Bioinformatics Advance Access published September 7, 2011

An empirical comparison of several recent epistatic interaction detection methods

Yue Wang[1]*, Guimei Liu[2], Mengling Feng[3], Limsoon Wong[2]

[1]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore
[2]School of Computing, National University of Singapore, Singapore
[3]Data Mining Department, Institute for Infocomm Research, Singapore

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Motivation: Many new methods have recently been proposed for detecting epistatic interactions in GWAS data. There is however no in-depth independent comparison of these methods yet.

Results: Five recent methods—TEAM, BOOST, SNPHarvester, SNPRuler, and Screen and Clean (SC)—are evaluated here in terms of power, type-1 error rate, scalability, and completeness. In terms of power, TEAM performs best on data with main effect and BOOST performs best on data without main effect. In terms of type-1 error rate, TEAM and BOOST have higher type-1 error rates than SNPRuler and SNPHarvester. SC does not control type-1 error rate well. In terms of scalability, we tested the five methods using a dataset with 100,000 SNPs on a 64-bit Ubuntu system, with Intel (R) Xeon(R) CPU 2.66GHz, 16G memory. TEAM takes ∼36 days to finish and SNPRuler reports heap allocation problems. BOOST scales up to 100,000 SNPs and the cost is much lower than that of TEAM. SC and SNPHarvester are the most scalable. In terms of completeness, we study how frequently the pruning techniques employed by these methods incorrectly prune away the most significant epistatic interactions. We find that, on average, 20% of datasets without main effect and 60% of datasets with main effect are pruned incorrectly by BOOST, SNPRuler, and SNPHarvester.

Availability: The software for the five methods tested are available from the URLs below. TEAM: `http://csbio.unc.edu/epistasis/download.php`. BOOST: `http://bioinformatics.ust.hk/BOOST.html`. SNPHarvester: `http://bioinformatics.ust.hk/SNPHarvester.html`. SNPRuler: `http://bioinformatics.ust.hk/SNPRuler.zip`. Screen and Clean: `http://wpicr.wpic.pitt.edu/WPICCompGen/`.

Contact: wangyue@nus.edu.sg

1 INTRODUCTION

A genome-wide association study (GWAS) examines the association between phenotypes and genotypes in a study group. The first exciting finding was on age-related macular degeneration (AMD) (Klein, 2005), which uncovers a disease allele (tyrosine-histidine polymorphism) with an effect size of 4.6 in ∼100,000 single-nucleotide polymorphisms (SNPs). Since then, over 600 GWAS's have been conducted for 150 diseases and traits; and ∼800 associated SNPs have been reported. The methodologies of these studies are similar: A quality control criteria is first defined to filter the genotype data; then the remaining genotypes are each tested for association with the disease phenotypes. Finally, the significant SNPs are reported after multiple-testing correction. Most of these GWAS's could only identify disease alleles with moderate effect size. Thus, single SNP association studies could explain very limited heritability of these diseases (Emahazion *et al.*, 2001).

Consequently, researchers have started exploring multi-SNP interactions in the hope of discovering more significant associations. Multi-SNP interactions are also called "epistatic interactions". This term originated from Bateson's definition of epistasis one hundred years ago (Bateson, 1909). It was defined as the change of segregation ratio and the interaction of genes. However, in the current literature, there is a debate on the exact definition of epistasis (Phillips, 1998, 2008). Our paper focuses on evaluating epistatic interaction detection methods in their computational aspect and all the experiments are based on simulation data. Thus, we consider epistatic interactions as the statistically significant associations of k-SNP interaction ($k \geq 2$) with phenotypes.

There are mainly two types of epistatic interaction detection methods: model-based methods and model-free methods. In general, model-based methods (Wu *et al.*, 2009; Yang *et al.*, 2009; Wan *et al.*, 2010a; Wu *et al.*, 2010) predefine a statistical model between phenotypes and genotypes; then they fit the data to the model; and finally they output the significant SNPs. They work well for only a small number of important and filtered candidate SNPs; but they often fail when the number of SNPs grows to hundreds of thousands. To make model-based methods more efficient, researchers have proposed a variety of heuristic and filtering techniques. For example, Wan *et al.* (2010a) develop an upper bound of the likelihood ratio test statistic for two-locus epistatic interaction to prune the search space and a boolean transformation of data to make collection of contingency table information faster. As another example, Wu *et al.* (2010) devise a two-stage analysis so that the overall analysis is more efficient. As a third example, Yang *et al.* (2009) use a stochastic search to identify only 40-50 (set by the user) groups of candidate epistatic interactions for follow-up model-fitting analysis.

*To whom correspondence should be addressed

In contrast, model-free methods (Ritchie *et al.*, 2001; Wan *et al.*, 2010b; Zhang *et al.*, 2010) have no prior assumption on the data and the model. Given the genotype data, these methods only examine the test statistic of each possible epistatic interaction with phenotypes. Zhang *et al.* (2010) propose a minimum spanning tree (MST) structure to represent the data; by traversing this MST, exhaustive search of every epistatic interaction is an order faster than that of brute-force search. Wan *et al.* (2010b) connect the epistatic interactions with predictive rules and use a rule mining strategy to find epistatic interactions.

Our evaluation study of epistatic interaction detection methods is different from earlier studies such as Motsinger-Reif *et al.* (2008a), Motsinger-Reif *et al.* (2008b) and Sucheston *et al.* (2010). Firstly, Motsinger-Reif *et al.* (2008b) compare only approaches based on neural networks while our selected methods cover both data mining and statistical methods. Secondly, Motsinger-Reif *et al.* (2008a) evaluate multifactor dimensionality reduction (MDR) (Ritchie *et al.*, 2001), grammatical evolution neural networks (GENN) (Motsinger-Reif *et al.*, 2006), focused interaction testing framework (FITF) (Millstein *et al.*, 2006), random forests (RF) (Breiman, 2001), and logistic regression (LR) (Hosmer and Lemeshow, 2000) methods. They show that MDR is superior in all settings. After two years of advancement, most methods selected in this paper have demonstrated that their performance is better than that of MDR; we therefore omit discussing methods mentioned in Motsinger-Reif *et al.* (2008a). Thirdly, Sucheston *et al.* (2010) compare AMBIENCE (Chanda *et al.*, 2008) with MDR, restricted partitioning method (RPM) (Culverhouse, 2007) and logistic regression. They conclude that the performance of AMBIENCE is equivalent to that of logistic regression for two-locus models and better than that of RPM and MDR. However, according to Wan *et al.* (2010a), the performance of BOOST is better than that of PLINK (Purcell *et al.*, 2007), which uses a pure logistic regression model. Therefore we omit the evaluation of AMBIENCE and RPM in our study. Lastly, Wan *et al.* (2010b) and Yang *et al.* (2009) have shown that their overall performance is much better than that of BEAM (Zhang and Liu, 2007). We thus omit BEAM.

In this paper, we give an independent empirical comparison of five methods for detecting epistatic interactions—namely, TEAM (Zhang *et al.*, 2010), BOOST (Wan *et al.*, 2010a), SNPRuler (Wan *et al.*, 2010b), SNPHarvester (Yang *et al.*, 2009), and Screen and Clean (Wu *et al.*, 2010)—to help users better understand which method is more suitable for their data, which method is good for detecting epistatic interactions with and without main effect, and which method is scalable to larger datasets. We also analyze why combining several of these methods cannot enhance power. Their basic characteristics are given in Table 1.

Table 1. Summary of the features of the five methods: BOOST (B), TEAM (T), SNPRuler (SR), SNPHarvester (SH), Screen and Clean (SC)

	B	T	SR	SH	SC
Exhaustive Search	×	√	×	×	×
Logit Model Assumed	√	×	×	√	√
Multi-Stage	×	×	×	×	√
Permutation Test	×	√	×	×	×
Bonferroni correction	√	×	√	√	√
Programming language	C	C++	Java	Java	R

The organization of this paper is as follows. We first formulation the problem in Section 2. Then we briefly introduce each of the five methods in Section 3. We describe how the evaluation data is simulated in Section 4 and the detailed setting of each experiment in Section 5. After that, we present the results under each setting in Section 6. Finally, we discuss the performance of each method and provide advice to users in Section 7.

2 PROBLEM FORMULATION

In a typical GWAS, researchers collect two types of data: genotype data that encodes the genetic information of each individual, and phenotype data that measures the quantitative traits of each individual. Here, we consider only bi-allelic SNPs. The allele that occurs more frequently is called the major allele, denoted as A. The allele that occurs less frequently is called the minor allele, denoted as a. The two alleles form three genotypes—AA, Aa and aa—and they are encoded as 0, 1 and 2 in raw data. For phenotype data, we consider the binary form (0 for control and 1 for case). With minor modification, current methods can handle other types of phenotype data, e.g., by discretizing a continuous phenotype.

The goal of each method is to identify k-SNP ($k \geq 2$) epistatic interactions significantly associated with the phenotype. Thus, each method outputs a list of epistatic interactions, each involving up to k SNPs (usually k is set to 2) and is accompanied by its P-value after correction for multiple testing.

There are two challenges. First, if we constrain k to be 1, then the number of statistical tests is equal to the number of SNPs in a dataset. When k increases by 1, the number of tests grows by n-fold (n is the number of SNPs in a dataset). Thus, the total number of tests grows quickly as k increases, resulting in the inability of current methods to test all the combinations. For example, to study a moderate size of 500,000 SNPs, we can test only two-locus epistatic interactions if we use the EPISNP program (Ma *et al.*, 2008) on a 2.66GHz single processor, as it may take 1.2 years to finish all the tests. Therefore, heavy computation cost is one of the challenges for current methods (Wang *et al.*, 2011). Second, since a huge number of possible combinations are tested, a large proportion of significant associations are expected to be false positives. Thus, reducing the number of false positives while retaining power is another challenge.

3 METHODS

3.1 SNPRuler

SNPRuler (Wan *et al.*, 2010b), MDR (Ritchie *et al.*, 2001), and a few other pattern-based methods (Li *et al.*, 2006; Long *et al.*, 2009) adopt data mining approaches for detecting epistatic interactions. These methods do not assume a model-fitting procedure but use some filtering methods to reduce the number of SNP combinations to be tested. SNPRuler (Wan *et al.*, 2010b) is a rule-based approach motivated by the fact that each epistatic interaction induces a set of rules. For example, $\mathrm{SNP}_1 \wedge \mathrm{SNP}_2 \Rightarrow$ Disease contains 9 rules, they are $\mathrm{SNP}_1 = i \wedge \mathrm{SNP}_2 = j \Rightarrow$ Disease, $i, j \in \{0, 1, 2\}$. In the paper, the quality of a rule is given by its χ^2 test value. We define $\mathrm{SNP}_1 \wedge \mathrm{SNP}_2 \Rightarrow$ Disease as a SNP-level epistatic interaction and $\mathrm{SNP}_1 = i \wedge \mathrm{SNP}_2 = j \Rightarrow$ Disease, $i, j \in \{0, 1, 2\}$ as allele-level epistatic interactions. To identify epistatic interactions that are significant, SNPRuler traverses a set enumeration tree where the nodes of the tree are the genotypes of the SNPs, the leaves of the tree are the phenotypes, and the path from the root to a leaf is an allele-level epistatic interaction. Exhaustive tree traversal is theoretically possible but practically impossible due to the explosive number of combinations as the tree grows. Therefore, the authors propose an upper bound on the χ^2 test statistic to prune the search space. After the search procedure, a post-processing step is used to get and rank SNP-level interactions. There are two hidden problems in this work. First, the upper bound they derived from the χ^2 formula is not a true upper bound and does not possess the anti-monotone property (Agrawal and Srikant, 1994). Although it helps prune a large search space, it also prunes many true-positive epistatic interactions. Second, the upper bound is based on the assumption that the number of cases should be larger than or equal to that of controls in a dataset; otherwise, the upper bound does not hold. This assumption is inconvenient since the number of controls is larger than that of cases in most GWAS datasets.

3.2 SNPHarvester

SNPHarvester (Yang *et al.*, 2009) is a stochastic search algorithm to identify epistatic interactions. It consists of two steps: a filtering and a model-fitting step. The filtering step is to identify m (40–50) significant SNP groups for the subsequent model-fitting step. In the filtering step, it first removes significant single SNPs according to their χ^2 test values, because this method is only interested in epistatic interactions that have weak marginal effect but significant joint effect. Then it randomly picks k SNPs. These form an active set $S = \{SNP_1, SNP_2, ..., SNP_k\}$. The rest of the SNPs form a candidate set S_c. After all these preparations are done, the nested *PathSeeker* algorithm is called to swap $SNP_i \in S$ with $SNP_j \in S_c$ to get the group with the highest χ^2 test value. A total of $k(n-k)$ combinations need to be tested to identify such a group. After this, the identified group is removed from the n SNPs. The next iteration continues to select k SNPs to form an active set and the remaining $n - 2k$ SNPs form a candidate set. The same procedure is repeated again. The complexity to identify m groups is $O(knm)$, which is affordable even when there are $> 100,000$ SNPs. In the second step, each of the m significant groups is fitted into the L$_2$ penalized logistic regression model; see (Park and Hastie, 2008) for details.

3.3 Screen and Clean

The Screen and Clean method (Wu *et al.*, 2010) uses a two-stage analysis; datasets from stage 1 for the screening and datasets from stage 2 for the cleaning. In the screening stage, it only considers tag SNPs and marginal significant SNPs. These SNPs are first fitted into the main effect lasso logistic regression model

$$g(E[Y|X]) = \beta_0 + \sum_{j=1}^{N} \beta_j X_j$$

where X_j is the encoded genotype value 0, 1 or 2, Y is the encoded phenotype value 0 or 1. This model first identifies a set of SNPs whose coefficients satisfy $\beta_j \neq 0$, $j \in \{1,2,...,n\}$; then it obtains the least square estimates $\hat{\beta_k}$, $k \in \{1,2,...,n\}$ of these SNPs. To test the significance of each regression coefficient, the t-test statistic value is calculated. Only the significant SNPs and their corresponding two-SNP combinations enter the interaction model

$$g(E[Y|X]) = \beta_0 + \sum_{j=1}^{N} \beta_j X_j + \sum_{i<j;i,j=1,...,N} \beta_{ij} X_i X_j.$$

A similar procedure applies to interaction model fitting. After this stage, the "surviving" SNP pairs go to the second cleaning stage for controlling type-1 error. T-test is used again to remove SNP pairs whose significance level is lower than a user specified threshold.

3.4 BOOST

BOOST (Wan *et al.*, 2010a) contributes to the epistatic detection problem in two aspects. Firstly, it provides a new boolean representation of the data. By transforming the data representation to the boolean type, BOOST uses established methods (Wegner, 1960) of logic operations to collect contingency table information, which is very efficient. Secondly, it proposes an upper bound for the likelihood ratio test statistic to prune insignificant epistatic interactions. The likelihood ratio test is originally based on the deviance of difference between the full logistic regression model

$$\log \frac{P(Y=1|X_{l_1}=i, X_{l_2}=j)}{P(Y=2|X_{l_1}=i, X_{l_2}=j)} = \beta_0 + \beta_i^{X_{l_1}} + \beta_j^{X_{l_2}} + \beta_{ij}^{X_{l_1}X_{l_2}},$$

X_{l_1} and X_{l_2} are genotype variables, i, $j \in \{0,1,2\}$, and the main logistic regression model

$$\log \frac{P(Y=1|X_{l_1}=i, X_{l_2}=j)}{P(Y=2|X_{l_1}=i, X_{l_2}=j)} = \beta_0 + \beta_i^{X_{l_1}} + \beta_j^{X_{l_2}}.$$

We denote the log likelihood of the full model under maximum likelihood estimate (MLE) as $\hat{L_F}$, the log likelihood of the main model under MLE as $\hat{L_M}$, the log likelihood of log-linear saturated model as $\hat{L_S}$, and the homogeneous model as $\hat{L_H}$. The likelihood ratio statistic between the main model and the full model is $-2(\hat{L_M} - \hat{L_F})$. The log-linear homogeneous association model corresponds to the main logistic regression model and the log-linear saturated model corresponds to the full logistic regression model (Agresti, 2002). This leads to an upper bound for the two log-linear models: $-2(\hat{L_S} - \hat{L_H})$. Matsuda (2000) uses Kirkwood Superposition Approximation to get a lower bound of the homogeneous association model ($\hat{L}_{KSA} \leq \hat{L_H}$). Therefore, the upper bound of the likelihood is established ($\hat{L_S} - \hat{L_H} \leq \hat{L_S} - \hat{L}_{KSA}$). This upper bound is tight and most nonsignificant interactions can be pruned. Its GPU version GBOOST (Yung *et al.*, 2011) provides 40-fold speedup compared with that of BOOST.

3.5 TEAM

TEAM (Zhang *et al.*, 2010) is an exhaustive algorithm to detect two-locus epistatic interactions in GWAS. It controls false positives by using permutation test. Permutation test is generally more accurate at finding the cut-off p-value threshold than direct adjustment methods like Bonferroni correction Benjamini and Hochberg (1995), but at a much higher cost. TEAM needs to compute the contingency table for every pair of SNPs on all the permutations to calculate p-values, which is very expensive. To reduce the computation cost, the authors observe that if two SNPs have the same genotype values on many individuals, then the computation of their contingency tables can be shared by considering only those individuals with different values. TEAM uses a Minimum Spanning Tree (MST), where nodes are SNPs and the weight of edges is the number of individuals with different values on the two SNPs, to maximize the sharing of contingency table computation. As the construction of MST can be costly, TEAM constructs an approximate MST instead. The performance of TEAM is faster than the brute-force approach by an order of magnitude. As TEAM does not presume any statistical model, it is applicable to any test statistic—e.g., χ^2 test, exact likelihood ratio test, and entropy-based test—based purely on contingency table information.

4 DATA SIMULATION

We simulate different types of datasets to evaluate the power, type-1 error rate, and scalability of each method.

4.1 Power

For each setting in both data with and without main effect below, 100 datasets are generated. In each dataset, we embed one ground-truth epistatic interaction. Power is defined as the fraction of the 100 datasets on which the top prediction matches the ground-truth.

Data with main effect The embedded epistatic interaction demonstrates both main effect and interaction effect. There are at least fifty different models that satisfy the constraints for two-locus epistatic interactions (Li and Reich, 2000). We consider the three commonly used models (Marchini *et al.*, 2005) given in Figure 1. We simulate the data based on these three models. For each model, we try two different minor allele frequencies (MAF) at 0.2 and 0.5, and three different main effect values at 0.2, 0.3 and 0.5; thus giving a total of six different settings. These values represent the low and high value for each parameter. We use 2,000 samples and 1,000 SNPs for each dataset, as per previous works. These datasets are available from `http://compbio.ddns.comp.nus.edu.sg/~wangyue/`.

Data without main effect This type of epistatic interaction demonstrates weak main effect but strong interaction effect. Finding such type of epistatic interactions is a challenging "dark area" which many methods fail to explore. We use data from Dartmouth Medical School. The website, `http://discovery.dartmouth.edu/epistatic_data`, provides 70 models, composed of combinations of the following parameter values. (1) Two MAF settings of 0.2 and 0.4. (2) Seven heritability settings of 0.4, 0.3, 0.2, 0.1, 0.05, 0.025 and 0.01. (3) Five different penetrance tables. Each model is simulated using four sample sizes of 200, 400, 800 and 1,600. The number of SNPs is 1,000 for each dataset.

4.2 Type-1 error rate

We simulate 1,000 datasets without embedding any epistatic interaction, each with 2,000 samples and 1,000 SNPs. The MAF of each SNP is uniformly distributed in [0.05, 0.5]. Type-1 error rate of the methods is defined as the proportion of the 1000 datasets on which the significance level of the top prediction satisfies the user-specified threshold.

4.3 Scalability

To test the scalability, we use datasets with 100, 1,000, 10,000 and 100,000 SNPs. Each of the 4 datasets has 2,000 samples.

5 EXPERIMENTAL SETTING

All the experiments are conducted on a 64-bit Ubuntu system, with Intel (R) Xeon(R) CPU 2.66GHz, 16G memory.

SNPRuler provides a Java program. The heap size is set to `-Xmx7000M`, giving the maximum memory for the program to use. The maximum number of rules is set as 50,000. The rule length is set to 2 since we focus on two-locus epistatic interactions. The pruning threshold is set as 0, to test as many combinations as possible.

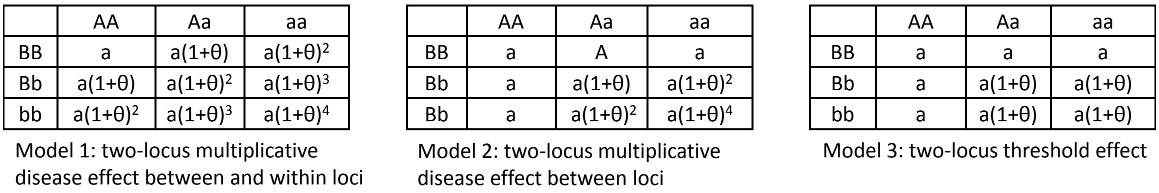
SNPHarvester also provides a Java program; it has two running modes. One is the "Threshold-Based" mode, where the user indicates the threshold significance level and the program outputs all results whose significance level is lower than the threshold. Another is the "Top-K Based" mode, where the program outputs the top K most significant results regardless of their significance level. The "Top-K Based" mode is used for our analysis.

TEAM provides a C++ program which consists of two sub programs: (1) to test all combinations and record the corresponding test statistic value and (2) to get the SNP pairs according to the user-specified False Discovery Rate (FDR). We use the default setting of other parameters and set the FDR value to 1.

BOOST provides a C program that only runs on Windows system. To let all programs run on the same hardware configuration, we use the `Wine` program (`http://www.wine.org`) which allows us to run a Windows program on a Unix system. There is no setting for BOOST; the output is the list of results whose likelihood ratio test statistic values are higher than 30 with 4 degrees of freedom.

Screen and Clean provides an R program; it has 4 running strategies, among which we choose the "Kitchen Sink". We set the P-value threshold to 0.1 and the number of pairs to be tested to 100.

BOOST filters out epistatic interactions with test statistic values less than 30 with 4 degrees of freedom. This corresponds to 0.1



	AA	Aa	aa
BB	a	a(1+θ)	a(1+θ)^2
Bb	a(1+θ)	a(1+θ)^2	a(1+θ)^3
bb	a(1+θ)^2	a(1+θ)^3	a(1+θ)^4

Model 1: two-locus multiplicative disease effect between and within loci

	AA	Aa	aa
BB	a	A	a
Bb	a	a(1+θ)	a(1+θ)^2
Bb	a	a(1+θ)^2	a(1+θ)^4

Model 2: two-locus multiplicative disease effect between loci

	AA	Aa	aa
BB	a	a	a
Bb	a	a(1+θ)	a(1+θ)
bb	a	a(1+θ)	a(1+θ)

Model 3: two-locus threshold effect

Fig. 1. Illustraion of three main models. For the two-locus problem, suppose the baseline odds of getting a disease is α, and having the disease allele (a or b) increases the odds by $1 + \theta$. A person with genotype Aa or Bb has an $\alpha(1 + \theta)$ odds of getting a disease, while one with genotype aa or bb has an odds of $\alpha(1 + \theta)^2$. Model 1 means the final odds is multiplied by the odds of two loci. Model 2 requires both of the loci to contain at least one disease allele before the odds can be multiplied within and between loci. For Model 3, the odds is kept the same if both loci contain the disease allele.

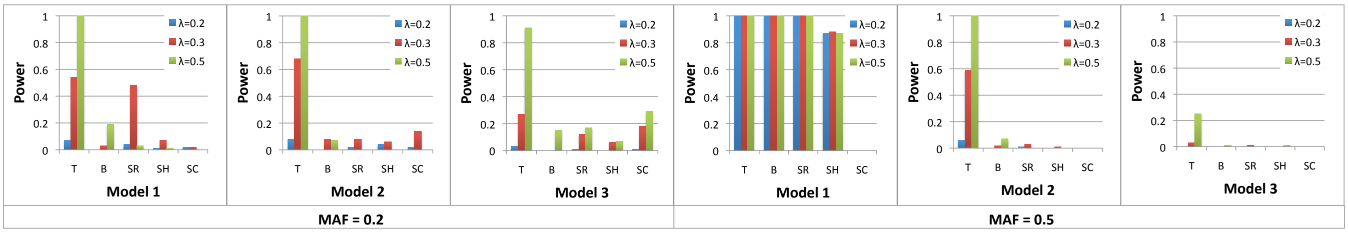


Fig. 2. Power comparison under three main effect models. Each model has two MAF settings and three λ settings which control the main effect of the ground-truth SNP. For each model, we generate 100 datasets. For each dataset, the sample size is 2,000 (1,000 cases and 1,000 controls) and the number of SNPs is 1,000. Abbreviations of the methods are: T (TEAM), B (BOOST), SR (SNPRuler), SH (SNPHarvester) and SC (Screen and Clean). The p-value for one-way ANOVA test is 0. 0009.

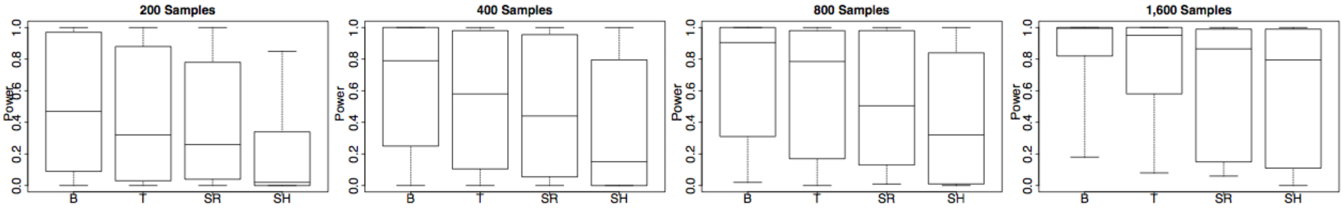


Fig. 3. Power comparison under 70 models without main effect. For each model, we simulate data using four different sample sizes. These sizes simulate the study design from small scale to large scale. Abbreviations of the methods are: T (TEAM), B (BOOST), SR (SNPRuler), and SH (SNPHarvester).

significance level. For fair comparison, we add a post-processing step to filter output with P-values higher than 0.1 for other methods.

6 EXPERIMENT RESULTS

6.1 Model with main effect

The results here are obtained by using data generated in the first part of Section 4.1. Figure 2 shows that in each setting, TEAM outperforms all other methods. For the other four methods, different model settings lead to different rankings. For example, in Model 1 with λ=0.3, SNPRuler is second; in Model 2 with λ=0.5, Screen and Clean is second. The different performance of TEAM and BOOST is due to a key difference in defining the interaction effect. TEAM uses the χ^2 test to measure the significance of two-locus interactions and thus makes no assumption about the data. BOOST uses a log likelihood ratio test to get the deviance difference between the log likelihood of the log-linear homogeneous association model and log-linear saturated model. BOOST performs well when the interaction terms contribute significantly to the model. However, when single SNP association terms fit the model well and interaction terms do not contribute significantly, BOOST may not be able to detect the ground-truth. This type of epistatic interactions is often referred as "statistical epistasis" (Cordell, 2002) and is widely accepted by the statistical community. SNPRuler is not an exhaustive method, but the test used is the same as that of TEAM. We set the pruning threshold to 0; thus it explores as many epistatic interactions as possible. Compared to TEAM, this method potentially misses true-positives. The result of SNPHarvester is expected as its randomization technique makes it difficult to perform better than exhaustive search. Screen and Clean performs poorly, due to its numerous filtering steps in the two-stage design. In the screening step, before the main-effect lasso procedure starts, it includes only marginally significant and tag SNPs. After that, it still only considers n (set by the user) pairs of SNPs instead of all the possible pairs to continue the interaction model fitting procedure. In the cleaning step, the filtering test is applied to only a small number of SNP pairs, resulting in little power to detect the ground-truth.

All five methods perform best on Model 1 compared to Model 2 and Model 3. This is because of the multiplicative effect between and within the two loci, making the epistatic interaction effect stronger and easier to detect. Model 2 only considers the multiplicative effect between two loci; the power to detect epistatic interactions drops obviously for all methods. The interaction effect of Model 3 is even weaker than Model 2, leading to the lowest power in all methods. It is also noted that the higher the main effect of the model, the easier it is for each method to detect epistatic interactions. However, SNPRuler and SNPHarvester do not follow this pattern because, when the main effect of the ground-truth pair is large, these two methods prune such main effect SNPs at the filtering stage. This leads to the missing detection of ground-truth.

6.2 Model without main effect

The results here are obtained using data generated in the latter part of Section 4.1. Screen and Clean is applicable only to data with main effect; thus we omit it here. Figure 3 gives an overall picture of the performance of the methods for each sample size, while Figure 4 gives the details. The median power of BOOST is the highest followed by TEAM. The performance of SNPRuler is close to that of an exhaustive method (TEAM) but is at a lower computational cost. BOOST performs the best in each setting and TEAM second; but the difference is not as obvious as that in data with main effect. SNPHarvester performs relatively poorly for each sample size. All methods perform well when heritability is high; when heritability reduces to 0.001, all methods have little power. Lescai and Franceschi (2010) point out in their study of neurological

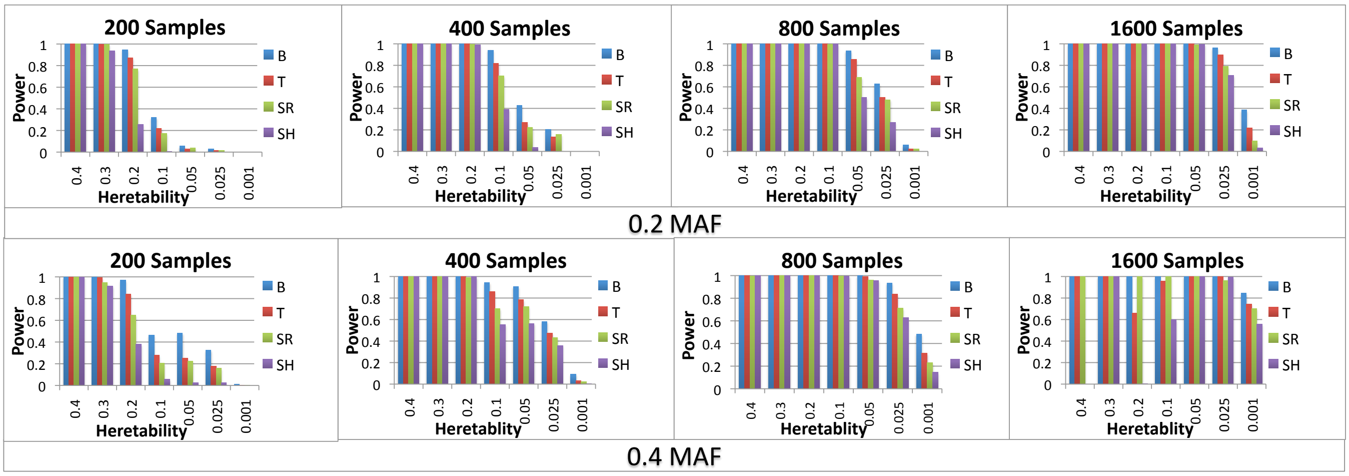


Fig. 4. Detailed results of four methods on data without main effect. In particular, for models with heritability 0.001, MAF 0.2 and sample size 200, the results of these datasets were not reported previously; all four methods have zero power on them. This shows the limitations of purely statistical methods. The p-value for one-way ANOVA test is 0. 0997.

cancers that low heritability caused by phenocopy level (PE) is the main reason for the methods to lose power. We also notice that increasing the sample size helps all these methods to improve their power in each heritability setting.

When we evaluate the four methods on data without main effect, we use all datasets that are publicly available. They include 70 models and 4 different sample sizes for each model. Part of these datasets are also used in BOOST, SNPRuler and SNPHarvester. BOOST does not include the results of 70 models for 200 samples. SNPRuler and SNPHarvester merely show results of 60 models and each model with 400 samples. Our reported results are consistent with previous reported results and are complementary to them. In particular, for those models with 0.001 heritability, 0.2 MAF and 200 samples, the results of these datasets were not reported previously; and all four methods have zero power (see Figure 4). This shows the limitations of purely statistical methods.

6.3 Scalability

We apply all methods to datasets with 100, 1,000, 10,000, and 100,000 SNPs. From Table 2, BOOST is the fastest under the first three settings. This is due to its fast Boolean operation to collect contingency tables and upper-bound-pruning technique. When the SNP size grows to 100,000, it is much slower than the two non-exhaustive methods SNPHarvester and Screen and Clean. TEAM is the slowest in all settings for two reasons. First, the overall running time is only an order faster than that of a brute-force approach. Second, the permutation procedure makes it even more expensive, although traversing MST helps reduce the cost. SNPRuler cannot execute on the dataset with 100,000 SNPs because we get the “out of memory” error, even though we have set the heap size to 12.8G for the Java virtual machine, which is the maximum on our PC. SNPHarvester and Screen and Clean only identify a fixed number of candidate epistatic interactions, and then fit them to a statistical model for follow-up analysis. Thus, their scalability is much better than the other three methods when SNP size grows.

Table 2. Running time comparison of the five methods. Abbreviations of the methods are: SR (SNPRuler), SH (SNPHarvester), SC (Screen and Clean).

# SNPs	TEAM	BOOST	SR	SH	SC
100	58.23s	0.16s	2.43s	2.29s	7.39s
1,000	353.20s	2.47s	21.73s	22.33s	55.48s
10,000	7,406.29s	156.16s	1,097.65s	224.24s	626.96s
100,000	∼36 days	15,010.42s	NA	6,616.65s	5,858.34s

6.4 Type-1 error

We define the type-1 error rate of a method as the proportion of datasets that the method reports the existence of significant epistatic interactions, out of the 1,000 datasets in which no epistatic interactions are actually embedded. The significance level is set to 0.05 after Bonferroni correction. The type-1 error rate for TEAM is 0.018, BOOST is 0.065, and SNPRuler and SNPHarvester both are 0.003. TEAM and BOOST have higher power thus higher type-1 error rates are reasonable. Screen and Clean has problems controlling type-1 error, as its type-1 error rate is as high as 0.86.

6.5 Completeness

SNPRuler, SNPHarvester and BOOST use some pruning techniques to speed up the search. Hence they have better scalability than TEAM as shown in Table 2. The side effect of using pruning techniques is the loss of power—the most significant SNP pairs may be thrown away. To study the magnitude of this side effect, we pick the most significant SNP pair on each dataset and study how many of them are pruned. For each method, the most significant SNP pair is the SNP pair with the lowest p-value calculated using the statistical test used by the method. Thus, for BOOST, the most significant SNP pair is the SNP pair with the lowest p-value calculated using likelihood-ratio test. For the other two methods, the most significant SNP pair is the SNP pair with the lowest p-value calculated using chi-square test. BOOST prunes away the most significant SNP pair on 4,195 out of the 26,860 datasets without main effect, and on 756 out of 1,698 datasets with main effect. Among these 4,195 datasets, the power of BOOST is 12.2% compared to 18.3% for

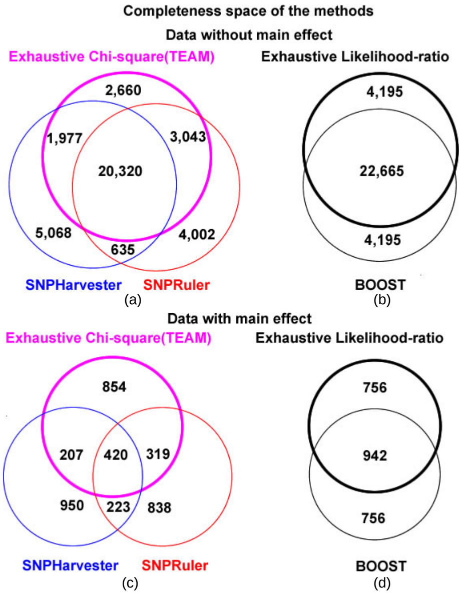


Fig. 5. The completeness space for the four methods. As there are two types of datasets and two types of test statistics, four venn diagrams are drawn respectively. In Part (a), all three methods—TEAM, SNPRuler and SNPHarvester—use χ^2 test. TEAM's outputs represent the 28,000 (20,320 + 1,977 + 2,660+ 3,043) top significant SNP pairs in 28,000 datasets. SNPHarvester can identify 22,297 (20,320+1977) of them. Among the 28,000 top SNP pairs, 20,320 of them can be identified by all three methods. Parts (b), (c) and (d) follow similar explanations.

the corresponding exhaustive method. Figure 5 also shows that the number of incorrectly pruned datasets of SNPRuler is smaller than that of SNPHarvester for both types of data. Correspondingly, the power of SNPRuler is higher than that of SNPHarvester.

7 DISCUSSION

The five methods all demonstrate respective utilities through the experiments results above. No single method is simultaneously the most powerful, the most scalable, and has the lowest type-1 error rate in every setting. When users want powerful results and are not concerned with computation cost, we recommend using TEAM and BOOST. Compared with TEAM, BOOST uses a model-fitting procedure. If the data fits the model well, the result is usually good; otherwise, a model-free method may be the alternative choice. When users expect moderate running time and power, we recommend using SNPRuler. Its pruning technique helps reduce running time albeit at the risk of losing power. If users are conscious of computation cost and have to run very large datasets, we recommend using SNPHarvester because it only identifies a small number (40–50) of groups for the model-fitting procedure.

Our evaluations are based on simulation results. In a real study, users usually have no idea of the "ground-truth" in the dataset. Hence it may not be sufficient to rely only on one method to obtain results. We suggest that, if time and computation resources permit, users try both the recommended model-free (i.e., TEAM) and model-fitting (i.e., BOOST) methods.

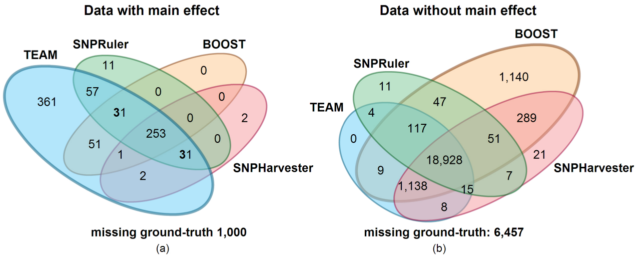


Fig. 6. The power space for the four methods on data with and without main effect. In part (a), there are in total 1,800 datasets for 18 settings of the simulated datasets, which corresponds to 1,800 ground-truth. Among these ground-truth, only 800 of them can be detected by at least one of the four methods, while the best method—TEAM—identifies 787 ground-truth out of 800. This explains why using ensemble methods cannot outperform TEAM. Similar observation is illustrated in Part (b).

It is tempting to consider taking a "majority vote" of the results of two or more methods. For example, let every algorithm report their top-3 predictions. A SNP pair receives k votes if it is reported by k methods. We select the one with the highest vote as the final prediction. When there is a tie, we choose the one with the lowest P-value. Unfortunately, for both types of data tested, we find that an ensemble using such a strategy cannot increase power over using solely BOOST or TEAM. In Figure 6, we see that for data without main effect, BOOST's ground-truth predictions highly overlap with the other three methods', so any ensemble cannot contribute a significant number of new ground-truth predictions. Specifically, the proportion of BOOST's ground-truth predictions that are not predicted by the other three methods is 4.1%, while the proportion of the other methods' ground-truth predictions not predicted by BOOST is 0.2%. Similarly, for data with main effect, no ensemble can outperform TEAM.

Our evaluations above only focus on two-locus epistatic interaction. Recently, Wang *et al.* (2010) and Liu *et al.* (2011) provide a general model that can be extended to n-locus epistasis. They also provide mathematical details of dissecting the χ^2 test into different epistatic components. For example, two-way epistatic interaction can be partitioned into four epistatic components: additive × additive, additive × dominant, dominant × additive and dominant × dominant. This helps characterize epistatic interactions in a more specific way and provides more physiological insights.

ACKNOWLEDGEMENT

Funding: This research is funded in part by a National University of Singapore NGS scholarship and by a Singapore Agency for Science Technology & Research grant SERC 102 101 0030.

REFERENCES

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. *Proceedings of 20th International Conference on Very Large Data Bases*, pages 487–499.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons.

Bateson, W. (1909). *Mendel's Principles of Heredity*. University Press.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Breiman, L. (2001). Random forest. *Machine Learning*, **45**(1), 5–32.

Chanda, P., Sucheston, L., Zhang, A., et al. (2008). AMBIENCE: A novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. *Genetics*, **180**(2), 1191–1210.

Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, **11**(20), 2463–2468.

Culverhouse, R. (2007). The use of the restricted partition method with case-control data. *Human Heredity*, **63**(2), 93–100.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley & Sons.

Emahazion, T., Feuk, L., Jobs, M., et al. (2001). SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. *Trends in Genetics*, **17**(7), 407–413.

Klein, R. J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**(5720), 385–389.

Lescai, F. and Franceschi, C. (2010). The impact of phenocopy on the genetic analysis of complex traits. *PLoS One*, **5**(7), e11876.

Li, W. and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human Heredity*, **50**(6), 334–349.

Li, Z., Zheng, T., Califano, A., and Floratos, A. (2006). Pattern-based mining strategy to detect multi-locus association and gene × environment interaction. *BMC Proceedings*, **1**(Suppl 1), S16–S16.

Liu, T., Thalamuthu, A., Liu, J., et al. (2011). Asymptotic distribution for epistatic tests in case-control studies. *Genomics*, **98**(2), 145–151.

Long, Q., Zhang, Q., and Ott, J. (2009). Detecting disease-associated genotype patterns. *BMC Bioinformatics*, **10**(Suppl 1), S75.

Ma, L., Runesha, H. B., Dvorkin, D., Garbe, J., and Da, Y. (2008). Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies. *BMC Bioinformatics*, **9**(1), 315.

Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, **37**(4), 413–417.

Matsuda, H. (2000). Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Physical Review E*, **62**(3), 3096.

Millstein, J., Conti, D. V., Gilliland, F. D., and Gauderman, W. J. (2006). A testing framework for identifying susceptibility genes in the presence of epistasis. *American Journal of Human Genetics*, **78**(1), 15–27.

Motsinger-Reif, A. A., Reif, D. M., Dudek, S. M., and Ritchie, M. D. (2006). Understanding the evolutionary process of grammatical evolution neural networks for feature selection in genetic epidemiology. *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 1–8.

Motsinger-Reif, A. A., Reif, D. M., Fanelli, T. J., and Ritchie, M. D. (2008a). A comparison of analytical methods for genetic association studies. *Genetic Epidemiology*, **32**(8), 767–778.

Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W., and Ritchie, M. D. (2008b). Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genetic Epidemiology*, **32**(4), 325–340.

Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, **9**(1), 30–50.

Phillips, P. C. (1998). The language of gene interaction. *Genetics*, **149**(3), 1167–1171.

Phillips, P. C. (2008). Epistasis-the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, **9**(11), 855–867.

Purcell, S., Neale, B., Todd-Brown, K., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**(3), 559–575.

Ritchie, M. D., Hahn, L. W., Roodi, N., et al. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics*, **69**(1), 138–147.

Sucheston, L., Chanda, P., Zhang, A., et al. (2010). Comparison of information-theoretic to statistical methods for gene-gene interactions in the presence of genetic heterogeneity. *BMC Genomics*, **11**, 487.

Wan, X., Yang, C., Yang, Q., et al. (2010a). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *American Journal of Human Genetics*, **87**(3), 325–340.

Wan, X., Yang, C., Yang, Q., et al. (2010b). Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics*, **26**(1), 30–37.

Wang, Z., Liu, T., Lin, Z., et al. (2010). A general model for multilocus epistatic interactions in case-control studies. *PLoS One*, **5**(8), e11384.

Wang, Z., Wang, Y., Tan, K. L., et al. (2011). eCEO: An efficient Cloud Epistasis cOmputing model in genome-wide association study. *Bioinformatics*, **27**(8), 1045–1051.

Wegner, P. (1960). A technique for counting ones in a binary computer. *Communication of ACM*, **3**(5), 322.

Wu, J., Devlin, B., Ringquist, S., et al. (2010). Screen and Clean: A tool for identifying interactions in genome-wide association studies. *Genetic Epidemiology*, **34**(3), 275–285.

Wu, T. T., Chen, Y. F., Hastie, T., et al. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**(6), 714–721.

Yang, C., He, Z., Wan, X., et al. (2009). SNPHarvester: A filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics*, **25**(4), 504–511.

Yung, L. S., Yang, C., Wan, X., and Yu, W. (2011). GBOOST: A GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. *Bioinformatics*, **27**(9), 1309–1310.

Zhang, X., Huang, S., Zou, F., and Wang, W. (2010). TEAM: Efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, **26**(12), i217–i227.

Zhang, Y. and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, **39**(9), 1167–1173.