Published in final edited form as:

Genet Epidemiol. 2014 February ; 38(2): 123–134. doi:10.1002/gepi.21786.

Testing Gene-Gene Interactions in Genome Wide Association Studies

Jie Kate Hu¹, Xianlong Wang², and Pei Wang^{2,*}

¹Department of Biostatistics, University of Washington, Seattle, WA, 98105

²Fred Hutchinson Cancer Research Center, Seattle, WA, 98109

Abstract

Detection of gene-gene interaction has become increasingly popular over the past decade in genome wide association studies (GWAS). Besides traditional logistic regression analysis for detecting interactions between two markers, new methods have been developed in recent years such as comparing linkage disequilibrium (LD) in case and control groups. All these methods form the building blocks of most screening strategies for disease susceptibility loci in GWAS. In this paper, we are interested in comparing the competing methods and providing practical guidelines for selecting appropriate testing methods for interaction in GWAS. We first review a series of existing statistical methods to detect interactions, and then examine different definitions of interactions to gain insight into the theoretical relationship between the existing testing methods. Lastly, we perform extensive simulations to compare powers of various methods to detect either interaction between two markers at two unlinked loci or the overall association allowing for both interaction and main effects. This investigation reveals informative characteristics of various methods that are helpful to GWAS investigators.

Keywords

gene-gene interaction; GWAS; linkage disequilibrium; penetrance; logistic regression

1 Introduction

Over the past two decades, many genome wide association studies (GWAS) have been conducted to understand the genetic architecture of complex diseases. These studies have revealed a repertoire of disease susceptibility loci (e.g., Kraft and Cox [2008]; Seng and Seng [2008]). However, these loci individually have small or, at best, moderate effect sizes, and account for only a small fraction of the genetic variation in complex diseases. For example, the seven susceptibility loci identified through GWAS together are estimated to explain 5.9% of the familial breast cancer risk [Seng and Seng, 2008]. For more reviews of the published GWAS, see Donnelly [2008]; Goldstein [2009]; Hardy and Singleton [2009]; Hirschhorn [2009]; Hirschhorn and Daly [2005]; Kraft and Hunter [2009]; Ku et al. [2010];

Manolio [2008, 2009]; McCarthy et al. [2008]; Teo [2008]; Wang et al. [2005]; Ziegler et al. [2008].

Where should we discover the genetic factors explaining the remaining variation? The polygenic nature of complex diseases suggests that the disease risk has to be a synergistic result of multiple genes in a complex biological network. Traditional GWAS tested one marker at a time, and thus ignored the complex genomic context [Moore and Williams, 2009]. Given the common-disease-common-variant hypothesis, which is supported by both evolution theory [Chakravarti, 1999; Reich and Lander, 2001] and studies, part of the unexplained genetic variation could very well be attributed to the interactions among genes. Endeavors aimed to detect and assess gene $\times$ gene ($G \times G$) interactions have surged up in both the statistical and biomedical research communities recently, and a variety of methods have been developed for detecting genetic interactions [Kooberberg et al., 2009; Moore et al., 2010; Moore and Williams, 2009]. Cordell [2009] provides informative review for a handful of methods on $G \times G$ interaction. Many of these methods make specific statistical/genetic model assumptions, and are derived from different definitions of $G \times G$ interactions. Hence, they test different hypotheses. To gain better understanding, we study the theoretical relationship between various definitions of interaction between two unlinked loci regarding a binary outcome. We clarify the null spaces of several tests for interaction effects, and derive the conditions under which these tests share the common null space. The studied tests include two log odds ratio tests T_{logit} , T_{OR} , three linkage disequilibrium (LD) based tests T_{LD} [Zhao et al., 2006], T_{LD^*} (χ^2_I in Yang C. et al. [2009]), T_{CLD} [Wu et al., 2008], and three case-only statistics T_{Pearson} , T_{LD_c} , T_{OR_c} . We then conduct extensive simulation studies with settings mimicking the real scenarios with or without main effects to assess the performance of these competing methods. Since the ultimate goal of GWAS is to detect the association between diseases and genetic markers, we also evaluate tests for detecting association signals of a pair of loci allowing for both interaction and main effects. For this purpose, in addition to the tests mentioned above, we also include in our investigation three overall association tests — $T_{\text{logisticall}}$, T_{χ^2} , and the logistic kernel machine method T_{kernel} [Liu et al., 2008]—, as well as two correlation based tests — T_{Perlman} [Larntz and Perlman, 1988] and T_{Kullback} [Kullback, 1967].

The remainder of this article is organized as follows. In section 2, we briefly review the existing statistical methods for detecting genetic interactions between two single markers and between two sets of markers, as well as methods for genome-wide screening for genetic interactions. In section 3, we evaluate the powers of tests for detecting interactions between two unlinked single markers and study their theoretical properties, since these tests serve as the building blocks of most interaction testing methods for GWAS. In section 4, we extend our research to a range of tests on association between a binary outcome and a pair of unlinked markers allowing for both interaction and main effects. Through these investigations, we will provide a practical guideline for selecting appropriate testing methods in GWAS, which is summarized in section 5.

2 Methods for testing $G \times G$ interactions in GWAS

Due to the lack of established biochemical mechanism behind $G \times G$ interaction, the form in which interaction is defined is inexplicable. This necessitates a review on the various interaction definitions behind existing tests. The concept of interaction was first introduced by Bateson [1909], in which epistasis was defined as a masking effect from the allele at one locus to the allele at another locus. Later a more quantitative definition was proposed by Fisher [1918, p.404] as the deviation from addition of superimposed effects from different factors. This makes it possible to define interaction through a mathematical model. Since then, testing an interaction has been translated to testing a statistical hypothesis concerning a specific quantity in a model that describes the relation between a phenotype and a genotype. Over the last decade, as GWAS have become a powerful tool for genetic studies of complex diseases, a variety of new methods have been developed for detecting genetic interactions. Particularly, some special features of GWAS data can be used to increase the statistical testing power and have given rise to numerous statistical methods. In this review, we collect and organize them in three categories and synopsise them in Supplemental Table 1-3. These lists are not intended to be fully inclusive but instead to serve as a guide to understanding the features of these methods.

Supplemental Table 1 summarizes the statistical methods for detecting interactions between two single markers. They are the most basic tools for detecting $G \times G$ interaction. In its fundamental form, the logistic regression method directly relates disease risks to two genetic markers in a prospective fashion. Whereas the other methods, including LD-based and Chi-square tests, retrospectively detect genetic interactions by comparing “association” between two markers in case and control populations. Although the two categories of methods appear to arise from different motivations, they are closely related analytically, as will be revealed in section 3.

Supplemental Table 2 briefly describes the statistical methods for testing interactions between multiple markers. There are two main types of approaches. The first one includes multi-factor dimensionality reduction (MDR), the canonical correlation based test and the covariance based test that all focus on characterizing the joint, or some aspects of the joint distributions of the genotype data. The second group is principal-component analysis and least absolute shrinkage and selection operator (PCA-LASSO) approach and the latent variable logistic regression that both detect interaction via regression models coupled with dimension reduction strategies.

In order to screen for genetic interactions genome-widely among hundreds of thousands of genetic markers, methodologies that are both statistically powerful and computationally feasible are in great demand. Numerous methods have been published in the literature to tackle the challenges, and we summarize some major ones in Supplemental Table 3. These methods are primarily based on stochastic screens and searches. About one third of them pre-screen genetic markers based on, for example, their main effects to reduce the dimensionality (number of markers) of the problem (e.g., method 4, 7, 10 and 11 in Table 3), while some others genome-widely search for candidate genetic interactions. Some methods aggregates results from a collection of models/searches to reap promising

interactions (e.g., method 2, 3, 7, 8, 10 and 11). The number of single-nucleotide polymorphisms (SNPs) of the real data sets described in the papers vary from a dozen to hundreds of thousands of genetic markers. While the majority of these methods directly screen for genetic interactions, some also detect the overall association in the presence of genetic interactions (e.g., method 2, 3, 4, 5, and 10). Genome-wide search for genetic interactions is bound to involve a large number of tests, and thus prohibitive computation is inevitable. What makes such testing truly feasible are recent advances in computation and strategic algorithm design. A partial list of such methods includes methods 9, 12 and 13 in Supplemental Table 3.

Since testing the interaction between two single markers forms the basis for virtually all the methods in Supplemental Table 2 and 3, in what follows, we will focus on investigating the methods in Supplemental Table 1 as well as their variants.

3 Tests on gene-gene interaction

In this section, a collection of statistical methods presented in Supplemental Table 1 and their variants with slight modifications will be investigated. These methods include two logit-based tests, three LD-based tests [Wu et al., 2008; Yang Y. et al., 2009; Zhao et al., 2006], and three case-only tests. They are designed to test for interaction between two single markers. We first study various definitions of interactions and the relations between them. Then we clarify the definition of interaction assumed by each test. Next we elaborate the scenarios under which these tests share the same null space, and lastly compare the tests in these settings. The investigation proceeds first with the dominant (recessive) disease model (section 3.1) and then extends to the general disease model (section 3.2).

3.1 Dominant (or recessive) disease model (2×2)

3.1.1 Notation and definition—When the disease model is dominant or recessive, we can reduce the commonly used nine-genotype model at two loci to the four-genotype model. Without loss of generality, we assume one allele is dominant to the other on both loci. Let G (or H) denote the genotypes AA (or BB) and Aa (or Bb); let g (or h) denote the genotype aa (or bb). Then the genotype information of a dataset can be summarized in a frequency table, Table 1, where $\mathbf{p} \equiv \{p_{00}, p_{01}, p_{10}, p_{11}\}$ and $\mathbf{q} \equiv \{q_{00}, q_{01}, q_{10}, q_{11}\}$ are the genotype frequencies for $\{GH, gH, Gh, gh\}$ in cases and controls. Let $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ be the observed frequencies of the corresponding genotypes. If haplotype at two loci can be inferred from diplotype without uncertainty, the same framework applies as well. We denote the case (control) status by $D = 1$ ($D = 0$) and let n_1 and n_0 be the sample sizes for the case and control groups. We consider three major approaches to model interaction effects: the penetrance based definition, the logistic regression model, and the LD based measurement. We focus on the exploration of these three concepts and how they are inter-related.

(a) Penetrance based definition: measures the inter-locus dependence in the penetrance table, whose elements represent the probability of developing a disease given certain genotypes [Cordell, 2002]. If we choose to measure the dependence on a multiplicative scale, we have

$$I_{penetrance} = P(D=1|GH)P(D|gh) - P(D=1|gH)P(D=1|GH) = \frac{P(D=1)^2 p_{10}p_{01}}{P(GH)P(gh)} \left(\frac{p_{00}p_{11}}{p_{10}p_{01}} - \frac{P(GH)P(gh)}{P(Gh)P(gH)} \right).$$

According to this definition, interaction is measured by the difference between the inter-locus genotype dependences in case and general populations. If we assume the two loci are

unlinked in the general population, i.e., $\frac{P(GH)P(gh)}{P(Gh)P(gH)} = 1$, the above definition examines only the inter-locus dependence in cases:

$$I_{penetrance} = \frac{P(D=1)^2}{P(G)P(g)P(H)P(h)} (p_{00}p_{11} - p_{10}p_{01}). \quad (1)$$

We can then apply case-only analysis and use Pearson's chi-squared statistic to test for interaction. Note that multiplicativity is only one of many possible scales we can use. For example, Risch [1990] used the deviation from additivity on penetrances as the measure of interaction.

(b) Logit based definition: stems from logistic regression, which is commonly used for testing for interaction between two factors. Let $X_1 = 1$ if G is present and 0 otherwise; let $X_2 = 1$ if H is present and 0 otherwise. Consider the following logistic regression model

$$\log \frac{P(D=1|X_1, X_2)}{P(D=0|X_1, X_2)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2. \quad (2)$$

We decide whether there is an interaction effect by testing if $\beta_3 = 0$, where

$$\beta_3 = \log \frac{P(D=1|GH)P(D=1|gh)}{P(D=1|Gh)P(D=1|gH)} - \log \frac{P(D=0|GH)P(D=0|gh)}{P(D=0|Gh)P(D=0|gH)}.$$

By Bayes' rule, β_3 as a function of penetrances can be replaced by genotype frequencies:

$$I_{logit} \equiv \beta_3 = \log \frac{p_{00}p_{11}}{p_{10}p_{01}} - \log \frac{q_{00}q_{11}}{q_{10}q_{01}}. \quad (3)$$

As shown in (3), when detecting interaction using logistic regression, we actually compare the differences of inter-locus dependences between the case and control groups on a log-odds scale and examine if these differences are equal.

(c) LD based definition: uses LD to measure inter-locus dependence, and defines interaction as

$$I_{LD} = \delta_1 - \delta_0 \quad (4)$$

where $\delta_1 = p_{00}p_{11} - p_{10}p_{01} = p_{11} - p_{1+}p_{+1}$; $\delta_0 = q_{00}q_{11} - q_{10}q_{01} = q_{11} - q_{1+}q_{+1}$. [Zhao et al., 2006]. By this definition, detecting interaction is equivalent to testing if the LDs are the same in case and control groups.

3.1.2 Relations between the three definitions—Suppose linkage equilibrium (LE) holds in the general population for the two loci we consider. According to (1), $I_{penetrance} = 0$ implies $p_{00}p_{11} = p_{10}p_{01}$. When a disease is rare, we expect that almost no linkage exists between the two loci in the control population: $q_{00}q_{11} = q_{10}q_{01}$. As a result, both $I_{logit} = 0$ and $I_{LD} = 0$ reduce to $p_{00}p_{11} = p_{10}p_{01}$. Therefore under the circumstance that the disease is rare and LE holds in the general population, the three definitions of no interaction by (1), (3), and (4) are equivalent.

If we do not assume the disease is rare, these three definitions are still equivalent if there is at most one marker that has a main effect and LE holds in the population.

Proposition 1: Under the assumption that at least one marker has no main effects and LE holds in the population,

$$I_{penetrance} = 0 \iff I_{logit} = 0 \iff I_{LD} = 0.$$

See the proof in Supplemental Proof A.

3.1.3 Interaction testing statistics—In this section, we describe two statistical tests for the interaction defined by I_{logit} , two tests for the interaction defined by I_{LD} and three tests for the interaction defined by $I_{penetrance}$.

(a) Logit based testing statistics: Deviance statistic $T_{logistic}$ is a commonly used statistic for testing the interaction term in logistic regression (2). Specifically let M_1 denote the full model (2) and M_0 denote the reduced model:

$$\log \frac{P(D=1|X_1, X_2)}{P(D=0|X_1, X_2)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

Let L_{M_1} and L_{M_0} be the maximum log likelihoods of the data under models M_1 and M_0 , respectively, then $T_{logistic} = -2(L_{M_0} - L_{M_1})$. Another test statistic is the Wald statistic based

on (3): $T_{OR} = \frac{\hat{I}_{OR}^2}{\hat{var}(\hat{I}_{OR})}$ where $\hat{I}_{OR} = \log \frac{\hat{p}_{00}\hat{p}_{11}}{\hat{p}_{10}\hat{p}_{01}} - \log \frac{\hat{q}_{00}\hat{q}_{11}}{\hat{q}_{10}\hat{q}_{01}}$, and $\hat{var}(\hat{I}_{OR}) = \sum_{i=0}^1 \sum_{j=1}^1 \left(\frac{1}{n_1 \hat{p}_{ij}} + \frac{1}{n_0 \hat{q}_{ij}} \right)$. Both $T_{logistic}$ and T_{OR} are asymptotically χ_1^2 distributed under the null $I_{logit} = 0$. Note our T_{OR} is the same as Ueki and Cordel [2012]'s

improved statistic $\mathbf{T}_{\text{AWU-cc}}$ under LE (See the supporting information Text S1 of Ueki and Cordel [2012]).

(b) LD based testing statistics: According to the definition of interaction in (4),

$$\begin{aligned}\mathbf{T}_{\text{LD}} &= \frac{(\hat{\delta}_1 - \hat{\delta}_0)^2}{\hat{V}_1 + \hat{V}_0}, [\text{Zhao et al., 2006}] \\ \text{where } \hat{\delta}_1 - \hat{\delta}_0 &= (\hat{p}_{11} - \hat{p}_{1+}\hat{p}_{+1}) - (\hat{q}_{11} - \hat{q}_{1+}\hat{q}_{+1}); \\ \hat{v}\hat{a}r(\hat{\delta}_1) &= \hat{V}_1 = \frac{\hat{p}_{1+}(1-\hat{p}_{1+})(1-\hat{p}_{+1}) + (1-2\hat{p}_{1+})(1-2\hat{p}_{+1})\hat{\delta}_1 - \hat{\delta}_1^2}{2n_1}, \\ \hat{v}\hat{a}r(\hat{\delta}_0) &= \hat{V}_0 = \frac{\hat{q}_{1+}(1-\hat{q}_{1+})(1-\hat{q}_{+1}) + (1-2\hat{q}_{1+})(1-2\hat{q}_{+1})\hat{\delta}_0 - \hat{\delta}_0^2}{2n_0}.\end{aligned}\quad (5)$$

Yang Y. et al. [2009] proposed another LD related statistic $\mathbf{T}_{\text{LD}^*}$ using a different estimator for the variance of $(\hat{\delta}_1 - \hat{\delta}_0)^2$.

$$\begin{aligned}\mathbf{T}_{\text{LD}^*} &= \frac{n_1 n_0}{n_1 + n_0} \sum_{i,j} \frac{(\hat{\delta}_{1ij} - \hat{\delta}_{0ij})^2}{\hat{r}_{ij}} \\ &= \frac{n_1 n_0}{n_1 + n_0} \sum_{i,j} \frac{\{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j}) - (\hat{q}_{ij} - \hat{q}_{i+}\hat{q}_{+j})\}^2}{\hat{r}_{ij}} \\ &= \frac{n_1 n_0}{n_1 + n_0} (\hat{\delta}_1 - \hat{\delta}_0)^2 \sum_{i,j} \frac{1}{\hat{r}_{ij}}, \text{ where } \hat{r}_{ij} \\ &= \frac{\hat{p}_{ij} n_1 + \hat{q}_{ij} n_0}{n_1 + n_0}.\end{aligned}\quad (6)$$

In $\mathbf{T}_{\text{LD}^*}$, the variance is obtained by conditioning on the margins of the genotype frequency table and the null hypothesis of no association, $p_{ij} = q_{ij}$ for $i, j = 0, 1$. Both $\mathbf{T}_{\text{LD}}$ and $\mathbf{T}_{\text{LD}^*}$ are asymptotically χ_1^2 distributed under the null $I_{\text{LD}} = 0$. In addition, $\mathbf{T}_{\text{LD}^*}$ was shown to be an interaction component partitioned from a two-locus total χ^2 -statistics [Yang Y. et al., 2009].

(c) Case-only testing statistics: When LE holds in the population, definition of interaction by (1) suggests a case-only analysis. Pearson's chi-squared statistic of independence between two loci readily serves here

$$\mathbf{T}_{\text{Pearson}} = n_1 \sum_{i,j} \frac{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}}. \quad (7)$$

Furthermore, $\mathbf{T}_{\text{OR}}$ and $\mathbf{T}_{\text{LD}}$ can be easily modified to case-only test statistics:

$$\begin{aligned}\mathbf{T}_{\text{ORc}} &= \frac{\hat{I}_{\text{ORc}}^2}{\hat{v}\hat{a}r(\hat{I}_{\text{ORc}})} \text{ and } \mathbf{T}_{\text{LDc}} = \frac{\hat{\delta}_1}{\hat{V}_1} \text{ where } \hat{I}_{\text{ORc}} = \log \frac{\hat{p}_{00}\hat{p}_{11}}{\hat{p}_{10}\hat{p}_{01}} \text{ and} \\ \hat{v}\hat{a}r(\hat{I}_{\text{ORc}}) &= \sum_{i=0}^1 \sum_{j=1}^1 \left(\frac{1}{n_1 \hat{p}_{ij}} \right). \text{ Note } \hat{v}\hat{a}r(\hat{I}_{\text{ORc}}) \text{ is always smaller than } \hat{v}\hat{a}r(\hat{I}_{\text{OR}}) \text{ in part (a),} \\ &\text{and thus is more efficient. Also } \mathbf{T}_{\text{ORc}} \text{ is equivalent to } \mathbf{T}_{\text{AWU-co}} \text{ in Ueki and Cordel [2012].}\end{aligned}$$

Test statistics $T_{\text{Pearson}}, T_{\text{LDc}}, T_{\text{ORc}}$ are all asymptotically central χ_1^2 distributed under the null hypothesis $I_{\text{penetrance}} = 0$.

3.1.4 Simulation Study—In section 3.1.2, we show theoretically that when LE between two loci holds in the general population, if one marker has no main effect or the disease is rare, then the parameter space corresponding to “no interaction” defined by $I_{\text{penetrance}} = 0$, $I_{\text{logit}} = 0$ and $I_{\text{LD}} = 0$ are the same. Consequently, under these circumstances, the seven test statistics $T_{\text{logistic}}, T_{\text{OR}}, T_{\text{LD}}, T_{\text{LD*}}, T_{\text{Pearson}}, T_{\text{ORc}}$, and T_{LDc} share the same null space and are comparable. In this section, we will first illustrate this result by examining the type I errors of various statistics. Different forms of different test statistics make them target on different alternative spaces. Whether one is superior to another depends on whether the alternative space of the statistic reflects the underlying mechanism of biological interaction, of which we often do not have a good understanding. Therefore we also conduct extensive simulation studies to compare the performances of these statistics under different settings mimicking various biological interaction patterns. The simulation result provides insights on the choices of interaction tests in practice.

We simulate data according to a penetrance model described in Table 2 where each cell specifies the penetrance of the given genotype. We set the minor allele frequencies for g and h at 0.1. Then we vary γ , the strength of interaction, and λ_1, λ_2 , the main effects of the two markers. When examining the type I errors, we set $\gamma = 1$ and change the prevalences through α , the penetrance for genotype GH . When comparing the powers of different tests, we increase γ from 1 to 3. For each simulation setting, we have 5000 replicates. For each dataset, 5000 cases and 5000 controls are generated to calculate each test statistic, but the case-only statistics $T_{\text{Pearson}}, T_{\text{ORc}}$ and T_{LDc} use only the 5000 cases.

Type I error: Table 6 summarizes the type I error results. When there is no main effect ($\lambda_1 = 1, \lambda_2 = 1$) or only one marker has a main effect ($\lambda_1 = 1, \lambda_2 = 3$), type I errors of all statistics are well controlled at 0.05 across different prevalences (Supplemental Table 4). This result is consistent with Proposition 1. When both markers have main effects, the type I errors of $T_{\text{logistic}}, T_{\text{OR}}, T_{\text{LD}}$, and $T_{\text{LD*}}$ are inflated as the disease becomes common. Specifically when $\lambda_1 = \lambda_2 = 1.5$, type I errors are not controlled after the disease prevalence becomes greater than 0.2; when $\lambda_1 = \lambda_2 = 3$, type I errors are not controlled as the prevalence is beyond 0.05. On the other hand, the type I errors of case-only statistics $T_{\text{Pearson}}, T_{\text{ORc}}, T_{\text{LDc}}$ remain at 0.05 across different disease prevalences. This simulation results also illustrate the range of disease risk, in which these different tests are comparable. Note the range depends on the strength of main effects.

Power: We consider three interaction settings in our simulations. In the first setting (Figure 1(A)), both defective alleles g and h are required to exhibit the trait ($\lambda_1 = \lambda_2 = 1$). An example of this phenomenon would be the development of Retinitis Pigmentosa (RP) in which only double heterozygotes of both genes ROM1 and peripherin/RDS develop the disease [Kajiwara et al., 1994]. In the second setting (Figure 1(B)), we set $\lambda_1 = 1$, and $\lambda_2 = 1.5$. This simulation mimics the scenario that one locus acts as a modifier of the other locus. For example, it is reported that a single-nucleotide polymorphism (SNP) in RAD51 gene

modified breast or ovarian cancer risk in BRCA2 carriers [Levy-Lahad et al., 2001]. The third setting, (Figure 1(C)), assumes a heterogeneity model. Predisposing markers at either locus will lead to the disease, but the presence of both predisposing markers further increases the disease risk ($\lambda_1 = \lambda_2 = 1.5$). According to Table 6 and Supplementary Table 4, we set the disease risk at 0.005, so that the type I errors of all the statistics are well controlled at the same level and we can perform the power comparison. As illustrated in Figure 1, the seven tests fall into three categories: (1) T_{logistic} and T_{OR} ; (2) T_{LD} and T_{LD^*} ; (3) T_{Pearson} , T_{ORc} and T_{LDc} . Performances of the tests within each category are close to each other. Comparing these three groups of tests, the logit based tests have the lowest power in all scenarios. LD based tests lie in between the case-only tests and the logit based tests. The powers of the LD based tests increase as main effects increase when comparing Figure 1(B) to Figure 1(A) and Figure 1(C) to Figure 1(B). Case-only statistics use less data than case-control statistics (5000 cases versus 5000 cases plus 5000 controls) but have better powers. This is due to the fact that $\hat{v}ar(\hat{I}_{\text{ORc}}) < \hat{v}ar(\hat{I}_{\text{OR}})$ as we demonstrated in section 3.1.3. Our result agrees with Yang et al. [1999]'s conclusion that the case-only design requires fewer cases than case-control design, and is an efficient and valid approach to measuring gene-gene interaction, assuming the genes under study are in LE.

3.2 General disease model (3 × 3)

3.2.1 Notation and definition—In this section, we consider more general disease models with a total of nine genotypes on two loci. Let $\mathbf{p} \equiv \{p_{00}, p_{01}, p_{02}, p_{10}, \dots, p_{22}\}$ denote the genotype frequencies of $\{AABB, AaBB, aaBB, AABb, \dots, aabb\}$ for cases and $\mathbf{q} = \{q_{00}, q_{01}, q_{02}, q_{10}, \dots, q_{22}\}$ for controls. In section 3.1.1 and 3.1.3, three types of interaction definitions and their motivated testing statistics are introduced. In the following, we will extend them to the general disease models.

(a) Penetrance based definition: Table 3 parameterizes different penetrance patterns we might observe. A natural interpretation of no interaction is that the penetrance for a given genotype in Table 3 is the product of the two corresponding “marginal penetrances”. We parameterize penetrance patterns such that $\lambda_1, \dots, \lambda_4$ represents the main effects and $\gamma_1, \dots, \gamma_4$ the interaction effects, and when $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 1$, the rows and the columns of Table 3 are independent. Interaction is thus defined as $I_{\text{penetrance}} = \sum_{i=1}^4 (\gamma_i - 1)^2$.

(b) Logit based definition: Let X_1 be the count of allele A on the first locus and X_2 be the count of allele B on the second locus. Consider the logistic regression model

$$\begin{aligned}
& \log \frac{P(D=1|X_1, X_2)}{P(D=0|X_1, X_2)} \\
&= \beta_0 + \beta_1 I(X_1 \\
&= 2) + \beta_2 I(X_1 \\
&= 1) + \beta_3 I(X_2 \\
&= 2) + \beta_4 I(X_2 \\
&= 1) + \beta_5 I(X_1 \\
&= 1) I(X_2 \quad (8) \\
&= 1) + \beta_6 I(X_1 \\
&= 2) I(X_2 \\
&= 1) + \beta_7 I(X_1 \\
&= 1) I(X_2 \\
&= 2) + \beta_8 I(X_1 \\
&= 2) I(X_2 = 2).
\end{aligned}$$

Interaction based on this model is defined as $I_{logit} = \sum_{i=5}^8 \beta_i^2$, such that no interaction ($I_{logit} = 0$) means $\beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$. Note parameters $\beta_5, \dots, \beta_8$ can be written as

$$\begin{aligned}
\beta_5 &= \log \frac{p_{00}p_{11}}{p_{01}p_{10}} - \log \frac{q_{00}q_{11}}{q_{01}q_{10}}, & \beta_6 &= \log \frac{p_{00}p_{21}}{p_{01}p_{20}} - \log \frac{q_{00}q_{21}}{q_{01}q_{20}}; \\
\beta_7 &= \log \frac{p_{00}p_{12}}{p_{02}p_{12}} - \log \frac{q_{00}q_{12}}{q_{02}q_{12}}, & \beta_8 &= \log \frac{p_{00}p_{22}}{p_{02}p_{20}} - \log \frac{q_{00}q_{22}}{q_{02}q_{20}}.
\end{aligned}$$

This again suggests that testing for interaction by logistic regression is equivalent to testing the differences of inter-locus dependences between cases and controls on a log-odds scale.

(c) LD based definition: To extend the LD based definition to general disease models, we first need to find a measure for genotypic disequilibrium for two loci with nine genotypes. Weir [1996] proposed a digenic disequilibrium composite measure called composite LD (CLD). In this definition, the CLD for cases is defined as $\delta_{AB}^1 = p_{AB} + p_{A/B} - 2p_A p_B$ where gametic frequencies are recovered from genotypic frequencies

$$p_{AB} = P_{AB} + \frac{1}{2}(P_{Ab}^{AB} + P_{aB}^{AB} + P_{ab}^{AB}); p_{A/B} = P_{AB} + \frac{1}{2}(P_{Ab}^{AB} + P_{aB}^{AB} + P_{Ab}^{aB}).$$

so that $p_{AB} + p_{A/B} = 2p_{00} + p_{01} + p_{10} + \frac{1}{2}p_{11}$. Since $p_A p_B = 2(p_{0+} + p_{1+})(p_{+0} + p_{+1})$, δ_{AB}^1 can be calculated by $\mathbf{p}$. Following the same step, δ_{AB}^0 can be also calculated for the controls. Wu et al. [2008] then proposed the following definition of interaction:

$$I_{CLD} = \delta_{AB}^1 - \delta_{AB}^0. \quad (9)$$

An alternative way to extend the concept of LD for two loci with nine genotypes is to consider each genotype individually. Yang Y. et al. [2009] detected interaction by comparing a vector of inter-locus disequilibrium measure $\delta_{ij}^1 = p_{ij} - p_{i+}p_{+j}$ from the cases to the controls $\delta_{ij}^0 = q_{ij} - q_{i+}q_{+j}$, $i, j = 0, 1, 2$. In this approach, the definition of interaction assumed is

$$I_{LD}^* = \sum_{i,j} (\delta_{ij}^1 - \delta_{ij}^0)^2 \text{ for } i, j = 0, 1, 2. \quad (10)$$

3.2.2 Relations between the three definitions—Although the above definitions come from different perspectives, correspondence among them still exists under certain circumstances as stated in Proposition 2 and 3.

Proposition 2: $I_{LD}^* = 0$ is a sufficient condition for $I_{CLD} = 0$.

Proof is provided in Supplemental Proof B. However, $I_{CLD} = 0$ does not imply $I_{LD}^* = 0$. To show this, we provide a counterexample in Supplemental Proof B.

Proposition 3: Under the assumption that linkage equilibrium holds in a population,

$$I_{penetrance} = 0 \iff I_{logit_{case}} = 0 \iff I_{LD_{case}}^* = 0 \quad (11)$$

Where $I_{logit_{case}} \equiv \log^2 \frac{p_{00}p_{11}}{p_{01}p_{10}} + \log^2 \frac{p_{00}p_{21}}{p_{01}p_{20}} + \log^2 \frac{p_{00}p_{12}}{p_{02}p_{12}} + \log^2 \frac{p_{00}p_{22}}{p_{02}p_{20}}$ and $I_{LD_{case}}^* = \sum_{i,j} (p_{ij} - p_{i+}p_{+j})^2$ for $i, j = 0, 1, 2$.

Proof is provided in Supplemental Proof C. If we further assume that the disease is rare, the meanings of no interaction based on $I_{penetrance}$, $I_{logit_{case}}$ and $I_{LD_{case}}^*$ are almost the same. This is because LE in the population implies that the LD or log odds ratio in the control population is approximately 0. As a result, $I_{LD}^* = 0$ reduces to $I_{LD_{case}}^* \approx 0$ and $I_{logit} = 0$ to $I_{logit_{case}} \approx 0$. In Proposition 3, we show the equivalence of them to $I_{penetrance} = 0$, therefore $I_{logit} = 0$, $I_{LD}^* = 0$ and $I_{penetrance} = 0$ are equivalent when the disease is rare and LE holds in the population.

3.2.3 Testing statistics

(a) Logit based test: For $H_0 : I_{logit} = 0$, a deviance test can be used. Let M_1 be the full model in (8) and M_0 be the reduced model:

$$\log \frac{P(D=1|X_1, X_2)}{P(D=0|X_1, X_2)} = \beta_0 + \beta_1 I(X_1=2) + \beta_2 I(X_1=1) + \beta_3 I(X_2=2) + \beta_4 I(X_2=1). \quad (12)$$

Let L_{M_1} and L_{M_0} be the estimated maximum likelihood of the data under model M_1 and M_0 . Then $\mathbf{T}_{\text{logistic}} = -2(L_{M_0} - L_{M_1})$ is a deviance test statistic that is asymptotically χ_4^2 distributed under the null: $I_{\text{logit}} = 0$.

(b) LD based test: Following Weir [1996], Wu et al. [2008] proposed the following statistic to detect interaction defined using I_{CLD} :

$$\mathbf{T}_{\text{CLD}} = \frac{(\hat{\delta}_{AB}^1 - \hat{\delta}_{AB}^0)^2}{V(\hat{\delta}_{AB}^1) + V(\hat{\delta}_{AB}^0)}. \quad (13)$$

The derivation for the variance is provided in Supplemental Section D. $\mathbf{T}_{\text{CLD}}$ is asymptotically distributed as central χ_1^2 under the null $I_{\text{CLD}} = 0$.

In addition, Yang Y. et al. [2009] proposed to detect interaction by testing whether I_{LD}^* in (10) equals 0. The proposed test statistic is in the same form as $\mathbf{T}_{\text{LD}^*}$ in (6) with $i, j = 0, 1, 2$ instead. $\mathbf{T}_{\text{LD}^*}$ is asymptotically distributed as a χ_4^2 under the null $I_{LD}^* = 0$.

(c) Penetrance based test: We extend the case-only testing statistic $\mathbf{T}_{\text{Pearson}}$ in (7) using $i, j = 0, 1, 2$. The resulting new $\mathbf{T}_{\text{Pearson}}$ is asymptotically distributed as a central χ_4^2 under the null $I_{\text{penetrance}} = 0$.

The statistics from the same statistical model/method are annotated in the same way as in section 3.1 and section 3.2, although the detailed calculations are different. In section 3.1.4, we illustrate that the performances of $\mathbf{T}_{\text{OR}}$ and $\mathbf{T}_{\text{logistic}}$ are close to each other, and the case-only tests $\mathbf{T}_{\text{OR}_c}$, $\mathbf{T}_{\text{LD}_c}$ and $\mathbf{T}_{\text{Pearson}}$ show similar powers. Thus in this section we only consider $\mathbf{T}_{\text{logistic}}$ among the logit based tests and $\mathbf{T}_{\text{Pearson}}$ among the penetrance based tests.

3.2.4 Simulation study—To evaluate different tests for interaction under the general disease model, we simulate data based on a multiplicative penetrance model illustrated in Table 4, where the strength of interaction is controlled by varying γ from 1 to 3. Penetrance parameter α is chosen so that the risk of the disease in the population is always kept at 0.01. Minor allele frequencies of a and b are both set at 0.3, so that the genotype frequency for $aabb$ is close to 0.1, the value we used in section 3.1.4. Sample sizes are 5000 for both case and control groups. We run 5000 simulations for each of the three settings: (a) both λ_1 and λ_2 are 1, i.e., neither marker has a main effect; (b) $\lambda_1 = 1$ and $\lambda_2 = 1.2$, i.e., one marker has a main effect; and (c) $\lambda_1 = \lambda_2 = 1.2$, i.e., both markers have main effects. Here we set λ_1, λ_2 at 1.2 so that the penetrance risk for $aabb$ in Table 4 is close to that for gh in Table 2 in section 3.1.4. The powers of the tests $\mathbf{T}_{\text{logit}}$, $\mathbf{T}_{\text{LD}^*}$, $\mathbf{T}_{\text{CLD}}$, and $\mathbf{T}_{\text{Pearson}}$ under different settings are plotted in Figure 2. Firstly, when $\gamma = 1$, i.e., there is no interaction effect, the power curves of all tests sit right at the 0.05 level, suggesting proper controls of type I errors for all methods under simulation settings considered. As to the power performances, the overall trends are similar as that shown in Figure 1. Case-only test $\mathbf{T}_{\text{Pearson}}$ has the highest power whereas the logistic regression method has the lowest. Between $\mathbf{T}_{\text{LD}^*}$ and $\mathbf{T}_{\text{CLD}}$, $\mathbf{T}_{\text{CLD}}$ gives more favorable results. Its performance is very close to the case-only test and

outperforms T_{logit} and T_{LD}^* by as large as 20% in power (Figure 2). The advantage of T_{CLD} over T_{LD}^* is partly due to the fact that it is a 1-df test and has a larger null space than the other test (see Proposition 3).

4 Detecting association allowing for interaction effects

Besides selecting a powerful test for interaction, researchers are often interested in detecting genetic effects while taking into account the possible interactions among markers. In GWAS, it has been reported that the power to detect association was improved by including interaction into consideration [Cordell, 2002]. This motivates us to conduct the following simulations and comparison studies for detecting the association between a binary outcome and a pair of loci with main and/or interaction effects. Again, we consider both the dominant (recessive) model and the general disease model. Besides the few testing statistics discussed in previous sections (T_{Pearson} , T_{LD} , T_{LD}^* and T_{CLD}), we include five additional test statistics— $T_{\text{logisticall}}$, T_{χ^2} , T_{Kullback} , T_{Pearlman} , and T_{kernel} —in our investigation. Some of these tests are for detecting the overall association, while others are for detecting interaction alone. We consider both groups of tests in the comparison, because we are interested in knowing whether interaction tests may be superior to overall association tests when interaction effects are much larger than main effects. The results from our simulation study will provide practical guidelines for selecting appropriate testing methods in GWAS.

4.1 Additional testing statistics

In this section, we describe a few additional testing statistics: deviance test $T_{\text{logisticall}}$ based on logistic regression, Pearson's chi-squared statistic T_{χ^2} , and T_{kernel} based on the logistic kernel machine regression method [Liu et al., 2008; Wu et al., 2010]. The first three statistics detect both interaction effects between two single markers and their main effects. We include kernel machine regression method because, by employing the flexible kernel framework, this method is able to model complicated interaction patterns, and has been demonstrated to be effective in association studies [Liu et al., 2008; Wu et al., 2010]. The two correlation based tests T_{Kullback} and T_{Pearlman} for detecting interaction are considered here because correlation coefficient r can be calculated conveniently without estimating phase [Wellek and Ziegler, 2008]. Specifically, T_{Pearlman} are based on Fisher's z transformation of r , which relates closely to the statistics proposed by Ueki and Cordell [2012]; Wellek and Ziegler [2008] and Kam-Thong et al. [2010].

(a) Deviance test of total association through logistic regression—Consider the full model M_1 in (8) and the reduced model M'_0 :

$$\log \frac{P(D=1|X_1, X_2)}{P(D=0|X_1, X_2)} = \beta_0. \quad (14)$$

We can reject the null that there is no association between the two markers and the disease if $T_{\text{logisticall}} = -2(L_{M'_0} - L_{M_1})$ is large. The significance level can be decided based on $T_{\text{logisticall}}$.

(b) Pearson's chi-squared test of association—We arrange the observed genotype frequencies in cases ($\mathbf{p}$) and in controls ($\mathbf{q}$) in a 2×9 table with the first row for cases, the second row for controls, and the columns representing specific genotypes. Then we apply Pearson's chi-squared test to test whether the two rows have the same distribution, which corresponds to the null hypothesis that the disease is not attributable to the two markers. The test statistics $\mathbf{T}_{\chi^2}$ is asymptotically central χ^2_8 distributed under the null.

(c) Logistic kernel machine method—According to Liu et al. [2008]; Wu et al. [2010], the logistic kernel machine regression model for two loci can be written as

$$\text{logit}P(Y_i=1|x_{i1}, x_{i2})=\beta_0+h(x_{i1}, x_{i2}), i=1, 2, \dots, n,$$

in which $h(x_{i1}, x_{i2})=h(\mathbf{x}_i)=\sum_{i'=1}^n \gamma_{i'} K(\mathbf{x}_i, \mathbf{x}_{i'})$ for some $\gamma_1, \dots, \gamma_n$, and $K(\mathbf{x}_i, \mathbf{x}_{i'})$ measures the similarities between $\mathbf{x}_i$ and $\mathbf{x}_{i'}$. Based on the link between this logistic kernel machine framework and the generalized linear mixed model, $h(\mathbf{x}_i)$ was treated as a subject specific random effect. Suppose $h(\mathbf{x})$ follows an arbitrary distribution F with mean 0 and variance $\tau \mathbf{K}$. Then $H_0: h(\mathbf{x}) = 0$ is equivalent to $H_0: \tau = 0$. The corresponding score statistic is

$$\mathbf{T}_{\text{kernel}} = \frac{(\mathbf{y} - \hat{\mu}_0)' \mathbf{K} (\mathbf{y} - \hat{\mu}_0)}{2}.$$

Under H_0 , $\mathbf{T}_{\text{kernel}}$ is asymptotically $\kappa \chi^2_\nu$ distributed. The calculation of κ and ν are described in Supplemental Section E. We use a quadratic kernel in our simulation studies but other kernel choices are also possible [Wu et al., 2010].

(d) Correlation based tests—We introduce two more statistics $\mathbf{T}_{\text{Perlman}}$ and $\mathbf{T}_{\text{Kullback}}$, which target on testing interaction of two markers on the disease through comparing the correlations between the two loci in cases and controls. These tests relate closely to LD tests in which the interaction effect is assessed by comparing the covariance between two markers in the cases and controls. Specifically,

$$\mathbf{T}_{\text{Perlman}} = \frac{1}{2} \log \left(\frac{1+r^1}{1-r^1} \right) - \frac{1}{2} \log \left(\frac{1+r^0}{1-r^0} \right)$$

where r^1 and r^0 are the correlation coefficients of two markers in the cases and controls.

$\mathbf{T}_{\text{Perlman}}$ is asymptotically normal distributed with a variance of $\frac{1}{n_1-3} + \frac{1}{n_0-3}$ [Larntz and Perlman, 1988].

$$\mathbf{T}_{\text{Kullback}} = (n^1 - 1) \log \frac{|R|}{|r^1|} + (n^0 - 1) \log \frac{|R|}{|r^0|}$$

where $R = \frac{(n^1 - 1)r^1 + (n^0 - 1)r^0}{n^1 + n^0 - 2}$. $\mathbf{T}_{\text{Kullback}}$ is asymptotically central χ^2_1 distributed [Kullback, 1967]. Note Wellek and Ziegler's statistics is the same as $\mathbf{T}_{\text{Perlman}}$ but the variance was developed based on delta method and $\text{var}(\hat{r})$, thus is more complicated to calculate.

4.2 Simulation study for testing joint association

We perform simulation studies to compare above statistics including $\mathbf{T}_{\text{Logisticall}}$, $\mathbf{T}_{\chi^2}$, $\mathbf{T}_{\text{LD}}$, $\mathbf{T}_{\text{LD}^*}$, $\mathbf{T}_{\text{CLD}}$, $\mathbf{T}_{\text{Perlman}}$, $\mathbf{T}_{\text{Kullback}}$, $\mathbf{T}_{\text{Pearson}}$ and $\mathbf{T}_{\text{kernel}}$ for testing the association of a phenotype and a pair of loci. We compare their powers under simple dominant (recessive) disease model (2×2) first and then general disease model (3×3). Simulation settings and results under dominant (recessive) disease models are summarized in Supplemental Figure 1. For general disease model (3×3), we adopt the six epistatic models (Table 5) introduced by Neuman et al. [1992] to generate data. The relative risk with predisposing genotypes versus other genotypes f^*/f varies from 1 to 2. Disease prevalences are kept at 0.01 by adjusting f . Allele frequencies for a and b are both set to 0.3 for the same reason described in section 3.2.4. The sample sizes of cases and controls are 5000. For each setting, 5000 independent replicates are generated, and $\mathbf{T}_{\text{Logisticall}}$, $\mathbf{T}_{\chi^2}$, $\mathbf{T}_{\text{LD}^*}$, $\mathbf{T}_{\text{CLD}}$, $\mathbf{T}_{\text{Pearson}}$, $\mathbf{T}_{\text{Perlman}}$, $\mathbf{T}_{\text{Kullback}}$ and $\mathbf{T}_{\text{kernel}}$ are compared. Note $\mathbf{T}_{\text{LD}}$ for the dominant (recessive) model (2×2) is replaced by $\mathbf{T}_{\text{CLD}}$.

Figure 3 shows the performance of each test. The null hypotheses of interest here is $f^*/f = 1$, i.e., there is no association between the disease and the pair of loci. Under this null, all the tests properly control their type I errors, as all lines in Figure 3 sit around 0.05 when $f^*/f = 1$. Two correlation based tests $\mathbf{T}_{\text{Kullback}}$ and $\mathbf{T}_{\text{Perlman}}$ have identical results, and so do $\mathbf{T}_{\text{Logisticall}}$ and $\mathbf{T}_{\chi^2}$ with $\mathbf{T}_{\chi^2}$ computationally much simpler. However, if we need to incorporate other covariates, the logistic regression provides a framework with such capability. For simulation settings (B)-(F), the tests for overall association $\mathbf{T}_{\text{Logisticall}}$, $\mathbf{T}_{\chi^2}$ and $\mathbf{T}_{\text{kernel}}$ give favorable results compared to the two LD based tests and the two correlation based tests. Opposite results are observed in simulation setting (A), in which $\mathbf{T}_{\text{kernel}}$ has the worst power and the case-only test $\mathbf{T}_{\text{Pearson}}$ has the best power. When the interaction effect is relatively larger than the main effect as in simulation settings (B), (D) and (E), the advantage of case only tests such as $\mathbf{T}_{\text{Pearson}}$ remains. It is worth noting that, in general, $\mathbf{T}_{\text{LD}^*}$ has a better performance than $\mathbf{T}_{\text{CLD}}$, especially in the simulation settings (C) and (D). As Figure 3 shows, which method is superior depends on the structure of the underlying disease model, which is often unknown. Overall $\mathbf{T}_{\text{Logisticall}}$, $\mathbf{T}_{\chi^2}$, $\mathbf{T}_{\text{Pearson}}$ perform well. When interaction effects are relatively strong $\mathbf{T}_{\text{Pearson}}$ surpasses $\mathbf{T}_{\text{Logisticall}}$, $\mathbf{T}_{\chi^2}$; when main effects are stronger, the opposite are observed. $\mathbf{T}_{\text{kernel}}$ performs well in most scenarios unless the interaction effect is more prominent than the marginal effect. In summary, based on our simulations under LE, we would recommend using $\mathbf{T}_{\text{Logisticall}}$, $\mathbf{T}_{\chi^2}$ and $\mathbf{T}_{\text{Pearson}}$. If some level of interaction effects is believed to be present, $\mathbf{T}_{\text{Pearson}}$ is the

first choice; if, on the other hand, the interaction effect is believed to be very small and there is no other factor needs to be adjusted in the analysis, T_{χ^2} is recommended due to its computational convenience over $T_{\text{Logisticall}}$; if other covariates need to be adjusted, $T_{\text{Logisticall}}$ is advocated.

5 Discussion

The strong interest of biologists to learn the underlying disease mechanism and the growing need to incorporate interaction testing in GWAS have motivated our investigation on existing gene-gene interaction tests and association tests allowing for interaction.

Since interaction is a concept without agreed mathematical definition, it is worthy to clarify how various interactions are defined when performing the statistical tests. Interactions based on penetrance tables model the disease risk directly and provide a straightforward interpretation for biologists, whereas LD and logistic regression model based definitions are introduced from statistical perspectives. In this paper, we first derive the theoretical relationships between these definitions. We demonstrate that, under LE, when there is at most one marker that has a main effect or when the disease is rare, the three definitions share the same or similar null spaces. Therefore, under these circumstances, tests designed based on one definition shall always have properly controlled type I errors, even if the true underlying interaction mechanism is based on another definition. However, when disease is more common and both loci have main effects, the null spaces of various tests stemmed from different interaction definitions become different. And thus, if the true disease mechanism follows one interaction model, the tests based on another interaction model may very likely encounter inflated type I errors. These theoretical conclusions are well supported by empirical studies conducted in the paper. Specifically, we simulate data under LE according to the penetrance risk models (Table 2). Under this setting, case-only statistics considered in the paper all correspond to $I_{\text{penetrance}}$, the penetrance based definition of interaction, and thus are always targeting at the right null hypothesis space. Other statistics using both Case and Control data, referred to as CC statistics, are either testing $I_{\text{logit}} = 0$ or $I_{\text{LD}} = 0$. According to Proposition 1, when disease is more common and both markers have main effects, $I_{\text{logit}} = 0$ or $I_{\text{LD}} = 0$ are not equivalent to $I_{\text{penetrance}} = 0$. Therefore, as the disease prevalence increases, type I errors of CC statistics are subjected to type I error inflation, as their null hypothesis space is mis-specified (see Table 6).

We then evaluate the power performance of a collection of interaction tests through extensive simulation experiments. We again generate data based on penetrance risk models, and focus on settings where these tests share the same null space and thus all have properly controlled type I errors. The results suggest that case-only tests have the most favorable power for testing gene-gene interaction under LE. The advantage of case-only tests comes from their use of the LE assumption. For example, in section 3.1.3, we show the variance of the case-only statistics that is based on log odds ratio is always smaller than the one that uses both case and control data. These results are consistent with Yang et al. [1999] on case-only design for measuring gene-gene interaction. We also performed extensive simulation studies to compare the power of several association tests targeting on both main and interaction

effects. Across the six different disease models evaluated, the overall performance of the logistic regression based test $T_{\text{logisticall}}$ and Pearsons chi-squared test T_{χ^2} are the best.

This work relates closely to a recent paper by Ueki and Cordel [2012], in which the authors also studied several interaction testing methods using case/control or case-only data. Our theories are consistent with, and provide helpful insights towards interpreting the empirical results of Ueki and Cordel [2012]: (1) when the disease is more common, the presence of main effects appears to have an impact on the type I errors of some methods; (2) some case/control statistics and their corresponding case-only statistics have different behaviors in term of the type I error inflation caused by main effects; and (3) some statistics are insensitive to main effects at both loci under a rare disease assumption or in the presence of a single main effect. It is worth noting that which statistics have well controlled type I errors depend on the “true” interaction model underlying the data. Since penetrance model and logistic model are used to generate data in our and Ueki and Cordel [2012]'s simulations respectively, case-only statistics and case-control statistics demonstrate well controlled type I errors accordingly.

All the investigation conducted in this paper is under the assumption of LE. However, in real applications, often LE may not hold. With regard to this limitation, on one hand, we want to emphasize that the neat theoretical connections between various statistics derived under LE provide deep insights on the differences of these statistics, and can help us to better understand the behaviors of these statistics under more complicated settings. On the other hand, to complete the picture, we also conduct small scale simulations under LD (data not shown due to space limitation). We find the logistic regression test T_{logit} and log odds ratio test T_{OR} are robust to LD, while the two LD based tests (T_{LD} and $T_{\text{LD}*}$), as well as the case-only tests are not. The inflation in type I errors of LD or penetrance based statistics caused by main effects can be substantial even when a disease is rare. Similar behavior for LD based case/control statistics were also observed in Ueki and Cordel [2012]. Thus, the LD based statistics and penetrance based statistics are not recommended if LD is present in the data. These results also imply that the relationship among various definitions of interaction is more complicated under LD, which warrants future research.

The development of gene-gene interaction tests helps to detect genetic effects in association studies, especially for complex diseases. However the inference of biological mechanism through statistical interaction is sometimes limited. Biological interaction is more of a result of physical interaction between molecules within a biochemical pathway. The two interactions are closely related if the statistical model reflects the biological model. This poses a dilemma since often scientists are interested in making inferences about biological interaction from statistical results. Without one-to-one correspondence between the two, our ability to make such inference is limited [Cordell, 2009]. However, the advancement in new technologies might provide us more information on genome, proteome and metabolome. Combined with this information, it is not infeasible to elucidate biological interactions through sensible statistical models in the future [Moore and Williams, 2005].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Support for this work was provided by the National Institute of Health grant R01CA138215 and R01GM082802.

References

- Bateson, W. Mendel's principles of heredity. Cambridge University Press; 1909.
- Chakravarti A. Population genetics making sense out of sequence. *Nat Genet.* 1999; 21(1 Suppl):56–60. [PubMed: 9915503]
- Cook N, Zee R, Ridker P. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat Med.* 2004; 23:1439–1453. [PubMed: 15116352]
- Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet.* 2002; 11:2463–2468. [PubMed: 12351582]
- Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet.* 2009; 10:392–404. [PubMed: 19434077]
- D'Angelo G, Rao D, Gu C. Combining least absolute shrinkage and selection operator (LASSO) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. *BMC Proceedings.* 2009; 3(Suppl 7):S62. [PubMed: 20018056]
- Donnelly P. Progress and challenges in genome-wide association studies in humans. *Nature.* 2008; 456:728–731. [PubMed: 19079049]
- Fisher R. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edin.* 1918; 52:399–433.
- Goldstein D. Common genetic variation and human traits. *N Engl J Med.* 2009; 360:1696–8. [PubMed: 19369660]
- Hardy J, Singleton A. Genomewide Association Studies and Human Disease. *N Engl J Med.* 2009; 360:1759–68. [PubMed: 19369657]
- Hirschhorn J. Genomewide association studies illuminating biologic pathways. *N Engl J Med.* 2009; 360:1699–701. [PubMed: 19369661]
- Hirschhorn J, Daly M. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics.* 2005; 6:95–108.
- Jiang R, Tang W, Wu X, Fu W. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC bioinformatics.* 2009; 10:S65. [PubMed: 19208169]
- Kajiwaru K, Berson EL, Dryja TP. Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. *Science.* 1994; 264:1604–1608. [PubMed: 8202715]
- Kam-Thong T, Czamara D, Tsuda K, Borgwardt K, Lewis CM, et al. EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. *European Journal of Human Genetics.* 2010; 19:465–471. [PubMed: 21150885]
- Kooperberg C, LeBlanc M, YJ D, Rajapakse I. Structures and Assumptions: Strategies to Harness Gene $\times$ Gene and Gene $\times$ Environment Interactions in GWAS. *Statistical Science.* 2009; 24:472–488. [PubMed: 20640184]
- Kooperberg C, Ruczinski I. Identifying interaction SNPs using Monte Carlo logic regression. *Genet Epidemiol.* 2005; 28:157–170. [PubMed: 15532037]
- Kraft P, Cox G. Study designs for genome-wide association studies. *Advances in Genetics.* 2008; 60:465–504. [PubMed: 18358330]
- Kraft P, Hunter D. Genetic risk prediction are we there yet? *N Engl J Med.* 2009; 360:1701–3. [PubMed: 19369656]
- Ku C, Loy E, Pawitan Y, Chia K. The pursuit of genome-wide association studies: where are we now? *J Hum Genet.* 2010; 55:195–206. [PubMed: 20300123]

- Kullback S. On testing correlation matrices. *Journal of the Royal Statistical Society Series C (Applied Statistics)*. 1967; 16(1):80–85.
- Larntz, K.; Perlman, MD. *A Simple Test for the Equality of Correlation Matrices IV*. Vol. 2. New York: Springer-Verlag; 1988. p. 289–298.
- Levy-Lahad E, Lahad A, Eisenberg S, Dagan E, Paperna T, et al. A single nucleotide polymorphism in the RAD51 gene modifies cancer risk in BRCA2 but not BRCA1 carriers. *Proc Natl Acad Sci U S A*. 2001; 98:3232–3236. [PubMed: 11248061]
- Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*. 2008; 9:292. [PubMed: 18577223]
- Manolio Tea. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest*. 2008; 118:1590–1605. [PubMed: 18451988]
- Manolio Tea. Finding the missing heritability of complex diseases. *Nature*. 2009; 461:747–753. [PubMed: 19812666]
- Marchini J, Donnelly P, Cardon L. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*. 2005; 37(4):413–417. [PubMed: 15793588]
- McCarthy M, Abecasis G, Cardon L, Goldstein D, Little J, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*. 2008; 9:356–369. [PubMed: 18398418]
- Millstein J, Conti D, Gilliland F, Gauderman W. A Testing Framework for Identifying Susceptibility Genes in the Presence of Epistasis. *Am J Hum Genet*. 2006; 78:1527.
- Moore J, Asselbergs F, Williams S. Bioinformatics Challenges for Genome-Wide Association Studies. *Bioinformatics*. 2010; 26(4):445–55. [PubMed: 20053841]
- Moore J, Williams S. Epistasis and Its Implications for Personal Genetics. *American Journal of Human Genetics*. 2009; 85(3):309–20. [PubMed: 19733727]
- Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays*. 2005; 27:637–646. [PubMed: 15892116]
- Neuman RJ, Rice JP, Chakravarti A. Two-Locus models of disease. *Genetic Epidemiology*. 1992; 9:347–365. [PubMed: 1427023]
- Peng Q, Zhao J, Xue F. A gene-based method for detecting gene-gene co-association in a case-control association study. *European Journal of Human Genetics*. 2010; 18:582–587. [PubMed: 20029457]
- Rajapakse I, Perlman MD, Martin PJ, Hansen JA, Kooperberg C. Multivariate Detection of Gene-Gene Interactions. *Genetic Epidemiology*. 2012; 36:622–630. [PubMed: 22782518]
- Reich D, Lander E. On the allelic spectrum of human disease. *Trends Genet*. 2001; 17:502–510.
- Risch N. Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet*. 1990; 46:222–228. [PubMed: 2301392]
- Ritchie M, Hahn L, Roodi N, Bailey L, Dupont W, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*. 2001; 69(1):138–147. [PubMed: 11404819]
- Seng K, Seng C. The success of the genome-wide association approach: a brief story of a long struggle. *Eur J Hum Genet*. 2008; 16:554–564. [PubMed: 18285837]
- Teo Y. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Curr Opin Lipidol*. 2008; 19:133–143. [PubMed: 18388693]
- Ueki M, Cordel HJ. Improved Statistics for Genome-Wide Interaction Analysis. *PLoS Genetics*. 2012; 8
- Wan X, Yang C, Yang Q, Xue H, Fan X, et al. BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *The American Journal of Human Genetics*. 2010a; 87:325–340.
- Wan X, Yang C, Yang Q, Xue H, Tang NL, et al. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics*. 2010b; 26:30–37. [PubMed: 19880365]

- Wang Y, Barratt J, Clayton G, Todd A. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*. 2005; 6:109–118.
- Weir, B. Genetic data analysis II. Vol. II. Sinauer Associates; 1996.
- Wellek S, Ziegler A. A genotype-based approach to assessing the association between single nucleotide polymorphisms. *Human Heredity*. 2008; 67:128–139. [PubMed: 19077429]
- Wu J, Devlin B, Ringquist S, Trucco M, Roeder K. Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic epidemiology*. 2010; 34:275–285. [PubMed: 20088021]
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet*. 2010; 86:929–942. [PubMed: 20560208]
- Wu X, Jin L, Xiong M. Composite measure of linkage disequilibrium for testing interaction between unlinked loci. *Eur J Hum Genet*. 2008; 16:644–651. [PubMed: 18212814]
- Yang C, He Z, Wan X, Yang Q, Xue H, et al. SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics*. 2009; 25:504–511. [PubMed: 19098029]
- Yang Q, Khoury MJ, Sun F, Flanders WD. Case-Only Design to Measure Gene-Gene Interaction. *Epidemiology*. 1999; 10:167–170. [PubMed: 10069253]
- Yang Y, He C, Ott J. Testing association with interactions by partitioning chi-squares. *Annals of human genetics*. 2009; 73:109–117. [PubMed: 18798840]
- Zhang X, Pan F, Xie Y, F Z, Wang W. FastChi: An Efficient Algorithm for Analyzing Gene-Gene Interactions. *Pacific Symposium on Biocomputing'09*. 2009:528–539.
- Zhang X, Pan F, Xie Y, Zou F, Wang W. COE: A General Approach for Efficient Genome-Wide Two-Locus Epistasis Test in Disease Association Study. *Journal of Computational Biology*. 2010; 17(3)
- Zhang Y, Liu J. Bayesian inference of epistatic interactions in case-control studies. *Nat Genet*. 2007; 39:1167–1173. [PubMed: 17721534]
- Zhao J, Jin L, Xiong M. Test for interaction between two unlinked loci. *American journal of human genetics*. 2006; 79:831–845. [PubMed: 17033960]
- Zheng T, Wang H, Lo S. Backward genotype-trait association (BGTA) - based dissection of complex traits in case-control design. *Hum Hered*. 2006; 62:196–212. [PubMed: 17114886]
- Ziegler A, Knig I, Thompson J. Biostatistical aspects of genome-wide association studies. *Biom J*. 2008; 50(1):8–28. [PubMed: 18217698]

Figure 1.

Comparison of testing statistics on interaction effects under the dominant (recessive) disease models. 5000 cases and 5000 controls are simulated according to Table 2. γ is the strength of interaction. The main effect parameters are set at various values to simulate different scenarios: (A) $\lambda_1 = 1, \lambda_2 = 1$; (B) $\lambda_1 = 1, \lambda_2 = 1.5$; (C) $\lambda_1 = 1.5, \lambda_2 = 1.5$. A, B allele frequencies are both set at 0.1. 5000 replicates are simulated for each setting.

Figure 2.

Comparison of testing statistics for interaction effects under the general disease models. 5000 cases and 5000 controls are simulated according to Table 4. γ is the strength of interaction. The main effect parameters are set at various values to simulate different scenarios: (A) $\lambda_1 = 1, \lambda_2 = 1$; (B) $\lambda_1 = 1, \lambda_2 = 1.2$; (C) $\lambda_1 = 1.2, \lambda_2 = 1.2$. A, B allele frequencies are both set at 0.3. 5000 replicates are simulated for each setting.

Figure 3.

Power comparison of tests on joint effects of two SNPs under general disease models. The data are simulated according to penetrance Table 5(A)-(F). Relative risk $RR = f^*/f$. A, B allele frequencies are both set at 0.3 across all settings. Sample sizes for cases and controls are both 5000. 5000 replicates are simulated for each setting.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Notations for genotype frequencies in cases and controls

D = 0		D = 1	
G		g	
q_{00}	H	P_{01}	H
q_{01}	h	P_{11}	h
q_{10}			
q_{11}			

Table 2
Penetrance table for the dominant (recessive) model

	G	g
H	a	$a\lambda_1$
h	$a\lambda_2$	$a\lambda_1 \lambda_2 \gamma$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3
Penetrance table for the general disease model

Genotype	AA	Aa	aa
BB	a	$a\lambda_1$	$a\lambda_2$
Bb	$a\lambda_3$	$a\lambda_1\lambda_3\gamma_1$	$a\lambda_2\lambda_3\gamma_2$
bb	$a\lambda_4$	$a\lambda_1\lambda_4\gamma_3$	$a\lambda_2\lambda_4\gamma_4$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4
Simulation settings to detect interaction in the general disease model

Genotype	AA	Aa	aa
BB	a	$a\lambda_1$	$\alpha\lambda_1^2$
Bb	$a\lambda_2$	$a\lambda_2\lambda_1\gamma^{1/4}$	$\alpha\lambda_2\lambda_1^2\gamma^{1/2}$
bb	$\alpha\lambda_2^2$	$\alpha\lambda_2^2\lambda_1\gamma^{1/2}$	$\alpha\lambda_2^2\lambda_1^2\gamma$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5
Simulation settings to detect association in the general disease model

Genotype			Genotype			Genotype		
AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
(A)								
BB	f	f	BB	f	f	BB	f	f*
Bb	f	f	Bb	f	f*	Bb	f	f*
bb	f	f*	bb	f	f*	bb	f*	f
(B)								
(C)								
(D)								
Genotype			Genotype			Genotype		
AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
(E)								
BB	f	f	BB	f	f	BB	f	f*
Bb	f	f*	Bb	f	f*	Bb	f	f*
bb	f	f*	bb	f	f*	bb	f	f*
(F)								

Table 6
Type I errors for different testing statistics across disease prevalences

Test Statistics	Disease Prevalence							
	0.001	0.005	0.01	0.05	0.1	0.2	0.3	
T_{logit}	0.0528	0.054	0.053	0.0516	0.0494	0.0698	0.1594	
T_{OR}	0.0518	0.0532	0.051	0.0502	0.0478	0.0646	0.1412	
T_{LD}	0.0576	0.0528	0.0542	0.0504	0.0482	0.0602	0.0872	
T_{LD}^{*}	0.057	0.0518	0.053	0.0496	0.0472	0.0574	0.0834	
$T_{Pearson}$	0.0594	0.0506	0.0526	0.054	0.0528	0.0516	0.0464	
T_{Logitc}	0.0582	0.0498	0.052	0.0536	0.0526	0.0512	0.046	
T_{LDc}	0.0604	0.051	0.0524	0.058	0.0556	0.0518	0.049	
T_{Logit}	0.056	0.0524	0.0522	0.1766	0.749	-	*	
T_{OR}	0.0548	0.05	0.05	0.16	0.7154	-	-	
T_{LD}	0.0582	0.0482	0.0462	0.0706	0.1688	-	-	
T_{LD}^{*}	0.05	0.0412	0.0402	0.0602	0.146	-	-	
$T_{Pearson}$	0.0586	0.0544	0.0504	0.0504	0.0516	-	-	
T_{Logitc}	0.0586	0.0544	0.0502	0.0502	0.0516	-	-	
T_{LDc}	0.06	0.055	0.0506	0.052	0.0514	-	-	

* The penetrance for genotype gh is greater than 1 when $\lambda_1 = \lambda_2 = 3$ and the disease prevalence becomes too large.