Statistical Power of Model Selection Strategies for Genome-Wide Association Studies

Zheyang Wu[1], **Hongyu Zhao**[1,2]*

[1] Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut, United States of America, [2] Department of Genetics, Yale University School of Medicine, New Haven, Connecticut, United States of America

Abstract

Genome-wide association studies (GWAS) aim to identify genetic variants related to diseases by examining the associations between phenotypes and hundreds of thousands of genotyped markers. Because many genes are potentially involved in common diseases and a large number of markers are analyzed, it is crucial to devise an effective strategy to identify truly associated variants that have individual and/or interactive effects, while controlling false positives at the desired level. Although a number of model selection methods have been proposed in the literature, including marginal search, exhaustive search, and forward search, their relative performance has only been evaluated through limited simulations due to the lack of an analytical approach to calculating the power of these methods. This article develops a novel statistical approach for power calculation, derives accurate formulas for the power of different model selection strategies, and then uses the formulas to evaluate and compare these strategies in genetic model spaces. In contrast to previous studies, our theoretical framework allows for random genotypes, correlations among test statistics, and a false-positive control based on GWAS practice. After the accuracy of our analytical results is validated through simulations, they are utilized to systematically evaluate and compare the performance of these strategies in a wide class of genetic models. For a specific genetic model, our results clearly reveal how different factors, such as effect size, allele frequency, and interaction, jointly affect the statistical power of each strategy. An example is provided for the application of our approach to empirical research. The statistical approach used in our derivations is general and can be employed to address the model selection problems in other random predictor settings. We have developed an R package *markerSearchPower* to implement our formulas, which can be downloaded from the Comprehensive R Archive Network (CRAN) or http://bioinformatics.med.yale.edu/group/.

Citation: Wu Z, Zhao H (2009) Statistical Power of Model Selection Strategies for Genome-Wide Association Studies. PLoS Genet 5(7): e1000582. doi:10.1371/journal.pgen.1000582

Editor: Bruce Walsh, University of Arizona, United States of America

Received April 8, 2009; **Accepted** June 30, 2009; **Published** July 31, 2009

Funding: This work was supported by NIH grant RR19895, NIH grant GM 59507, and NSF grant DMS 0714817. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: hongyu.zhao@yale.edu

Introduction

In genome-wide association studies (GWAS), hundreds of thousands of markers are genotyped to identify genetic variations associated with complex phenotypes of interest. The detection of truly associated markers can be framed as a model selection problem: a group of statistical models are considered to assess how well each model predicts the phenotype, and the selected models are expected to include all or some of the truly associated genetic markers and few, if any, markers not associated with the phenotype. In the literature, three model-selecting procedures have been advocated: marginal search, exhaustive search, and forward search.

Marginal search analyzes markers individually and is the simplest and computationally least expensive among these three search methods. Under certain assumptions, such as no interactions among covariates (or markers in the GWAS context), Fan and Lv [1] proved that the truly associated covariates will be among those having the highest marginal correlations. However, Fan and Lv acknowledged that marginal search may suffer when an important covariate is jointly associated as a group but marginally unassociated as individuals with the response (phenotype). In GWAS, the phenotypes are likely associated with multiple genes, their gene-gene interactions (i.e. epistases), and gene-environment interactions. Therefore, marginal search may not be optimal for the analysis of GWAS data.

In contrast to marginal search, exhaustive search and forward search simultaneously consider multiple markers in the model. Exhaustive search examines all possible models within a given model dimension, and forward search identifies markers in a stepwise fashion. As they consider interactions, they may gain statistical power compared to marginal search [2–5]. In practice, exhaustive search bears a much larger computational burden because the number of models that need to be explored is an exponential function of the number of markers jointly considered. For example, if 500,000 markers are genotyped, an exhaustive search of all marker pairs would study around 10^{11} candidate models. This requires significant computational resources, especially when permutations are needed to establish overall significance levels, e.g. for the purpose of appropriately accounting for dependencies among markers. Because of this computational burden, it is difficult or even impossible to assess the power of exhaustive search through simulation studies.

Based on limited simulations and real data analysis, conflicting results exist in the literature on the relative merit of exhaustive search and forward search. Because exhaustive search considers many more models, it may increase the probability that the truly

Author Summary

Almost all published genome-wide association studies are based on single-marker analysis. Intuitively, joint consideration of multiple markers should be more informative when multiple genes and their interactions are involved in disease etiology. For example, an exhaustive search among models involving multiple markers and their interactions can identify certain gene–gene interactions that will be missed by single-marker analysis. However, an exhaustive search is difficult, or even impossible, to perform because of the computational requirements. Moreover, searching more models does not necessarily increase statistical power, because there may be an increased chance of finding false positive results when more models are explored. For power comparisons of different model selection methods, the published studies have relied on limited simulations due to the highly computationally intensive nature of such simulation studies. To enable researchers to compare different model search strategies without resorting to extensive simulations, we develop a novel analytical approach to evaluating the statistical power of these methods. Our results offer insights into how different parameters in a genetic model affect the statistical power of a given model selection strategy. We developed an R package to implement our results. This package can be used by researchers to compare and select an effective approach to detecting SNPs.

associated markers do not rise to the top as more models involving unrelated markers may outperform the true models simply due to chance. Forward search explores a smaller model space, allowing a less stringent threshold for significance. However, forward search may miss the markers that have a strong interaction effect but weak marginal effect. Through limited simulation studies, Marchini and colleagues [4,5] concluded that exhaustive search is more powerful in finding truly associated markers in the presence of epistasis. On the contrary, based on the analysis of a real data set for yeast, Storey and colleagues [2,3] recommended sequential forward search. They reported that exhaustive search suffers from lower power because a substantial increase in the number of models. By analytically demonstrating the conditions under which exhaustive search is better than forward search, and the reverse, our research systematically explains these contradictory results.

It is clear that the optimal model selection strategy depends on the underlying genetic model, which is unknown to researchers. In the most extreme case, if the underlying genetic model has no marginal association, an exhaustive search is the only way to find influential genes. On the other hand, for a model with purely additive genetic effects, marginal or forward search will be the most effective. For the cases between these two extremes, the optimal model selection strategy should achieve a delicate balance between computational efficiency, statistical power, and a low false positive rate. Without the knowledge of underlying models, it is necessary to evaluate the different methods by thoroughly comparing them across a large genetic model space, in which both computationally intensive simulations and limited real data analysis are difficult to fully explore.

In this article, we derive the analytical results for statistical power of marginal search, exhaustive search, and forward search. These formulas can significantly reduce the computational burden in power estimation. To implement the formulas, we developed an R package *markerSearchPower*. We demonstrate through simulations that our results are accurate. Through our results, we can systematically assess different SNP search methods across a large model space and efficiently identify the optimal one. Our derivation approaches are general and can be applied to the model selection procedures in other random predictor settings.

The rest of this article is organized as follows: in the Results section, we present the model set-up, the validation of our analytical results through simulations, and the comparisons among three model selection strategies; in the Discussion section, we summarize the power comparison results and discuss our methodological contributions; and in the Methods section, we outline the derivations of asymptotic distributions and power calculations. The Text S1 available online gives statistical details of proofs and derivations, extended power comparisons, and relevant formulas for the estimates of distribution parameters.

Results

Model Setup

A genetic model relates phenotype to genotypes, and this relationship can be rather complex. In general, statistical power depends on the effects of risk alleles, allele frequencies in the population, epistasis, as well as environmental risk factors and their interactions with genetic factors. We focus on a model commonly used in the literature, which offers valuable insights into the relative performance of model selection methods.

Assume that genotype data are available from p independent single nucleotide polymorphisms (SNPs). Our results can be generalized to other types of markers. We use $X_{i1}, \ldots, X_{ip}$, $i=1, \ldots, n$, to denote the genotypes for the ith sampled individual, for SNPs $1, \ldots, p$, respectively. Let the alleles at the jth SNP be M_j and m_j with frequencies p_j and $q_j=1-p_j$, respectively. Under the assumption of Hardy-Weinberg equilibrium and additive allelic effects, we use the following coding for this SNP:

$$X_{ij}=\begin{cases} 1 & \mathit{Genotype}=M_jM_j, \text{ with probability } p_j^2 \\ 0 & \mathit{Genotype}=M_jm_j, \text{ with probability } 2p_jq_j\,. \\ 1 & \mathit{Genotype}=m_jm_j, \text{ with probability } q_j^2 \end{cases} \quad (1)$$

We focus on the scenario that two of these SNPs, indexed by 1 and 2, are truly associated with a quantitative outcome Y through the following genetic model

$$Y_i=b_0+b_1X_{i1}+b_2X_{i2}+b_3X_{i1}X_{i2}+\varepsilon_i, \quad (2)$$

where $\varepsilon_i \sim N(0, \sigma^2)$ is independent of the genotypes. The interaction term represents the epistatic effect, and its coefficient b_3 measures the direction and magnitude of this effect.

Based on the observed data, we fit the following models using Ordinary Least Squares (OLS) involving one or two SNPs:

$$\hat{Y}_{i(j)}=\hat{\beta}_{0(j)}+\hat{\beta}_{1(j)}X_{ij}, \quad (3)$$

$$\hat{Y}_{i(jk)}=\hat{\beta}_{0(jk)}+\hat{\beta}_{1(jk)}X_{ij}+\hat{\beta}_{2(jk)}X_{ik}+\hat{\beta}_{3(jk)}X_{ij}X_{ik}. \quad (4)$$

The subscripts in the above models index the SNP(s) included in these models. Based on models (3) and (4), three model selection methods seek candidate markers according to the corresponding test statistics. In marginal search, we fit simple linear model (3) and compare the T-statistics [6] T_j for $j=1, \ldots, p$. A model, and thus its involved SNP, is selected if the corresponding T-statistic is among the largest from all tests. In two-dimensional exhaustive search, we fit regression model (4) for all SNP pairs and compare the F-statistics [6] F_{jk} for all $j<k$ where $j, k \in \{1, \ldots, p\}$. The models with

the highest values of the F-statistics are selected. In forward search, we first conduct a marginal model selection through model (3) and select the jth SNP if $|T_j|$ is the largest. With X_j, we then add another SNP X_k ($k \neq j$) for different SNPs, and choose models in format (4) which generate the highest F-statistics.

Two criteria are adopted to decide if the chosen models are correct. On one hand, we could be rather stringent and call a model correct only if it matches the true underlying genetic model. This is consistent with the concept of "joint significance" in Storey et al. [2]. On the other hand, we could be more generous and call a model correct if it contains at least one of the truly associated markers. This is consistent with the null hypothesis used in some published simulation studies [4,5]. Accordingly, we consider two definitions of power for a model selection procedure:

(A) the probability of identifying exactly the true model (in marginal search, it is the probability of detecting both true SNPs);

(B) the probability of detecting at least one of the true SNPs.

Under power definition (A), the null model is any model other than the true genetic model; under power definition (B), the null model is any model containing neither true SNP.

Comparison between Analytical and Simulation Results

We evaluated the accuracy of the asymptotic results derived in the Methods section by comparing the analytical results with those from simulations. To estimate power through simulation studies, we generated 1,000 data sets with n subjects and p candidate SNPs assuming Hardy-Weinberg equilibrium, as indicated in (1). The quantitative trait values were generated through true model (2) involving two true SNPs. We then used marginal search, exhaustive search, and forward search to identify SNPs associated with the trait. Under power definition (A), the target model(s) were the true model (or models with one true SNP in marginal search), and the other models were considered null models. Under power definition (B), the target models were those containing at least one true SNP, and the rest were considered null models. The empirical power estimated from these simulations was the proportion of that datasets that we were able to successfully find the target model(s) through model selection procedures, under the control of a pre-specified number (R) of falsely discovered null models. Such control offers a fair comparison of power among the three model selection methods and is numerically equal to the detection probability (DP) control [7], which is the probability of including a "correct model" when selecting R (or R+1 in marginal search under power definition (A)) of the most significant models.

In the first set-up for model (2), we considered $n = 100$ subjects, $p = 300$ SNPs, genetic effects $b_1 = b_2 = 0.1$, $b_3 = 2.4$, allele frequency of each SNP $q_j = 0.3$, $j = 1, \ldots, p$, and variance $\sigma^2 = 3$. Table 1 summarizes the calculated power and the simulated power under definitions (A) and (B). The second set-up is the same as the first except $b_3 = 1.4$. For this set-up, Table 2 shows the results under definitions (A) and (B). The two values of b_3 represent large and small interaction terms with which the simulation generated a broad spectrum of power values. In both set-ups, the analytical power is very close to the empirical power based on simulations.

We chose these two set-ups in which the power was reasonably large to approximate most practical settings. The chosen value of p is much smaller than that in GWAS (in the 100,000's) for the feasibility of simulation. As discussed in the Methods section, the asymptotic results are derived by assuming a large p. Therefore, we expect better approximations if p has a value similar to those in a real GWAS.

Power Comparisons of Model Selection Methods

The simulation results shown in Table 1 and Table 2 demonstrate that our analytical results provide good approximations to the true power, which is the basis for comparing the performance of these model search methods in a practical GWAS. We now consider a more realistic setting with a sample size of 1000 individuals (n) and a total of 300,000 SNPs (p). We assumed a genetic model of form (2) with $\sigma^2 = 3$ and varied the values of $b_1 = b_2$ as well as that of b_3 from -1 to 1 by a step size of 0.1. To simplify the discussion, we assumed all SNPs had the same allele frequency of $q_j = 0.3$, $j = 1, \ldots, p$. Note that this setting can be changed without affecting the qualitative nature of the comparison results.

Figure 1 gives the 3D plots of statistical power over the genetic model space for different model selection methods (in columns)

Table 1. The probability of detecting the exact true model (or both true SNPs in marginal search) under power definition A, and the probability of detecting at least one of the true SNPs under power definition B, with the false discovery number R varying. $b_1 = b_2 = 0.1$, $b_3 = 2.4$.

Category	Strategy	Source	$R = 1$	$R = 5$	$R = 10$	$R = 15$	$R = 20$	$R = 30$
Definition A	Marginal search	simulation	0.268	0.556	0.683	0.754	0.790	0.851
		calculation	0.279	0.552	0.673	0.738	0.781	0.836
	Exhaustive search	simulation	0.987	0.998	1.000	1.000	1.000	1.000
		calculation	0.978	0.995	0.998	0.998	0.998	1.000
	Forward search	simulation	0.780	0.788	0.789	0.789	0.789	0.789
		calculation	0.795	0.800	0.800	0.800	0.801	0.801
Definition B	Marginal search	simulation	0.790	0.950	0.980	0.985	0.993	0.995
		calculation	0.806	0.958	0.982	0.990	0.993	0.997
	Exhaustive search	simulation	0.993	0.999	1.000	1.000	1.000	1.000
		calculation	0.985	0.999	0.999	1.000	1.000	1.000
	Forward search	simulation	0.843	0.910	0.944	0.961	0.974	0.986
		calculation	0.828	0.906	0.938	0.952	0.966	0.983

doi:10.1371/journal.pgen.1000582.t001

Table 2. The probability of detecting the exact true model (or both true SNPs in marginal search) under power definition A, and the probability of detecting at least one of the true SNPs under power definition B, with the false discovery number R varying. $b_1 = b_2 = 0.1$, $b_3 = 1.4$.

Category	Strategy	Source	$R = 1$	$R = 5$	$R = 10$	$R = 15$	$R = 20$	$R = 30$
Definition A	Marginal search	simulation	0.055	0.180	0.291	0.359	0.425	0.512
		calculation	0.053	0.179	0.289	0.355	0.424	0.515
	Exhaustive search	simulation	0.394	0.586	0.667	0.706	0.715	0.753
		calculation	0.399	0.567	0.638	0.681	0.707	0.728
	Forward search	simulation	0.242	0.308	0.331	0.340	0.343	0.348
		calculation	0.238	0.308	0.331	0.342	0.349	0.354
Definition B	Marginal search	simulation	0.394	0.695	0.802	0.869	0.899	0.935
		calculation	0.406	0.698	0.807	0.862	0.894	0.932
	Exhaustive search	simulation	0.533	0.757	0.823	0.850	0.880	0.910
		calculation	0.569	0.738	0.809	0.848	0.874	0.906
	Forward search	simulation	0.422	0.561	0.654	0.731	0.769	0.841
		calculation	0.433	0.554	0.647	0.711	0.758	0.821

doi:10.1371/journal.pgen.1000582.t002

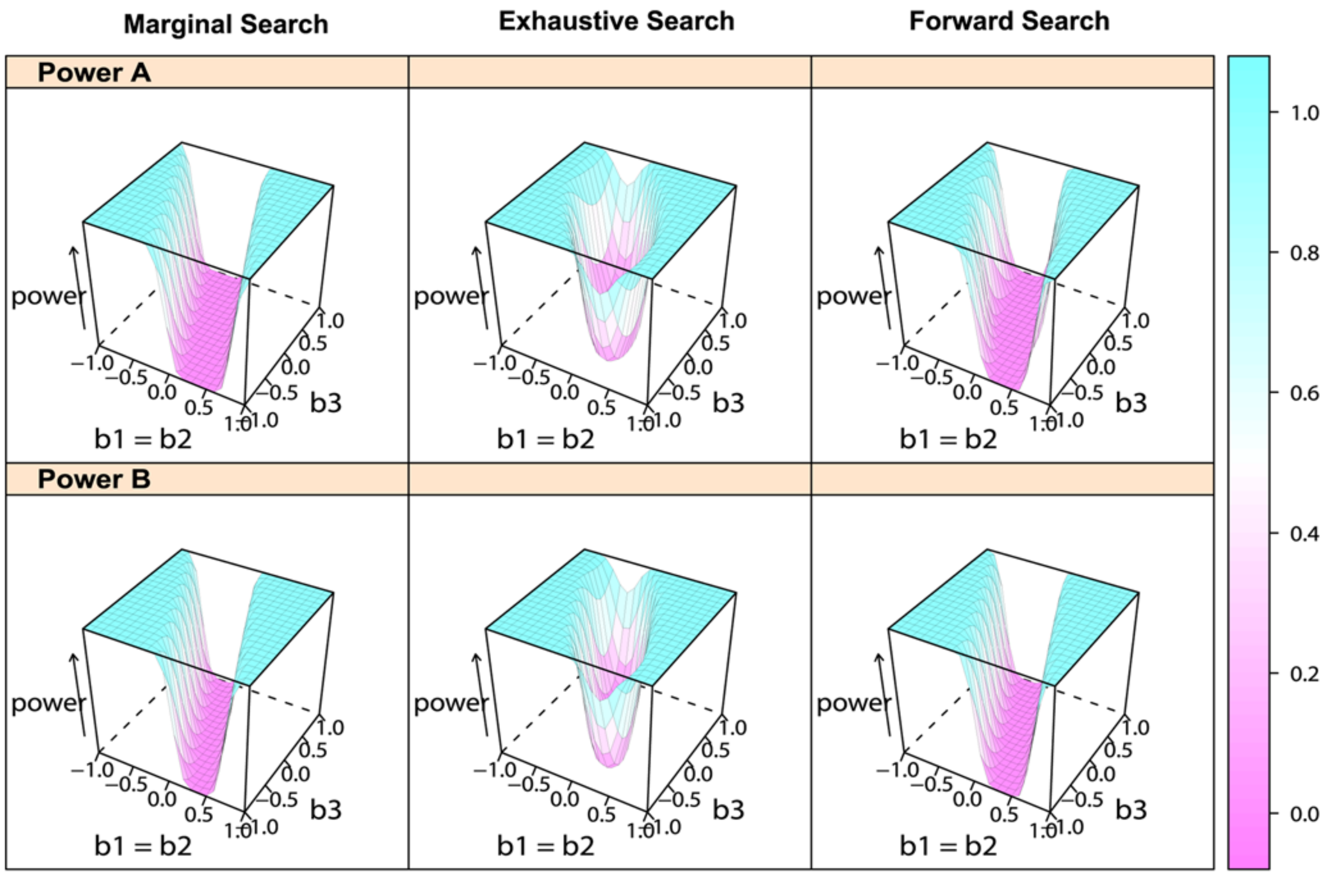

Figure 1. 3D plots of statistical power over genetic model space. The results of power for the three model selection methods: marginal search in the left column, exhaustive search in the middle column and forward search in the right column. Two definitions of power (A) for detecting the true model or both true SNPs in marginal search in row 1, and (B) for detecting either true SNP in row 2 are considered. We consider genetic models with the main effects $b_1 = b_2$ varying from −1 to 1 and the epistatic effect b_3 varying from −1 to 1. The allele frequency $q_j = 0.3$, $j = 1, \ldots, p$, and the false discovery number R is set to be 10.

doi:10.1371/journal.pgen.1000582.g001

under two power definitions (A) and (B) (in rows), when controlling the number of false discoveries to be $R = 10$. These figures illustrate that marginal search and forward search cannot detect the marginal association of the influential SNP 1 or 2 in a certain region of the model space, while exhaustive search can. This portion of the model space is represented by the region where the power of marginal search and that of forward search are very close to 0, no matter how large the genetic effect is. According to formulas (8) and (16) in the Methods section, the marginally non-detectable region for SNP 1, where $b_1+b_3(p_2-q_2)=0$, depends on the additive genetic effect b_1, epistatic effect b_3, and the allele frequency p_2 of SNP 2. The non-detectable region for SNP 2 is analogous by symmetry. In exhaustive search, such region does not exist, as indicated by formula (12). So, exhaustive search can better identify the signals when they are counterbalanced.

In order to better visualize the difference of model selection methods, we show the power differences between different methods. The left, middle, and right columns of Figure 2 and Figure 3 present the power difference between marginal search and exhaustive search, between marginal search and forward search, and between forward search and exhaustive search, respectively. For a specific comparison, the red areas represent negative values, indicating the former method has lower power, and the green areas represent positive values, indicating the former method has higher power. The dashed contours in these plots represent the heritability of the genetic model, i.e., the proportion of the total variation due to genetic effects, which is defined as

$$H^2 = \frac{\text{genetic variance}}{\text{total variance}}.$$

Under our model set-up,

$$\begin{aligned}\text{genetic variance} = {} & 2b_1^2 p_1 q_1 + 2b_2^2 p_2 q_2 + 4b_1 b_3 (p_2 - q_2) p_1 q_1 \\ & + 4b_2 b_3 (p_1 - q_1) p_2 q_2 \\ & + \left[\left(p_1^2 + q_1^2\right)\left(p_2^2 + q_2^2\right) - (p_1 - q_1)^2 (p_2 - q_2)^2\right] b_3^2, \\ \text{total variance} = {} & \text{genetic variance} + \sigma^2.\end{aligned}$$

In each plot, there are two areas in which the difference of power is close to 0. First, in the central area where the signal is weak (small H^2), all model selection procedures have low power and tend to fail to pick up the true SNPs. Second, in the edge areas where the signals are strong, all model selection procedures have similarly good power. The light colored areas represent these two special situations in which there is little difference in power among model selection methods.

To compare marginal search and exhaustive search, the left columns of Figure 2 and Figure 3 exhibit the power difference under power definitions (A) and (B), respectively. Exhaustive search has significant advantage in the red areas where the interaction effect b_3 is large or $b_1+b_3(p_2-q_2)$ is small. Such advantage is more

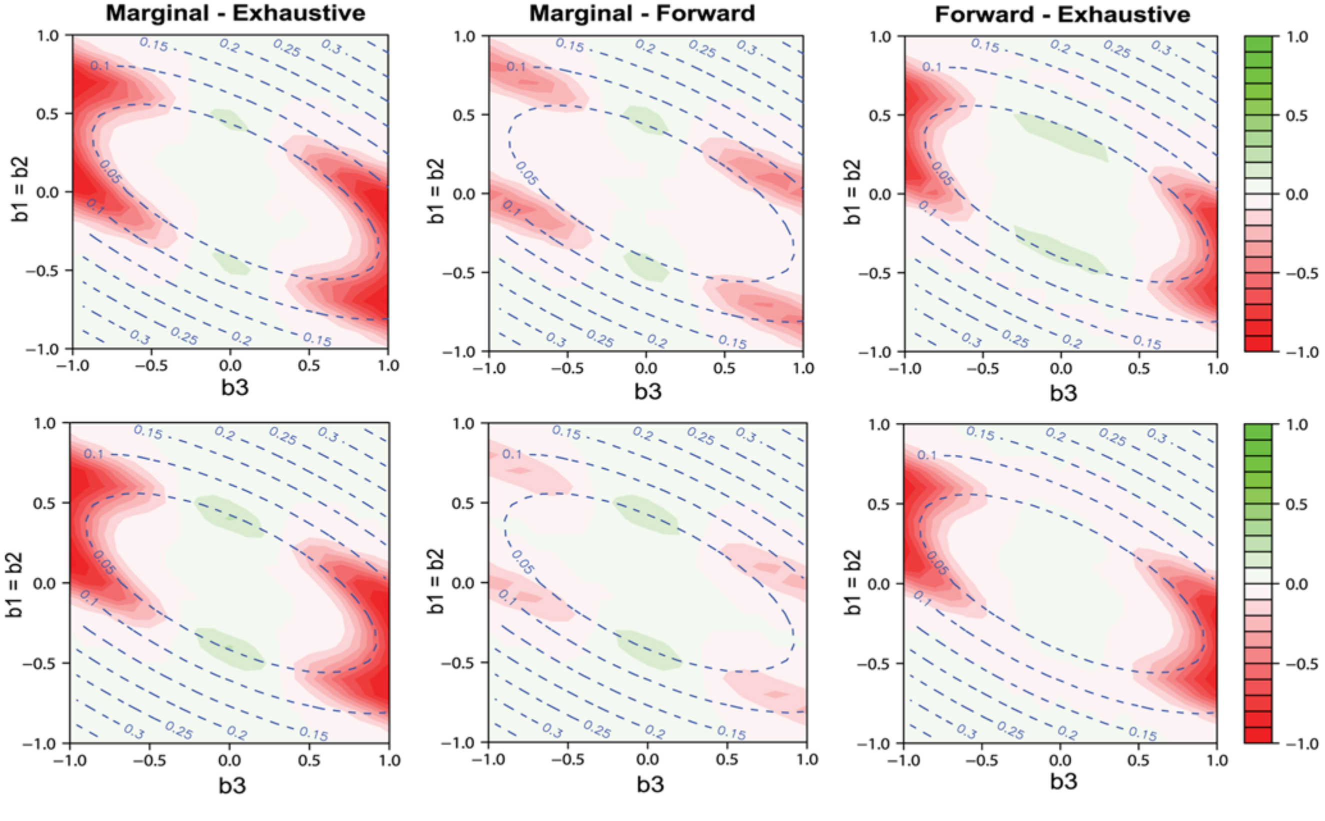

Figure 2. Comparisons among model selection power for detecting the true model or both true SNPs in marginal search over genetic model space. The power differences between marginal search and exhaustive search in the left column, between marginal search and forward search in the middle column, and between forward search and exhaustive search in the right column. Green areas indicate positive values of difference, and red areas indicate negative values of difference. We consider genetic models with the main effects $b_1 = b_2$ varying from −1 to 1 and the epistatic effect b_3 varying from −1 to 1. The allele frequency $q_j = 0.3$, $j = 1, \ldots, p$, and the false discovery number R is set to be 1 in row 1 and 10 in row 2.

doi:10.1371/journal.pgen.1000582.g002

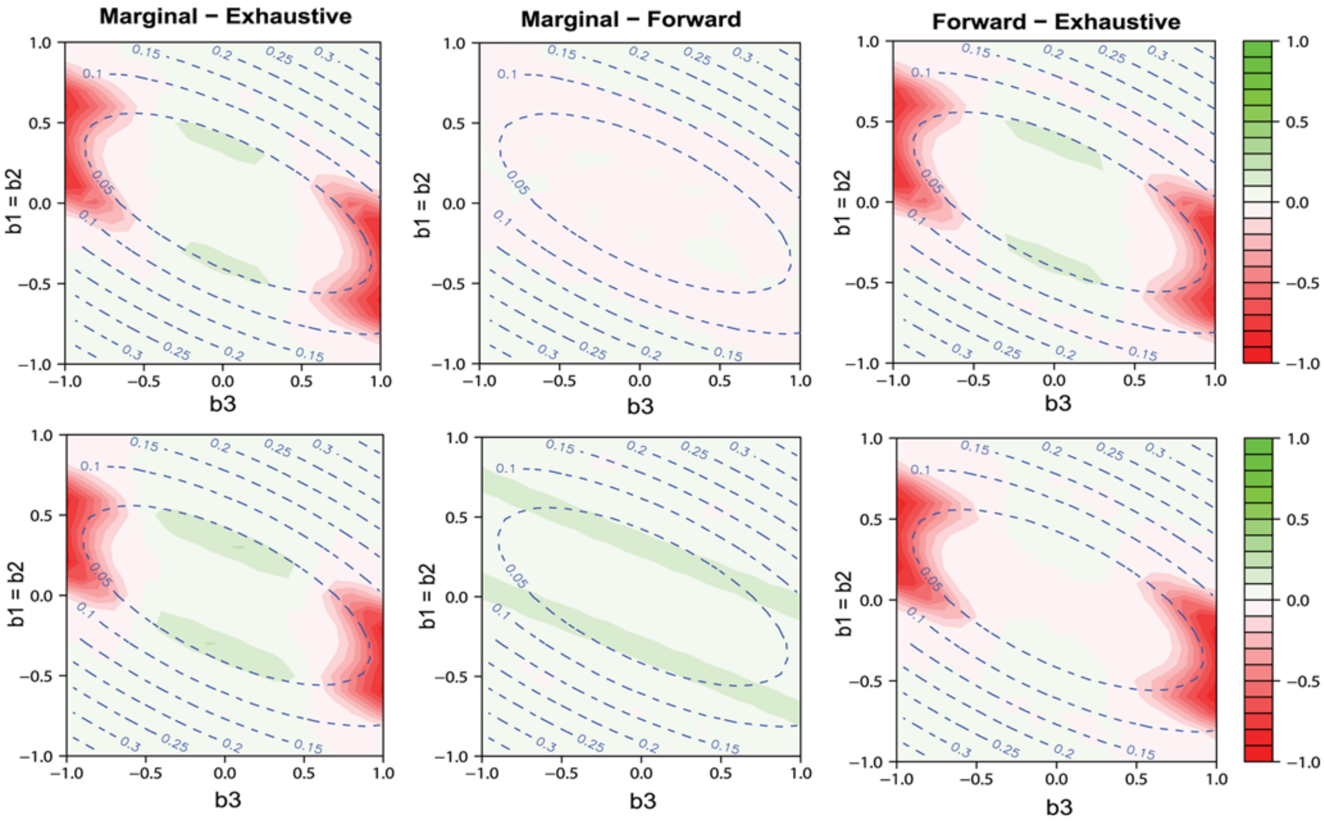

Figure 3. Comparisons among model selection power for detecting either true SNP over genetic model space. The power differences between marginal search and exhaustive search in the left column, between marginal search and forward search in the middle column, and between forward search and exhaustive search in the right column. Green areas indicate positive values of difference, and red areas indicate negative values of difference. We consider genetic models with the main effects $b_1 = b_2$ varying from −1 to 1 and the epistatic effect b_3 varying from −1 to 1. The allele frequency $q_j = 0.3$, $j = 1, \ldots, p$, and the false discovery number R is set to be 1 in row 1 and 10 in row 2.
doi:10.1371/journal.pgen.1000582.g003

pronounced under power definition (A) than under power definition (B). Marginal search performs better in the green areas where b_3 is small and b_1 and b_2 are both moderate. There are two reasons for the better performance of marginal search. First, with a small interaction term b_3 in these green areas, marginal search well detects the signals when the two-marker genetic effects are projected onto a marginal space through the simple regression of form (3). At the same time, with moderate b_1 and b_2, the power for these two methods is not close to 0 or 1, so that they are distinguishable. Second, marginal search considers fewer models so that the desired models are more likely to be found from the models with the best fit.

Under different power definitions, the performance of forward search relative to that of marginal search can change. Capable of including interaction terms, forward search has an advantage over marginal search in finding the full correct model under power definition (A), as shown by the red areas in the middle column of Figure 2. Based on the analytical formulas in the Methods section, there is a positive correlation between the test statistics in the first and second steps of forward search. Therefore, if one of the associated SNPs can be picked up in the first step, the contribution of the epistatic term makes forward search more powerful to identify the second correct SNP. Under power definition (B), the middle column of Figure 3 shows that marginal search always has similar or slightly better power than forward search, because forward search is less likely than marginal search to pick up a true SNP if an incorrect SNP is chosen first. The power of forward search will not improve greatly even if the number of false discoveries R increases.

As shown in the right column of Figure 2, exhaustive search under power definition (A) always has a similar or higher power to detect the true model when compared to forward search. Although forward search can also detect the interaction terms through joint analysis, its ability to capture the interaction terms is restricted, especially when marginal effect is small in the deep red areas of $b_1 + b_3(p_2 - q_2) \approx 0$. Under power definition (B), forward search is more powerful than exhaustive search when R, the number of controlled false discoveries, is small, but is less powerful when R is large. With small R (e.g. $R = 1$), forward search benefits from considering fewer models and is better than exhaustive search in the green areas of Figure 3. This benefit is reduced for larger R and will eventually be dominated by the advantage of exhaustive search. Since the first step of forward search is essentially a marginal search, the advantage of exhaustive search over marginal search also applies to forward search. This is reflected in the right columns of Figure 2 and Figure 3, where the red areas are similar to those in the left columns.

As reflected by the change of red/green areas between the first and the second rows in both Figure 2 and Figure 3, if we raise the number of allowed false discoveries R, the power of marginal search will increase the most, followed by the power of exhaustive search, and then the power of forward search. With the same increase in R, marginal search includes a much higher proportion of the models with true SNPs than exhaustive search. For forward search, the increase of power is smaller because it is more difficult to identify a correct SNP in the second step when an incorrect SNP is more likely to be selected in the first step.

We also explored additional model set-ups in Text S1 Section 3 with $n = 100$, $p = 1000$, $R = 1$, 5, and 10, $q_j = 0.3$ and 0.5, $j = 1, \ldots, p$, and $\sigma^2 = 3$. The values of the genetic effects $b_1 = b_2$ and b_3 varied from -2 to 2 by a step size of 0.2. When $q_j = 0.5$, the graphs are symmetric about $b_1 = b_2 = 0$ and $b_3 = 0$. In general, the patterns are similar to those shown in Figure 2 and Figure 3.

An Example of Power Comparisons Motivated from Real GWAS

In the following we provide an example to show how to apply our approach to calculating and comparing the power of model selection methods in empirical analysis. Because there are no consistently replicated interaction effects from real studies, we constructed hypothetical interaction models based on real data so that the marginal associations between traits and markers were matched, while allowing the interaction term to vary. Specifically, we calculated power based on a set of genetic models derived from a genome-wide association study of adult height by Weedon et al. [8]. Based on the reported 20 loci that putatively influence adult height, we set up a two-marker genetic model composed of SNPs rs11107116 and rs10906982, each of which showed moderate marginal effect. According to the Supplementary Table 4 in the original publication, the estimated marginal effects of rs11107116 and rs10906982 are respectively 0.045s.d. and 0.046s.d. with a sample standard deviation (s.d.) of height of 6.82 cm. Assuming different levels of interaction between the two SNPs (quantified by b_3), we estimated the parameters b_1, b_2, and σ^2 using model (2) so that the marginal effects matched the observed values. The Methods section gives the details of how these parameters were estimated. We used the set-up of Weedon's study: sample size $n = 16{,}482$, number of candidate SNPs $p = 402{,}951$, and the frequencies of the height-increasing allele for rs11107116 and rs10906982 $p_1 = 0.77$ and $p_2 = 0.48$, respectively.

Figure 4 shows the comparisons among the power of the three model selection methods over different values of b_3. For the detection of both SNPs, graphs A ($R = 1$) and C ($R = 20$) indicate that if the magnitude of epistasis b_3 is large, exhaustive search (red dashed curve) has significant advantage over forward search (green dotted curve), which is better than marginal search (black solid curve). If b_3 is small, marginal search has higher power than the other two. For the detection of at least one of the two SNPs, graphs B ($R = 1$) and D ($R = 20$) indicate that marginal search is similar or better than forward search; both methods are not affected by the variation of b_3. The relative performance of exhaustive search strongly depends on the magnitude of epistasis. Comparing graphs B ($R = 1$) and D ($R = 20$), it is clear that marginal search is superior over a larger region when a larger false discovery number R is tolerated.

With $R = 20$, graphs C and D indicate that exhaustive search is better than marginal search to find both or at least one of the SNPs when the magnitude of $b_3 > 0.3$ or 0.6, respectively. We studied the statistical significance of the interaction terms with the simulated data (1,000 runs) when b_3 equals these two cutoffs. When $b_3 = 0.3$, 11.4% of the simulations had the Bonferroni p-values (adjusted by the number of all possible pairs of the 20 found loci) that exceeded the significant threshold at 0.05. Therefore, a small epistatic effect, rarely showing significance from the observed data, can still make an exhaustive search more powerful than a marginal search under power definition (A). Under power definition (B), when $b_3 = 0.6$, 87.3% of the Bonferroni adjusted p-values were significant. That is, to make exhaustive search more powerful than marginal search for finding either SNP, a true epistatic effect needs to be large enough to often identify a statistically significant interaction.

This example demonstrates that the value of the interaction term and the number of false discoveries affect the relative performance of model selection methods, which can be one of the reasons for the conflicting results about the power of model selection methods in the existing literature [2,4]. Therefore, the suspected values of parameters such as epistatic effects can affect the researchers' choice of model selection methods.

Discussion

In this article, we have derived rigorous analytical results for the statistical power of three common model selection methods, and applied these results to compare the methods' performance for GWAS data. These results not only make the computationally expensive simulations unnecessary, but also systematically reveal how different genetic model parameters affect the power.

The comparison results among the three model selection methods illustrate the trade-off between searching the full model space and a reduced space. In one extreme, exhaustive search explores the full 2-dimensional space covering all possible epistatic effects, but it may reduce the probability that the true model(s) ranks among the top models because many more models are considered. In the other extreme, marginal search casts the true 2-dimensional model onto a 1-dimensional space without considering epistasis at all. However, we have a better chance to find more true positives when the marginal association is retained in the 1-dimenisonal space, because fewer models are examined and the false positive control appears comparatively liberal. Between these two extremes, forward search first considers marginal projection, and then partially searches the 2-dimensional space via residual projection given the chosen predictor in the first step. Thus, forward search has the partial benefit of joint analysis which considers epistatic effects conditionally. The stringency of its false positive control exists between those of exhaustive search and of marginal search.

The relative performance of these model selection methods also depends on the definition of power. Based on definition (A), exhaustive search performs the best in finding the true underlying genetic model in most of the model space considered. Under power definition (B), marginal search is a good choice: it is not much worse than exhaustive search for a large proportion of the model space, and it is always better than the classic forward search through which only one SNP is picked up in the first step. For most geneticists, finding at least one of the truly associated SNPs under power definition (B) is a primary concern, especially in the first stage of GWAS. Because we do not have prior information about the true genetic model in the beginning, marginal search, which is easy to compute, is a good start in the first stage of GWAS to find one or some of the main genetic effects. In the later stage(s), if the promising SNP candidates are limited, exhaustive search can be applied with less demanding computation, especially when epistasis among loci is of interest. Our conclusions based on the analytical studies justify this multi-stage strategy in GWAS.

Difference between Our Methods and Traditional Power Calculation and Simulations

Our power calculation for model selection strategies is different from a traditional power calculation for multiple regression models [9]. The traditional approach is to calculate the probability of accepting a specific multiple-regression model and rejecting the null hypothesis that the response and the covariates have no association, when controlling the type I error rate. This power calculation focuses on models instead of model selection methods, as it does not address any procedure of model selection. In

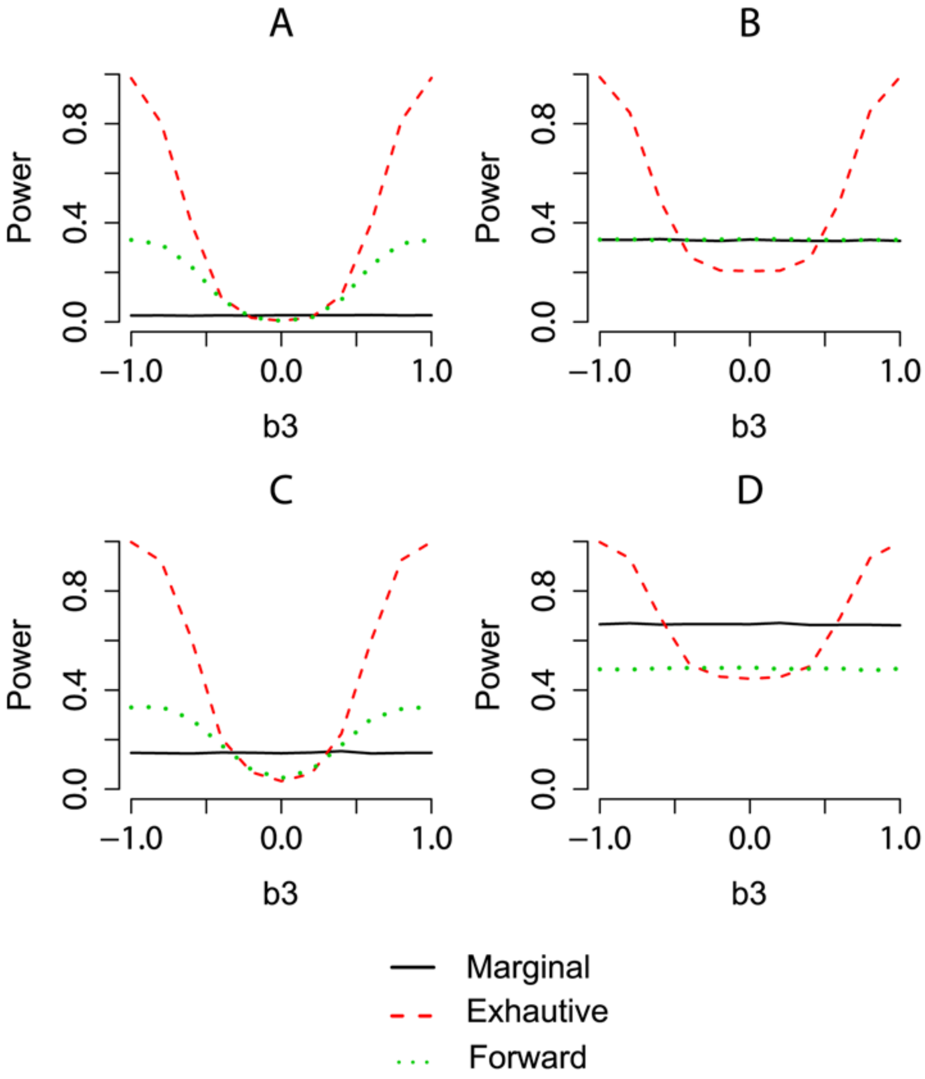

Figure 4. Plots of model selection power with given observed marginal effects. Power comparisons of three model selection procedures over a sequence of epistatic effect b_3: marginal search by black solid curve, exhaustive search by red dashed curve, and forward search by green dotted curve. We assume the true SNPs to be rs11107116 and rs10906982, which influence adult height with their marginal effects set to be the same as those observed in Weedon et al. 2008. Graphs A with $R = 1$ and C with $R = 20$ indicate the power of finding both SNPs; graphs B with $R = 1$ and D with $R = 20$ indicate the power of finding at least one of the two SNPs.
doi:10.1371/journal.pgen.1000582.g004

contrast, our analytical approach is to calculate the probability that a model selection method can pick up the models that contain the true covariates (true SNPs in GWAS).

Our analytical approach leads to new insights into model selection methods than simulations and limited real data analysis. Furthermore, our approach addresses a critical limitation of prior studies [4,5] that do not distinguish the models with all correct predictors from those with only a subset of the correct predictors. In those studies, the null distribution assumes the test statistic is from a model without any of the true predictors, and the alternative distribution assumes the statistic is from any model containing at least one true predictor (or, when considering the power for finding both true loci, the models with either true locus are ignored from the null distribution). This is a common problem of traditional multiple testing for model selection method, as pointed out by Storey et al. [2], who stated that "there is no statistically rigorous method to test for joint linkage, which exists only if both loci have nonzero terms in the full model." To address this issue, all involved models (including true, partially true, and wrong models) are considered and ranked by how well they fit the observed data. Our power calculation distinguishes the case that model selection procedures find the true model based on power definition (A) from the case that the procedures find a partially true model based on definition (B). We have derived the null and alternative distributions for each case, and thus provide the basis for model performance comparisons.

To compare the power of model selection methods, our approach explicitly considers the correlation structures among the test statistics for the null and alternative hypotheses, which achieves more accurate assessment of model selection methods than Bonferroni-corrected type I error control that is commonly used in the literature [4,5]. Bonferroni-based control is usually a conservative control when the test statistics are dependent on each other. As illustrated by both simulations (results not shown) and the theoretical derivations in the Methods section, the considered models and their test statistics usually exhibit complex correlation structures. Therefore Bonferroni-based control is not optimal as it only considers the number of models evaluated (that is, the number of hypothesis tests) and ignores correlation structures generated by different search strategies. The adequacy of our approach has been demonstrated through a good agreement between the analytical and the simulation results shown in Table 1 and Table 2. Furthermore, our study of correlation structures improves the understanding of the mechanism of different search strategies discovering genetic signals. For example, in forward search, the failure of the first stage is likely to cause the failure of the second stage even if there is a large epistatic effect, because the test statistics for the true predictors are positively correlated between the two stages.

Control Related to Type I Error Rate and False Discovery Proportion

To obtain the significance threshold, we control the number of false discoveries at R depending on how the power is defined. This control is practically meaningful and equals to the detection probability (DP) control [7] as discussed in the Results section. Furthermore, controlling the number of false discoveries is related to controlling the type I error rate. Since the type I error rate is defined as the probability of rejecting a hypothesis given it is a true null, with the definition of null models corresponding to the power definition (A) or (B), the estimation of component-wise type I error rate could be considered as

$$\hat{\alpha} = \frac{R}{\text{total } \# \text{ of null models}}.$$

The model selection problem is also a large-scale simultaneous hypothesis testing problem. A widely applied significance control criterion in this scenario is the false discovery rate (FDR) [10]. The false discovery number control in our study is also related to the control of the false discovery proportion (FDP), which is an estimate of FDR. Under power definition (A)

$$FDP = \frac{R}{R + i \times power(R)},$$

where $power(R)$ denotes the power calculated based on the number of selected null models R, and i indicates the number of correct models: $i=2$ for marginal search, and $i=1$ for exhaustive search and forward search.

On the Derivation of Asymptotic Distributions

Through the simulations in the Results section, our derivation of asymptotic distributions is shown to be accurate for moderately small genetic effects when the sample size $n=100$. Since the asymptotic derivation assumes large sample size, the power calculation results should provide accurate approximations for reasonably smaller genetic effects in GWAS which have a much larger number of observations in general. The asymptotic derivation has several benefits. First, we can derive the results for the models with random predictors. Because genotypes are randomly observed in genetic studies, it is necessary to consider such models. Traditional methods for deriving the non-central F distributions for the test statistics are based on fixed predictors [7,11,12]. As functions of predictor variables, these non-central parameters are not statistically consistent when genotypes are random. Although one may integrate the power over all possible configurations of markers [13], it is very cumbersome unless n is small. Our method, based on asymptotic theorems, provides a satisfactory solution for models with random predictors. Our novel approach presented here can be applied to derive the distributions of such models' test statistics. Second, the derived asymptotic multivariate normal distributions for theoretical null and alternative hypotheses allow us to incorporate complex correlations among the test statistics into power calculation based on population parameters. For a given GWAS data set, the correlations presented in the data may also be addressed by empirical estimation of the null hypothesis [14,15]. Third, the ideas behind the asymptotic derivation can be applied to study the distributions for hypothesis testing and power calculation in general as long as the statistics have certain functions of random variables.

On Simplifying Assumptions

We have assumed that the markers are independent in this paper. There may be linkage disequilibrium (LD) among SNPs. However, LD in general is weak among tagging SNPs [16–18]. Furthermore, simulations based on real GWAS data (results not shown) indicate that even in the presence of LD, our analytical results are quite accurate when more false positives are acceptable, i.e. a large R value. In addition, the analytical power approximations are more accurate for power definition (B) than for definition (A). In general, when the dependency among true SNPs and the ensemble of unrelated SNPs is weak or moderate, our power calculation provides acceptable approximations.

In reality, the underlying true model could be more complicated than model (2) with more related SNPs and interactions. Our analytical results of power calculation can be extended through the approaches similar to the one we developed here. Although the genetic models studied are simple, our results provide insights into the relative performance of different model selection procedures.

Methods

Asymptotic Distribution Results

To calculate the power of model selection procedures shown in the Results section, we first derive general results on the asymptotic distributions. Let $\mathbf{Z}_i = (\mathcal{Z}_{i1}, \ldots, \mathcal{Z}_{is})$, $i=1, \ldots, n$, be n independent and identically distributed (iid) random vectors of dimension s. Assume the mean vector is $\boldsymbol{\theta} = E(\mathbf{Z}_i) = (\theta_1, \ldots, \theta_s)$ with $\theta_j = E(\mathcal{Z}_{ij})$ and the variance-covariance matrix is $\Sigma = Cov(\mathbf{Z}_i)$ with $(\Sigma)_{jk} = Cov(\mathcal{Z}_{ij}, \mathcal{Z}_{ik})$, $j, k=1, \ldots, s$. Let $\bar{\mathbf{Z}} = (\bar{Z}_1,\ldots,\bar{Z}_s)$, where $\bar{Z}_j = \frac{1}{n}\sum_{i=1}^{n} Z_{ij}$. Considering a real valued function $h(\bar{\boldsymbol{Z}})$ of $\bar{\boldsymbol{Z}}$, if $\nabla h(\boldsymbol{\theta}) = \left(\frac{\partial h(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial h(\boldsymbol{\theta})}{\partial \theta_s}\right)' \neq 0$, we have

$$\sqrt{n}[h(\bar{\boldsymbol{Z}}) - h(\boldsymbol{\theta})] \xrightarrow{L} N(0,\tau^2), \tag{5}$$

where $\tau^2 = [\nabla h(\boldsymbol{\theta})]'\Sigma[\nabla h(\boldsymbol{\theta})]$ and $\xrightarrow{L}$ denotes the convergence in law [19].

We extend the above result in two ways to suit our needs of deriving the distribution of test statistics that are examined in model selection procedures (the proofs are given in Text S1 Section 1.1). First, we consider two real valued functions $h_1(\bar{\boldsymbol{Z}})$ and $h_2(\bar{\boldsymbol{Z}})$ of the same sample mean $\bar{\boldsymbol{Z}}$. If $\nabla h_1(\boldsymbol{\theta}) \neq 0$ and $\nabla h_2(\boldsymbol{\theta}) \neq 0$, we have the convergence in probability that:

$$Cov\left(\sqrt{nh_1}(\bar{\boldsymbol{Z}}), \sqrt{nh_2}(\bar{\boldsymbol{Z}})\right) \xrightarrow{P} [\nabla h_1(\boldsymbol{\theta})]'\Sigma[\nabla h_2(\boldsymbol{\theta})]. \tag{6}$$

Secondly, if $\nabla h(\theta) = 0$, the asymptotic distribution of $nh(\bar{\boldsymbol{Z}})$ is:

$$n[h(\bar{\boldsymbol{Z}}) - h(\boldsymbol{\theta}) \xrightarrow{L} c\chi_d^2, \tag{7}$$

where if $A \equiv D^2 h(\boldsymbol{\theta})\boldsymbol{\Sigma}$, with $D^2 h(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \theta^2} h(\boldsymbol{\theta})$ being the Hessian matrix of $h(\boldsymbol{\theta})$, we have

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

Power Calculations

With the results above, we derive the relevant distributions of T- and F-statistics associated with three types of regression models, which will be used for calculating the power of model selection methods. Specifically, F_{12} is the F statistics for the correct model in which both SNPs are true. T_i and F_{ij}, $i = 1, 2, j = 3, \ldots, p$, are test statistics for "half" correct models in which only one SNP is truly associated. T_j and F_{lk}, $3 \leq l < k \leq p$, are the statistics for incorrect models in which neither SNP in the models is associated with the phenotype. Complex correlations exist among the models even with the assumption of independence among SNPs. The correlations come from two sources. First, since the quantitative trait is associated with both SNPs 1 and 2, the fitted regression models containing either of these SNPs have correlated test statistics. Second, models sharing a common SNP (no matter it is true or wrong) also have correlated test statistics. To allow correlations, we therefore explore the marginal and the joint distributions of various test statistics for different models, and then derive how likely a "half" correct model would stand out from incorrect models, as well as how likely a correct model would outperform "half" correct models or incorrect models.

Marginal Search

Statistics and asymptotic distributions. To calculate the power of marginal search, we need to obtain the distributions of the involved test statistics. We first derive the T-statistic for the two true SNPs in the marginal model. In the simple regression model involving the first true SNP (SNP 1), i.e. $\hat{Y}_{i(1)} = \hat{\beta}_{0(1)} + \hat{\beta}_{1(1)} X_{i1}$, the corresponding T-statistic has the following asymptotic distribution (see Text S1 Section 2.1 for proof):

$$T_1 - \sqrt{n h_1(\boldsymbol{\theta})} \xrightarrow{L} N(0, \tau_1^2),$$

where

$$h_1(\boldsymbol{\theta}) = \frac{\sqrt{2p_1 q_1}(b_1 + b_3(p_2 - q_2))}{\sqrt{2p_2 q_2(b_2^2 + 2b_2 b_3(p_1 - q_1) + b_3^2(p_1^2 + q_1^2)) + \sigma^2}}, \quad (8)$$

and the formula of τ_1^2 (a constant of n) is given in Text S1 Section 4.1. For the marginal model of the second SNP (SNP 2), the asymptotic distribution of T_2 is gotten by symmetry between indices 1 and 2.

Based on the asymptotic mean of T_1 derived above, we can quantify the influence of genetic parameters of SNP 2 and epistasis on the power of marginal search to pick up SNP 1. As for some genetically interesting observations, when there is no epistatic effect (i.e. $b_3 = 0$), we have $h_1(\boldsymbol{\theta}) = \frac{b_1\sqrt{2p_1 q_1}}{\sqrt{2p_2 q_2 b_2^2 + \sigma^2}}$, which means the magnitude of marginal association of X_1 and thus the power of marginal search to find X_1 are decreasing functions of the main effect of X_2, the minor allele frequency (MAF) of X_2, and the random error variance σ^2, with the decreasing rate specifically given by $\sqrt{2p_2 q_2 b_2^2 + \sigma^2}$. When epistasis exists (i.e. $b_3 \neq 0$) but $b_1 = 0$, $h_1(\boldsymbol{\theta})$ reflects the marginally projected signal of epistasis, which is still a decreasing function of the MAF of X_2. The influence of b_2 depends on the allele frequencies p_1 and q_1. On the other hand, if $b_1 \neq 0$, it is possible that $b_1 + b_3(p_2 - q_2) = 0$ when the main effect b_1 and interaction effects b_3 have opposite directions (assuming q_2 is the MAF). With such epistatic pattern, marginal detection surely fails to detect the true genetic variants no matter how strong the true genetic effects are.

Now we derive the joint distribution of T_1 and T_2. Since Y is a function of both X_1 and X_2 in the underlying true model (2), T_1 and T_2 are correlated even when X_1 and X_2 are independent and do not interact, i.e. $b_3 = 0$. The correlation between T_1 and T_2 can be substantial in certain genetic models. The asymptotic joint distribution of $(T_1, T_2)'$ is

$$(T_1, T_2)' - \boldsymbol{\mu}_{T_1,T_2} \xrightarrow{L} MVN(\mathbf{0}, \boldsymbol{\tau}_{\mathrm{T}_1,\mathrm{T}_2}), \quad (9)$$

where $\mu_{T_1,T_2} = \sqrt{n}(h_1(\boldsymbol{\theta}), h_2(\boldsymbol{\theta}))'$, $\boldsymbol{\tau}_{T_1,T_2} = \begin{pmatrix} \tau_1^2 & \tau_{1,2} \\ \tau_{1,2} & \tau_2^2 \end{pmatrix}$, $\tau_i^2 = Var(T_i)$, $i = 1, 2$, and $\tau_{1,2} = Cov(T_1, T_2)$. The covariance $\tau_{1,2}$ is gotten based on the result in (6), and its formula (as a constant of n) is given in Text S1 Section 4.1.

Let T_j, $j = 3, \ldots, p$, be the T-statistic from model (3) for a wrong SNP j, according to the asymptotic result in (5),

$$T_j \xrightarrow{L} N(0,1), \quad (10)$$

which holds regardless of the allele frequencies and the underlying true genetic model. The proof for T_3 as an example is provided in Text S1 Section 2.2. It can be shown that T_j is also independent of T_1 and T_2 according to the result in (6). Under the assumption of fixed design matrix, T_j has a T distribution with $n-2$ degrees of freedom based on a traditional linear model analysis [6,12]. This null distribution is still asymptotically valid for random predictors since the T distribution converges to the standard normal as $n \rightarrow \infty$.

Power of marginal search procedure. Based on the above results for the distributions of T-statistics, we first calculate the power of marginal search under power definition (A). If the marginal search is allowed to contain R wrong SNPs, i.e. the number of false discoveries is controlled by R, the power of identifying both true SNPs is just the probability that both $|T_1|$ and $|T_2|$ are greater than the Rth largest value in the set $\{T_j, j \geq 3\}$:

$$P(|T_1| \wedge |T_2| \geq |T|_{(r)}) = \iint P(|T|_{(r)} \leq |t_1| \wedge |t_2|) g(t_1, t_2) dt_1 dt_2$$

where $|T_1| \wedge |T2| = \min\{|T_1|, |T_2|\}$, $r = p - 2 - R + 1$, $|T|_{(r)}$ is the rth smallest (or the Rth largest) order statistics of $|T_j|$, $j = 3, \ldots, p$, and $g(t_1, t_2)$ is the joint probability density function (PDF) of $(T_1, T_2)'$ given in (9). Let $\Phi(\cdot)$ be the cumulative distribution function (CDF) of $N(0, 1)$, then

$$P\left(|T|_{(r)} \leq |t_1| \wedge |t_2|\right) = (1 - 2\Phi(-|t_1| \wedge |t_2|))^r \sum_{l=0}^{p-2-r} \binom{r+l-1}{l} (2\Phi(-|t_1| \wedge |t_2|))^l.$$

To get the power of marginal search under definition (B) that either SNP 1 or SNP 2 is selected, we calculate the probability that either $|T_1|$ or $|T_2|$ is larger than the random cutoff point: $P(|T_1|\bigvee|T_2|\geq|T|_{(r)})$, where $|T_1|\bigvee|T_2|=\max\{|T_1|,\ |T_2|\}$.

Exhaustive Search

Statistics and asymptotic distributions. The distributions of the relevant test statistics are derived first for calculating the power of exhaustive search. We first get the joint distribution of the test statistics involving true SNPs 1 and 2: T_1, T_2, and F_{12}. Define $T_{12}\equiv\sqrt{F_{12}}$. Based on the asymptotic distribution result in (5) (see Text S1 Section 2.3 for details of derivation), we have

$$(T_{12},T_1,T_2)'-\boldsymbol{\mu}_{T_{12},T_1,T_2}\xrightarrow{L}MNV(\mathbf{0},\boldsymbol{\tau}_{T_{12},T_1,T_2}), \tag{11}$$

where $\boldsymbol{\mu}_{T_{12},T_1,T_2}=\sqrt{n}(h_{12}(\boldsymbol{\theta}),h_1(\boldsymbol{\theta}),h_2(\boldsymbol{\theta}))'$ and $\boldsymbol{\tau}_{T_{12},T_1,T_2}=\begin{pmatrix}\tau_{12}^2 & \tau_{12,1} & \tau_{12,2}\\ \tau_{12,1} & \tau_1^2 & \tau_{1,2}\\ \tau_{12,2} & \tau_{1,2} & \tau_2^2\end{pmatrix}$.

The formula of $h_1(\boldsymbol{\theta})$ is given in (8), and

$$h_{12}(\boldsymbol{\theta})=\left(\frac{2}{3\sigma^2}(b_1^2p_1q_1+b_3^2p_1q_1(p_2^2+q_2^2)+2b_1b_3p_1q_1(p_2-q_2)\right.$$
$$\left.+p_2q_2(b_2+b_3(p_1-q_1))^2\right)^{1/2}. \tag{12}$$

The formulas of $\tau_{12}^2=Var(T_{12})$ and $\tau_{12,i}=Cov(T_{12},\ T_i)$, $i=1,\ 2$, are independent of n and are given in Text S1 Section 4.1.

We then derive the F-statistics for the incorrect models in form (4) to fit Y with X_j and X_k, $3\leq j<k\leq p$. Following the result in (7), F_{jk} has a common marginal asymptotic distribution:

$$F_{jk}\xrightarrow{L}\frac{1}{3}\chi_3^2. \tag{13}$$

With F_{34} as an example, the detailed proof is given in Text S1 Section 2.4.

Based on the traditional power calculation for regression models, the null model is the incorrect model with neither SNP associated with the phenotype. When the design matrix is fixed, the null distribution of F_{jk} is an F distribution with degrees of freedom (3, $n-4$) [12]. Result (13) indicates the F distribution for null is also valid when the genotypes are treated as random variables, because F(3, $n-4$) converges to $\frac{1}{3}\chi_3^2$ when n is large.

In order to calculate the power of model selection methods, we need to address the correlation structures among involved statistics. The statistics are correlated when two epistatic models in form (4) share a common SNP. Also, F-statistics involving X_1 and those involving X_2 are correlated because the true underlying model includes both SNPs. Consequently, the elements in the set $\{F_{12},\ F_{ij},\ i=1,\ 2, j=3,\ldots,p\}$ are all correlated with each other. To capture the important dependency, we decompose F-statistics as follows:

$$F_{ij}\rightarrow\frac{1}{3}F_i+\frac{2}{3}\left(h_i^2(\boldsymbol{\theta})+1\right)F_{j|i}, \tag{14}$$

when $i=1$, $h_1(\boldsymbol{\theta})$ is given by equation (8). The detailed proof for decomposing F_{13} as an example is shown in Text S1 Section 1.2. Through this decomposition, the correlation between F_{ij} and F_{ik} can be explained by F_i while we treat $F_{j|i}$ and $F_{k|i}$ to be independent. Furthermore, with the result (14) we can use the joint distribution (11) to capture the correlation between F_{12} and F_{ij}.

Based on the asymptotic distribution in (7), we have

$$F_{j|i}\overset{(d)}{\cong}c\chi_d^2, \tag{15}$$

where $i=1,\ 2, j=3,\ \ldots,\ p$, $c=v/2e$, and $d=2e^2/v$, with $E(F_{j|i})\rightarrow e$ and $Var(F_{j|i})\rightarrow v$. Text S1 Section 2.5 shows the detailed proof for $F_{3|1}$. The formulas of e and v are given in Text S1 Section 4.2. Based on our numerical studies (results not shown), c is close to $1/2$ and d is close to 2 in a large proportion of the parameter space of $\{q_i,\ q_j,\ \mathbf{b},\ \sigma^2\}$ (e.g. when allele frequencies q_i and q_j do not converge to 0 or 1, genetic effect $\mathbf{b}=(b_1,\ b_2,\ b_3)'$ and random error variance σ^2 are not too large). When $c=1/2$ and $d=2$, $c\chi_d^2$ is asymptotically equivalent to an F distribution with degrees of freedom 2 and $n-4$. F(2, $n-4$) is the distribution of $F_{j|i}$ when X is fixed [6,12]. Our results demonstrate that for the random design matrix, the weighted chi-square distribution (15) is more appropriate.

Power of exhaustive search procedure. With the distribution of test statistics derived above, we first calculate the probability of exhaustive search to identify the exact true model. Under power definition (A), the test statistic F_{12} for the exact true model corresponds to the "alternative" distribution, whereas the F-statistics for all other models such as totally incorrect models and "half" correct models are combined together to generate a mixed "null" distribution. Let $S_1\equiv\{F_{ij},\ i=1,\ 2,\ j=3,\ldots,p\}$, $S_2\equiv\{F_{jk},\ 3\leq j<k\leq p\}$, and $F_{S,[R]}$ denote the Rth largest variable in a set S. When controlling the false discovery number by R, the probability of detecting the exact true model (2) is

$$P\left(F_{12}\geq F_{S_1\cup S_2,[R]}\right)=\iiint P\left(t_{12}^2\geq F_{S_1{}'\cup S_2,[R]}\right)g(t_{12},t_1,t_1)d(t_{12},t_1,t_1),$$

where $g(t_{12},t_1,t_2)$ is the PDF of (11), $S_1'=\left\{\frac{t_i^2}{3}+\frac{2\left(1+h_i^2(\boldsymbol{\theta})\right)F_{j|i}}{3},i=1,2;\ j=3,\ldots,p\right\}$ from the decomposition (14), and

$$P\left(t_{12}^2\geq F_{S'_1\cup S_2,[R]}\right)=\sum_{r=0}^{R-1}\sum_{\{r_1,r_2,r\}\in S_r}P_1P_2P_3,$$

in which

$$S_r=\left\{\{r_1,r_2,r_3\}:\sum r_i=r,\ 0\leq r_1,\ r_2\leq p-2,\ 0\leq r_3\leq N\right\},$$

$$P_1\binom{p-2}{r_1}\left[1-G_{11}\left(\frac{3t_{12}^2-t_1^2}{2(1+h_1^2(\boldsymbol{\theta}))}\right)\right]^{r_1}G_{11}\left(\frac{3t_{12}^2-t_1^2}{2(1+h_1^2(\boldsymbol{\theta}))}\right)^{p-2-r_1},$$

$$P_2\binom{p-2}{r_2}\left[1-G_{12}\left(\frac{3t_{12}^2-t_2^2}{2(1+h_2^2(\boldsymbol{\theta}))}\right)\right]^{r_2}G_{12}\left(\frac{3t_{12}^2-t_2^2}{2(1+h_2^2(\boldsymbol{\theta}))}\right)^{p-2-r_2},$$

$$P_3\binom{N}{r_3}\left[1-G_2\left(t_{12}^2\right)\right]^{r_3}G_2\left(t_{12}^2\right)^{N-r_3},$$

$N=\binom{p-2}{2}$ is the number of variables in S_2, $G_{1i}(\bullet)$ is the CDF of distribution (15) for $i=1,\ 2$, and $G_2(\bullet)$ is the CDF of distribution (13). The test statistics within the sets $S^*\equiv\{F_{j|1},\ F_{j|2},\ j=3,\ldots,p\}$ and

$S_2 \equiv \{F_{jk},\ 3 \le j < k \le p\}$ are treated as asymptotically independent as $p \longrightarrow \infty$ (see Text S1 Section 1.3 for details).

According to the power definition (B), the probability of exhaustive search to detect at least one of the associated SNPs is
$$P\left(\max(\{F_{12}\} \cup S_1) \ge F_{S_2,[R]}\right) = 1 - \iiint P_{t_{12},t_1,t_2} g(t_{12},t_1,t_2) d(t_{12},t_1,t_2),$$

where

$$\begin{aligned} P_{t_{12},t_1,t_2} &= P\left(\max\left(\{t_{12}^2\} \cup S'_1\right) < F_{S_2,[R]}\right) \\ &= \int P\left(\max\left(\{t_{12}^2\} \cup S'_1\right) < c\right) g_{2(N-R+1)}(c) dc \\ &= \int_{t_{12}^2}^{\infty} \left[G_{11}\left(\frac{3c - t_1^2}{2(1+h_1^2(\boldsymbol{\theta}))}\right) G_{12}\left(\frac{3c - t_2^2}{2(1+h_2^2(\boldsymbol{\theta}))}\right)\right]^{p-2} g_{2(N-R+1)}(c) dc, \end{aligned}$$

$g_{2(N-R+1)}(\bullet)$ is the PDF of the $(N-R+1)$th order statistics distribution with the following density function:

$$g_{2(N-R+1)}(c) = \frac{N!}{(N-R)!(R-1)!} G_2(c)^{N-R} [1 - G_2(c)]^{R-1} g_2(c),$$

$G_2(\bullet)$ and $g_2(\bullet)$ are the CDF and PDF of the distribution of (13) respectively.

If R is neither too small nor too large, i.e. $R/N \longrightarrow c$, $0<c<1$, as $N \longrightarrow \infty$, we can use quantiles to replace the order statistics in order to simplify the calculation [20], i.e., $F_{S_2,[R]} \rightarrow G_2^{-1}\left(\frac{N-R+0.5}{N}\right) \equiv Q$. So for a given (t_{12},t_1,t_2), we can approximately replace the integrand P_{t_{12},t_1,t_2} with

$$I_{\{t_{12}^2 \le Q\}}(t_{12}) \left[G_{11}\left(\frac{3Q - t_1^2}{2(1+h_1^2(\boldsymbol{\theta}))}\right) G_{12}\left(\frac{3Q - t_2^2}{2(1+h_2^2(\boldsymbol{\theta}))}\right)\right]^{p-2},$$

where $I_A(x)$ denotes the indicator function of set A. Simulations (results not shown) illustrate that the approximation of integrand is reasonably accurate for the integration.

Forward Search

Statistics and asymptotic distributions. For forward search, first we derive the distributions of test statistics, which will be used to calculate the corresponding statistical power. Here we need to handle the comparison between two models: the model with SNPs 1 and j, $j = 3, \ldots, p$, taking form (4), and the model with SNP j taking form (3). Let $F_{1|j}$ be the F statistic measuring the significance of the extra terms in the bigger model over the smaller model [6]. Define $T_{1|j} \equiv \sqrt{F}_{1|j}$. When $b_1 + b_3(p_2 - q_2) \ne 0$, following the asymptotic result in (5), we can derive

$$T_{1|j} - \mu_{T_{1|j}} \xrightarrow{L} N\left(0.\tau_{i|j}^2\right),$$
$$\mu_{T_{1|j}} = \left(\frac{np_1q_1(b_1 + b_3(p_2 - q_2))^2}{2p_2q_2\left((b_2 + b_3p_1)^2 - 2b_2b_3q_1 + b_3^2q_1^2\right) + \sigma^2}\right)^{1/2}, \qquad (16)$$

and the formula for $\tau_{T_{1|j}}^2$ with $j = 3$ as an example is provided in Text S1 Section 4.3. Both $\mu_{T_{1|j}}$ and $\tau_{T_{1|j}}^2$ do not depend on the allele frequency p_j. When $b_1 + b_3(p_2 - q_2) = 0$, $T_{1|j}^2$ has a $\frac{1}{2}\chi_2^2$ distribution by (7). Similarly we can get the asymptotic distribution of $T_{2|j} \equiv \sqrt{F_{2|j}}$ when comparing the model having SNPs 2 and j in form (4) with the model having SNP j in form (3). The covariance $Cov(T_{1|j}, T_{2|j})$ can also be calculated. As an example the formula of $Cov(T_{1|3}, T_{2|3})$ is given in Text S1 Section 4.3.

Moreover, the statistics $(T_1, T_2, T_{1|j}, T_{2|j})'$ involving true SNPs have a multivariate normal distribution:

$$\left(T_1, T_2, T_{1|j}, T_{2|j}\right)' - \boldsymbol{\mu}_{T_1,T_2,T_{1|j},T_{2|j}} \xrightarrow{L} MNV(\mathbf{0}, \boldsymbol{\tau}_{T_1,T_2,T_{1|j},T_{2|j}}). \quad (17)$$

When $j = 3$, the details of the calculation and the formulas for $\boldsymbol{\mu}_{T_1,T_2,T_{1|j},T_{2|j}}$ and $\boldsymbol{\tau}_{T_1,T_2,T_{1|j},T_{2|j}}$ are given in Text S1 Sections 2.6 and 4.6.

Through result (6), we have proved that T_j and $F_{1|j}$ are asymptotically independent (refer to Text S1 Section 2.6 for details), i.e.

$$Cov(T_j, T_{1|j}) \rightarrow 0, j = 3, \ldots, p. \qquad (18)$$

When comparing the model having two incorrect SNPs j and k ($3 \le j < k \le p$) in form (4) with the model having SNP j in form (3), the corresponding F-statistic $F_{k|j}$ has the asymptotic distribution

$$F_{k|j} \xrightarrow{L} \frac{1}{2}\chi_2^2. \qquad (19)$$

Based on the result in (7), Text S1 Section 2.7 shows the proof for (19) with $j = 3$ and $k = 4$ as an example. This distribution is consistent with $F(2, n-4)$ which can be derived with the fixed design matrix and is routinely used for $F_{k|j}$ in the traditional model comparison [6,12].

Power of forward search procedure. In the forward search procedure, we first apply marginal search to find the most significant SNP among models (3). Based on the selected SNP, we then fit models (4) in the second step to find the SNPs that have strong joint effects, while controlling for R false discoveries. Under power definition (A) for finding the exact true model, we need to calculate the probability of forward search to choose SNP 1 or 2 in the first step, and then pick up the true model in the second step. Define $i^* \equiv \text{argmax}_{i=1,2}\{|T_i|\}$, $S_{i^*} \equiv \{F_{i^*3}, \ldots, F_{i^*p}\}$, as $p \longrightarrow \infty$, we can write the power as

$$\begin{aligned} &P\left(|T_{i^*}| > |T|_{(p-2)} \cap F_{12} \ge F_{S_{i^*},[R]}\right) \\ &= \iiint P\left(|t_{i^*}| > |T|_{(p-2)} \cap t_{12}^2 \ge F_{S'_{i^*}[R]}\right) g(t_{12},t_1,t_2) d(t_{12},t_1,t_2) \\ &\rightarrow \iiint P\left(|t_{i^*}| > |T|_{(p-2)}\right) P\left(t_{12}^2 \ge F_{S'_{i^*}[R]}\right) g(t_{12},t_1,t_2) d(t_{12},t_1,t_2) \end{aligned}$$

where $g(t_{12},t_1,t_2)$ is the PDF of $(T_{12},T_1,T_2)'$ given in (11), $|T|_{(p-2)} = \max_{j \ge 3}\{|T_j|\}$, $S'_{i^*} = \left\{\frac{1}{3}t_{i^*}^2 + \frac{2}{3}\left(1 + h_{i^*}^2(\boldsymbol{\theta})\right) F_{j|i^*}, j=3,\ldots p\right\}$ by F-statistic decomposition (14), and

$$P\left(|t_{i^*}| > |T|_{(p-2)}\right) = (1 - 2\Phi(-(|t_1| \vee |t_2|)))^{p-2},$$

$$P\left(t_{12}^2 \ge F_{S'_{i^*},[R]}\right) = G_{1i^*}(u)^r \sum_{l-0}^{p-2-r} \binom{r+l-1}{l} [1 - G_{1i^*}(u)]^l,$$

where $u = \frac{3t_{12}^2 - t_{i^*}^2}{2(1+h_{i^*}^2(\theta))}$, $h_{i^*}(\boldsymbol{\theta})$ is given in (8) for $i^* = 1$, $r = p-2-R+1$, and $G_{1i^*}(\cdot)$ is the CDF of the distribution for $F_{j|i^*|}$ given in (15). i^* is fixed for an observed value $(t_1,t_2)'$ of random vector $(T_1,T_2)'$, so it is easy to implement the power calculation with Monte Carlo integration.

Note that $|T|_{(p-2)}$ and $F_{S'_{i^*},[R]}$ are asymptotically independent. This is because $corr\left(\left|T_j\right|,F_{j|i^*}\right)<1$ for each $j\geq 3$, so with $p\rightarrow\infty$, $P\left(j^*\neq k^*:\left|T_{j^*}\right|=|T|_{(p-2)},F_{k^*|i^*}=F_{S'_{i^*},[R]}\right)\rightarrow 1$. But when $j^*\neq k^*$, $\left|T_{j^*}\right|$ and $F_{k^*|i^*}$ are always independent.

When R and p are large, we can simplify the formula of $P\left(t_{12}^2\geq F_{S'_{i^*},[R]}\right)$ by approximating the Rth largest variable in set $\left\{F_{j|i^*},j=3,\ldots,p\right\}$ with $G_{1i^*}^{-1}\left(1-\frac{R-0.5}{p-2}\right)$, where $G_{1i^*}^{-1}$ is the quantile function of $F_{j|i^*}$. So we can approximately replace $P\left(t_{12}^2\geq F_{S'_{i^*},[R]}\right)$ with $I_{\left\{u>G_{1i^*}^{-1}\left(1-\frac{R-0.5}{p-2}\right)\right\}}(u)$ for calculating the integration.

Under power definition (B), the power of forward model selection method is the sum of P_A: the probability to detect SNP 1 or 2 in the 1st step, and P_B: the probability that step 1 fails but step 2 picks up at least one correct SNP, while controlling for R incorrect models as false positives. Specifically,

$$\begin{aligned}P_A&=P\left((|T_1|\vee|T_2|)>|T|_{(p-2)}\right)\\&=\iint(1-2\Phi(-|t_1|\vee|t_2|))^{p-2}g(t_1,t_2)dt_1dt_2,\end{aligned}$$

where $g(t_1,t_2)$ is the PDF of joint distribution of $(T_1,T_2)'$ given in (9). Defining $j^*\equiv\arg\max_{k\geq 3}\{|T_k|\}$ and $S_{j^*}\equiv\left\{F_{k|j^*},k\geq 3,k\neq j^*\right\}$, we have

$$P_B=P\left(\left\{\left|T_{j^*}\right|>(|T_1|\vee|T_2|)\right\}\cap\left\{\left(F_{1|j^*}\vee F_{2|j^*}\right)\geq F_{S_{j^*},[R]}\right\}\right).$$

For each $k\geq 3$, $F_{i|k}$ and T_k are independent, so $F_{i|j^*}$ and T_{j^*} are independent. Given the results in (16) and (19), the distribution of $F_{i|j^*}$ does not depend on j^*. Hence, $F_{i|j^*}$ has the same distribution as $F_{i|j}$, $j=3,\ldots,p$. We then have

$$P_B=\oint P_{t_1t_2}P_{t_{1|j}t_{2|j}}g(t_1,t_2,t_{1|j},t_{2|j})\ d(t_1,t_2,t_{1|j},t_{2|j}),$$

where $g(t_1,t_2,t_{1|j},t_{2|j})$ is the PDF of $(T_1,T_2,T_{1|j},T_{2|j})'$ given in (17),

$$P_{t_1t_2}=P\left(|T|_{(p-2)}>(|t_1|\vee|t_2|)=1-(1-2\Phi(-|t_1|\vee|t_2|))\right)^{p-2},$$

$$\begin{aligned}P_{t_{1|j}t_{2|j}}&=P\left(\left(t_{1|j}^2\vee t_{2|j}^2\right)\geq F_{S_j,[R]}\right)\\&=G\left(t_{1|j}^2\vee t_{2|j}^2\right)^r\sum_{l=0}^{p-3-r}\binom{r+l-1}{l}\left[1-G\left(\left(t_{1|j}^2\vee t_{2|j}^2\right)\right)\right]^l,\end{aligned}$$

in which $r=p-3-R+1$, and $G(\bullet)$ is the CDF of $F_{k|j}$, $3\leq j<k\leq p$, given in (19). We can approximate $F_{S_j,[R]}$ through the quantile function $G^{-1}\left(1-\frac{R-0.5}{p-3}\right)$ to simplify the calculation of integration.

Calculating Post-Hoc Power with a Given Marginal Model

To demonstrate how to evaluate the power of model selection methods in the empirical analysis, we have applied our approach in a real study example. In this example, the simple regression model on X_1, $\hat{Y}_{i(1)}=\hat{\beta}_{0(1)}+\hat{\beta}_{1(1)}X_{i1}$, is an estimate of marginal model

$$E(Y_i|X_{i1})=b_0+b_2EX_{i2}+(b_1+b_3EX_{i2})X_{i1},$$

based on the full model (2). So the estimator of main effect is $\hat{b}_1=\hat{\beta}_{1(1)}-b_3EX_{i2}=\hat{\beta}_{1(1)}-b_3(p_2-q_2)$. Similarly $\hat{b}_1=\hat{\beta}_{1(1)}-b_3(p_1-q_1)$, where $\hat{\beta}_{1(2)}$ is given in the simple regression model on X_2. To estimate the variance of random error, note that

$$\sigma^2=Var(Y_i)-Var(b_0+b_1X_{i1}+b_2X_{i2}+b_3X_{i1}X_{i2}).$$

Therefore,

$$\begin{aligned}\hat{\sigma}^2=&sd(Y)^2-2b_1^2p_1q_1-2b_2^2p_2q_2-4b_1b_3(p_2-q_2)p_1q_1\\&-4b_2b_3(p_1-q_1)p_2q_2-\left[\left(p_1^2+q_1^2\right)\left(p_2^2+q_2^2\right)-(p_1-q_1)^2\right]b_3^2\end{aligned}$$

With an assumed value of b_3 and the corresponding estimators $\hat{b}_1$, $\hat{b}_2$, and $\hat{\sigma}^2$, we can apply the above calculation to obtain the power of model selection strategies.

Supporting Information

Text S1 Supplementary Note for proofs and arguments, distributions of test statistics, extended comparisons of power for model selection methods, and formulas for distribution parameters of test statistics.
Found at: doi:10.1371/journal.pgen.1000582.s001 (0.91 MB PDF)

Acknowledgments

We are grateful to Yale University Biomedical High Performance Computing Center for computation support. We thank Dr. Joshua Sampson and Dr. Yedan Zhang for their comments on the paper.

Author Contributions

Conceived and designed the experiments: ZW HZ. Analyzed the data: ZW. Contributed reagents/materials/analysis tools: ZW HZ. Wrote the paper: ZW HZ.

References

1. Fan J, Lv J (2008) Sure independence screening for ultra-high dimensional feature space. J R Statist Soc B 70: 849–911.
2. Storey JD, Akey JM, Kruglyak L (2005) Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biolgoy 3(8): 1380–1390. doi:10.1371/journal.pbio.0030267.
3. Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436(7051): 701–703.
4. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37(4): 413–417.
5. Evans DM, Marchini J, Morris AP, Cardon LR (2006) Two-stage two-locus models in genome-wide association. PLoS Genet 2(9): e157. doi:10.1371/journal.pgen.0020157.
6. Kutner MH, Nachtsheim CJ, Li W, Neter J (2005) Applied linear statistical models. 5th ed. New York: McGraw-Hill Irwin. 1396p.
7. Gail MH, Pfeiffer RM, Wheeler W, Pee D (2008) Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. Biostatistics 9(2): 201.
8. Weedon MN, et al. (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet 40: 575–583.

9. Cohen J (1988) Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Association. 572p.
10. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Statist Soc B 57: 289–300.
11. Scheffé H (1959) The analysis of variance. New York: John Wiley & Sons Inc. 477p.
12. Searle SR (1971) Linear models. New York: John Wiley & Sons Inc. 532p.
13. Ambrosius WT, Lange EM, Langefeld CD (2004) Power for genetic association studies with random allele frequencies and genotype distributions. Am J Hum Genet 74(4): 683–693.
14. Efron B (2004) Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. J Am Stat Assoc 99(465): 96–104.
15. Efron B (2007) Correlation and large-scale simultaneous significance testing. J Am Stat Assoc 102(477): 93–103.
16. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, et al. (2003) The international HapMap project. Nature 426(6968): 789–796.
17. Ke X, Cardon LR (2003) Efficient selective screening of haplotype tag SNPs. Bioinformatics 19(2): 287–288.
18. Weale ME, Depondt C, Macdonald SJ, Smith A, Lai PS, et al. (2003) Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: Implications for linkage-disequilibrium gene mapping. Am J Hum Genet 73(3): 551–565.
19. Lehmann EL, Casella G (1998) Theory of point estimation, Second Edition. New York: Springer Verlag. 589p.
20. David HA, Nagaraja HN (2003) Order statistics, Third Edition. New York: J. Wiley. 488p.