Closed-form Wald tests for genome-wide analysis of gene-gene interactions

Zhaoxia Yu[1*], Michael Demetriou[2,3], Daniel L. Gillen[1]

[1]Department of Statistics, University of California, Irvine, CA 92697, USA

[2]Department of Neurology, University of California, Irvine, CA 92697, USA

[3]Department of Microbiology & Molecular Genetics, University of California, Irvine, CA 92697, USA

*Corresponding author: yu.zhaoxia@ics.uci.edu

Abstract

Despite the successful discovery of hundreds of variants for complex human traits using genome-wide association studies, the degree to which genes jointly affect disease risk is largely unknown. One obstacle toward this goal is that the computational effort required for testing gene-gene interactions is enormous. As a result, numerous computationally efficient tests were recently proposed, such as PLINK [1], Boost [2], and a joint test [3]. However, the validity of these methods often relies on unrealistic assumptions such as additive main effects, main effects at only one SNP, and no linkage disequilibrium. Here we propose to use closed-form Wald tests. The Wald tests are asymptotically equivalent to the corresponding likelihood ratio tests, largely considered to be the gold standard tests but generally too computationally demanding for genome-wide interaction analysis. Simulation studies show that the Wald tests have very similar performance with their computationally intensive counterparts. Applying the proposed tests to a genome-wide study of multiple sclerosis, we identify interactions within the major histocompatibility complex region. In this application, we found that (1) focusing on pairs where both single nucleotide polymorphisms (SNPs) are marginally significant leads to more significant interactions when compared to focusing on pairs where at least one SNP is marginally significant; and (2) parsimonious parameterization of interaction effects might decrease, rather than increase, statistical power.

Keywords: epistasis; gene-gene interaction; genome-wide; Wald test

Author Summary

Testing gene-gene interactions requires a significant amount of computational effort, which is a major hurdle toward the understanding of how genes jointly affect disease risk. As a result, numerous computationally efficient tests were recently proposed. However, the validity of these tests relies on various model assumptions that are untestable and may often be violated in practice. Here we propose to use closed-form Wald tests. Using both simulated and real data, we show that the proposed tests are statistically valid and computationally feasible for genome-wide analysis of gene-gene interactions.

Introduction

Genome-wide association studies have led to the discovery of hundreds of common variants for complex human traits. However, the identified variants to date only explain a small fraction of heritability, leaving the majority of genetic determinants yet to be discovered. Emerging evidence suggests that gene-gene interactions, also known as epistasis, may explain part of the missing heritability [4-6]. There is little doubt that testing gene-gene interactions is of fundamental importance for defining the etiology of diseases. The knowledge of gene-gene interactions may also be used to further the goal of personalized medicine [7]. For example, if a treatment alters the function of a gene, then patients with varying genetic background are expected to respond differently to the treatment if the targeting gene interacts with other genes.

In the literature, a variety of definitions for gene-gene interactions have been used by investigators in different disciplines. For example, Bateson [8] first defined a gene-gene interaction as a phenomenon in which one gene masks the effect of another. If we limit the definitions to those used by statisticians, there is no unique definition either, as the presence and magnitude of a gene-gene interaction is model and scale dependent. Some excellent discussions about varying definitions can be found in [7, 9-11].

In this article, we define a gene-gene interaction as the departure from a main-effects linear model on the log odds of disease. More specifically, we model the disease risk for a subject with genotype *g* at SNP 1 and genotype *h* at SNP 2 using the following logistic regression

$$\begin{aligned}\text{logit}(\pi_{gh}) = \mu + \alpha_1 I(g=1) + \alpha_2 I(g=2) + \beta_1 I(h=1) + \beta_2 I(h=2) + \\ \lambda_{11} I(g=1)I(h=1) + \lambda_{12} I(g=1)I(h=2) + \lambda_{21} I(g=2)I(h=1) + \lambda_{22} I(g=2)I(h=2).\end{aligned} \quad (1)$$

Here we code genotypes at each SNP using the number 0 for the common homozygote, 1 for the heterozygote, and 2 for the rare homozygote. In model (1), the parameter μ is not related to genetic effects, α_1 and α_2 are the main effects of SNP 1, β_1 and β_2 are the main effects of SNP 2, and $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are the interaction effects between the two SNPs.

To test the interaction between two SNPs, one often uses the likelihood ratio test (LRT) stemming from a binomial probability model [9, 12]. In the LRT, the maximized likelihood under the full model (model 1) is compared to that under the reduced model (model 2) given by

$$\text{logit}(\pi_{gh}) = \mu + \alpha_1 I(g=1) + \alpha_2 I(g=2) + \beta_1 I(h=1) + \beta_2 I(h=2). \quad (2)$$

In general, the maximum likelihood estimates (MLEs) for logistic regression parameters do not have a closed-form solution, and iterative algorithms are generally employed to obtain the MLEs for the reduced model. This implies that use of the LRT can be quite computationally demanding for a genome-wide analysis of gene-gene interactions. In the past few years, a considerable number of computationally efficient tests have been proposed [3, 13-25]. These tests are often used as a screening tool to identify promising pairs and the LRT is then used to formally test the identified pairs.

The vast majority of computationally feasible methods for testing gene-gene interactions are correlation-based and have one degree of freedom (1 df). In each of these tests, a chosen linkage disequilibrium (LD) measure is compared between cases and controls. For example, Wu et al. [21] use a measure calculated from haplotype data; the "fast-epistasis" option in PLINK [1] and the test of [26] are based upon genotype data. However, as reported in [3], they are invalid for testing gene-gene interactions in the presence of both main effects and LD. Ueki and Cordell [3] recently propose a joint test and illustrated that the test is valid for testing gene-gene interactions for rare diseases. We will show that the joint test is invalid when both SNPs have marginal effects.

Tests with four degrees of freedom (4 *df*) have also been considered. Yang et al. [15] proposed to test interactions by partitioning chi-squares. However, Plackett [27] has pointed out that the test statistic used in the chi-square partitioning method does not necessarily follow a chi-squared distribution in the absence of interactions. Because the MLEs of the parameters under the reduced model do not exist in closed-form, Wan et al. [2] approximated the MLEs using marginal frequencies of low orders. However, the approximation performs poorly except in "perfect" contingency tables, which require several

constraints on the marginal probabilities [28, 29]. Thus, neither of the two four-degree-freedom tests is valid for testing gene-gene interactions in general.

Motivated by the limitations of existing computationally efficient tests of gene-gene interactions, we propose to use the Wald test for genome-wide interaction analysis. The Wald test enjoys the same asymptotic properties as the LRT, including being optimal under standard regularity conditions. On the other hand, compared to the LRT, the Wald test is much more computationally feasible for genome-wide interaction analysis. The remainder of the article is organized as follows. We derive the Wald test with 4 *df* and a modified Wald test with 1 *df*, demonstrate feasibility using simulations and then apply the method to a genome-wide study of multiple sclerosis. A discussion is provided at the end of the article.

Methods

In this section we first propose two closed-form Wald tests and then describe several existing tests.

The Wald test

The data for testing the interaction between a pair of SNPs when modeling the probability of a binary disease indicator can be summarized using a 3x3x2 table, as each SNP has three levels (0, 1, and 2) and the disease status has two levels (0 for normal and 1 for diseased). In categorical analysis, the gene-gene interactions that we are interested are known as the three-factor interaction or second-order interaction in a 3-way contingency table [28]. Roy and Kastenbaum [28] tested three-factor interactions using Pearson chi-squared statistic, which compares observed and expected cell counts under the null hypothesis of no interaction. Darroch [29] further suggested use of the LRT from log-linear models. Both the Pearson chi-squared statistic and the LRT require computation of the MLEs under the reduced model in (2). While the MLEs under the reduced model exist and are unique under mild regularity conditions, they do not have a closed-form [30]. Therefore, iterative algorithms are required in order to implement the two tests. Alternatively, Plackett [27] proposed to make inference using the MLEs estimated from the full

model. Bhapker and Koch [31] observed that the test used by Plackett is a Wald test. The advantage of the Wald test is that it only requires the MLEs of the full model, which can be estimated using the observed proportions. Therefore, it is computationally easier and faster than the Pearson chi-squared test or the LRT. This motivated us to derive the Wald test statistic for testing gene-gene interactions.

Let n_{ghk} denote the number of subjects with genotype *g* at SNP 1, genotype *h* at SNP 2, and disease status *k*. Here *g*=0,1,2; *h*=0,1,2 and *k*=0 or 1. Similarly, let p_{ghk} denote the probability of observing a subject with genotype *g* at SNP 1, genotype *h* at SNP 2, and disease status *k*. For simplicity, we use

$\vec{\theta} = (\mu, \alpha_1, \alpha_2, \alpha_1, \beta_1, \beta_1, \lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22})^T$ to denote the vector of all parameters in the full model and $\vec{\theta}_\lambda = (\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22})^T$ to denote the vector of interaction parameters. It is not difficult to verify that the MLE of $\vec{\theta}$ has a closed-form. In particular, the MLEs of the interaction parameters are given by

$$\hat{\lambda}_{gh} = \log\left(\frac{n_{gh1} n_{001}}{n_{g1} n_{0h1}} / \frac{n_{gh0} n_{000}}{n_{g00} n_{0h0}} \right), g = 1,2; h = 1,2.$$

Let $\hat{\vec{\theta}}, \hat{\vec{\theta}}_\lambda$ and $I(\hat{\vec{\theta}})$ denote the MLE of the vector of all the parameters, the MLE of the vector of interaction parameters, and the observed Fisher information, respectively. When all p_{ghk} 's are greater than 0 and the sample size is large, it is easily shown that the joint distribution of the MLEs are approximately multivariate normal, i.e.,

$$\hat{\vec{\theta}} \sim N_9(\vec{\theta}, I^{-1}(\hat{\vec{\theta}})).$$

Using symbolic matrix inverse in Mathematica, we calculate the asymptotic covariance of $\hat{\vec{\theta}}_\lambda$, which is denoted by $\hat{\text{cov}}(\hat{\vec{\theta}}_\lambda)$ (see the Appendix for details). From this, the Wald test statistic for testing $H_0 : \lambda_{11} = \lambda_{12} = \lambda_{21} = \lambda_{22} = 0$ is

$$Wald = \hat{\vec{\theta}}_\lambda^T [\hat{\text{cov}}(\hat{\vec{\theta}}_\lambda^T)]^{-1} \hat{\vec{\theta}}_\lambda^T.$$

Under the null hypothesis, the above quadratic form then follows the chi-squared distribution with 4 *df* asymptotically.

When a SNP has a low minor allele frequency (MAF), the number of subjects with the rare homozygote is small and using it as a separate genotype category is likely to reduce power. In our study, when the rare homozygote has less than 20 subjects, we collapse it with the heterozygote. The degrees of freedom of the Wald test are then reduced accordingly.

The modified Wald test with 1 *df*

One concern of using the 4 *df* Wald test or LRT is that the power could be low due to the large number of *df* [32]. As a result, several tests with 1 *df* have been proposed [33-36], including Tukey's 1 *df* test [37]. These tests usually consider some parsimonious functional form when modeling interactions. One particularly interesting model is the single interaction-parameter model used in [38], which incorporates an additive interaction term with unconstrained main effects:

$$\text{logit}(\pi_{gh}) = \mu + \alpha_1 I(g=1) + \alpha_2 I(g=2) + \beta_1 I(h=1) + \beta_2 I(h=2) + \lambda gh. \qquad (3)$$

By assuming an additive interaction of the total number of alleles, focus lies in testing a single interaction parameter in model (3) and hence reduces the *df* to 1. Note that this model allows flexible main effects, thereby avoiding the potential bias in testing interaction that could be caused by mis-specifying the main effects [38, 39].

To generalize the idea of [38] , we assume that the four interaction parameters satisfy some constraints such that we can rewrite the vector $\vec{\theta}_\lambda$ as $\vec{\theta}_\lambda = \lambda A\mathbf{1}$, where A is a 4x4 diagonal matrix, λ is a univariate interaction parameter, and $\mathbf{1} = (1,1,1,1)^T$. For example, with A=diag(1,2,2,4), the full logistic regression model in (1) reduces to the single interaction-parameter model in (3), and with A=diag(1,1,1,1), the parameterization of interactions is the same as that of [36]. The conventional Wald test requires the MLE of λ , which does not have a closed-form solution. To derive a computationally feasible test, similar to [3], we combine the information in the MLEs of $\vec{\theta}_\lambda$. Note that the vector

$A^{-1}\hat{\bar{\theta}}_\lambda \xrightarrow{p} A^{-1}\bar{\theta}_\lambda$ as $n \to \infty$ and hence provides four consistent estimates of λ. We consider using a weighted average to estimate λ, where the weight *w* is chosen to minimize the variance of $w^T A^{-1}\hat{\bar{\theta}}_\lambda$. Using the Lagrange multiplier, it is not difficult to see that

$$w = cA[\hat{\text{cov}}(\hat{\bar{\theta}}_\lambda)]^{-1}A\mathbf{1},$$

where *c* is a constant. Therefore, the corresponding test statistic is given by

$$Wald_1 = \frac{(w^T A^{-1}\hat{\bar{\theta}}_\lambda)^2}{w^T A^{-1}\hat{\text{cov}}(\hat{\bar{\theta}}_\lambda)A^{-1}w} = \frac{(\mathbf{1}^T A[\hat{\text{cov}}(\hat{\bar{\theta}}_\lambda)]^{-1}\hat{\bar{\theta}}_\lambda)^2}{\mathbf{1}^T A[\hat{\text{cov}}(\hat{\bar{\theta}}_\lambda)]^{-1}A\mathbf{1}}.$$

Under the null hypothesis of no interaction, i.e, $\lambda = 0$, the test statistic $Wald_1$ follows the chi-squared distribution with 1 *df* asymptotically. In this article, we define $Wald_1$ as the test that corresponds to additive interaction, i.e., A=diag(1,2,2,4).

The likelihood ratio tests

The likelihood ratio test compares the maximized likelihood under the full model (model 1) to that under the full model (model 2). Let $\{\hat{\pi}_{gh}\}_{g=0,1,2,h=0,1,2}$ denote the MLEs under the full model. The MLEs exist in closed-form, as provided in *Text S1*. Then the maximized likelihood is

$$L_1 = \prod_{g=0}^{2}\prod_{h=0}^{2}(\hat{\pi}_{gh})^{n_{gh1}}(1-\hat{\pi}_{gh})^{n_{gh0}}.$$

The reduced model assumes that there is no gene-gene interaction. Iterative methods, such as the Newton-Raphson method, are needed to obtain the MLEs, as the MLEs have no closed-form solution. Let $\{\hat{\pi}^0_{gh}\}_{g=0,1,2,h=0,1,2}$ denote the MLE under the reduced model, then the maximized likelihood is

$$L_2 = \prod_{g=0}^{2}\prod_{h=0}^{2}(\hat{\pi}^0_{gh})^{n_{gh1}}(1-\hat{\pi}^0_{gh})^{n_{gh0}}.$$

Under the null hypothesis of no interaction, the test statistic $LRT = -2\log(L_2 / L_1)$ follows the chi-squared distribution with 4 *df* asymptotically.

We can also construct a 1 *df* likelihood ratio test for gene-gene interaction. To do so, we use an iterative algorithm to maximize the likelihood under model (3). Let L_3 denote the maximized likelihood. Under the null hypothesis of no interaction, the test statistic $LRT_1 = -2\log(L_3 / L_2)$ follows the chi-squared distribution with 1 *df* asymptotically.

Boost (Wan et al. 2010)

Wan et al. [2] proposed an approximation for the likelihood ratio test. The logistic regression model in model (1) is equivalent to a log-linear model via re-parameterization. The likelihood function of the log-linear model is

$$L(\mu) = \prod_{g=0}^{2}\prod_{h=0}^{2}\prod_{k=0}^{1}\frac{e^{-\mu_{ghk}}\mu_{ghk}^{n_{ghk}}}{n_{ghk}!}$$

The full model is saturated, i.e., the number of parameters is the same as the number of observations. It is well known that the MLEs under the full model exist in closed-form, with $\hat{\mu}_{ghk} = n_{ghk}$. Similar to the logistic models, under the null hypothesis of no gene-gene interaction, the MLEs of the log-linear model have no closed-form. Wan et al. [2] proposed to use the Kirkwood superposition approximation [40]:

$$\hat{\mu}_{ghk}^{0} = c\frac{n_{ij+}n_{i+k}n_{+jk}}{n_{i++}n_{+j+}n_{++k}},$$

where *c* is a constant such that the sum of $\hat{\mu}_{ghk}^{0}$'s equals *n*.

While the approximation is attractive in that it does not require iterative computation, it is valid only when the marginal probabilities satisfy some conditions [29, 41]. Darroch [13] referred to tables satisfying these conditions as "perfect" tables. When two pairs of the three factors (SNP1, SNP2, and disease status) are independent, the resulting table is "perfect". The tables for testing gene-gene interactions are unlikely to be perfect, as it is impractical to assume both no LD and no marginal effects.

PLINK's fast-epistasis

The fast-epistasis test in PLINK [42] was described and discussed by Ueki and Cordell [3] in detail. Briefly speaking, the test is based upon collapsing the 3x3 genotype counts table for cases and that for controls into a 2x2 table for cases and a 2x2 table for controls. The log-odds ratio calculated from the collapsed case table and that from the control table are then compared. Ueki and Cordell [43] showed that the original variance formula underestimates the variance and proposed an adjusted test statistic. In this article, we use the adjusted test statistic. This test assumes random mating, additive main and interaction effects, and no LD. Violations of these assumptions can lead to inflated false positives.

The joint test of Ueki and Cordell (2012)

Motivated by the inflated type I error rates of existing methods, Ueki and Cordell [44] proposed a new test and found that it is a valid test of gene-gene interaction for rare diseases. The main idea is to compare the log-odds ratios for measuring haplotypic correlation between the cases and controls. Specifically, we define

$$\delta_k = \log(\frac{\psi_{11}^k \psi_{00}^k}{\psi_{10}^k \psi_{01}^k}), k = 0,1,$$

where is ψ_{ij}^k is the frequency of the haplotype with allele i at SNP1 and allele j at SNP2 in cases ($k = 1$) or controls ($k = 0$). Let

$$\hat{\vec{\xi}} = (\hat{\xi}_1^k, \hat{\xi}_2^k, \hat{\xi}_3^k, \hat{\xi}_4^k)^T = \left(\log\left(2\frac{n_{11k} n_{00k}}{n_{10k} n_{01k}} - 1 \right), \log\left(\frac{n_{12k} n_{00k}}{n_{10k} n_{02k}} \right), \log\left(\frac{n_{21k} n_{00k}}{n_{20k} n_{01k}} \right), \log\left(\frac{n_{22k} n_{00k}}{n_{20k} n_{02k}} \right) / 2 \right)^T, k = 0,1.$$

Ueki and Cordell used $\hat{\vec{\xi}}^k = (\hat{\xi}_1^k, \hat{\xi}_2^k, \hat{\xi}_3^k, \hat{\xi}_4^k)^T$ as four estimates for δ_k . From $\hat{\vec{\xi}}^k$, a weighted sum is calculated, where the weights minimize the variance. Let w_k denote the vector of optimal weights and V_k denote the resulting variance. Then the following test statistic was used

$$\frac{\left(w_0^T\hat{\tilde{\xi}}^0 - w_1^T\hat{\tilde{\xi}}^1\right)^2}{V_0 + V_1}.$$

Ueki and Cordell [3] showed that the test is valid for testing gene-gene interactions for rare disease. In *Text S2* we prove that it is valid as long as only one SNP has main effects.

Simulations

To examine the performance of the proposed methods, we conduct simulations in six scenarios (Table 1). The null hypothesis of no gene-gene interaction holds in scenarios 1-4. In scenario 1, there is no LD between the two SNPs. In scenarios 2-4, the two SNPs are in LD, with the Pearson's correlation coefficients varying from 0.1 to 0.9. These three scenarios are used to examine how LD and other violations of model assumptions jointly affect type I error rates. Specifically, scenario 2 assumes additive main effects at SNP1 and no main effect at SNP2; scenario 3 assumes dominant main effects at SNP1 and no main effect at SNP2; and scenario 4 assumes non-additive main effects at both SNP1 and SNP2. In scenarios 5-6, the SNPs are interacting with each other on disease risk. Scenario 5 assumes that the interaction is in an additive manner, i.e., the value of the interaction parameters depend on the number of risk alleles. Scenario 6 assumes that the four interaction parameters have the same value.

We fix the MAF of SNP1 at 0.5 and vary the MAF of SNP2 from 0.1 to 0.5 for scenarios 1, 5, and 6. For scenario 2-4, the MAF at both SNPs is 0.5. In each of the scenarios 1-3 and 5-6, the parameter μ is chosen such that the disease prevalence is approximately 0.1 when the MAF of both SNPs is 0.5 and r=0. In scenario 4, both the effect sizes and population prevalence are unrealistically large. We use this scenario to identify methods that work well in general but poorly in the presence of severe violation of model assumptions. For each scenario, 1,000 simulations are used; in each simulation, 1,000 cases and 1,000 controls are sampled from a homogenous population.

To assess the performance of the proposed methods, we compared the performance of three 4 *df* tests and four 1 *df* tests. The three 4 *df* tests are the Wald test (*Wald*), the likelihood ratio test (*LRT*), and the

test uses MLE approximation in "perfect" tables (*Boost*). The four 1 *df* tests are the modified Wald test with 1 *df* (*Wald*), the likelihood ratio test with 1*df* (LRT_1), the fast-epistasis test in *PLINK*, and the joint test of Ueki and Cordell [3] (*UC*). The proportion of simulations with p-values less than 0.05 are used to estimate the Type I error rates (scenarios 1-4) or power (scenarios 5 and 6) of the seven tests.

A real study

We apply the *Wald* test to a genome-wide study of multiple sclerosis [MIM 126200]. In this study, the control group includes 1,476 subjects enrolled in the 1958 British Birth Cohort Study and the case group includes 994 patients diagnosed with multiple sclerosis. Subjects in both groups were self-identified as white and of European descent [45]. Approximately 15k nonsynonymous SNPs across the genome were genotyped. We exclude SNPs with a low MAF (<1%) or filtered out by the Wellcome Trust Case Control Consortium due to quality concerns. This leaves approximately 11k SNPs for interaction analysis.

Results

Simulation results

The estimated Type I error rates under scenario 1 (Figure 1a) indicate that all tests have reasonable control of false positives in the absence of LD. Among the seven tests, *Wald* is slightly conservative and *Boost* is anti-conservative when the MAF of SNP2 is small. These results are not surprising, as *Boost* gives an upper bound of the LRT statistic [2] and the Wald test is known to be conservative under some situations [46]. In the presence of LD (Figure 1b-d), *Boost* has inflated false positives. This is because the conditions for "perfect" do not hold in the presence of both main effects and LD; as a result, the approximation of MLE in *Boost* is incorrect. It is also clear that the inflation increase with LD.

PLINK has an appropriate control of false positives when the true main effect is additive (scenarios 1-2 and Figure a-b) but not in the presence of both LD and non-additive main effects (scenarios 3-4 and

Figure 1c-d). This can be explained by the fact that the fast-epistasis test of PLINK is an alternative to tests based upon the logistic regression that assumes both additive main effects and interactions [3]. When the true underlying main effects are non-additive, forcing them to be additive corresponds to under fitting. As a result, when the two SNPs are in LD, we have biased estimates of gene-gene interaction parameters.

Figure 1 also indicates that the inflation of the joint test (*UC*) is ignorable in most situations (scenarios 1-3 and Figure 1a-c) except when both the disease prevalence and main effects are quite large (scenario 4 and Figure 1d). Ueki and Cordell [3] illustrated that under the assumption of rare diseases, the joint test is valid for testing gene-gene interactions. Here we found that it has acceptable control of type I error rates even when the disease prevalence is 10% (scenarios 1-3 and Figure 1a-c). In scenario 4, the prevalence varies from 45% to 50% and the main effects are very large. Compared to *Boost* and *PLINK*, *UC* is only moderately inflated, suggesting that the type I error rate of the joint test probably is not a concern for mostly encountered situations.

The estimated power under additive and dominant interactions is summarized in Figure 2 and Figure 3, respectively. In general, *Wald* agrees well with *LRT*, and $Wald_1$ agrees well with LRT_1. This demonstrates that we can obtain accurate results while avoiding iterative algorithms. When the true interactions are additive (scenario 5 and Figure 2), the 1 *df* tests, namely $Wald_1$, LRT_1, *PLINK*, and *UC*, are more powerful than *Wald* or *LRT*. On the other hand, when the true underlying interaction model is dominant (scenario 6 and Figure 3), the 1 *df* tests are less powerful, among which *PLINK* has the lowest power, then followed by *UC*.

Depending on association parameters, the power of *Boost* can be higher or lower than the other tests. However, since *Boost* results in inflated Type I error rates in the presence of LD (Figure 2), we do not recommend it as a test for interactions.

The WTCCC study of multiple sclerosis

The Bonferroni criterion is used for bounding the family-wise type I error rate at 0.05. Because there are about $6x10^7$ pairs from the 11k SNPs, pairs with nominal p-values less than $0.05/(6x10^7)$ are

considered significant. According to this standard, three SNP pairs are significant, as shown in the first three rows of Table 1. For a complex trait like multiple sclerosis, a large number of interactions are likely present. However, applying the conservative Bonferroni correction, only very strong interactions survive the stringent p-value cutoff.

To address the burden of multiple testing, several multi-stage approaches have been considered and discussed [16, 41-44, 47-58]. In most of these approaches, a subset of pairs was focused on in order to reduce the number of tests. Here we utilize two of these approaches, namely the conditional approach [42] and the simultaneous approach [47]. In the conditional approach, at least one of the two SNPs in the pair is marginally significant; whereas in the simultaneous approach both SNPs in the pair are significant. To compare the two strategies, we provide QQ plots with varying nominal p-value cutoffs (Figure 4). The p-values for marginal effects are calculated using the Armitage trend test [59]. The numbers of significant interactions are also given in Figure 5. It is obvious that the proportion of significant interactions is enriched with the simultaneous approach, but not with the conditional approach. As a result, the simultaneous approach identifies a greater number of significant interactions when compared to the conditional approach. In particular, using a nominal p-value cutoff of 0.001, we identified 67 interactions. It is worth pointing out that we also examined the two two-stage strategies using the false discovery rate criterion [60]. Again, the simultaneous approach identifies more interactions than does the conditional approach (data not shown).

We next examined whether different tests agree with each other for the 67 pairs detected by the simultaneous approach. Detailed information about the involved SNPs, interactions, and estimated parameters are given in Table 2. Most of the SNPs are highly significant marginally. The MAFs are relatively large, ranging from 0.19 to 0.50. The Pearson correlation ranges from near zero to about 0.30, indicating most of the identified pairs are not in strong LD. Comparing the p-values resulting from the considered tests echoes what was observed in the simulation study. First, *Wald* gives similar results to *LRT* and $Wald_1$ gives similar results to LRT_1 (Table 2 and the top panel of Figure 5). Second, the comparison between *Boost* and *LRT* (lower left panel of Figure 5) shows that *Boost* is anti-conservative,

especially when SNPs are in LD. In the two remaining tests, *UC* is more similar to LRT_1 than *PLINK*. Note that *PLINK* is an alternative to the logistic regression that assumes additive main effects. The difference between *PLINK* and LRT_1 indicates that the main effects of the SNPs in the top 67 pairs are unlikely to be additive.

Both Table 2 and the lower panel of Figure 5 indicate that $Wald_1$ and LRT_1 are lower in magnitude than *Wald* and *LRT* in most situations. This difference suggests that the additive interaction model in (3) does not fit well for many of the 67 pairs. Indeed, the estimated interaction parameters (Table 2) for most SNP pairs do not follow the additive interaction model. To assess the degree to which the additive interaction assumption is violated, we conduct a test. Specifically, define the following matrix

$$B = \begin{pmatrix} 1 & -0.5 & 0 & 0 \\ 1 & 0 & -0.5 & 0 \\ 1 & 0 & 0 & -0.25 \end{pmatrix}.$$

The test statistic for additive interaction is

$$(B\hat{\hat{\theta}}_\lambda)^T [B\hat{\text{cov}}(\hat{\hat{\theta}}_\lambda)]^{-1} B\hat{\hat{\theta}}_\lambda.$$

For large sample size and under the null hypothesis of additive interaction, the test statistic is approximately distributed as a chi-squared random variable with 3 *df*. Note that when one homozygote group of a SNP is rare, we collapse it with the heterozygote; as a result, the *B* matrix and the *df* will be modified accordingly. The resulting p-values (the last column of Table 2) reveal that the additive assumption does not hold for many of the SNP pairs.

All of the significant interactions are from SNPs in the major histocompatibility complex (MHC) region on Chromosome 6. This is not surprising, as it is already known that the MHC region plays a major role in the function of the immune system and numerous variants in the region have been associated with autoimmune diseases [61]. In fact, gene-gene interactions in this region have been identified for other autoimmune diseases, such as type 1 diabetes and rheumatoid arthritis [2, 52, 62]. Figure 6 visualizes the positions of the SNPs of the 67 identified interactions. All except two interactions are within or between class II and Class III. Lincoln et al. also reported interactions within the MHC class

II region in multiple sclerosis [63]. Thus, the interactions identified here are likely to be true. Interpreting the interactions requires biological experiments, as a significant p-value could be due to SNP pairs that are flanking a functional variant. Recently, a functional assay in humanized mice suggested functional epistasis between two MHC class II alleles in multiple sclerosis [64]. Indeed, six of the identified interactions are within the MHC class II region (Figure 6).

To compare the speeds of the seven methods, we implemented all of them using C++ and tested their speed using the WTCCC study of multiple sclerosis. We found that *Wald*, $Wald_1$, *Boost*, and *PLINK* have roughly the same speed and that they are three times as fast as *UC*, five times as fast as *LRT*, and ten times as fast as LRT_1.

Discussion

In this article we proposed a Wald test and a modified Wald test for genome-wide analysis of gene-gene interactions. Note that neither of the tests places constraints on the main effects. When a constraint on the main effects disagrees with the underlying true model, the main effects are mis-specified. Similar to the effect of under fitting in regression models, mis-specifying main effects could lead to invalid statistical inference for the interaction parameters [38]. By allowing full flexibility in modeling main effects, our tests are valid even if both main effects and LD are present. Here we used logistic regression to derive test statistics. The test statistics can be equivalently derived using a log-linear model, as logistic and log-linear models are identical with appropriate parameterizations.

The strategies presented here were motivated by the problem of testing gene-gene interactions. They are readily applicable to testing gene-environment (G-by-E) interactions when the environmental factor is a categorical variable or can be converted to a categorical variable in a scientifically meaningful way. For example, for an environmental factor with K categories, we can derive a Wald test with $2(K\text{-}1)$ *df*. This test can be modified to construct a test with K-1 *df* if we assume that the genetic factor plays an additive

role in the G-by-E interactions. The degrees of freedom may be further reduced when meaningful numerical scores can be assigned to the environmental factor.

Existing computationally efficient tests for gene-gene interactions are invalid when some of their model assumptions, such as additive main effects, main effects at no more than one SNPs, and no LD, are violated. Among them, the joint test of Ueki and Cordell [3] has the best control of false positives under various violations of assumptions. Although their test was motived by the haplotype-based correlation of [21], it is insensitive to non-additive main effects or LD. Instead of calculating correlations from reconstructed haplotypes, collapsed tables, or numerical genotypes, they calculated four correlations from sub two-by-two tables. This explains why their test is insensitive to LD and non-additive main effects. Ueki and Cordell [3] showed that the joint test is valid for testing gene-gene interactions for rare diseases and we proved that the test is valid as long as there is at least SNP does not have main effects. Our simulations indicate that its inflation in the type I error rates is ignorable even the disease prevalence is as high as 10%. Compared to PLINK and Boost, the inflation of the joint test is still moderate when the prevalence is over 45%. Therefore, the inflation in false positives of the joint test is unlikely to be a concern in practice. Compared to the methods in Ueki and Cordell [3], ours are derived from a formal statistical framework, are valid under all circumstances, and are more general in several perspectives. First, we derived Wald-type tests from MLEs, which ensures asymptotic efficiency of the resulted tests. Second, our tests cover both constrained and unconstrained parameterizations of interaction effects, making them useful under multiple modeling assumptions.

From the interaction study of multiple sclerosis we concluded that (1) the simultaneous approach, where both SNPs in a SNP pair show marginal effects, identified a greater number of significant interactions when compared to the conditional approach, where either of the SNPs has a marginal effect; and (2) additive interactions are unlikely to be common. One needs to be cautious when generalizing these conclusions to other situations. It is well understood that the MHC region plays an important role in autoimmune diseases, including multiple sclerosis. Genes in this region often have very large effect sizes. It is unclear whether the two conclusions hold for other complex diseases. Additive interactions might be

more common in interactions with small to moderate effect sizes than those with large effect sizes. For example, we have observed additive interactions between *MGAT*1 and other genes in both multiple sclerosis and type 1 diabetes [65, 66]. Similarly, when there are a large number of interactions with small to moderate sizes, the conditional approach might also be useful. It is of great future interest to utilize and evaluate these strategies in more GWAS for complex diseases.

Acknowledgements

The research was supported in part by grant R01AI082266 from the National Institute of Allergy and Infectious Diseases to M.D, and by grant P30CA062203 from the National Cancer Institute to D.L.G. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

The C++ source code that implements the proposed tests is available from http://www.ics.uci.edu/~zhaoxia/.

Tables

Table 1: Parameters used in the simulation study. We consider six scenarios, with the first four under the null hypothesis (absence of gene-gene interaction) and the last two under the alternative hypothesis (presence of gene-gene interaction).

Table 2: Results of the top 67 pairs. The four columns under each SNP are the SNP name, position (in bp) on chromosome 6, MAF, and marginal p-value, respectively. The column with the name "r^2" shows the Pearson's correlation of the two SNPs using 0, 1, and 2 coding. The four columns under "P-values for G-G Interactions" are the p-values using *Wald*, *LRT*, $Wald_1$, and LRT_1, respectively. The first four columns under "Estimated parameters for interactions" and the estimates of $\theta_{11}, \theta_{12}, \theta_{21}$ and θ_{22}; the last column is the p-value for testing the null hypothesis of multiplicative interactions.

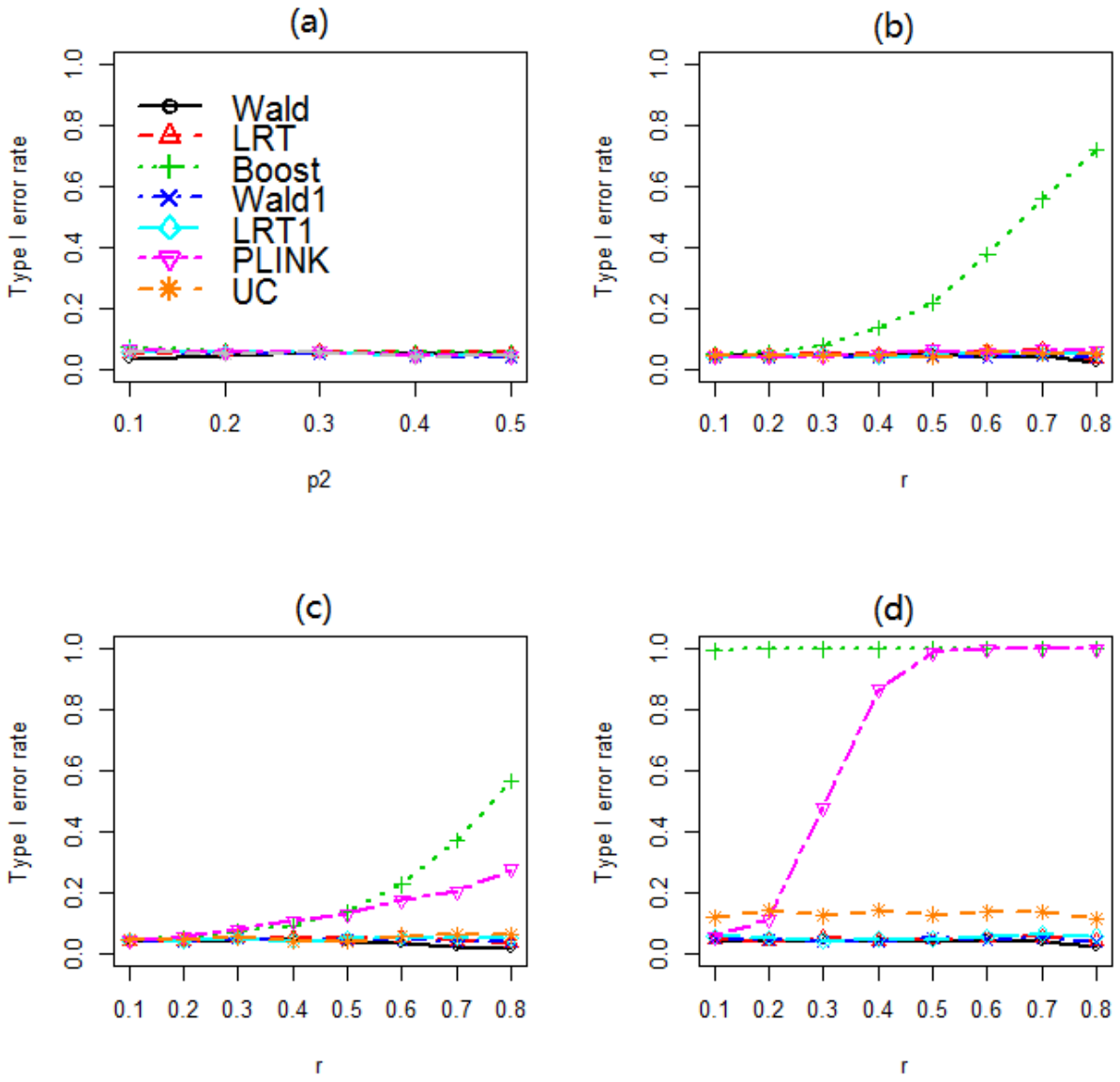

Figure 1: Estimated type 1 error rates. (a) Scenario 1, i.e., additive main effects at both SNPs and no LD. The x-axis shows the MAF of SNP2. (b) Scenario 2, i.e., additive main effects at SNP1 and LD between the two SNPs. (c) Scenario 3, i.e., dominant main effects at SNP1 and LD between the two SNPs. (d) Scenario 4, i.e., non-additive main effects at both NSPs and LD between the two SNPs. For (b)-(d), the x-axis shows *r*, i.e, the Pearson's correlation coefficient based on haplotypes.

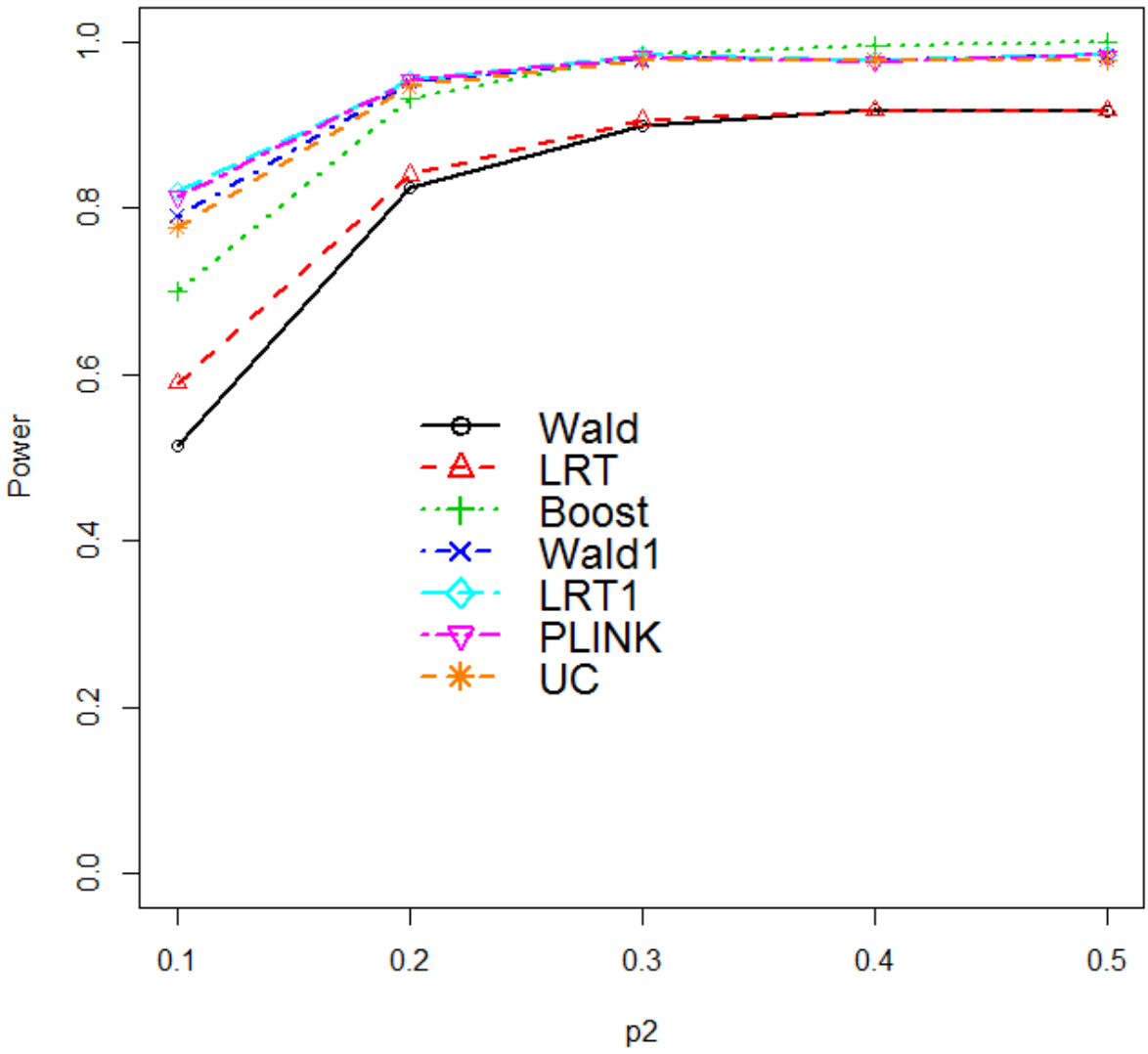

Figure 2: Estimated power for scenario 5, i.e., additive interactions. The x-axis is the MAF of SNP2.

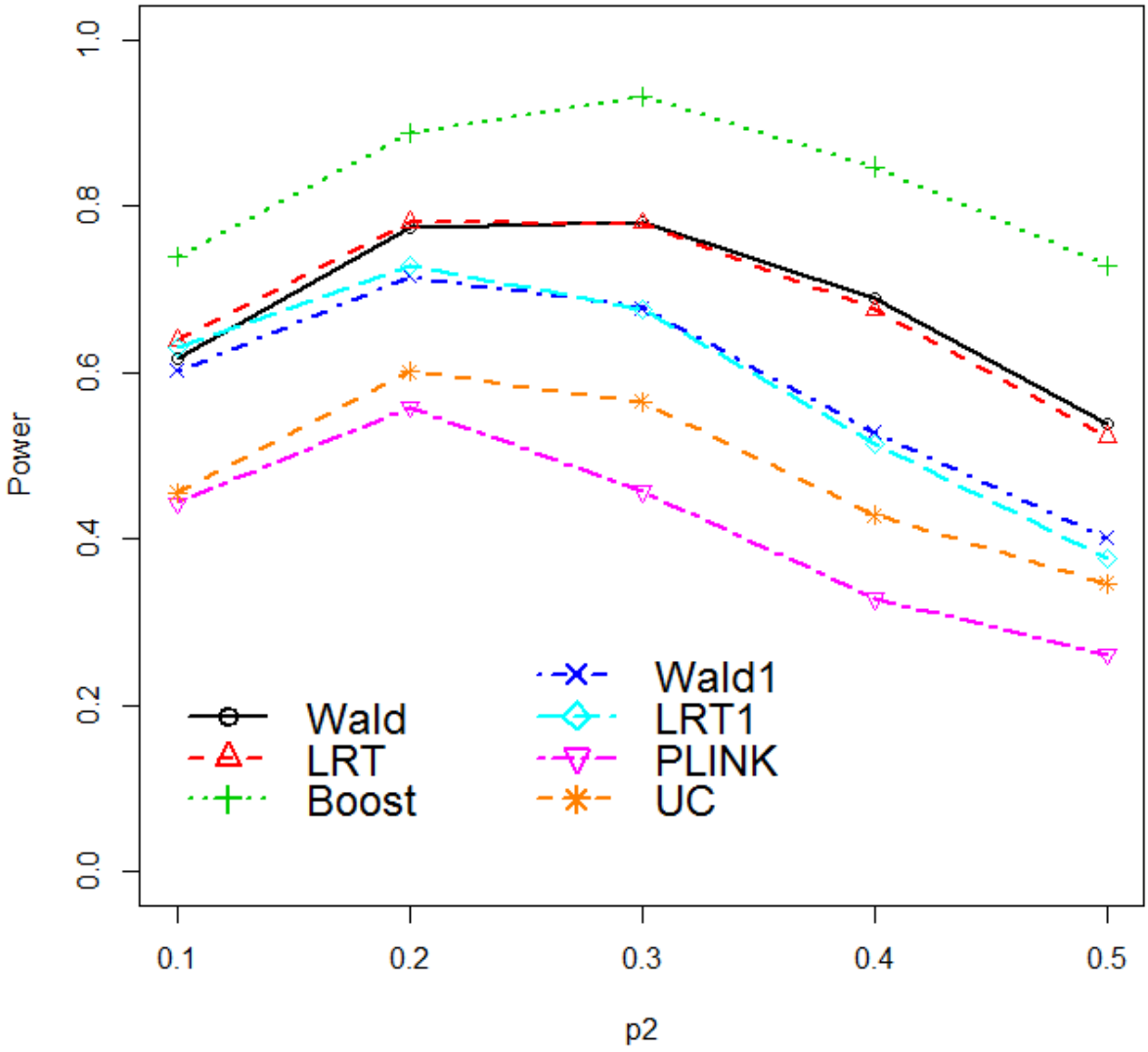

Figure 3: Estimated power for scenario 6, i.e., dominant interactions. The x-axis is the MAF of SNP2.

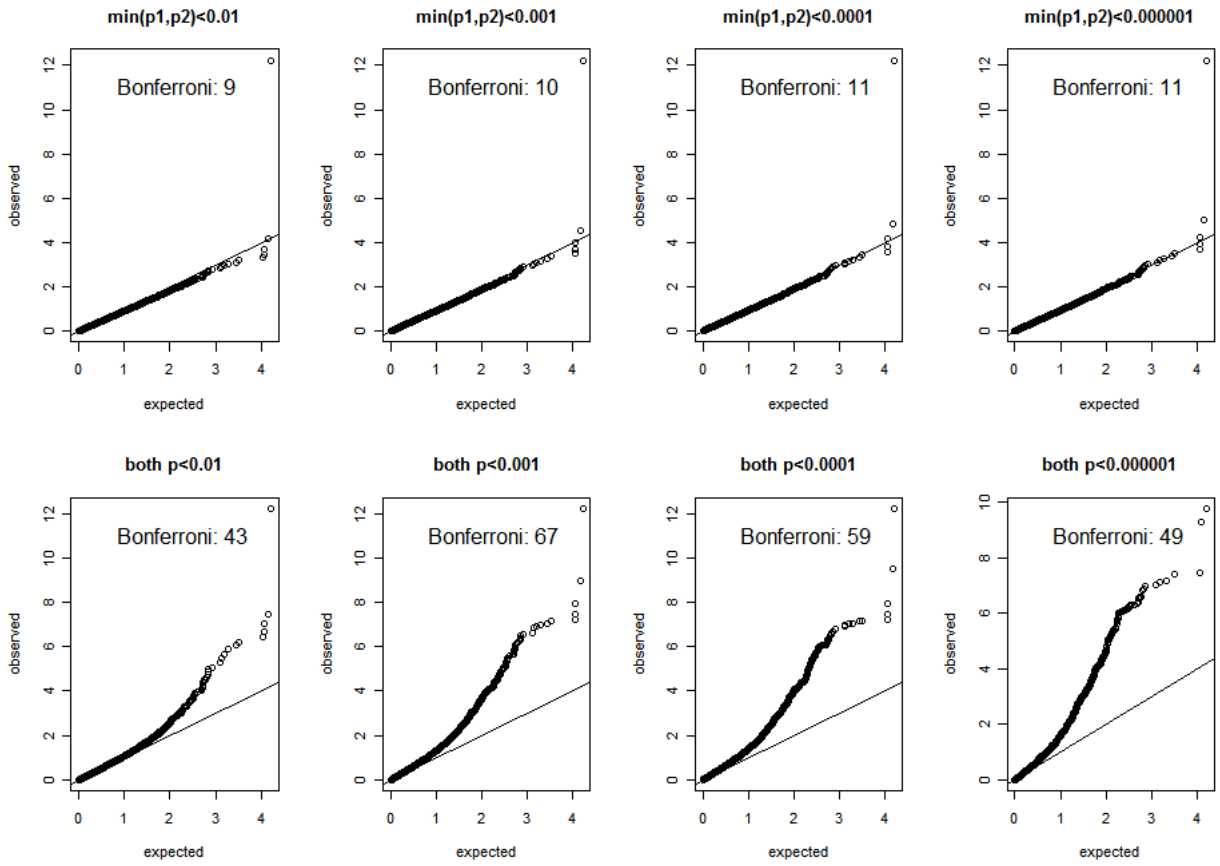

Figure 4: The QQ plots of –log10 p-values from subsets of SNP pairs. In each QQ plot, the observed p-value quantiles were plotted versus those expected under the null hypothesis of no interaction. The top panel: the QQ plots using the conditional approach; the bottom panel: the QQ plots using the simultaneous approach. The numbers of significant interactions for different strategies and nominal p-value cutoffs are given in the plots.

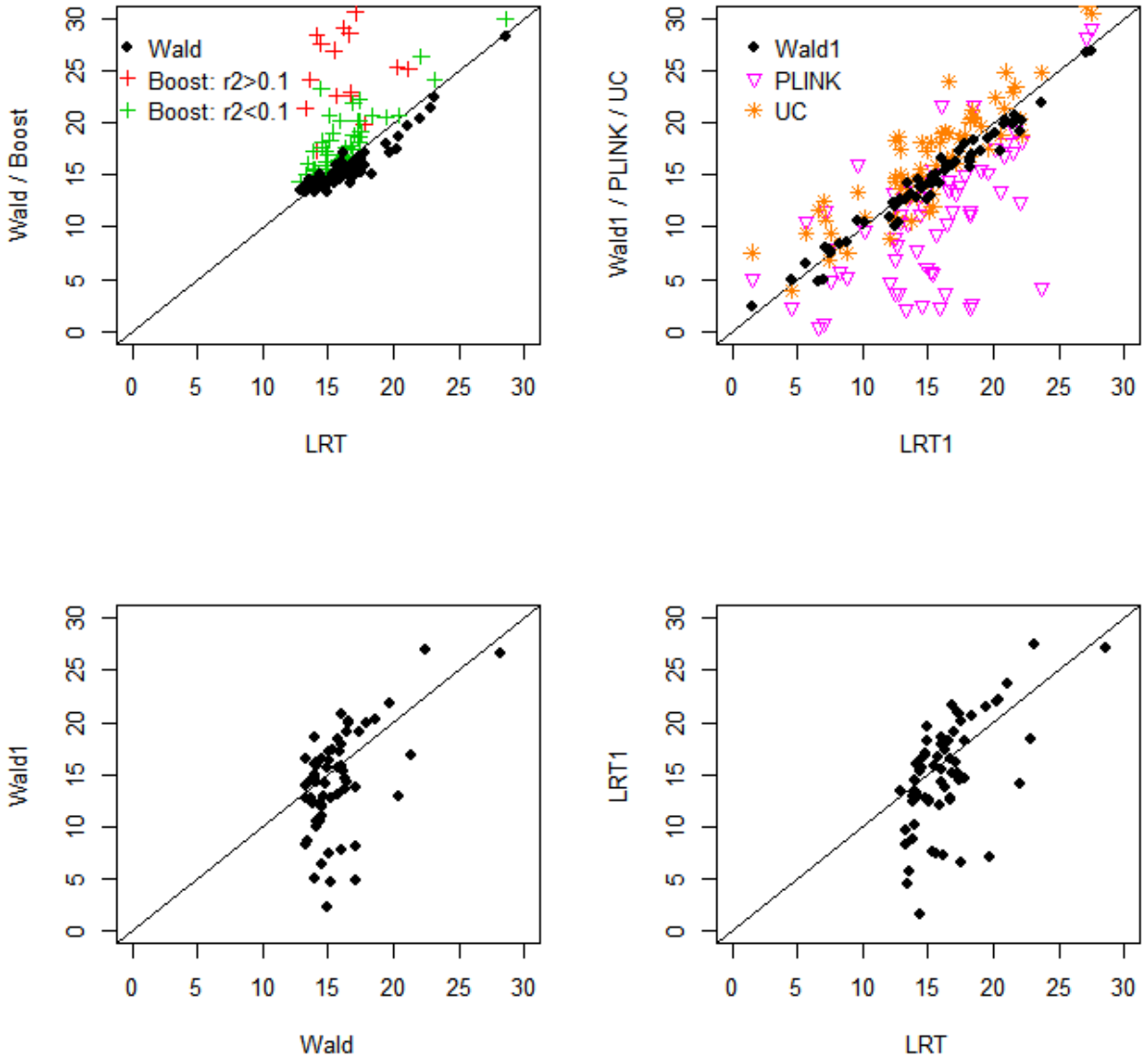

Figure 5: The -log10 p-values for the 67 pairs selected using the simultaneous approach. Upper left: *Wald* and *Boost* vs *LRT*; upper right: $Wald_1$, *PLINK* and *UC* vs LRT_1; lower left: $Wald_1$ vs *Wald*; lower right: LRT_1 vs *LRT*.

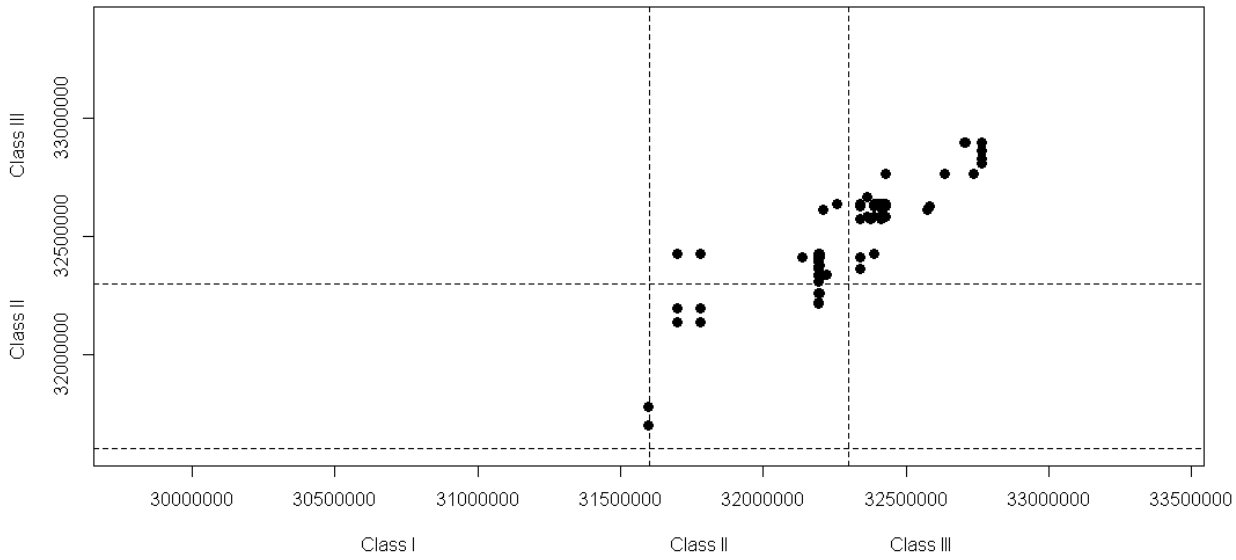

Figure 6: The 67 interactions identified by the simultaneous approach. All the interactions are located in the MHC region.

References

1. Purcell, S., et al., *PLINK: A tool set for whole-genome association and population-based linkage analyses.* American Journal of Human Genetics, 2007. **81**(3): p. 559-575.
2. Wan, X.A., et al., *BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies.* American Journal of Human Genetics, 2010. **87**(3): p. 325-340.
3. Ueki, M. and H.J. Cordell, *Improved Statistics for Genome-Wide Interaction Analysis.* Plos Genetics, 2012. **8**(4): p. 141-159.
4. Carlborg, O. and C.S. Haley, *Epistasis: too often neglected in complex trait studies?* Nature Reviews Genetics, 2004. **5**(8): p. 618-U4.
5. Eichler, E.E., et al., *VIEWPOINT Missing heritability and strategies for finding the underlying causes of complex disease.* Nature Reviews Genetics, 2010. **11**(6): p. 446-450.
6. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-753.
7. Moore, J.H. and S.M. Williams, *Epistasis and Its Implications for Personal Genetics.* American Journal of Human Genetics, 2009. **85**(3): p. 309-320.
8. Bateson, W., *Facts limiting the theory of heredity.* Science, 1907. **26**: p. 649-660.
9. Cordell, H.J., *Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans.* Human Molecular Genetics, 2002. **11**(20): p. 2463-2468.
10. Phillips, P.C., *Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems.* Nature Reviews Genetics, 2008. **9**(11): p. 855-867.
11. Wang, X.F., R.C. Elston, and X.F. Zhu, *The Meaning of Interaction.* Human Heredity, 2010. **70**(4): p. 269-277.
12. North, B.V., D. Curtis, and P.C. Sham, *Application of logistic regression to case-control association studies involving two causative loci.* Human Heredity, 2005. **59**(2): p. 79-87.
13. Kam-Thong, T., et al., *EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units.* European Journal of Human Genetics, 2011. **19**(4): p. 465-471.
14. Fan, R., et al., *Entropy-Based Information Gain Approaches to Detect and to Characterize Gene-Gene and Gene-Environment Interactions/Correlations of Complex Diseases.* Genetic Epidemiology, 2011. **35**(7): p. 706-721.
15. Yang, Y., C. He, and J. Ott, *Testing Association with Interactions by Partitioning Chi-Squares.* Annals of Human Genetics, 2009. **73**: p. 109-117.
16. Millstein, J., et al., *A testing framework for identifying susceptibility genes in the presence of epistasis.* American Journal of Human Genetics, 2006. **78**(1): p. 15-27.
17. Zhao, J.Y., L. Jin, and M.M. Xiong, *Test for interaction between two unlinked loci.* American Journal of Human Genetics, 2006. **79**(5): p. 831-845.
18. Brinza, D., et al., *RAPID detection of gene-gene interactions in genome-wide association studies.* Bioinformatics, 2010. **26**(22): p. 2856-2862.
19. Yuan, Z.S., et al., *From Interaction to Co-Association -A Fisher r-To-z Transformation-Based Simple Statistic for Real World Genome-Wide Association Study.* Plos One, 2013. **8**(7).
20. Wu, X.S., L. Jin, and M.M. Xiong, *Composite measure of linkage disequilibrium for testing interaction between unlinked loci (vol 16, pg 644, 2008).* European Journal of Human Genetics, 2008. **16**(9): p. 1160-1160.
21. Wu, X.S., et al., *A Novel Statistic for Genome-Wide Interaction Analysis.* Plos Genetics, 2010. **6**(9).
22. Hu, J.K., X.L. Wang, and P. Wang, *Testing Gene-Gene Interactions in Genome Wide Association Studies.* Genetic Epidemiology, 2014. **38**(2): p. 123-134.

23. Bhattacharya, K., M.I. McCarthy, and A.P. Morris, *Rapid testing of gene-gene interactions in genome-wide association studies of binary and quantitative phenotypes.* Genetic Epidemiology, 2011. **35**(8): p. 800-808.
24. Dong, C.Z., et al., *Exploration of gene-gene interaction effects using entropy-based methods.* European Journal of Human Genetics, 2008. **16**(2): p. 229-235.
25. Chanda, P., et al., *The interaction index, a novel information-theoretic metric for prioritizing interacting genetic variations and environmental factors.* European Journal of Human Genetics, 2009. **17**(10): p. 1274-1286.
26. Wellek, S. and A. Ziegler, *A Genotype-Based Approach to Assessing the Association between Single Nucleotide Polymorphisms.* Human Heredity, 2009. **67**(2): p. 128-139.
27. Plackett, R.L., *A Note on Interactions in Contingency-Tables.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1962. **24**(1): p. 162-166.
28. Roy, S.N. and M.A. Kastenbaum, *On the Hypothesis of No Interaction in a Multi-Way Contingency Table.* Annals of Mathematical Statistics, 1956. **27**(3): p. 749-757.
29. Darroch, J.N., *Interactions in Multi-Factor Contingency-Tables.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1962. **24**(1): p. 251-263.
30. Birch, M.W., *Maximum-Likelihood in 3-Way Contingency-Tables.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1963. **25**(1): p. 220-233.
31. Bhapker, V.P. and G.G. Koch, *Hypotheses of No Interaction in Multidimensional Contingency Tables.* Technometrics, 1968. **10**(1): p. 107-&.
32. Song, M.S. and D.L. Nicolae, *Restricted Parameter Space Models for Testing Gene-Gene Interaction.* Genetic Epidemiology, 2009. **33**(5): p. 386-393.
33. Barhdadi, A. and M.P. Dube, *Testing for Gene-Gene Interaction with AMMI Models.* Statistical Applications in Genetics and Molecular Biology, 2010. **9**(1).
34. Hoffmann, T.J., et al., *Gene-Environment Interaction Tests for Dichotomous Traits in Trios and Sibships.* Genetic Epidemiology, 2009. **33**(8): p. 691-699.
35. Chatterjee, N., et al., *Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions.* American Journal of Human Genetics, 2006. **79**(6): p. 1002-1016.
36. Jiao, S., et al., *Genome-Wide Search for Gene-Gene Interactions in Colorectal Cancer.* Plos One, 2012. **7**(12).
37. Tukey, J.K., *One degree of freedom for nonadditivity.* Biometrics, 1949. **5**: p. 232-242.
38. VanderWeele, T.J. and N.M. Laird, *Tests for Compositional Epistasis under Single Interaction-Parameter Models.* Annals of Human Genetics, 2011. **75**: p. 146-156.
39. Yu, Z.X., *Testing Gene-Gene Interactions in the Case-Parents Design.* Human Heredity, 2011. **71**(3): p. 171-179.
40. Matsuda, H., *hysical nature of higher-order mutual information: Intrinsic correlations and frustration* Physical Review E, 2000. **62**: p. 3096-3102.
41. Ionita, I. and M. Man, *Optimal two- stage strategy for detecting interacting genes in complex diseases.* Bmc Genetics, 2006. **7**.
42. Daly, M.J. and D. Altshuler, *Partners in crime.* Nature Genetics, 2005. **37**(4): p. 337-338.
43. Lewinger, J.P., et al., *Efficient Two-Step Testing of Gene-Gene Interactions in Genome-Wide Association Studies.* Genetic Epidemiology, 2013. **37**(5): p. 440-451.
44. Zhang, Y. and J.S. Liu, *Bayesian inference of epistatic interactions in case-control studies.* Nature Genetics, 2007. **39**(9): p. 1167-1173.
45. Wellcome Trust Case Control Consortium and The Australo-Anglo-American Spondylitis Consortium (TASC), *Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants.* Nature Genetics, 2007. **39**(11): p. 1329-1337.

46. Newcombe, R.G., *Two-sided confidence intervals for the single proportion: Comparison of seven methods.* Statistics in Medicine, 1998. **17**(8): p. 857-872.
47. Marchini, J., P. Donnelly, and L.R. Cardon, *Genome-wide strategies for detecting multiple loci that influence complex diseases.* Nature Genetics, 2005. **37**(4): p. 413-417.
48. Evans, D.M., et al., *Two-stage two-locus models in genome-wide association.* Plos Genetics, 2006. **2**(9): p. 1424-1432.
49. Musani, S.K., et al., *Detection of gene x gene interactions in genome-wide association studies of human population data.* Human Heredity, 2007. **63**(2): p. 67-84.
50. Liu, Y., et al., *Genome-Wide Interaction-Based Association Analysis Identified Multiple New Susceptibility Loci for Common Diseases.* Plos Genetics, 2011. **7**(3).
51. Emily, M., et al., *Using biological networks to search for interacting loci in genome-wide association studies.* European Journal of Human Genetics, 2009. **17**(10): p. 1231-1240.
52. Wu, J., et al., *Screen and Clean: A Tool for Identifying Interactions in Genome-Wide Association Studies.* Genetic Epidemiology, 2010. **34**(3): p. 275-285.
53. Wang, Y., et al., *AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm.* BMC research notes, 2010. **3**: p. 117.
54. Yang, C., et al., *SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies.* Bioinformatics, 2009. **25**(4): p. 504-511.
55. Calle, M.L., et al., *Improving strategies for detecting genetic patterns of disease susceptibility in association studies.* Statistics in Medicine, 2008. **27**(30): p. 6532-6546.
56. De Lobel, L., et al., *A screening methodology based on Random Forests to improve the detection of gene-gene interactions.* European journal of human genetics : EJHG, 2010. **18**(10): p. 1127-32.
57. Moore, J.H., *Genome-wide analysis of epistasis using multifactor dimensionality reduction: feature selection and construction in the domain of human genetics*, in *Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data*, X. Zhu and I. Davidson, Editors. 2007, IGI press: Hershey. p. 17-30.
58. Kooperberg, C. and M. LeBlanc, *Increasing the power of identifying gene x gene interactions in genome-wide association studies.* Genetic Epidemiology, 2008. **32**(3): p. 255-263.
59. Sasieni, P.D., *From genotypes to genes: Doubling the sample size.* Biometrics, 1997. **53**(4): p. 1253-1261.
60. Storey, J.D., *The positive false discovery rate: A Bayesian interpretation and the q-value.* Annals of Statistics, 2003. **31**(6): p. 2013-2035.
61. Fernando, M.M.A., et al., *Defining the role of the MHC in autoimmunity: A review and pooled analysis.* Plos Genetics, 2008. **4**(4).
62. Lippert, C., et al., *An Exhaustive Epistatic SNP Association Analysis on Expanded Wellcome Trust Data (vol 3, pg 1099, 2013).* Scientific Reports, 2013. **3**.
63. Lincoln, M.R., et al., *Epistasis among HLA-DRB1, HLA-DQA1, and HLA-DQB1 loci determines multiple sclerosis susceptibility.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(18): p. 7542-7547.
64. Gregersen, J.W., et al., *Functional epistasis on a common MHC haplotype associated with multiple sclerosis.* Nature, 2006. **443**(7111): p. 574-577.
65. Mkhikian, H., et al., *Genetics and the environment converge to dysregulate N-glycosylation in multiple sclerosis.* Nature Communications, 2011. **2**.
66. Yu, Z., et al., *Family studies of type 1 diabetes reveal additive and epistatic effects between MGAT1 and three other polymorphisms.* Genes and immunity, 2014.
67. Wald, A., *Tests of statistical hypotheses concerning several parameters when the number of observations is large.* Transactions of the American Mathematical Society, 1943. **54**(1-3): p. 426-482.