_{i,AB} \\ \sum x_{i,Ab} \\ \sum x_{i,aB} \end{pmatrix} \begin{pmatrix} \sum x_{i,AB} & \sum x_{i,Ab} & \sum x_{i,aB} \end{pmatrix} \\ = N \frac{R}{N} \left(1 - \frac{R}{N} \right) \begin{pmatrix} P_A P_B (1 - 0.5P_a)(1 - 0.5P_b) & P_A P_B P_b (1 - 0.5P_a) & P_A P_B P_a (1 - 0.5P_b) \\ P_A P_B P_b (1 - 0.5P_a) & P_A P_b (1 - 0.5P_a)(1 - 0.5P_b) & P_A P_B P_a P_b \\ P_A P_B P_a (1 - 0.5P_b) & P_A P_B P_a P_b & P_a P_B (1 - 0.5P_a)(1 - 0.5P_b) \end{pmatrix} - \frac{R}{N} \left(1 - \frac{R}{N} \right) \begin{pmatrix} P_A P_B \\ P_A P_b \\ P_a P_B \end{pmatrix} \begin{pmatrix} P_A P_B & P_A P_b & P_a P_B \end{pmatrix}.$$

Appendix C

We multiply both sides by $X = (X_{AB}, X_{Ab}, X_{aB}, X_{ab})$, where $X_{AB} = (X_{1,AB}, X_{2,AB}, \dots, X_{n,AB})^T$ and take the expectation in order to derive regression coefficients in equation (3)

$$E \begin{pmatrix} X_{AB} y \\ X_{Ab} y \\ X_{aB} y \\ X_{ab} y \end{pmatrix} = E \begin{pmatrix} X_{AB}^2 & X_{AB} X_{Ab} & X_{AB} X_{aB} & X_{AB} X_{ab} \\ X_{Ab} X_{AB} & X_{Ab}^2 & X_{Ab} X_{aB} & X_{Ab} X_{ab} \\ X_{aB} X_{AB} & X_{aB} X_{Ab} & X_{aB}^2 & X_{aB} X_{ab} \\ X_{ab} X_{AB} & X_{ab} X_{Ab} & X_{ab} X_{aB} & X_{ab}^2 \end{pmatrix} \times \begin{pmatrix} \beta_{AB} \\ \beta_{Ab} \\ \beta_{aB} \\ \beta_{ab} \end{pmatrix}$$

The expectation of each element in matrix $E(X'X)$ is calculated with the assumption of no covariates as shown in Appendix A.

$$\begin{aligned}
E(X_{AB}y) &= \sum_{klmn} X_{AB} \times y \times P(y=1, D_1^* = D_k D_l, D_2^* = D_m D_n, M_1, M_2) \\
&= \sum_{klmn} x_{AB} \times P(y=1|D_1^*, D_2^*, M_1, M_2) \times P(D_1^*, D_2^*, M_1, M_2) \\
&= (D_{AD_1} + P_A P_{D_1}) \times (D_{BD_2} + P_B P_{D_2}) f_{D_1 D_2} + \dots + (D_{AD_1} + P_A P_{D_1}) \times (D_{BD_2} + P_B P_{D_2}) f_{d_1 d_2} \\
&= f P_A P_B + (f_{D_1 D_2} - f_{D_1 d_2} - f_{d_1 D_2} + f_{d_1 d_2}) D_{AD_1} D_{BD_2} + (f_{D_1} - f_{d_1}) D_{AD_1} P_B + (f_{D_2} - f_{d_2}) D_{BD_1} P_A \\
\beta &= \begin{pmatrix} \alpha \\ \beta_{AB} \\ \beta_{Ab} \\ \beta_{aB} \end{pmatrix} = K^{-1} \left[\frac{R}{N} \begin{pmatrix} 1 \\ P_A P_B \\ P_A P_b \\ P_a P_B \end{pmatrix} + (f_{D_1 D_2} - f_{D_1 d_2} - f_{d_1 D_2} + f_{d_1 d_2}) \begin{pmatrix} 0 \\ D_{AD_1} D_{BD_2} \\ D_{AD_1} D_{bD_2} \\ D_{aD_1} D_{BD_2} \end{pmatrix} + (f_{D_1} - f_{d_1}) \begin{pmatrix} 0 \\ D_{AD_1} P_B \\ D_{AD_1} P_b \\ D_{aD_1} P_B \end{pmatrix} + (f_{D_2} - f_{d_2}) \begin{pmatrix} 0 \\ D_{BD_2} P_A \\ D_{bD_2} P_A \\ D_{BD_2} P_a \end{pmatrix} \right], \quad (2)
\end{aligned}$$

where

$$K = \begin{pmatrix} 1 & P_A P_B & P_A P_b & P_a P_B \\ P_A P_B & P_A P_B (1 - 0.5 P_a) (1 - 0.5 P_b) & 0.5 P_A P_B P_b (1 - 0.5 P_a) & 0.5 P_A P_B P_a (1 - 0.5 P_b) \\ P_A P_b & 0.5 P_A P_B P_b (1 - 0.5 P_a) & P_A P_b (1 - 0.5 P_a) (1 - 0.5 P_b) & 0.25 P_A P_B P_a P_b \\ P_a P_B & 0.5 P_A P_B P_a (1 - 0.5 P_b) & 0.25 P_A P_B P_a P_b & P_a P_B (1 - 0.5 P_a) (1 - 0.5 P_b) \end{pmatrix}.$$

References

- Ritchie MD, Hahn LW, Roodi N, Bailey R, Dupont WD, Parl FF, Moore J: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;69:138–147.
- Cook NR, Zee RY, Ridker PM: Tree and spline based association analysis of gene $\times$ gene interaction models for ischemic stroke. *Stat Med* 2004;23:1439–1453.
- Chen SH, Sun J, Dimitrov L, Turner AR, Adams TS, Meyers DA, Chang BL, Zheng SL, Grönberg H, Xu J, Hsu FC: A support vector machine approach for detecting gene-gene interaction. *Genet Epidemiol* 2008;32:152–167.
- Kooperberg C, Ruczinski I: Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol* 2005;28:157–170.
- Jung J, Sun B, Kwon D, Koller D, Foroud T: Allelic based gene-gene interaction association with quantitative trait loci. *Genet Epidemiol*, 2009, in press.
- Pillips PC: Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev* 2008;9:855–867.
- Boomsma DI, Willemsen G, Sullivan PF, Heutink P, Meijer P, Sondervan D, Kluft C, Smit G, Nolen WA, Zitman FG, Smit JH, Hoogendijk WJ, van Dyck R, de Geus EJ, Penninx BW: Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. *Eur J Hum Genet* 2008;16:335–342.
- Anderson JA: Separate sample logistic discrimination. *Biometrika* 1972;59:19–35.
- Prentice RL, Pyke R: Logistic disease incidence models and case-control studies. *Biometrika* 1979;66:403–411.
- Scott AJ, Wild CJ: Fitting regression models to case-control data by maximum likelihood. *Biometrika* 1997;84:57–71.
- Self SG, Mauritsen RH: Power/Sample size calculations for generalized linear models. *Biometrics* 1988;44:79–86.
- Falconer DS, Mackay TFC: Introduction to quantitative genetics, ed 4. London, Longman, 1996.
- Armitage P: Tests for linear trends in proportions and frequencies. *Biometrics* 1955; 11:375–386.
- Cochran WG: Some methods for strengthening the common tests. *Biometrics* 1954;10: 417–451.
- McAleer M: Exact tests of a model against nonnested alternatives. *Biometrika* 1983;70: 285–288.
- Nothnagel M: Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-control data by supervised learning methods. *Am J Hum Genet* 2002; 71(suppl):A2363.
- Millstein J, Conti DV, Gilliland FD, Gauderman WJ: A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Gene* 2006;78:15–27.
- Alfimova MV, Alfimova MV, Lezheiko TV, Golimbet VE, Korovaitseva GI, Lavrushkina OM, Kolesina Niu, Frolova LP, Muratova AA, Abramova LI, Kaleda VG: Investigation of association of the brain-derived neurotrophic factor (BDNF) and a serotonin receptor 2A (5-HTR2A) genes with voluntary and involuntary attention in schizophrenia. *Zh Nevrol Psikhiatr Im S S Korsakova* 2008; 108:62–69.
- Kotte A, McQuaid JR, Kelsoe JR: Psychotherapeutic mechanisms of change: the role of genes in depression treatment outcome, Abstract, *Am Soc Hum Genet*, 2007.
- Lazary J, Lazary A, Gonda X, Benko A, Molnar E, Juhasz G, Bagdy G: New evidence for the association of the serotonin transporter gene (SLC6A4) haplotypes, threatening life events, and depressive phenotype. *Biol Psychiatry* 2008;64:498–504.