Statistica Sinica **22** (2012), 1041-1074
doi:http://dx.doi.org/10.5705/ss.2010.183

# ON MODEL SELECTION STRATEGIES TO IDENTIFY GENES UNDERLYING BINARY TRAITS USING GENOME-WIDE ASSOCIATION DATA

Zheyang Wu and Hongyu Zhao

*Worcester Polytechnic Institute and Yale University*

*Abstract:* For more fruitful discoveries of disease genes in genome-wide association studies, it is important to know whether joint analysis of multiple markers is more powerful than the commonly used single-marker analysis, especially in the presence of gene-gene interactions. The existing literature has different, even conflicting, arguments about the power of the common model selection strategies: marginal search, exhaustive search, and forward search. Here we analytically calculate the power of these strategies and two-stage screen search to detect binary trait loci. Our approach incorporates linkage disequilibrium, random genotypes, and correlations among test statistics, which are critical characteristics of model selection that are often ignored for simplicity in the existing literature. We derive analytical results for the power of the methods to find all the associated markers, and the power to find at least one associated marker. We also consider two types of widely applied error controls: the discovery number control and the Bonferroni type I error rate control. After demonstrating the accuracy of our analytical results by simulations, we apply them to investigate the relative performance of various model selection methods in a broad genetic model space. Our research demonstrates the significant differences in power calculation and power comparison between the selection methods for binary trait and the methods for quantitative trait. Our analytical study provides rapid computation as well as insights into the statistical mechanism of capturing genetic signals under different genetic models including gene-gene interactions. We develop an R package to implement our analytical methods. Even though we focus on genetic association analysis, our results on the power of model selection procedures are general, and applicable to other studies.

*Key words and phrases:* Gene-gene interaction, genome-wide association studies, model selection, random predictors, statistical power.

## 1. Introduction

Marker-by-marker analyses in genome-wide association studies (GWAS) have unraveled many genetic variants associated with a variety of complex traits. However, current progress is still limited in two aspects. First, for such diseases as asthma and coronary heart disease, fewer novel loci have been found than those for other diseases (McCarthy et al. (2008)). Second, the discovered genes only

account for a small proportion of genetic risk in most diseases (Kraft and Hunter (2009)). As a result, developing more sophisticated methods to better identify genetic variants associated with diseases has become a main focus of GWAS data analysis after initial scanning through single marker analysis.

Because of the genetic complexity of common diseases, a joint consideration of multiple markers is intuitively more informative when multiple genes and their interactions are involved in disease etiology. However, joint methods often lead to a sharp increase in the computational burden and in the stringency of statistical significance control that can weaken their statistical power due to the large amount of candidate models considered. An optimal marker selection strategy should achieve a delicate balance between computational efficiency, satisfactory statistical power, and low error rates. To recognize the optimal marker selection strategy for a certain GWA study, researchers look for techniques to quickly evaluate possible strategies, marginal versus joint, for a variety of interesting genetic models.

There are three fundamental marker search strategies: marginal search chooses the best fitted single-marker models separately; exhaustive search selects the best fitted multiple-marker models from all possible combinations of predictors; forward search looks for the preferred models conditional on the best fitted marker(s). Other marker search strategies are mostly extensions of these three. For example, in a marginal-exhaustive two-stage search strategy, one can first screen the marker candidates through marginal search, and then choose the best multiple-marker models within the previously selected marker set. In the literature, the power evaluations of these methods have been explored by either simulations or data analyses (Marchini, Donnelly, and Cardon (2005); Evans et al. (2006); Storey, Akey, and Kruglyak (2005); Brem et al. (2005)). Various, even conflicting, opinions exist about the performance of different model selection methods. Through limited simulation studies, Marchini and colleagues concluded that exhaustive search is more powerful. On the contrary, based on the analysis of a data set for yeast, Storey and colleagues recommended sequential forward search. They reported that exhaustive search suffers from lower power because of a substantial increase in the number of models. As real data are too specific and cannot be used in experimental design, and simulations are time-consuming and less insightful about the statistical mechanism of how genetic signals are captured, it is desirable to have analytical results. Our theoretical results demonstrate the conditions under which one method is better than the others.

The analytical power calculation methods for quantitative trait have been developed (Wu and Zhao (2009)). However, since GWAS mostly focus on binary disease outcomes, there is a need to derive results for binary traits. Statistically,

genetic models for quantitative and binary traits are quite different. The genetic model for a quantitative trait is generally a linear regression model with a random error component, whereas the genetic model for a binary trait specifies the disease risk for each possible genotype. Searching quantitative trait loci is commonly performed through fitting linear regression models, and the F-statistic is usually used to measure the model goodness-of-fit. On the other hand, searching binary trait loci is commonly performed through fitting logistic regression models, and the log-likelihood ratio test (LRT) or a score test statistic is generally applied for model comparisons. The distributions of the F-statistic are quite different from the LRT or a score statistic. So it is necessary to rigorously explore whether the marker search methods behave differently for binary traits in comparison to quantitative traits. In the results and discussion sections, we demonstrate the common and distinctive patterns of power comparison of model selection methods for binary and quantitative traits.

The application of different error control criteria affects the relative performance of different model search strategies. We investigate such effects by comparing and contrasting two types of error controls that have been widely applied in practice: discovery number control and type I error rate control. With the former, one collects a pre-specified number of models after ordering all candidate models. With the later, one selects models that have test statistics exceeding a critical value based on a genome-wide type I error rate. The analytic power calculations help us to demonstrate and explain the power comparison for each type of control.

Previous work (Wu and Zhao (2009)) assumes that markers are independent, but it is not unusual that linkage disequilibrium (LD) exists between the genotyped markers and the causative but unobserved markers. To study the influence of LD on the power of model selection methods, we take into account such LD in our method. Besides the three basic search methods, we also consider a marginal-exhaustive two-stage search strategy that is practically appealing because of its computational efficiency.

Our analytical research reveals how the magnitude of interactions, research goals to seek true model or to detect some associated markers, and the application of different error controls systematically change the power comparison of different model selection methods. These findings also explain the inconsistent conclusions in the existing literature about the relative effectiveness of model selection methods. We have implemented our analytical methods in an R package *markerSearchPower* that provides researchers with a convenient tool to find proper sample size in experimental design, decide suitable strategies in data analysis, and increase the chance of true findings. Our statistical technique can also be used to address model selection problems of binary responses with general random predictor settings in other application areas.

The rest of this article is organized as follows. Section 2 sets up the genetic models to be studied, and defines marker search strategies as well as statistical power. In Section 3, we introduce the score test statistics, derive asymptotic distributions for score tests, and develop the power calculation formulas for four model selection strategies. In Section 4, the accuracy of our analytical results is demonstrated by simulation, and the power comparisons among search strategies are illustrated in a large space of genetic parameters. In Section 5, we discuss the advantages of the analytical approaches, compare the results for quantitative traits and binary traits, and summarize the distinction between the performance under discovery number control and that under Bonferroni control.

## 2. Genetic Model and Marker Search

### 2.1. Model setup

We assume that the odds of a binary trait (or hereafter described as "disease") are specified by two loci which may or may not be directly genotyped. A genotype data set contains $n$ independent individuals indexed by $i = 1, \ldots, n$, and $L$ candidate markers indexed by $j, k = 1, \ldots, L$. Each marker has two alleles $A$ and $a$. The random variable of the genotype of the $j$th marker in the $i$th individual is

$$G_{ji} = \begin{cases} 2 & \text{Genotype} = A_jA_j, \text{ with probability } p_j^2, \\ 1 & \text{Genotype} = A_ja_j, \text{ with probability } 2p_j\left(1-p_j\right), \\ 0 & \text{Genotype} = a_ja_j, \text{ with probability } \left(1-p_j\right)^2, \end{cases}$$

where $p_j$ is the disease minor allele frequency (MAF). The most general way to specify the underlying genetic model is through a 3-by-3 table of disease odds. Without loss of generality, we assume the first two markers, indexed by $j = 1$ or 2, are the associated markers. The conditional odds of disease is

$$O\left(g_1, g_2\right) = \frac{p\left(D|g_1, g_2\right)}{p\left(\bar{D}|g_1, g_2\right)},$$

where $g_1$ and $g_2$ are the given genotype values, $D$ and $\bar{D}$ denote disease and non-disease, respectively. If SNPs 1 and 2 are causative factors in the disease, the following three tables represent three commonly studied genetic models specifying the disease odds under the combination of the genotypes (Marchini, Donnelly, and Cardon (2005)):

Model 1:

| | $A_2A_2\,(g_2=2)$ | $A_2a_2\,(g_2=1)$ | $a_2a_2\,(g_2=0)$ |
|---|---|---|---|
| $A_1A_1\,(g_1=2)$ | $\alpha\left(1+\theta_1\right)^2\left(1+\theta_2\right)^2$ | $\alpha\left(1+\theta_1\right)^2\left(1+\theta_2\right)$ | $\alpha\left(1+\theta_1\right)^2$ |
| $A_1a_1\,(g_1=1)$ | $\alpha\left(1+\theta_1\right)\left(1+\theta_2\right)^2$ | $\alpha\left(1+\theta_1\right)\left(1+\theta_2\right)$ | $\alpha\left(1+\theta_1\right)$ |
| $a_1a_1\,(g_1=0)$ | $\alpha\left(1+\theta_2\right)^2$ | $\alpha\left(1+\theta_2\right)$ | $\alpha$ |

, (2.1)

Model 2:

| | $A_2A_2\,(g_2=2)$ | $A_2a_2\,(g_2=1)$ | $a_2a_2\,(g_2=0)$ |
|---|---|---|---|
| $A_1A_1\,(g_1=2)$ | $\alpha\,(1+\theta)^4$ | $\alpha\,(1+\theta)^2$ | $\alpha$ |
| $A_1a_1\,(g_1=1)$ | $\alpha\,(1+\theta)^2$ | $\alpha\,(1+\theta)$ | $\alpha$ |
| $a_1a_1\,(g_1=0)$ | $\alpha$ | $\alpha$ | $\alpha$ |

, (2.2)

Model 3:

| | $A_2A_2\,(g_2=2)$ | $A_2a_2\,(g_2=1)$ | $a_2a_2\,(g_2=0)$ |
|---|---|---|---|
| $A_1A_1\,(g_1=2)$ | $\alpha\,(1+\theta)$ | $\alpha\,(1+\theta)$ | $\alpha$ |
| $A_1a_1\,(g_1=1)$ | $\alpha\,(1+\theta)$ | $\alpha\,(1+\theta)$ | $\alpha$ |
| $a_1a_1\,(g_1=0)$ | $\alpha$ | $\alpha$ | $\alpha$ |

, (2.3)

where, $\alpha$ and $\theta$'s are parameters for baseline and additional genotypic effects, respectively. In Model 1, the additional genetic effect of SNP $j$, $j = 1$ or 2, is $(1+\theta_j)^{g_j}$, which is multiplicative on the genotype value $g_j$. Model 2 assumes $\theta_1 = \theta_2$, and describes a thresholding rule – the genetic effect remains at the baseline $\alpha$, unless both $g_1$ and $g_2$ are non-zero. Model 3 specifies a similar threshold except there are only two levels of disease odds. Models 2 and 3 describe two typical gene-gene interactions, where the genetic effect of one marker depends on the status of the other marker.

Based on the odds of disease, the genotypic disease risks are $p\,(D|g_1,g_2) = O\,(g_1,g_2)/(1+O\,(g_1,g_2))$, and the joint distribution of genotypes in diseased individuals is

$$p\,(g_1,g_2|D) = \frac{p\,(D|g_1,g_2)\,p\,(g_1,g_2)}{\sum_{g_1,g_2} p\,(D|g_1,g_2)\,p\,(g_1,g_2)}.$$

Similarly, we can get the joint distribution of genotypes in controls $p\left(g_1,g_2|\bar{D}\right)$. For any marker-pair involving one non-associated marker $j$, it is clear that $p\,(g_1,g_j|D) = p\,(g_1|D)\,p\,(g_j)$ with $p\,(g_1|D) = \sum_{g_2} p\,(g_1,g_2|D)$. For any non-associated marker-pair, we have $p\,(g_j,g_k|D) = p\,(g_j)\,p\,(g_k)$, $k > j \geq 3$.

The odds of disease can be defined through a logistic model

$$\log\,(O\,(g_1,g_2)) = b_0 + b_1g_1 + b_2g_2 + b_3g_1g_2. \tag{2.4}$$

With various values of $b_1$, $b_2$, and $b_3$, (2.4) defines a flexible genetic interaction model. For example, the genetic model in (2.2) can be rewritten as in (2.4) with $b_0 = \log\,(\alpha)$, $b_1 = b_2 = 0$, $b_3 = \log\,(1+\theta)$.

The genetic model can also be defined through a disease prevalence $p\,(D)$ together with genetic relative risks $GRR\,(g_1,g_2) = p\,(D|g_1,g_2)\,/p\,(D|0,0)$. We

can get the corresponding odds of disease through

$$p(D|g_1,g_2) = GRR(g_1,g_2)\,p(D|0,0) = \frac{GRR(g_1,g_2)\,p(D)}{\sum_{g_1,g_2} GRR(g_1,g_2)\,p(g_1,g_2)}.$$

It is not unusual that the causative markers are not observed but they have linkage disequilibrium (LD, a genetic measure of correlation) with the genotyped markers that are not causative. In this situation, we asssume the genotyped markers 1 and 2 are non-causative but are correlated with the causative but unobserved markers indexed by, say -1 and -2, respectively. The negative index represents unobserved markers. Since the observed markers 1 and 2 are indirectly *associated* with the disease through the LD, they still are the targets to be identified through analyzing available GWAS data. The odds of disease at markers 1 and 2 is

$$O = \frac{p(D|g_1,g_2)}{p\left(\bar{D}|g_1,g_2\right)} = \frac{\sum_{g_{-1},g_{-2}} p(D|g_{-1},g_{-2})\,p(g_1|g_{-1})\,p(g_2|g_{-2})\,p(g_{-1},g_{-2})}{\sum_{g_{-1},g_{-2}} p\left(\bar{D}|g_{-1},g_{-2}\right)p(g_1|g_{-1})\,p(g_2|g_{-2})\,p(g_{-1},g_{-2})}. \tag{2.5}$$

$p(D|g_{-1},g_{-2})$ is defined by a genetic model described above. The method of calculating $p(g_1|g_{-1})$ and $p(g_2|g_{-2})$ is illustrated in the supplementary material, which follows the LD models in the literature (Marchini, Donnelly, and Cardon (2005)).

### 2.2. Model fitting and selection procedures

The above genetic models of disease susceptibility are defined by specifying the disease odds from a perspective data-generating point of view. In case-control retrospective studies, the samples are collected from a random sample of $n_1$ cases and $n_0$ controls. For a given GWAS data set, marker search requires fitting the following one-marker or two-marker logistic regression models:

$$\text{logit}\left(\hat{P}_j\left(Y_i=1|g_{ji}\right)\right) = \hat{\beta}_{0j} + \hat{\beta}_{1j}g_{ji}, \tag{2.6}$$

$$\text{logit}\left(\hat{P}_{jk}\left(Y_i=1|g_{ji},g_{ki}\right)\right) = \hat{\beta}_{0jk} + \hat{\beta}_{1jk}g_{ji} + \hat{\beta}_{2jk}g_{ki} + \hat{\beta}_{3jk}g_{ji}g_{ki}, \tag{2.7}$$

where $\text{logit}(p) = \log\left(p/\left(1-p\right)\right)$, $g_{ji}$ and $g_{ki}$ are the observed genotype values of markers $j$ and $k$ in individual $i$, and $Y_i = 1$ or 0 indicating disease or non-disease status. The marginal search method looks for the best fitted models (2.6) over all single markers. The exhaustive search method seeks the best fitted models (2.7) over all marker-pairs. The forward search method first selects the best fitted model (2.6), and then picks the best two-marker models (2.7) given

the previously chosen marker. Extended from these approaches, a marginal-exhaustive two-stage search method applies marginal search at the first stage and the exhaustive search through the chosen set of markers at the second stage. After any search procedure, the markers contained in the selected models are treated as the putative disease-associated markers.

The significance of statistical associations, or equivalently the goodness of model fitting, relies on either the log-likelihood ratio test (LRT) or the score test. As shown in the simulations below, the two tests are similar for the purpose of selecting markers when the sample size is moderately large. Because the LRT statistic has no closed form, we used the score test statistic to calculate power.

As a model can contain none, one, or two associated markers, it is necessary to consider both stringent and relaxed criteria to decide whether any chosen model is what we seek. Accordingly, the following two definitions of power in genetics literature (Marchini, Donnelly, and Cardon (2005); Storey, Akey, and Kruglyak (2005)) are considered in any selection procedure:

(A) Power is the probability of identifying the true genetic association model (in marginal search, it is the probability of detecting both associated markers, as marginal search does not consider the interaction terms).

(B) Power is the probability of detecting at least one associated marker.

Note that these power definitions for model selection strategies are different from the traditional power definition for a specific model. The power studied here measures the effectiveness of a model selection method, while the power for a specific model refers to the probability that this model is to be found significant in a hypothesis test.

We study two criteria for significance level control. The first is a discovery number control, in which one selects the top $R$ most significant models. The power under this control is a generalization of detection probability (DP) (Gail et al. (2008)) into the context of model selection. The second is a type I error rate control at a genome-wide significance level $\alpha$, which applies the Bonferroni correction according to the number of models to be compared. Specifically, in marginal search, the correction is $\alpha/L$ for comparing a total of $L$ one-marker models, and the null distribution is $\chi_1^2$. In exhaustive search, the correction is $\alpha/\binom{L}{2}$ for comparing a total of $\binom{L}{2}$ pairwise-marker models, and the null distribution is $\chi_3^2$. In the forward search, the correction of the first step is $\alpha/L$ with the null distribution $\chi_1^2$, while the correction of the second step is $\alpha/(L-1)$ with the null distribution $\chi_2^2$. In Section 3, we construct the score test statistics for logistic regression, derive the null and alternative asymptotic distributions

of relevant test statistics in marker search, and present the formulae of power calculation under significance control criteria.

## 3. Model Selection Power

### 3.1. Score test statistic

We adopt a score test (Zhang (2006)) in the context of genome-wide association case-control studies. In general, let $\mathbf{T}_i^1 = \left(T_{1i}^1, \ldots, T_{mi}^1\right)'$, $i = 1, \ldots, n_1$, be a random sample of $m$ covariates of logistic regression in cases, $\mathbf{T}_i^0 = \left(T_{1i}^0, \ldots, T_{mi}^0\right)'$, $i = 1, \ldots, n_0$, be a random sample of $m$ covariates in controls. $\mathbf{T}_i^1$ and $\mathbf{T}_i^0$ have distributions $p\left(t_1, \ldots, t_m | D\right)$ and $p\left(t_1, \ldots, t_m | \bar{D}\right)$, respectively. Let $\mathbf{T}_{m\times n}$=($\mathbf{T}_1$, $\ldots, \mathbf{T}_n$) = $\left(\mathbf{T}_1^1, \ldots, \mathbf{T}_{n_1}^1, \mathbf{T}_1^0, \ldots, \mathbf{T}_{n_0}^0\right)$ represent the combined variables with total sample size $n = n_1+n_0$. To test the null hypothesis that there is no association between the outcome and the covariates, the score statistic is

$$S = n\mathbf{U}'\mathbf{\Gamma}^{-1}\mathbf{U},$$

where $\mathbf{U}_{m\times 1} = (n_0 n_1/n^2)\left(\bar{\mathbf{T}}^1 - \bar{\mathbf{T}}^0\right)$ and $\mathbf{\Gamma}_{m\times m} = (n_0 n_1/n^2)((1/n)\sum_{i=1}^n \mathbf{T}_i\mathbf{T}_i'$ $-\bar{\mathbf{T}}\bar{\mathbf{T}}')$, with $\bar{\mathbf{T}}^1$=$\sum_{i=1}^{n_1} \mathbf{T}_i^1/n_1$, $\bar{\mathbf{T}}^0$=$\sum_{i=1}^{n_0} \mathbf{T}_i^0/n_0$, and $\bar{\mathbf{T}} = \sum_{i=1}^n \mathbf{T}_i/n$ being the vectors of sample averages.

In GWAS, let $\mathbf{Z}_j = (Z_{j1}, \ldots, Z_{jn_1})$ and $\mathbf{X}_j = (X_{j1}, \ldots, X_{jn_0})$ be samples of genotype values for the $j$th marker in cases and controls, respectively. That is, the vector of random genotype of marker $j$ is $\mathbf{G}_j = (G_{j1}, \ldots, G_{jn}) = (Z_{j1}, \ldots, Z_{jn_1}, X_{j1}, \ldots, X_{jn_0})$. In general, the elements of $\mathbf{T}_i^1$ and $\mathbf{T}_i^0$ are functions of random genotypes corresponding to the form of the logistic regression model. Particularly, for a single-marker model (2.6) of marker $j$, $j = 1, \ldots, L$, there is $m = 1$ covariate such that $\mathbf{T}_i^1 = (Z_{ji})$, $\mathbf{T}_i^0 = (X_{ji})$, and $\mathbf{T} = \mathbf{G}_j$. So the score test statistic is

$$S_j = \frac{n}{2}\left(\frac{2r\left(1-r\right)\left(\bar{Z}_j - \bar{X}_j\right)^2}{r\overline{Z_j^2} + (1-r)\,\overline{X_j^2} - \left(r\bar{Z}_j + (1-r)\,\bar{X}_j\right)^2}\right), \tag{3.1}$$

where $r = n_1/n$, $\bar{Z}_j = \sum_{i=1}^{n_1} Z_{ji}/n_1$, $\bar{X}_j = \sum_{i=1}^{n_0} X_{ji}/n_0$, $\overline{Z_j^2} = \sum_{i=1}^{n_1} Z_{ji}^2/n_1$ and $\overline{X_j^2} = \sum_{i=1}^{n_0} X_{ji}^2/n_0$. For a two-marker model (2.7) of markers $j$ and $k$, there are $m = 3$ covariate such that $\mathbf{T}_i^1 = (Z_{ji}, Z_{ki}, Z_{ji}Z_{ki})'$, $\mathbf{T}_i^0 = (X_{ji}, X_{ki}, X_{ji}X_{ki})'$, and $\mathbf{T}_{3\times n}$=$(\mathbf{G}_j, \mathbf{G}_k, \mathbf{G}_j * \mathbf{G}_k)'$, where $*$ denotes the element-wise cross-product of two vectors. Thus the score test statistic is

$$S_{jk} = n\mathbf{U}_{jk}'\left(\mathbf{\Gamma}_{jk}\right)^{-1}\mathbf{U}_{jk}, \tag{3.2}$$

where

$$\mathbf{U}_{jk} = \frac{n_1 n_0}{n}\left(\bar{Z}_j - \bar{X}_j, \bar{Z}_k - \bar{X}_k, \overline{Z_j Z_k} - \overline{X_j X_k}\right)',$$

$$\mathbf{\Gamma}_{jk} = \frac{n_1 n_0}{n} \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} \end{pmatrix}, \tag{3.3}$$

with $\mathbf{\Gamma}_{jk}$ being symmetric and

$$\begin{aligned}
\gamma_{11} &= \frac{n_1\overline{Z_j^2} + n_0\overline{X_j^2}}{n} - \left(\frac{n_1\overline{Z_j} + n_0\overline{X_j}}{n}\right)^2, \\
\gamma_{12} &= \frac{n_1\overline{Z_j Z_k} + n_0\overline{X_j X_k}}{n} - \left(\frac{n_1\overline{Z_j} + n_0\overline{X_j}}{n}\right)\left(\frac{n_1\overline{Z_k} + n_0\overline{X_k}}{n}\right), \\
\gamma_{13} &= \frac{n_1\overline{Z_j^2 Z_k} + n_0\overline{X_j^2 X_k}}{n} - \left(\frac{n_1\overline{Z_j Z_k} + n_0\overline{X_j X_k}}{n}\right)\left(\frac{n_1\overline{Z_j} + n_0\overline{X_j}}{n}\right), \\
\gamma_{22} &= \frac{n_1\overline{Z_k^2} + n_0\overline{X_k^2}}{n} - \left(\frac{n_1\overline{Z_k} + n_0\overline{X_k}}{n}\right)^2, \\
\gamma_{23} &= \frac{n_1\overline{Z_j Z_k^2} + n_0\overline{X_j X_k^2}}{n} - \left(\frac{n_1\overline{Z_j Z_k} + n_0\overline{X_j X_k}}{n}\right)\left(\frac{n_1\overline{Z_k} + n_0\overline{X_k}}{n}\right), \\
\gamma_{33} &= \frac{n_1\overline{Z_j^2 Z_k^2} + n_0\overline{X_j^2 X_k^2}}{n} - \left(\frac{n_1\overline{Z_j Z_k} + n_0\overline{X_j X_k}}{n}\right)^2.
\end{aligned}$$

Note that $\overline{Z_j Z_k} = \sum_{i=1}^{n_1} Z_{ji} Z_{ki}/n_1$ and $\overline{X_j X_k} = \sum_{i=1}^{n_0} X_{ji} X_{ki}/n_0$; the other averages functions of cross-product terms are analogously defined.

### 3.2. Asymptotic distributions

In GWAS, the marker genotypes of individuals are not controllable but are randomly observed. It is crucial to consider the genotype predictors as random variables. We apply a generalization of the Delta method to derive the null and the alternative distributions of the score test statistics that are functions of random predictors. Let $\mathbf{W}_i = (W_{1i}, \ldots, W_{mi})$, $i = 1, \ldots, n$, be $n$ independent and identically distributed random vectors of dimension $m$. The corresponding mean vector is $\theta = (\theta_1, \ldots, \theta_m)$ with $\theta_s = E(W_{si})$, and the covariance matrix is $\Sigma = Cov(\mathbf{W}_i)$ with $(\Sigma)_{st} = Cov(W_{si}, W_{ti})$, $s, t = 1, \ldots, m$. Let $\bar{\mathbf{W}} = \left(\bar{W}_1, \ldots, \bar{W}_m\right)$ be the vector of the sample means, $\bar{W}_s = (1/n)\sum_{i=1}^{n} W_{si}$. Consider a real valued function $h\left(\bar{\mathbf{W}}\right)$ of $\bar{\mathbf{W}}$. If $\nabla h(\theta) \equiv (\partial h(\theta)/\partial\theta_1, \ldots, \partial h(\theta)/\partial\theta_m)' \neq 0$,

$$\sqrt{n}\left[h\left(\bar{\mathbf{W}}\right) - h(\theta)\right] \xrightarrow{L} N\left(0, \tau^2\right), \tag{3.4}$$

where $\tau^2 = [\nabla h(\theta)]' \Sigma [\nabla h(\theta)]$ and $\xrightarrow{L}$ denotes convergence in law. If $\nabla h(\theta) = 0$,

$$n\left[h\left(\bar{\mathbf{W}}\right) - h(\theta)\right] \xrightarrow{L} c\chi_d^2. \tag{3.5}$$

Let $A \equiv D^2h(\theta)\Sigma$, with $D^2h(\theta) = \frac{\partial^2}{\partial\theta^2}h(\theta)$ be the Hessian matrix of $h(\theta)$. We then have

1. $c = 1/2$, $d = rank(A)$, if $A$ is idempotent,
2. $c \approx trace\left(A^2\right)/2trace(A)$, $d \approx trace(A)^2/trace\left(A^2\right)$, if $A$ is not idempotent.

Furthermore, if $\nabla h_1(\theta) \neq 0$ and $\nabla h_2(\theta) \neq 0$,

$$Cov\left(\sqrt{n}h_1\left(\bar{\mathbf{W}}\right), \sqrt{n}h_2\left(\bar{\mathbf{W}}\right)\right) \xrightarrow{P} \left[\nabla h_1(\theta)\right]' \Sigma \left[\nabla h_2(\theta)\right]. \tag{3.6}$$

Clearly, the score tests in (3.1) and (3.2) are functions of the genotypic sample means. The distribution of $\bar{\mathbf{W}}$ can be derived from the genotypic distributions in cases or in controls determined by genetic models. For the score tests involved in each marker search method, the following sections specify the distribution of $\bar{\mathbf{W}}$ for the given markers involved in the model fittings. When the causative loci are not genotyped but markers 1 and 2 are in LD with them, by (2.5) we obtain the conditional joint genotypic distribution $p(g_{A_1}, g_{A_2}|D)$ in cases and $p\left(g_{A_1}, g_{A_2}|\bar{D}\right)$ in controls. We then derive the distribution of relevant score test statistics based on the mean vector and covariance matrix involving the associated markers 1 and 2: $\theta_{A_1A_2} = E\left(\mathbf{W}_{A_1A_2}\right)$ and $\mathbf{\Sigma}_{A_1A_2} = Var\left(\mathbf{W}_{A_1A_2}\right)$.

### 3.3. Marginal search

#### Asymptotic Distribution of Test Statistic

The relevant tests and the corresponding distributions for marginal search are as follows. For the $j$th single marker, $j = 1, \ldots, L$, let $T_j \equiv \sqrt{S_j}$, where $S_j$ is the score test in (3.1). We can write $T_j = \sqrt{n/2}h\left(\bar{\mathbf{W}}_j\right)$ with $\bar{\mathbf{W}}_j$=$(\bar{Z}_j, \bar{X}_j, \overline{Z_j^2}, \overline{X_j^2})$ being a vector of sample averages over cases and controls. Let $\mathbf{W}_j = (Z_j, X_j, Z_j^2, X_j^2)$ represent the random genotypic vector of any observation. For an associated marker $j = 1$ (similarly for $j = 2$), $Z_1$ has the distribution $p(g_1|D) = \sum_{g_2} p(g_1, g_2|D)$, and $X_1$ has the distribution $p\left(g_1|\bar{D}\right)$. For the non-associated markers $j = 3, \ldots, L$, $Z_j$ and $X_j$ have the same distribution $p(g_j)$. The mean vectors and variance matrices are $\theta_j$=$E\left(\mathbf{W}_j\right)$ and $\mathbf{\Sigma}_j = Cov\left(\mathbf{W}_j\right)$, respectively. When $n_1 = n_0$, by (3.4),

$$T_j - \sqrt{\frac{n}{2}}h(\theta_j) \xrightarrow{L} N\left(0, \tau_j^2\right),$$

where $\tau_j^2 = \left[\nabla h(\theta_j)\right]' \mathbf{\Sigma}_j \left[\nabla h(\theta_j)\right]$.

To calculate the alternative distributions, note that $T_1$ and $T_2$ are correlated because the odds of disease is a function of both markers 1 and 2. The joint distribution of $(T_1, T_2)'$ is asymptotically multivariate normal

$$(T_1, T_2)' - \mu_{T_1,T_2} \xrightarrow{L} MVN\left(\mathbf{0}, \tau_{T_1,T_2}\right), \tag{3.7}$$

and is a part of the joint distribution in (3.8). Corresponding to the non-associated markers $j = 3, \ldots, L$, the null distribution for marginal search is $T_j \xrightarrow{L} N(0,1)$ by (3.4), or consistently, $S_j \xrightarrow{L} \chi_1^2$ by (3.5). Further, by (3.6), the correlations between $T_j$, $j = 3, \ldots, L$, and $T_1$ (or $T_2$) are asymptotically 0.

**Power under Discovery Number Control**

For the discovery number control, the power of detecting alternative model(s) in the top $R$ most significantly fitted models is the probability that an alternative model is better fitted than the $R$th (or in marginal search under power definition (A), the $(R-1)$th) best fitted null models. Specifically, when the number of discoveries is controlled by $R$, the power of marginal search under definition (A) for detecting both associated markers 1 and 2 is

$$P\left(S_1 \wedge S_2 \geq S_{(r)}\right) = \iint P\left(S_{(r)} \leq t_1^2 \wedge t_2^2\right) dG\left(t_1, t_2\right),$$

where $S_1 \wedge S_2 = \min\{S_1, S_2\}$, $r = L - 2 - R + 1$, $S_{(r)}$ is the $r$th smallest (or the $R$th largest) order statistics in the set $\{S_j,\ j = 3, \ldots, L\}$, and $G(t_1, t_2)$ is the cumulative distribution function (CDF) of $(T_1, T_2)'$ in (3.7). Let $G_1(\cdot)$ be the CDF of $\chi_1^2$. Then

$$P\left(S_{(r)} \leq x\right) = G_1^r(x) \sum_{l=0}^{L-2-r} \binom{r+l-1}{l} \left(1 - G_1(x)\right)^l.$$

To get the power of marginal search under definition (B), that either associated marker 1 or marker 2 is selected, we calculate the probability that either $S_1$ or $S_2$ is larger than the cutoff point: $P\left(S_1 \vee S_2 \geq S_{(r)}\right)$, where $S_1 \vee S_2 = \max\{S_1, S_2\}$.

**Power under Bonferroni Control**

Since the null distribution of a score statistic used in marginal search is $\chi_1^2$, the cutoff under the Bonferroni corrected type I error rate control is $c = G_1^{-1}(1 - \alpha/L)$, where $G_1^{-1}$ is the quantile function of $\chi_1^2$ and $\alpha$ is the genome-wide significance level. Under power definition (A) or (B), the probability of finding both or either associated marker is $P(S_1 \wedge S_2 \geq c)$ or $P(S_1 \vee S_2 \geq c)$, respectively.

### 3.4. Exhaustive search

**Asymptotic Distribution of Test Statistic**

The relevant test statistic distributions for exhaustive search are based on the score test statistics in (3.1) and (3.2). For the statistics involving associated

markers 1 and 2, let $T_{12} \equiv \sqrt{S_{12}} = \sqrt{n/2}h_{12}\left(\bar{\mathbf{W}}_{12}\right)$, $T_i \equiv \sqrt{S_i} = \sqrt{n/2}h_i\left(\bar{\mathbf{W}}_{12}\right)$, $i = 1, 2$, where

$$\bar{\mathbf{W}}_{12} = \begin{pmatrix} \overline{Z_1}, \overline{X_1}, \overline{Z_2}, \overline{X_2}, \overline{Z_1Z_2}, \overline{X_1X_2}, \overline{Z_1^2}, \overline{X_1^2}, \overline{Z_2^2}, \overline{X_2^2}, \\ \overline{Z_1^2Z_2}, \overline{X_1^2X_2}, \overline{Z_1Z_2^2}, \overline{X_1X_2^2}, \overline{Z_1^2Z_2^2}, \overline{X_1^2X_2^2} \end{pmatrix}$$

is a vector of sample averages over cases and controls. Let $\mathbf{W}_{12}$ be the corresponding genotypic vector of one random observation. We have the mean vector $\theta_{12}=E\left(\mathbf{W}_{12}\right)$ and the variance matrix $\mathbf{\Sigma}_{12} = Var\left(\mathbf{W}_{12}\right)$. Based on the asymptotic distribution results in (3.4) and (3.6), when $n_1 = n_0$,

$$\left(T_{12}, T_1, T_2\right)' - \mu_{T_{12},T_1,T_2} \xrightarrow{L} MVN\left(\mathbf{0}, \tau_{T_{12},T_1,T_2}\right), \tag{3.8}$$

where

$$\mu_{T_{12},T_1,T_2} = \sqrt{\frac{n}{2}}\left(h_{12}\left(\theta_{12}\right), h_1\left(\theta_{12}\right), h_2\left(\theta_{12}\right)\right)',$$
$$\tau_{T_{12},T_1,T_2} = \mathbf{D}'\mathbf{\Sigma}_{12}\mathbf{D},$$

with $\mathbf{D} = \left(\nabla h_{12}\left(\theta_{12}\right), \nabla h_1\left(\theta_{12}\right), \nabla h_2\left(\theta_{12}\right)\right)$.

Because $S_{12}$ has 3 degrees of freedom, the convergence in (3.8) is relatively slower than that in (3.7). In the following we describe an approximation for the mean of $T_{12}$, in the case that sample size is small (e.g. $n_1 < 1,000$) and genetic effect is weak (e.g. $\theta < 0.2$ in model (2.2)). Note that if we consider an observed (and thus fixed) data design matrix $\mathbf{t} = (\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_1 * \mathbf{g}_2)$ of a logistic regression in the form (2.7), where $\mathbf{g}_j = (z_{j1}, \ldots, z_{n_1 j}, x_{j1}, \ldots, x_{jn_0})'$, $j = 1$ or 2, $*$ represents pair-wise product, it has been shown that (Zhang (2006))

$$S_{12} \sim \chi^2_{3,\delta_n(\mathbf{t})}$$

where $\delta_n(\mathbf{t}) = n\mathbf{b}'\mathbf{\Gamma}_{12}\left(\mathbf{t}\right)\mathbf{b}$, with $\mathbf{\Gamma}_{12}\left(\mathbf{t}\right)$ given in the form (3.3) and $\mathbf{b} = (b_1, b_2, b_3)'$ being the vector of coefficients in (2.4). Now we define $\delta_n = \delta_n\left(E\left(\mathbf{T}\right)\right)$ in our set-up for random genotype $\mathbf{T} = (\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_1 * \mathbf{G}_2)$ with $\mathbf{G}_j = (Z_{j1}, \ldots, Z_{jn_1}, X_{j1}, \ldots, X_{jn_0})'$, $j = 1$ or 2. We can use a weighted Chi-square with one degree of freedom to approximate $S_{12}$, i.e. $S_{12} \cong a_n\chi^2_{1,\lambda_n}$. Solving the equations assuming equal mean and variance (Scheffé (1959)) $a_n\left(1 + \lambda_n\right) = 3 + \delta_n$ and $a_n^2\left(2 + 4\lambda_n\right) = 6 + 4\delta_n$, we get $a_n = (3 + \delta_n)/(1 + \lambda_n)$, $\lambda_n = 2t_n - 1 + \sqrt{4t_n^2 - 2t_n}$, and $t_n = (3 + \delta_n)^2/(6 + 4\delta_n)$. So we can apply the approximation for weak genetic effects

$$E\left(T_{12}\right) \cong \sqrt{a_n\lambda_n}.$$

By (3.5), the score test statistic $S_{jk}$ for model (2.7) of two non-associated markers $j$ and $k$, $3 \leq j < k \leq L$, has an asymptotic distribution

$$S_{jk} \xrightarrow{L} \chi^2_3. \tag{3.9}$$

Let $S_{k|j}$ denote the score test statistic for the extra terms in model (2.7) over model (2.6). The following is a useful decomposition

$$S_{jk} = S_j + S_{k|j}. \tag{3.10}$$

So the correlation between the score test statistics $S_{jk_1}$ and $S_{jk_2}$, sharing the same marker $j$, can be captured by $S_j$, while $S_{k_1|j}$ and $S_{k_2|j}$ can be treated independently. Furthermore, by (3.5), for $k \geq 3$,

$$S_{k|j} \xrightarrow{L} \chi^2_2. \tag{3.11}$$

**Power under Discovery Number Control**

With test statistic distributions derived, we can calculate the probability of identifying the whole associated genetic model through exhaustive search under power definition (A). Let $A_1 \equiv \{S_{ij},\, i = 1, 2,\, j = 3, \ldots, L\}$ and $A_2 \equiv \{S_{jk},\, 3 \leq j < k \leq L\}$. Let $S_{A,[R]}$ denote the $R$th largest score test statistics in a set $A$. When controlling the false discovery number by $R$, the probability of detecting the associated marker-pair is

$$P\left(S_{12} \geq S_{A_1 \cup A_2,[R]}\right) = \iiint P\left(t^2_{12} \geq S_{A'_1 \cup A_2,[R]}\right) dG\left(t_{12}, t_1, t_2\right),$$

where $G\left(t_{12}, t_1, t_2\right)$ is the CDF of (3.8), $A'_1 = \left\{t^2_i + S_{j|i},\, i = 1, 2,\, j = 3, \ldots, L\right\}$ is from the decomposition (3.10), and

$$P\left(t^2_{12} \geq S_{A'_1 \cup A_2,[R]}\right) = \sum_{r=0}^{R-1} \sum_{\{r_1, r_2, r_3\} \in S_r} P_1 P_2 P_3,$$

where

$$\begin{aligned}
S_r &= \left\{\{r_1, r_2, r_3\} : \sum r_i = r, 0 \leq r_1, r_2 \leq (L-2), 0 \leq r_3 \leq N\right\}, \\
P_1 &= \binom{L-2}{r_1} \left[1 - G_1\left(t^2_{12} - t^2_1\right)\right]^{r_1} G_1\left(t^2_{12} - t^2_1\right)^{L-2-r_1}, \\
P_2 &= \binom{L-2}{r_2} \left[1 - G_1\left(t^2_{12} - t^2_2\right)\right]^{r_2} G_1\left(t^2_{12} - t^2_2\right)^{L-2-r_2}, \\
P_3 &= \binom{N}{r_3} \left[1 - G_3\left(t^2_{12}\right)\right]^{r_3} G_3\left(t^2_{12}\right)^{N-r_3},
\end{aligned}$$

$N = \binom{L-2}{2}$ is the number of variables in $S_2$, $G_1(\cdot)$ is the CDF of (3.11), and $G_3(\cdot)$ is the CDF of (3.9). With the same argument given in the literature (Wu and Zhao (2009)), the test statistics within the sets $A^* \equiv \left\{S_{j|1}, S_{j|2},\ j = 3, \ldots, L\right\}$

and $S_2 \equiv \{S_{jk}, 2 < j < k \leq L\}$ can be treated as asymptotically independent as $L \to \infty$.

To simplify the heavy computation needed above, we can use the following approximations of the integrands. Simulations (results not presented for space limit) illustrated that these approximations are fairly accurate in the context of Monte Carlo integration. Let $m = 2(L-2) + N$ be the total number of the elements in $A_1' \cup A_2$. $Q$ denotes the quantile function of mixed distribution of these elements. For a given $(t_{12}, t_1, t_2)$, $P\left(t_{12}^2 \geq S_{A_1' \cup A_2, [R]}\right)$ can be approximately replaced with

$$I\left\{t_{12}^2 > Q\left(\frac{m - R + 0.5}{N}\right)\right\} \cong I\left\{t_{12}^2 > R\text{th largest value in } A_3\right\},$$

where $I\{E\}$ denotes the indicator function of event $E$, and the set

$$A_3 = \left\{Q_1(r) + t_1^2, \quad Q_1(r) + t_2^2, \quad Q_3(r), \quad r = 1, \ldots, R\right\},$$

with $Q_1(r) = G_1^{-1}(1 - (r - 0.5)/(L-2))$ and $Q_3(r) = G_3^{-1}((N - r + 0.5)/N)$.

According to power definition (B), the probability for exhaustive search to detect either associated marker is

$$P\left(\max\left(\{S_{12}\} \cup A_1\right) > S_{A_2,[R]}\right) = 1 - \iiint P_{t_{12},t_1,t_2} dG(t_{12}, t_1, t_2),$$

where

$$\begin{aligned} P_{t_{12},t_1,t_2} &= P\left(\max\left(\left\{t_{12}^2\right\} \cup A_1'\right) \leq S_{A_2,[R]}\right) \\ &= \int P\left(\max\left(\left\{t_{12}^2\right\} \cup A_1'\right) \leq x\right) g_{3_{(N-R+1)}}(x)\, dx \\ &= \int_{t_{12}^2}^{\infty} \left[G_1\left(x - t_1^2\right) G_1\left(x - t_2^2\right)\right]^{L-2} g_{3_{(N-R+1)}}(x)\, dx, \end{aligned}$$

$g_{3_{(N-R+1)}}(\cdot)$ is the PDF of the $(N-R+1)$th order statistics distribution with the density function

$$g_{3_{(N-R+1)}}(x) = \frac{N!}{(N-R)!\,(R-1)!} G_3(x)^{N-R} \left[1 - G_3(x)\right]^{R-1} g_3(x),$$

$G_3(c)$ and $g_3(\cdot)$ are the CDF and PDF of (3.9), respectively.

If $R/N \to c$, $0 < c < 1$, as $N \to \infty$, we can use quantiles to replace the order statistics in order to simplify the calculation (David and Nagaraja (2003, Chap. 4.6)), i.e. $S_{A_2,[R]} \to G_3^{-1}((N - R + 0.5)/N) \equiv Q$. So for given $(t_{12}, t_1, t_2)$, we can approximate the integrand $P_{t_{12},t_1,t_2}$ with

$$I\left\{t_{12}^2 \leq Q\right\} \left[G_1\left(Q - t_1^2\right) G_1\left(Q - t_2^2\right)\right]^{L-2}.$$

**Power under Bonferroni Control**

Traditional type I error control for exhaustive search does not consider the models with one associated and one non-associated markers, so the null distribution is from non-associated two-marker models (Marchini, Donnelly, and Cardon (2005); Storey, Akey, and Kruglyak (2005)). Let $G_3^{-1}$ be the quantile functions of $\chi_3^2$ in (3.9). The cutoff is $c = G_3^{-1}\left(1 - \alpha/\binom{L}{2}\right)$. The probability of finding the whole associated genetic model under power definition (A) is

$$P\left(S_{12} \geq c\right) = \int P\left(t_{12}^2 \geq c\right) dG\left(t_{12}\right).$$

With a similar argument for the power under discovery number control, the probability of finding either associated marker under power definition (B) is

$$P\left(\max\left(\{S_{12}\} \cup A_1\right) \geq c\right) = 1 - \iiint P_{t_{12},t_1,t_2}\left(c\right) dG\left(t_{12}, t_1, t_2\right)$$

where

$$\begin{aligned} P_{t_{12},t_1,t_2}\left(c\right) &= P\left(\max\left(\left\{t_{12}^2\right\} \cup A_1'\right) \leq c\right) \\ &= I\left\{t_{12}^2 \leq c\right\}\left[G_1\left(c - t_1^2\right) G_1\left(c - t_2^2\right)\right]^{L-2}, \end{aligned}$$

with $G_1\left(\cdot\right)$ being the CDF of distribution (3.11).

### 3.5. Forward search

**Asymptotic Distribution of Test Statistic**

For forward search, first we derive the distributions of test statistics that are used to calculate the power of this search procedure. For the score tests involving the associated markers 1 and 2, let $T_{i|j} \equiv \sqrt{S_{i|j}}$, where $S_{i|j}$ follows (3.10), $i = 1, 2$, $j = 3, \ldots, L$. We can rewrite $T_{i|j} = \sqrt{n/2} h_{i|j}\left(\bar{\mathbf{W}}_{12j}\right)$, $T_i = \sqrt{n/2} h_i\left(\bar{\mathbf{W}}_{12j}\right)$, where

$$\bar{\mathbf{W}}_{12j} = \begin{pmatrix} \overline{Z_1}, \overline{X_1}, \overline{Z_2}, \overline{X_2}, \overline{Z_1 Z_2}, \overline{X_1 X_2}, \overline{Z_1^2}, \overline{X_1^2}, \overline{Z_2^2}, \overline{X_2^2}, \overline{Z_1^2 Z_2}, \overline{X_1^2 X_2}, \\ \overline{Z_1 Z_2^2}, \overline{X_1 X_2^2}, \overline{Z_1^2 Z_2^2}, \overline{X_1^2 X_2^2}, \overline{Z_j}, \overline{X_j}, \overline{Z_1 Z_j}, \overline{X_1 X_j}, \overline{Z_j^2}, \overline{X_j^2}, \overline{Z_1^2 Z_j}, \\ \overline{X_1^2 X_j}, \overline{Z_1 Z_j^2}, \overline{X_1 X_j^2}, \overline{Z_1^2 Z_j^2}, \overline{X_1^2 X_j^2}, \overline{Z_2 Z_j}, \overline{X_2 X_j}, \overline{Z_2^2 Z_j}, \overline{X_2^2 X_j}, \\ \overline{Z_2 Z_j^2}, \overline{X_2 X_j^2}, \overline{Z_2^2 Z_j^2}, \overline{X_2^2 X_j^2} \end{pmatrix}$$

is a vector of sample averages over cases and controls. Let $\mathbf{W}_{12j}$ be the corresponding random genotypic vector of any observation. The mean vector is $\theta_{12j}{=}E\left(\mathbf{W}_{12j}\right)$ and the variance matrix is $\mathbf{\Sigma}_{12j} = Var\left(\mathbf{W}_{12j}\right)$. Following (3.4) and (3.6), we have the asymptotic joint distribution

$$\left(T_1, T_2, T_{1|j}, T_{2|j}\right)' - \mu_{T_1,T_2,T_{1|j},T_{2|j}} \xrightarrow{L} MVN\left(\mathbf{0}, \tau_{T_1,T_2,T_{1|j},T_{2|j}}\right), \tag{3.12}$$

where

$$\mu_{T_1,T_2,T_{1|j},T_{2|j}} = \sqrt{n/2}\left(h_1\left(\theta_{12j}\right), h_2\left(\theta_{12j}\right), h_{1|j}\left(\theta_{12j}\right), h_{2|j}\left(\theta_{12j}\right)\right)',$$
$$\tau_{T_1,T_2,T_{1|j},T_{2|j}} = \mathbf{D}'\boldsymbol{\Sigma}_{12j}\mathbf{D},$$

with $\mathbf{D} = \left(\nabla h_1\left(\theta_{12j}\right), \nabla h_2\left(\theta_{12j}\right), \nabla h_{1|j}\left(\theta_{12j}\right), \nabla h_{2|j}\left(\theta_{12j}\right)\right)$. Through calculation (Wolfram (1999)), we have $\nabla h_{i|j}\left(\theta_{12j}\right) = \nabla h_i\left(\theta_{12j}\right)$, $i = 1, 2$. By (3.4) and (3.6) it is clear that

$$\begin{aligned} Var\left(T_i\right) &= \left(\nabla h_i\left(\theta_{12j}\right)\right)' \boldsymbol{\Sigma}_{12j} \nabla h_i\left(\theta_{12j}\right) \\ &= \left(\nabla h_i\left(\theta_{12j}\right)\right)' \boldsymbol{\Sigma}_{12j} \nabla h_{i|j}\left(\theta_{12j}\right) = Cov\left(T_i, T_{i|j}\right). \end{aligned}$$

So $T_i$ and $T_{i|j}$ have correlation coefficient converging to 1. This explains why forward selection has similar power as marginal search for detecting either associated marker: if a genetic effect cannot stand out in a marginal scan, it does not likely show a strong signal in the following step either. Furthermore, $T_j$ and $T_{i|j}$ are asymptotically independent

$$Cov\left(T_j, T_{i|j}\right) \to 0,\ i = 1, 2,\ j = 3, \ldots, L.$$

When comparing a model involving two incorrect markers $j$ and $k$ ($3 \le j < k \le L$) in (2.7) with a model for marker $j$ in (2.6), by (3.5), the corresponding score test statistic $S_{k|j}$ has the asymptotic chi-square distribution:

$$S_{k|j} \xrightarrow{L} \chi^2_2. \tag{3.13}$$

**Power under Discovery Number Control**

In the forward search procedure, we first apply marginal search to find the most significant marker among models in (2.6). Based on the selected marker, we then fit models in (2.7) in the second step to find the markers that have strong joint association. When controlling for $R$ total discoveries, under power definition (A) for finding the whole associated model, we need to calculate the probability that the forward search chooses marker 1 or 2 in the first step, and then picks the genetic model in the second step. Define $i^* \equiv \arg\max_{i=1,2}\{S_i\}$, $A_{i^*} \equiv \{S_{i^*3}, \ldots, S_{i^*L}\}$, as $L \to \infty$. The power can be written as

$$\begin{aligned} &P\left(S_{i^*} \ge S_{(L-2)} \ \cap\ S_{12} > S_{A_{i^*},[R]}\right) \\ &= \iiint P\left(t^2_{i^*} > S_{(L-2)} \cap t^2_{12} > S_{A'_{i^*},[R]}\right) dG\left(t_{12}, t_1, t_2\right) \\ &\to \iiint P\left(t^2_{i^*} > S_{(L-2)}\right) P\left(\ t^2_{12} \ge S_{A'_{i^*},[R]}\right) dG\left(t_{12}, t_1, t_2\right), \end{aligned}$$

where $G(t_{12}, t_1, t_2)$ is the CDF of $(T_{12}, T_1, T_2)'$ given in (3.8), $S_{(L-2)} = \max_{j\geq 3}\{S_j\}$, $A'_{i^*} = \left\{t_{i^*}^2 + S_{j|i^*},\ j = 3, \ldots, L\right\}$ by the decomposition (3.10), and

$$P\left(t_{i^*}^2 > S_{(L-2)}\right) = \left(G_1\left(t_1^2 \vee t_2^2\right)\right)^{L-2},$$
$$P\left(\ t_{12}^2 \geq S_{A'_{i^*},[R]}\right) = G_2(u)^r \sum_{l=0}^{L-2-r} \binom{r+l-1}{l} \left[1 - G_2(u)\right]^l,$$

where $u = t_{12}^2 - t_{i^*}^2$, $r = L-2-R+1$, $G_1(\cdot)$ is the CDF of $\chi_1^2$ for the distribution of $S_j$, and $G_2(\cdot)$ is the CDF of $\chi_2^2$ for the distribution of $S_{j|i^*}$. $i^*$ is fixed for an observed value $(t_1, t_2)'$ of the random vector $(T_1, T_2)'$, so it is easy to implement the power calculation with Monte Carlo integration.

Note that $S_{(L-2)}$ and $S_{A'_{i^*},[R]}$ are asymptotically independent. This is because $corr\left(S_j, S_{j|i^*}\right) < 1$ for each $j \geq 3$. So as $L \to \infty$,

$$P\left(j^* \neq k^* : S_{j^*} = S_{(L-2)},\ S_{k^*|i^*} = S_{A'_{i^*},[R]}\right) \to 1.$$

When $j^* \neq k^*$, $S_{j^*}$ and $S_{k^*|i^*}$ are always independent.

When $R$ and $L$ are large, we can simplify the formula for $P(t_{12}^2 \geq S_{A'_{i^*},[R]})$ by approximating the $R$th largest variable in $\{S_{j|i^*}, j = 3, \ldots, L\}$ with $G_2^{-1}(1 - (R-0.5)/(L-2))$, where $G_2^{-1}$ is the quantile function of $S_{j|i^*}$. So, we can approximately replace $P(t_{12}^2 \geq S_{A'_{i^*},[R]})$ with $I\{u > G_2^{-1}(1-(R-0.5)/(L-2))\}$ to calculate the integration.

Under power definition (B) for finding either associated marker, the power of the forward search is the sum of $P_A$: the probability of detecting marker 1 or 2 in the 1st step, and $P_B$: the probability that step 1 fails but step 2 picks up at least one associated marker. When controlling for $R$ total discovered models, it is straightforward that

$$P_A = P\left(S_{i^*} > S_{(L-2)}\right) = \iint \left(G_1\left(t_1^2 \vee t_2^2\right)\right)^{L-2} dG(t_1, t_2),$$

where $G(t_1, t_2)$ is the CDF of the joint distribution of $(T_1, T_2)'$ given in (3.7). Define $j^* \equiv \arg\max_{k\geq 3}\{S_k\}$, $A_{j^*} \equiv \left\{S_{k|j^*}, k \geq 3, k \neq j^*\right\}$. The second probability is

$$P_B = P\left((S_1 \vee S_2) < S_{j^*}\ \cap\ \left(S_{1|j^*} \vee S_{2|j^*}\right) \geq S_{A_{j^*},[R]}\right).$$

For any $k \geq 3$, $S_{i|k}$ and $S_k$ are independent, so $S_{i|j^*}$ and $S_{j^*}$ are independent. By the results in (3.12) and (3.13), the distribution of $S_{i|j^*}$ does not depend on $j^*$. Hence, $S_{i|j^*}$ has the same distribution of $S_{i|j}$, $j = 3, \ldots, L$. Then

$$P_B = \oint P_{t_1 t_2} P_{t_{1|j} t_{2|j}} dG\left(t_1, t_2, t_{1|j}, t_{2|j}\right),$$

where $G\left(t_1, t_2, t_{1|j}, t_{2|j}\right)$ is the CDF of $\left(T_1, T_2, T_{1|j}, T_{2|j}\right)'$ given in (3.12), and

$$\begin{aligned} P_{t_1 t_2} &= P\left(S_{(L-2)} > \left(t_1^2 \vee t_2^2\right)\right) = 1 - \left(G_1\left(t_1^2 \vee t_2^2\right)\right)^{L-2}, \\ P_{t_{1|j} t_{2|j}} &= P\left(\left(t_{1|j}^2 \vee t_{2|j}^2\right) \geq S_{A_{j^*},[R]}\right) \\ &= G_2\left(t_{1|j}^2 \vee t_{2|j}^2\right)^r \sum_{l=0}^{L-3-r} \binom{r+l-1}{l} \left[1 - G_2\left(t_{1|j}^2 \vee t_{2|j}^2\right)\right]^l, \end{aligned}$$

with $r = L - 3 - R + 1$, $G_2(\cdot)$ is the CDF of $S_{k|j}$, $k \geq 4$, given in (3.13). We can approximate $S_{A_{j^*},[R]}$ through the quantile function $G_2^{-1}\left(1 - (R - 0.5)/(L-3)\right)$ to simplify the calculation.

**Power under Bonferroni Control**

When we utilize the Bonferroni control in forward search, the first step selects the most significant single marker only if the test is larger than the cutoff $c_1 = G_1^{-1}(1 - \alpha/L)$, where $G_1^{-1}$ is the quantile function of $\chi_1^2$. In the second step, the null distribution is always $\chi_3^2$ no matter which marker is first selected. Let the cutoff be $c_2 = G_2^{-1}\left(1 - \alpha/\left(L-1\right)\right)$, where $G_2^{-1}$ is the quantile function of $\chi_2^2$. For finding the true genetic model under power definition (A), the analytical power calculation is

$$\begin{aligned} &P\left(S_{i^*} > S_{(L-2)} \cap S_{i^*} > c_1 \cap S_{12} - S_{i^*} > c_2\right) \\ &= \iiint \left(G_1\left(t_1^2 \vee t_2^2\right)\right)^{L-2} \left\{\left(t_1^2 \vee t_2^2\right) > c_1\right\} \left\{t_{12}^2 - \left(t_1^2 \vee t_2^2\right) > c_2\right\} dG\left(t_{12}, t_1, t_2\right). \end{aligned}$$

As in the calculation under discovery number control, the power of forward search for finding either associated marker is $P_A + P_B$, where

$$\begin{aligned} P_A &= P\left(S_{i^*} > S_{(L-2)} \cap S_{i^*} > c_1\right) \\ &= \iint \left(G_1\left(t_1^2 \vee t_2^2\right)\right)^{L-2} \left\{\left(t_1^2 \vee t_2^2\right) > c_1\right\} dG\left(t_1, t_2\right), \\ P_B &= P\left(S_{j^*} > S_{i^*} \cap S_{j^*} > c_1 \cap \left(S_{1|j^*} \vee S_{2|j^*}\right) \geq c_2\right) \\ &= \oint P_{t_1 t_2 t_{1|j} t_{2|j}} dG\left(t_1, t_2, t_{1|j}, t_{2|j}\right), \end{aligned}$$

and

$$P_{t_1 t_2 t_{1|j} t_{2|j}} = P\left(S_{j^*} > \left(t_1^2 \vee t_2^2\right) \cap \left(t_{1|j}^2 \vee t_{2|j}^2\right) \geq c_2 \cap S_{j^*} > c_1\right).$$

As shown above, $T_i$ and $T_{i|j}$ have large correlation coefficient converging to 1. Furthermore, $S_{j^*} \cong G_1^{-1}\left(1 - (1 - 0.5)/(L-2)\right) < c_2$ , so

$$P\left(S_{j^*} > \left(t_1^2 \vee t_2^2\right) \cap \left(t_{1|j}^2 \vee t_{2|j}^2\right) \geq c_2\right) \cong 0,$$

and thus $P_B \cong 0$.

### 3.6. Marginal-exhaustive two-stage search

We also study the power of a marginal-exhaustive two-stage method (Marchini, Donnelly, and Cardon (2005)). In the first stage for screening, it selects a set of single markers $I_1 \in \{1, 2, \ldots, L\}$ with a liberal type I error-rate cutoff $c_1 = G_1^{-1}(1-\alpha_1)$, where $G_1^{-1}$ is the quantile function of $\chi_1^2$. Then, in the second stage, it applies exhaustive search to the selected markers set. In the context of score test statistics, we adopt the approach of Marchini, Donnelly, and Cardon (2005) to define the statistic in the second stage. Specifically, the statistic is $S'_{lm} = S_{lm} - 2c_1$, where $S_{lm}$ is the score test statistic for the marker-pair $(l, m) \subset I_1$. We select markers $l$ and $m$ if $S'_{lm} \geq c_2 \equiv G_3^{-1}\left(1 - \alpha / \binom{\alpha_1 L}{2}\right)$, where $G_3^{-1}$ is the quantile function of $\chi_3^2$. To find the associated genetic model, the first stage has to find both associated markers marginally. So under definition (A), the power of the two-stage search is

$$\begin{aligned} &P\left(S_1 \wedge S_2 \geq c_1 \ \cap \ S'_{12} > c_2\right) \\ &\quad = \iiint I\left(\left(t_1^2 \wedge t_2^2\right) > c_1 \cap \left(t_{12}^2 - 2c_1\right) > c_2\right) dG\left(t_{12}, t_1, t_2\right). \end{aligned}$$

## 4. Results

### 4.1. Comparison between analytical and simulation results

In order to demonstrate the accuracy of our analytical power calculation, we compared the power values from calculations with those from simulations. For the feasibility of simulation, we considered $L = 300$ candidate markers with minor allele frequency (MAF) $p_j = 0.3$, $j = 1, \ldots, L$, $n_1 = 1,000$ cases, $n_0 = 1,000$ controls, the baseline effect $\alpha = 0.007$ and the genotypic effect $\theta = 0.3$ for genetic model in (2.2). These set-ups lead to a disease prevalence close to 0.01. Table 1 shows the power of marker search procedures under discovery number control. Table 2 shows the power under the Bonferroni corrected type I error rate control with similar set-ups as in Table 1, except that sample size $n_1 = n_0 = 5,000$ and genotypic effect $\theta = 0.2$ in (2.2). These parameters were chosen to get the power values that are in a spectrum of values of practical interests. We simulated $1,000$ data sets and ran search procedures for each. The empirical power is the proportion of successful detections. In simulations, we used both the score and the log likelihood ratio tests for model comparisons. With various parameter set-ups, more comparisons between the simulated and the calculated power values can be found in the supplementary material. The consistent closeness between the analytical and simulation results demonstrates

Table 1. Under the control of the discovery number $R$, the comparisons of the power values between simulations (based on the score test and the log-likelihood ratio test) and analytical calcualtions (based on the score test). Power definitions (A) and (B) are considered. $n_1 = n_0 = 1,000$, $\alpha = 0.007$, $\theta = 0.3$. (*$R = 2$ in marginal search under power definition (A)).

| **Strategy** | **Source** | $R = \mathbf{1}^*$ | $R = \mathbf{5}$ | $R = \mathbf{10}$ | $R = \mathbf{15}$ | $R = \mathbf{20}$ | $R = \mathbf{30}$ |
|---|---|---|---|---|---|---|---|
| *Power definition* (A) | | | | | | | |
| Marginal search | Score Simu. | 0.14 | 0.38 | 0.51 | 0.58 | 0.64 | 0.71 |
| | LRT Simu. | 0.14 | 0.38 | 0.51 | 0.58 | 0.64 | 0.71 |
| | Score Calcu. | 0.15 | 0.38 | 0.50 | 0.58 | 0.65 | 0.72 |
| Exhaustive search | Score Simu. | 0.32 | 0.50 | 0.57 | 0.61 | 0.65 | 0.70 |
| | LRT Simu. | 0.33 | 0.52 | 0.61 | 0.64 | 0.69 | 0.72 |
| | Score Calcu. | 0.32 | 0.50 | 0.59 | 0.63 | 0.65 | 0.69 |
| Forward search | Score Simu. | 0.28 | 0.42 | 0.48 | 0.50 | 0.52 | 0.55 |
| | LRT Simu. | 0.29 | 0.43 | 0.48 | 0.50 | 0.53 | 0.55 |
| | Score Calcu. | 0.28 | 0.44 | 0.48 | 0.51 | 0.52 | 0.54 |
| *Power definition* (B) | | | | | | | |
| Marginal search | Score Simu. | 0.55 | 0.83 | 0.90 | 0.92 | 0.95 | 0.96 |
| | LRT Simu. | 0.55 | 0.83 | 0.90 | 0.92 | 0.95 | 0.96 |
| | Score Calcu. | 0.55 | 0.83 | 0.91 | 0.94 | 0.95 | 0.98 |
| Exhaustive search | Score Simu. | 0.57 | 0.80 | 0.87 | 0.91 | 0.93 | 0.95 |
| | LRT Simu. | 0.59 | 0.81 | 0.87 | 0.81 | 0.93 | 0.95 |
| | Score Calcu. | 0.62 | 0.77 | 0.84 | 0.88 | 0.90 | 0.93 |
| Forward search | Score Simu. | 0.64 | 0.76 | 0.83 | 0.87 | 0.91 | 0.94 |
| | LRT Simu. | 0.64 | 0.76 | 0.83 | 0.87 | 0.91 | 0.94 |
| | Score Calcu. | 0.64 | 0.74 | 0.82 | 0.88 | 0.90 | 0.94 |

that our power calculation methods perform well, and that the score test and the LRT have similar performance for model selection.

### 4.2. Power comparisons of marker search methods

We applied the analytical power calculations to compare different marker search methods in a hypothetical GWAS that contains $n_1 = 1,000$ cases, $n_0 = 1,000$ controls, and $L = 300,000$ candidate markers with minor allele frequency $p_j = 0.3$, $j = 1, \ldots, L$. Assume the true genetic model is a logistic model of form (2.4) with the baseline intercept $b_0 = \log(0.007)$. Let the main effect $b_1 = b_2$ and the interaction effect $b_3$ both vary from -1 to 1 by a step size of 0.1. Figures 1 and 2 show the 3-D plots of statistical power over a set of main and interaction effects under the discovery number control $R = 20$ and the Bonferroni corrected type I error rate $\alpha = 0.05$, respectively. As demonstrated by the pink "trenches" in Figures 1 and 2, marginal search and forward search will unavoidably fail when a disease susceptibility is controlled by interactions that show no marginally detectable signal. Exhaustive search, on the other hand,

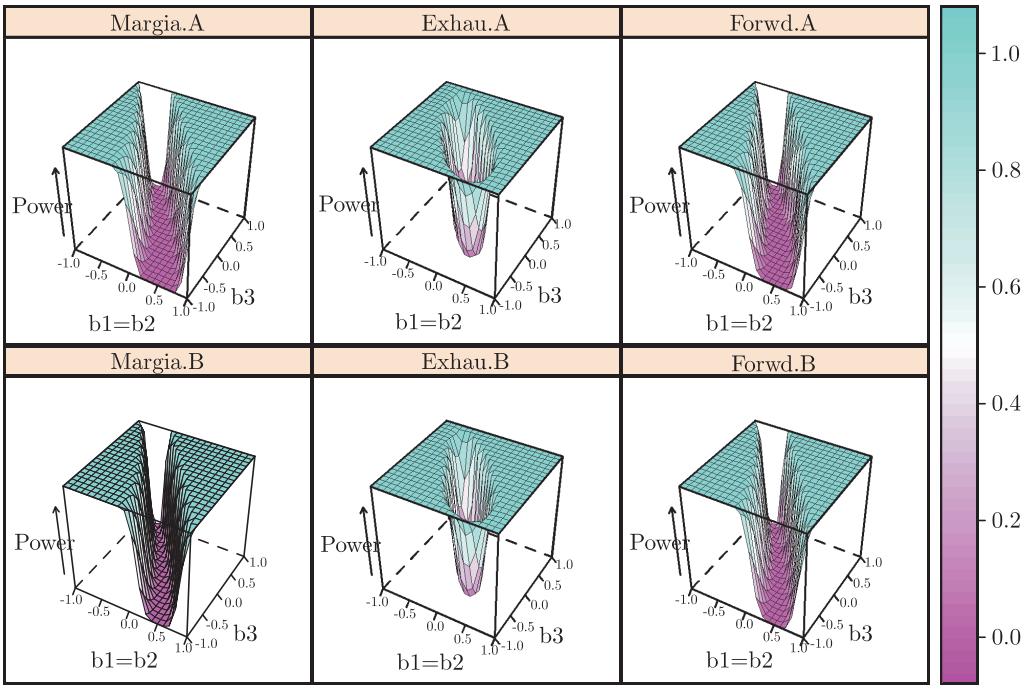


Figure 1. 3-D plots of statistical power under discovery number control over genetic effect space. Model selection methods: marginal search in the left column, exhaustive search in the middle column, and forward search in the right column. Two definitions of power: (A) detecting the joint association (or both associated markers in marginal search) in row 1, and (B) detecting either associated marker in row 2. The genetic models are logistic with main effect $b_1 = b_2$ and epistatic effect $b_3$, both varying from -1 to 1. The MAF $p_j = 0.3$, $j = 1, \ldots, L$. The total discovery number $R$ is set to be 20.

can avoid this problem and detect the full signal in two dimensions as long as the effect size is large enough. Researchers can benefit from exhaustive search to discover new genes that are missed by marginal search and forward search, as these genes are associated with the diseases only through gene-gene interactions.

In order to contrast one marker search method with another, we subtracted the power values of one method from those of another method. The differences between the power values of two methods are plotted in Figures 3 and 4. Shown as the left column of Figure 3, the marginal search is more effective than the exhaustive search to find both associated markers in areas where the interaction is weak with small $b_3$, but the main effects are modest with moderate $b_1$ and $b_2$. The disadvantage of exhaustive search in considering more false models overwhelms its advantage when the interaction effect is not large enough. This

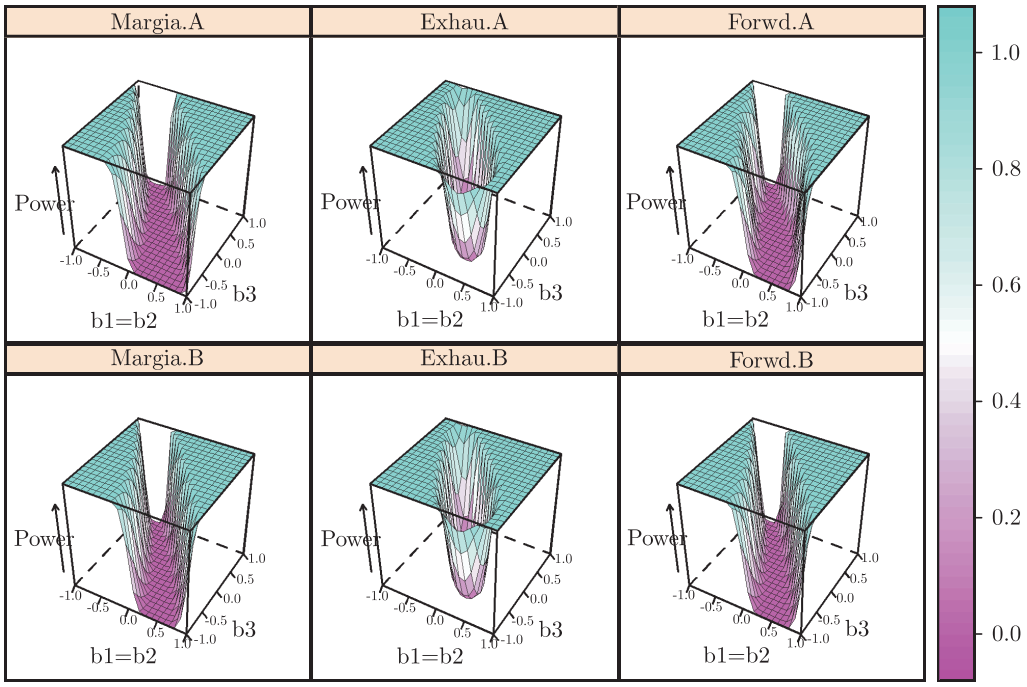


Figure 2. 3-D plots of statistical power under Bonferroni control over genetic effect space. Model selection methods: marginal search in the left column, exhaustive search in the middle column and forward search in the right column. Two definitions of power: (A) detecting the joint association (or both associated markers in marginal search) in row 1, and (B) detecting either associated marker in row 2. The genetic models are logistic with main effect $b_1 = b_2$ and epistatic effect $b_3$, both varying from -1 to 1. The MAF $p_j = 0.3$, $j = 1, \ldots, L$. The genome-wide significance level $\alpha$ is set to be 0.05.

superiority of marginal search over exhaustive search is even enhanced when only one marker is required to be found, which is shown by the bigger superior areas of marginal search in the lower left panel. Based on the middle column, forward search is more powerful in finding both markers than marginal search by modeling interactions that are not close to zero. However, forward search is uniformly beaten by marginal search in finding at least one marker through the whole genetic model space. This is because its restricted first step (selecting only one marker) constrains the probability to find a correct marker at this step while, in the second step, a wrongly chosen marker highly reduces the probability of finding a correct marker. From the right column, we can see that exhaustive search is always better than or similar to forward search in finding true epistatic models; it performs worse than the forward search in finding at least one marker by con-

Table 2. Under the Bonferroni corrected type I error with family-wise significance level $\alpha$, the comparisons of the power values between simulations (based on the score test and the log-likelihood ratio test) and analytical calcualtions (based on the score test). Power definitions (A) and (B) are considered. $n_1 = n_0 = 5,000$, $\alpha = 0.007$, $\theta = 0.2$.

| **Strategy** | **Source** | $\alpha = \mathbf{0.01}$ | $\alpha = \mathbf{0.05}$ | $\alpha = \mathbf{0.10}$ | $\alpha = \mathbf{0.15}$ |
|---|---|---|---|---|---|
| *Power definition* (A) | | | | | |
| Marginal search | Score Simu. | 0.20 | 0.37 | 0.46 | 0.51 |
| | LRT Simu. | 0.20 | 0.37 | 0.46 | 0.51 |
| | Score Calcu. | 0.19 | 0.34 | 0.42 | 0.48 |
| Exhaustive search | Score Simu. | 0.86 | 0.92 | 0.94 | 0.95 |
| | LRT Simu. | 0.87 | 0.92 | 0.94 | 0.94 |
| | Score Calcu. | 0.86 | 0.92 | 0.93 | 0.95 |
| Forward search | Score Simu. | 0.55 | 0.73 | 0.79 | 0.82 |
| | LRT Simu. | 0.56 | 0.73 | 0.80 | 0.83 |
| | Score Calcu. | 0.51 | 0.70 | 0.76 | 0.80 |
| *Power definition* (B) | | | | | |
| Marginal search | Score Simu. | 0.68 | 0.83 | 0.88 | 0.91 |
| | LRT Simu. | 0.68 | 0.83 | 0.88 | 0.91 |
| | Score Calcu. | 0.66 | 0.82 | 0.87 | 0.89 |
| Exhaustive search | Score Simu. | 0.87 | 0.93 | 0.94 | 0.96 |
| | LRT Simu. | 0.87 | 0.92 | 0.94 | 0.95 |
| | Score Calcu. | 0.87 | 0.93 | 0.95 | 0.96 |
| Forward search | Score Simu. | 0.69 | 0.82 | 0.88 | 0.89 |
| | LRT Simu. | 0.69 | 0.82 | 0.88 | 0.89 |
| | Score Calcu. | 0.67 | 0.82 | 0.86 | 0.89 |

sidering many more false models, the weak interaction effect close to zero does not play a dominant role in improving its power. Type I error control with the Bonferroni correction leads to notably different patterns of power comparisons. By examining the left and right columns of Figure 3 together with those of Figure 4, we can see that exhaustive search increases its advantage so as to uniformly beat both marginal search and forward search in finding the true genetic model over the whole genetic model space. Under Bonferroni control, the forward search improves its performance to match over that of marginal search to find at least one marker, which is demonstrated by comparing the middle column of Figure 3 with that of Figure 4. For both control criteria, when we relax the control level of $R$ or $\alpha$, marginal search becomes relatively more powerful than exhaustive and forward searches. This is shown in the supplementary material which contains more maps of comparisons under different genetic parameter set-ups.

### 4.3. Power comparisons when marginal effects are fixed

As illustrated above, the interaction effect is crucial for the statistical power

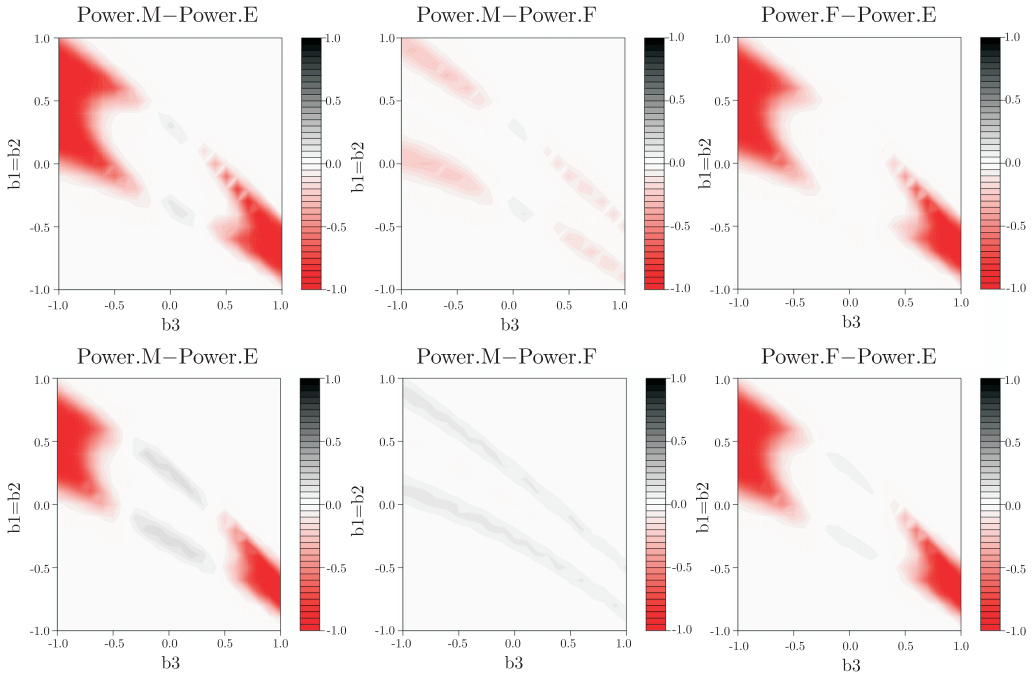


Figure 3. Power differences with the discovery number $R = 10$, and with varying main and interaction effects. Row 1 illustrates power comparison results under power definition (A), row 2 illustrates the results under power definition (B). Left column: marginal search vs. exhaustive search; middle column: marginal search vs. forward search; right column: forward search vs. exhaustive search. Main effect $b_1 = b_2$ and epistatic effect $b_3$ both vary from -1 to 1. The allele frequency $p_j = 0.3$, $j = 1, \ldots, L$.

of marker selection. However, because there is usually a lack of knowledge of interaction effects from real studies, it is meaningful to compare the search methods when the marginal association, possibly revealed from different interaction patterns, is fixed. Assume the genetic models in (2.1)−(2.3) have the same marginal association, represented by the heterozygote odds ratio $\lambda$ at each causative marker. When the values of $\lambda$ and the population disease prevalence $p(D)$ are fixed, we can calculate $\alpha$ and $\theta$ (letting $\theta_1 = \theta_2$ in model (2.1)). It is interesting to study the influence of LD when the true disease-causing loci are not observed but are linked with genotyped markers. We adopted the squared correlation coefficient $r^2$ to measure LD (Pritchard and Przeworski (2001)) while deriving the analytical power calculation to find the linked markers. The assumptions for the fixed marginal effect and the constraints of LD follow Marchini, Donnelly, and Cardon (2005). Technical details are given in the supplementary material.

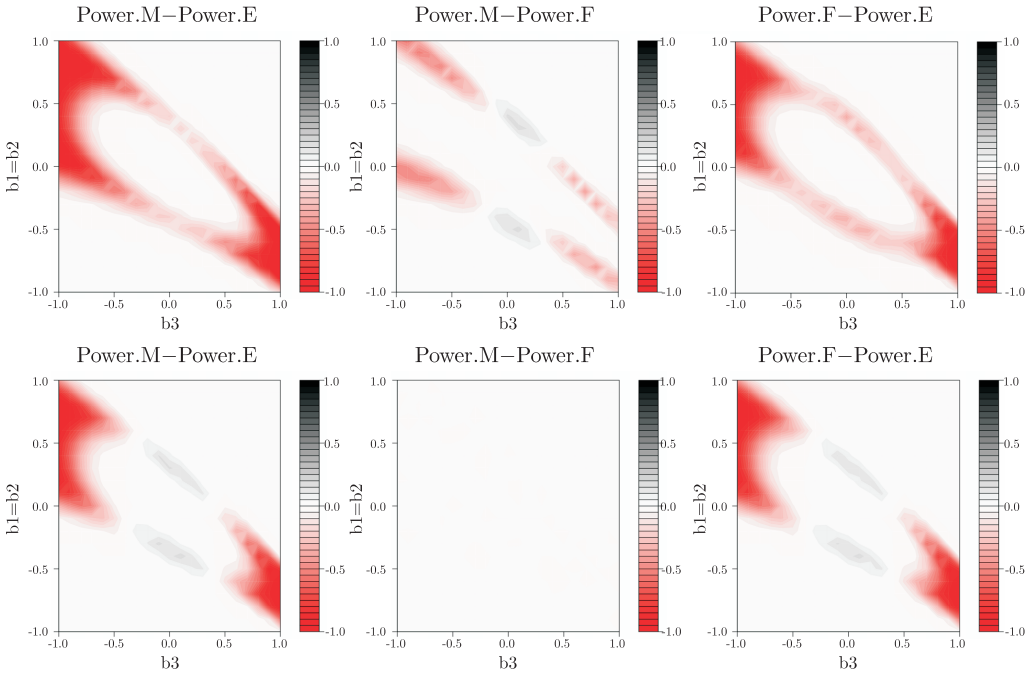


Figure 4. Power differences with Bonferroni type I error rate $\alpha = 0.05$, and with varying main and interaction effects. Power definition (A) is applied in row 1 and (B) is applied in row 2. Left column: marginal search vs. exhaustive search; middle column: marginal search vs. forward search; right column: forward search vs. exhaustive search. Main effect $b_1 = b_2$ varies from -1 to 1 with epistatic effect $b_3$. The allele frequency $p_j = 0.3$, $j = 1, \ldots, L$.

Let $\lambda = 1.5$, $p(D) = 0.01$, $n_1 = n_0 = 2,000$, the MAF $p_j = 0.05$, 0.1, 0.2, and 0.5, the LD strength $r^2 = 0.5$, 0.7, and 1, the LD constraint $p(A_i|A_{-i}) = 1$ and $p(A_i|a_{-i}) = q$, $i = 1, 2$, where $A_{-i}$ is the disease-causing allele at the unobserved locus indexed by $-i$, $A_i$ is the disease allele at the genotyped locus of marker $i$, which is in LD with the causative locus $-i$.

In general, the power definition, genetic model, allele frequency, and sample size influence the relative performance of search strategies. Under the discovery number control $R = 5$, Figure 5 shows the power comparisons for finding the joint association. Here, marginal search is the best for detecting Model (2.1). For detecting Models (2.2) and (2.3), exhaustive search is the best and marginal selection is the worst. The forward search is similar (for detecting Model (2.1)) or better than marginal search (for detecting Models (2.2) and (2.3)). This is not surprising because Model (2.1) is additive in the log scale of odds, whereas Models (2.2) and (2.3) are interactive, and accommodate those strategies facili-

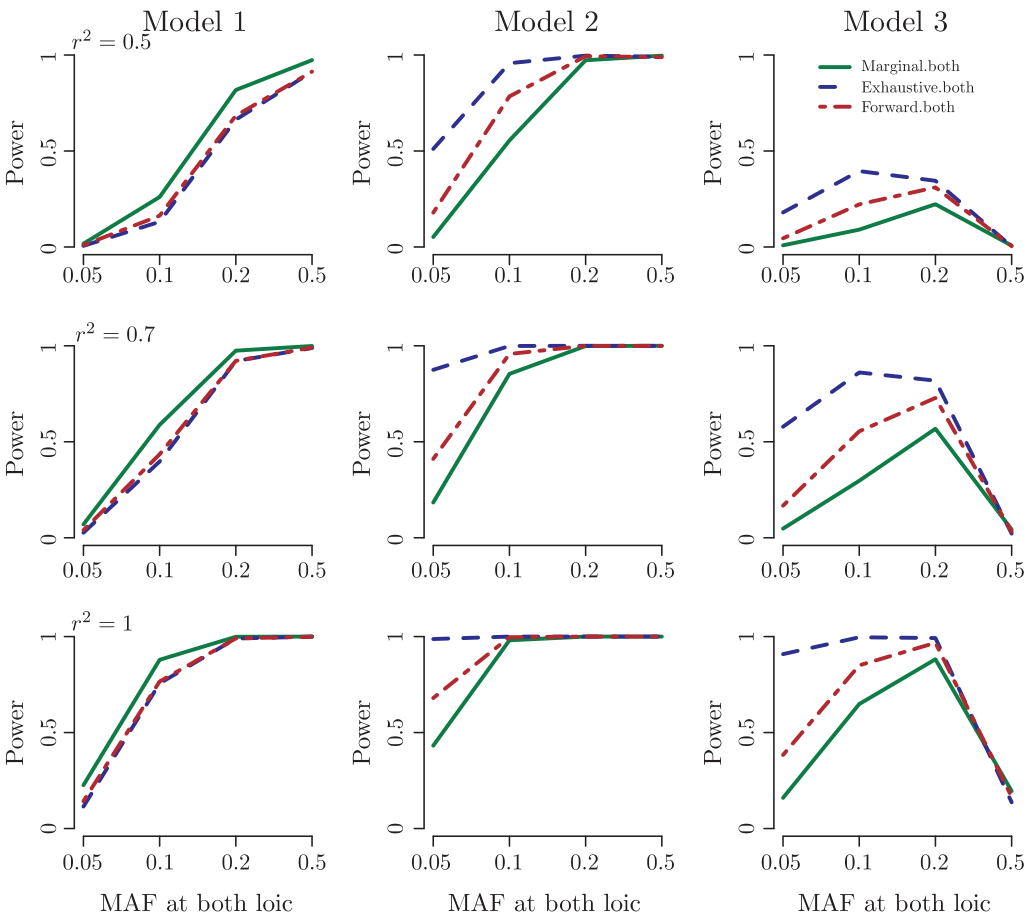


Figure 5. Power of finding the joint association with the discovery number $R = 5$. Solid lines, marginal search; dashed lines, exhaustive search; dot-dashed lines, forward search. The marginal odds ratio at both loci is 1.5, disease prevalence is 0.01, case and control numbers are both 2,000. Columns of panels show genetic Models 1, 2, 3, respectively; rows show LD strength $r^2 = 0.5$, 0.7, and 1. The minor allele frequencies are 0.05, 0.1, 0.2, and 0.5 on the x-axis of each penal.

tated by interaction effects. The increase of MAF shrinks the differences in the performance of the three model selection methods.

Figure 6 shows the power comparisons in finding either associated marker. For Model (2.1), marginal selection still performs the best, and forward search is better than exhaustive search. For Model (2.2), exhaustive search is the most powerful, while marginal search is better than forward search. Such performance differences also apply to Model (2.3) with small MAF. However, the rise of MAF improves the power of marginal search and forward search faster than the power

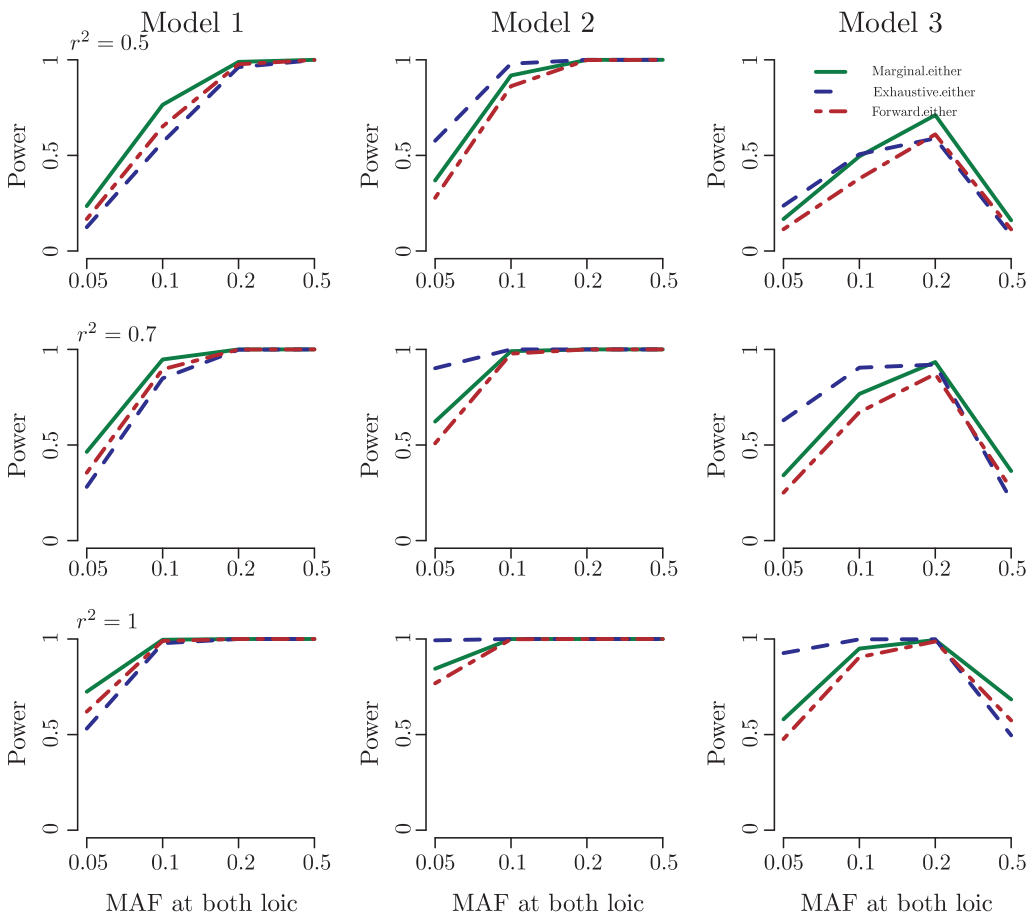


Figure 6. Power of finding either associated marker with the discovery number $R = 5$. Solid lines, marginal search; dashed lines, exhaustive search; dot-dashed lines, forward search. The marginal odds ratio at both loci is 1.5, disease prevalence is 0.01, case and control numbers are both 2,000. Columns of panels show genetic Models 1, 2, 3, respectively; rows show LD strength $r^2 = 0.5$, 0.7, and 1. The minor allele frequencies are 0.05, 0.1, 0.2, and 0.5 on the x-axis of each penal.

of exhaustive search, which is the worst with large MAF for Model (2.3). Furthermore, with sample size increasing, exhaustive search increases its power faster than marginal search and forward search for the interaction Models (2.2) and (2.3) (see the supplementary material for the figures of a smaller sample size $n_1 = n_0 = 1,000$). This indicates that exhaustive search has a more stringent requirement for sample size, but it provides greater potential to detect a small interactive effect when one has enough observations.

Under the Bonferroni corrected type I error rate control, Figure 7 (or 8)

shows the power comparisons for finding the joint association Model (or either associated marker). The genetic parameter set-up is the same as that in Figures 5 and 6. To find the joint association model, Figure 7 shows that exhaustive search is uniformly the best for all models, over all allele frequencies. Compared with forward search, marginal search is a more favored method for Model (2.1), but not for Models (2.2) and (2.3). As shown in Figure 8, marginal search and forward search are very similar for finding either associated marker. To find either marker, exhaustive search is the worst for Model (2.1), but the best for Model (2.2), and for Model (2.3) with small MAF.

We calculated the statistical power of a marginal-exhaustive two-stage method for finding joint association. The first stage screens single markers at a liberal type I error control level $\alpha_1$. The second stage then carries out exhaustive search within the selected set of markers, at a Bonferroni corrected type I error level $\alpha/\binom{\alpha_1 L}{2}$. This method is appealing for its potential of reducing the computational burden in exhaustive search, while still with high power. Figure 9 shows the comparisons among the powers of marginal search in finding either or both associated markers, the power of exhaustive search in finding the joint association, and the power of the two-stage method in finding the joint association. The genetic parameters are the same as those for Figures 5−8. The two-stage method performs similarly as, or even slightly better than, exhaustive search. However, this result is valid only under the moderate marginal association that guarantees a high probability of picking the associated markers in the screening stage. As shown in Figure 2, it is possible, at least in theory, that the marginal association from particular interactions may totally vanish. In this case, the two-stage method is certainly not able to surpass exhaustive search. Note that we only used minutes of computational time to analytically calculate power for generating Figure 9, which reproduces the similar comparison patterns as that shown by heavy simulations (Figure 2 in Marchini, Donnelly, and Cardon (2005)).

## 5. Discussion

### 5.1. Analytical power calculation

To measure the performance of model selection methods, we define power as the probability of model selection methods to find the models that contain all or some of associated markers. Therefore, our power definition differs from the traditional power of a specific model, the latter calculates the probability to reject the null hypothesis that the covariates have no association with the response.

The analytical power calculation for marker search strategies offers valuable tools for GWAS. Because the underlying genetic model is often not known, it is important to efficiently evaluate power so that researchers can explore a wide

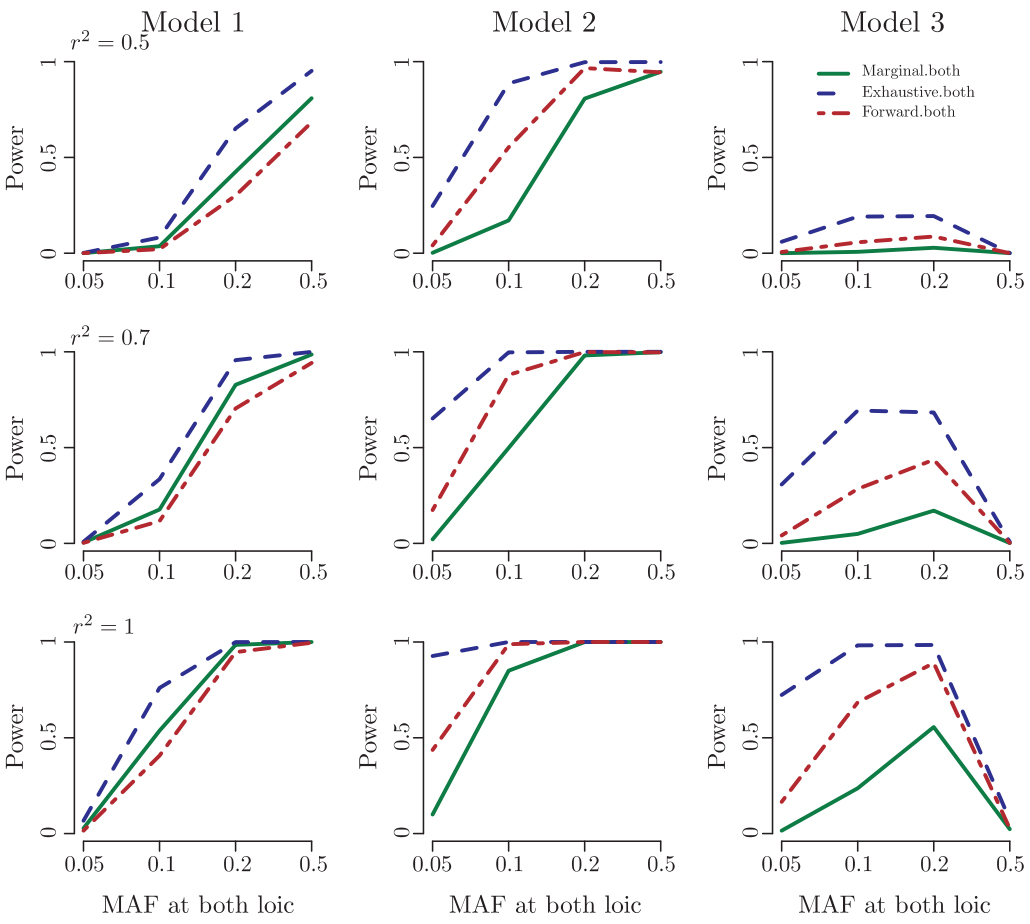


Figure 7. Power of finding the joint association with Bonferroni type I error rate $\alpha = 0.05$. Solid lines, marginal search; dashed lines, exhaustive search; dot-dashed lines, forward search. The marginal odds ratio at both loci is 1.5, disease prevalence is 0.01, case and control numbers are both 2,000. Columns of panels show genetic Models 1, 2, 3, respectively; rows show LD strength $r^2 = 0.5$, 0.7, and 1. The minor allele frequencies are 0.05, 0.1, 0.2, and 0.5 on the x-axis of each penal.

range of possibilities. Our analytical power calculation significantly reduces the computational burden of simulations. With sophisticated consideration of LD and flexible genetic models with or without interactions, the R package enables researchers to calculate the proper sample size and the statistical power at the experimental design stage, and investigate the performance of different marker search strategies at the data analysis stage.

It is generally hypothesized that complex diseases are jointly influenced by multiple markers with potential interactions. Therefore, we view the underlying

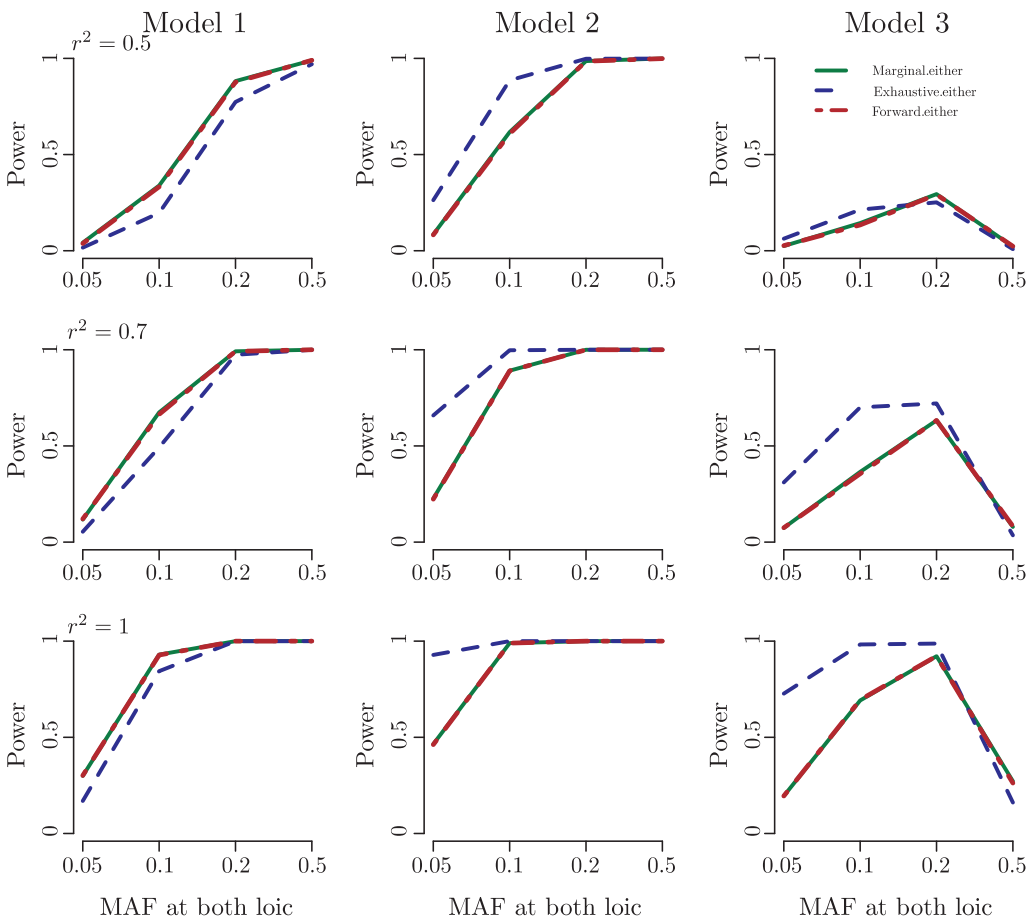


Figure 8. Power of finding either associated marker with Bonferroni type I error rate $\alpha = 0.05$. Solid lines, marginal search; dashed lines, exhaustive search; dot-dashed lines, forward search. The marginal odds ratio at both loci is 1.5, disease prevalence is 0.01, case and control numbers are both 2,000. Columns of panels show genetic Models 1, 2, 3, respectively; rows show LD strength $r^2 = 0.5$, 0.7, and 1. The minor allele frequencies are 0.05, 0.1, 0.2, and 0.5 on the x-axis of each penal.

genetic models as multivariate models with joint effects and interactions, not as the oversimplified single marker models studied in the literature (Gail et al. (2008)). The genetic model studied here can be extended to more complex models with higher order interactions. Moreover, for the multivariate joint marker models, our study of the distributions and the correlation structures among test statistics provides the understanding of how joint genetic signals can be picked up by various statistical model selection procedures.

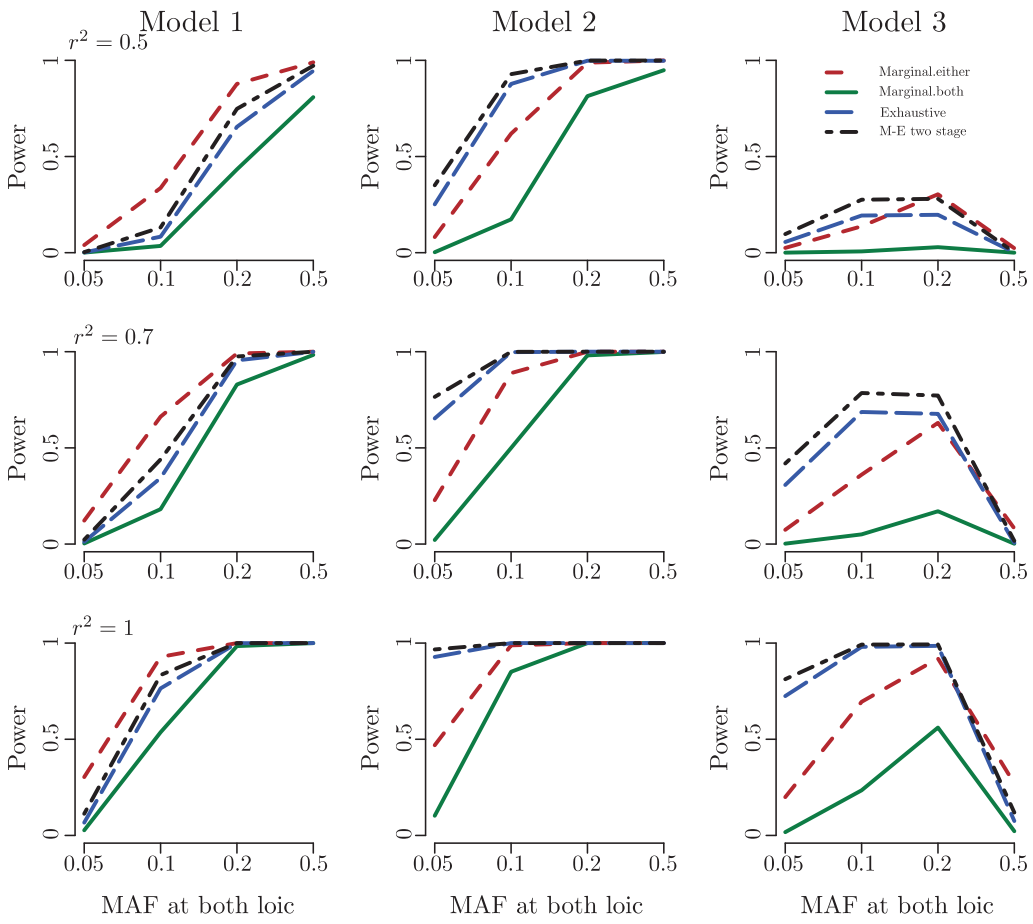


Figure 9. Power of four marker search methods controlled by Bonferroni type I error rate $\alpha = 0.05$. Dashed lines, marginal search for either association; Solid lines, marginal search for joint association; long-dashed lines, exhaustive search for joint association; dot-dashed lines, two-stage search for joint association. The marginal odds ratio at both loci is 1.5, disease prevalence is 0.01, case and control numbers are both 2,000. Columns of panels show genetic Models 1, 2, 3, respectively; rows show LD strength $r^2 = 0.5$, 0.7, and 1. The minor allele frequencies are 0.05, 0.1, 0.2, and 0.5 on the x-axis of each penal.

### 5.2. Binary trait versus quantitative trait

Comparing the results obtained here for the binary trait with those for the quantitative trait (Wu and Zhao (2009)), the patterns of the relative performances of marginal search, exhaustive search, and forward search have both common and distinguishable characteristics. In common, the strength of interaction effect is a key factor to the performance. Both marginal and forward searches

fail when markers have interactions that mask the marginal effects. Our calculation shows that this happens when the expectation of marginal test statistics (such as these in (3.1)) is 0. To find joint genetic association, the strong interaction effect benefits exhaustive and the forward search, as they are carried out in a higher dimensional model space. Also in common is the practical context in which each marker-search method is preferred. Forward search is not the best choice in most circumstances, as it is usually matched or outperformed either by exhaustive search in finding the genetic model, or by marginal search in finding one of the genetic factors. Exhaustive search is generally recommended for finding the genetic model when interactions are believed to exist. Marginal search is recommended for finding at least one genetic factor in the preliminary study of a new trait, because of its simplicity as well as its acceptable power in most parts of the genetic parameter spaces.

As to distinctions, exhaustive search is relatively more powerful for binary traits than for quantitative traits, especially in finding the joint association. Forward search is relatively less powerful for binary traits than for quantitative traits, especially in finding at least one associated marker. Based on the analytical study, this distinction is caused by the correlation between the test statistics of single-marker model fittings in (2.6), and the test statistics of the extra terms in (2.7) over (2.6). The correlation is stronger among score test statistics for binary traits than that among F test statistics for quantitative traits. Thus an exhaustive search that finds one associated marker has a greater chance to find the other associated marker for binary traits than for quantitative traits. On the other hand, if the selected marker at the first step is not associated, it is much harder for forward search in the second step to discover the associated markers for binary traits than for quantitative traits.

Another difference is the symmetry of the influence of genetic effects on power. If the total genetic effect can be represented by $b_1g_1 + b_2g_2 + b_3g_1g_2$, the search power is the same for quantitative traits when the main effects $b_1 = b_2 = 0$ and the interaction effect $b_3 = \pm c$. This is because the underlying quantitative trait genetic model is linear (Wu and Zhao (2009)), and the regression models have the same goodness-of-fit for signals with opposite directions but the same magnitude. This is no longer true for binary traits because the total genetic effects with the same magnitude but opposite directions lead to different disease risks, as shown in (2.4). Thus, the heat maps in Figures 3-4 are not symmetric.

### 5.3. The R-control versus the Bonferroni control

We studied two different types of statistical significance controls: the total discovery number $R$-control, related to false discovery number or false discovery

proportion control, and the type I error rate $\alpha$-control with the Bonferroni adjustment. These two have different crucial effects on the power of model selection methods. First, the $\alpha$-control is more stringent than $R$-control, and $R$-control leads to higher statistical power. So even if we do not have many "significant" results with Bonferroni control, we can still include the top ranked genetic variations into the validation stage with $R$-control, as long as resources permit. Second, because GWAS are mostly used to screen for candidate genes, the type I error rate control is not the number one aim. $R$-control is more commonly adopted in phased designs by researchers who want to control the number of markers for the follow-up validation study. Third, with Bonferroni control, we expect more joint associations to be found when applying exhaustive scan in GWAS data analysis, since Bonferroni control enhances the power superiority of exhaustive search relative to marginal search, especially for finding joint associations. Lastly, for finding either associated marker, $R$-control favors forward selection less than marginal search, whereas $\alpha$-control makes the two methods very similar.

The widely applied Bonferroni control procedure provides an intuitively simple rule for model selection. Nevertheless, it does not provide an accurate type I error rate control for the whole model selection process. For example, if exhaustive search aims to find all associated markers (Marchini, Donnelly, and Cardon (2005); Evans et al. (2006); Storey, Akey, and Kruglyak (2005); Brem et al. (2005)), the models with partially associated markers should contribute to the null distribution. However, the Bonferroni correction procedure only applies $\chi^2_3$ as the null distribution, which ignores these "wrong" models. As for forward selection, applying type I error rate control in each step does not necessarily lead to the overall family-wise type I error rate control.

Our power calculation can be extended to more general situations where more than two genetic factors are involved, with potentially higher-order interactions. For a given genetic model, the asymptotic distributions of the relevant test statistics can be easily derived by computer because of the general formula for the score test statistic in (3.2) and the asymptotic results given in Section 3.2. Nevertheless, the power calculation would be more tedious because of more complicated correlation structures among the test statistics, caused by the various overlappings of fitted models. Furthermore, more situations need be considered for the "partially correct" models that contain one or more true SNPs.

**Acknowledgement**

We are grateful to the Yale University Biomedical High Performance Computing Center and the WPI Computing and Communications Center for computational support.

## References

Brem, R. B., Storey, J. D., Whittle, J. and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436**, 701–703.

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics*. Third Edition. Wiley, New York.

Evans, D. M., Marchini, J., Morris, A. P. and Cardon, L. R. (2006). Two-stage two-locus models in genome-wide association. *PLoS Genet* **2**, e157.

Gail, M. H., Pfeiffer, R. M., Wheeler, W. and Pee, D. (2008). Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. *Biostatistics* **9**, 201-215.

Kraft, P. and Hunter, D. J. (2009). Genetic Risk Prediction – Are We There Yet? *N. Engl. J. Med.* **360**,1701-1703.

Marchini, J., Donnelly, P. and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* **37**, 413-417.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A. and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**, 356-369.

Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**,1-14.

Scheffé, H. (1959). The Analysis of Variance. Wiley, New York.

Storey, J. D., Akey, J. M. and Kruglyak, L. (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biolgoy* **3**, 1380-1390.

Wolfram, S. (1999). *The Mathematica Book*. Cambridge University Press.

Wu, Z. and Zhao, H. (2009). Statistical power of model selection strategies for genome-wide association studies. *PLoS Genet* **5** e1000582. 10.1371/journal.pgen.1000582

Zhang, B. (2006). A score test under logistic regression models based on case-control data. *Statist. Neerlandica* **60**, 477-496.

Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA, 01609-2280, USA.

E-mail: zheyangwu@wpi.edu

Yale University.Department of Epidemiology and Public Health, Yale University, School of Medicine, 60 College Street New Haven, CT 06520-8034, USA.

E-mail: hongyu.zhao@yale.edu

(Received August 2010; accepted May 2011)