Tsinghua Science and Technology

Volume 19 | Issue 6

Article 5

2014

Genome-Wide Interaction-Based Association of Human Diseases — A Survey

Xuan Guo
Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA.

Ning Yu
Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA.

Feng Gu
Department of Computer Science, College of Staten Island, Staten Island, NY 10314, USA.

Xiaojun Ding
Central South University, Changsha 410083, China.

Jianxin Wang
Central South University, Changsha 410083, China.

See next page for additional authors

Recommended Citation

Xuan Guo, Ning Yu, Feng Gu et al. Genome-Wide Interaction-Based Association of Human Diseases — A Survey. Tsinghua Science and Technology 2014, 19(06): 596-616.

TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 05/10 pp596-616
Volume 19, Number 6, December 2014

Genome-Wide Interaction-Based Association of Human Diseases — A Survey

Xuan Guo, Ning Yu, Feng Gu, Xiaojun Ding, Jianxin Wang*, and Yi Pan*

Abstract: Genome-Wide Association Studies (GWASs) aim to identify genetic variants that are associated with disease by assaying and analyzing hundreds of thousands of Single Nucleotide Polymorphisms (SNPs). Although traditional single-locus statistical approaches have been standardized and led to many interesting findings, a substantial number of recent GWASs indicate that for most disorders, the individual SNPs explain only a small fraction of the genetic causes. Consequently, exploring multi-SNPs interactions in the hope of discovering more significant associations has attracted more attentions. Due to the huge search space for complicated multi-locus interactions, many fast and effective methods have recently been proposed for detecting disease-associated epistatic interactions using GWAS data. In this paper, we provide a critical review and comparison of eight popular methods, i.e., BOOST, TEAM, epiForest, EDCF, SNPHarvester, epiMODE, MECPM, and MIC, which are used for detecting gene-gene interactions among genetic loci. In views of the assumption model on the data and searching strategies, we divide the methods into seven categories. Moreover, the evaluation methodologies, including detecting powers, disease models for simulation, resources of real GWAS data, and the control of false discover rate, are elaborated as references for new approach developers. At the end of the paper, we summarize the methods and discuss the future directions in genome-wide association studies for detecting epistatic interactions.

Key words: Single Nucleotide Polymorphism (SNP); genome-wide association; epistasis; epistatic interaction; complex disease

1 Introduction

Genome-Wide Association Studies (GWASs) have been proven to be a powerful tool for investigating the genetic architecture of human disease over the last ten years. It is a genomic and statistical inference study that involves statistical tests to measure and analyze DNA sequence variations in different individuals to see if any variant is associated with a trait. The ultimate goal of GWAS is to use genetic risk factors to predict who is at risk and to identify the biological underpinnings of disease susceptibility for developing new preventions and treatment strategies[1]. The first exciting finding of GWAS is on Age-related Macular Degeneration (AMD), which identifies the *Complement Factor H* gene as a major risk factor[2]. The associated *Complement Factor H* gene demonstrates not only the DNA sequence variations but also the biological basis for the effect. Another successful application

• Xuan Guo, Ning Yu, and Yi Pan are with the Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA. E-mail: yipan@gsu.edu.
• Feng Gu is with Department of Computer Science, College of Staten Island, Staten Island, NY 10314, USA.
• Xiaojun Ding and Jianxin Wang are with Central South University, Changsha 410083, China. E-mail: jxwang@mail.csu.edu.cn.
∗ To whom correspondence should be addressed.
Manuscript received: 2014-06-16; accepted: 2014-06-23

of GWAS is in the area of pharmacology, which heavily depends on understanding the biological basis of genetic effects. The goal of pharmacogenetics is to identify associated DNA sequence variations with drug metabolism and efficacy as well as adverse effects. Take warfarin for example, which is a blood-thinning drug that helps prevent blood clots in patients. A recent validation studies reveal that warfarin dosing can be largely influenced by the DNA sequence variations in several genes[3]. Accordingly, GWAS is increasingly used to identify biological pathways and underlying networks of complex diseases[4]. It has led to the exciting era of personalized medicine and personal genetic testing aiming to tailor healthcare for individual patients based on their genetic background and other biological features.

The modern unit of genetic variation is the Single Nucleotide Polymorphism (SNP) which refers to a single base change in a DNA sequence with an usual alternative of two possible nucleotides at a given position[5]. SNPs are the most common and abundant form of genetic variation amongst the human population. It is estimated that on average there is an SNP per every 300 bp of DNA, and about 11 million SNPs are on the whole genome of human species. However, due to a genetic phenomenon called Linkage Disequilibrium (LD) and the large majority of them with minor impacts on biological systems, it is not necessary to study all the SNPs[6,7]. The basic functional consequence of SNPs is amino acid change, which leads to the fluctuation of mRNA transcript stability and the transcription factor binding affinity through the central dogma[8]. In general, there are two commonly occurring base-pair possibilities for the same sequence location in a population. In this case, we say that the SNP has two alleles. An SNP is assigned a minor allele frequency or a frequency of less common allele when it is observed less frequently in a particular population. For example, if 20% of a population has the Cytosine allele versus the more common allele or the major allele, which takes up 80% of the population, then this SNP has a minor allele (C) with frequency of 0.2.

In the last two decades, extensive computational efforts have been provided to study the functional and structural consequences of the SNPs[9]. High-throughput chip based microarray technology has made GWAS possible for assaying one million or more SNPs. Currently, two primary platforms from Illumina (San Diego, CA) and Affymetrix (Santa Clara, CA) have been used for most GWASs. The Illumina platform uses a bead-based technology, which features better specificity at a little high cost, with slightly longer DNA sequences to detect alleles. The Affymetrix platform prints short DNA sequences as a spot on the chip that recognizes a specific SNP allele by differential hybridization of the sample DNA. More details of these two competing technologies can be found in Ref. [10]. By far, over 600 GWASs have been launched for 150 diseases and traits. In addition, aiming to examine the relationship between human genome sequence variation and the associated disease phenotypes, various international consortia are collecting information about variations in human genome, including HapMap consortium, Human Variation Project, 1000 Genomes Project, and Wellcome Trust Case Control Consortium (WTCCC).

The phenotypes in GWAS can be classified as either categorical (often binary case/control) or quantitative. Although quantitative traits are preferred and they improve power of detecting a genetic effect from the statistical perspective, well established quantitative measures are not available for many disease traits. Consequently, individuals in standard studies are usually categorized to binary variable. In this review, we focus on the methods for genome-wide case/control studies. Note that existing methods can handle quantitative traits by discretizing a continuous phenotype with minor modification. A routine in GWAS is the comparison between two groups of individuals: One has a higher prevalence of susceptibility alleles for interested trait, and another has a lower prevalence of such alleles[11]. The primary analysis paradigm for GWAS is dominated by the analysis on the susceptibility of individual SNPs, which can only explain a small part of genetic causal effects for complex diseases[12]. As a matter of fact, single locus-based approaches are insufficient to detect all interacting genes, especially for those with small marginal effects. For better understanding underlying causes of complex disease traits, identifying joint genetic effects (epistasis) across the whole genome has attracted more attentions[13]. The concept of epistasis[14] was introduced around 100 years ago. It was referred as an extension of the concept of dominance for alleles within the same allelomorphic pair[15]. In recent literatures, epistasis has been defined as the interaction among different genes (SNPs)[16]. Many studies have demonstrated that the epistasis is an important contributor to genetic variation in complex diseases, such as asthma, breast cancer[17], diabetes, coronary heart disease[18], and obesity[19]. In this paper, we also refer epistasis as a gene-gene interaction. Genome-wide association studies provide an enormous opportunity to identify high-order epistatic interactions among genetic variants throughout the genome. Without loss of generality, we consider high-order epistatic interactions or epistasis as the statistically significant associations of k-SNP modules ($k \geqslant 2$) with phenotypes.

Two challenges arise from finding high-order epistatic interactions associated with an interested trait among a large number of SNPs. The first comes from the combinatorial nature of the problem that the number of SNP combinations exponentially increases as the order goes up. Given a GWAS dataset with hundreds of thousands of SNPs, using brute-force approaches to examine all combinations of SNPs is computationally challenging, and even requires specialized hardwares[20-22]. For example, in order to detect pairwise interactions from 500 000 SNPs, which is a typical size of data generated by the Affymetrix platform, with thousands of samples genotyped, about 1.25×10^{11} statistical tests require to be proceeded. The second challenge concerns the statistical power for high-order SNP combination search. Since the huge number of hypothesis tests are often conducted on limited sample sizes with high degree of freedom, many false epistases are significantly associated with a disease trait by random chance[23,24]. Many computational algorithms have been proposed to overcome these two challenges. Based on their search strategies, existing approaches for searching epsitasis can be grouped into four broad categories: exhaustive search, stepwise search, stochastic search, and heuristic search. The naive solution to tackle the problem is exhaustive search by enumerating all possible combinations of multiple loci and performing desired interaction tests for each combination. Marchini et al.[13] showed that it is computationally possible to test two-locus associations allowing for interactions in GWAS based on current computation resources. Instead of explicitly enumerating all possible combinations of k-locus, stepwise search strategies select a subset of SNPs or combinations of SNPs based on some low-order statistic tests (or measures), then extend them to higher order interactions if it is statistically possible. Similar to stepwise searching, stochastic methods use random sampling procedures to search the space of interactions, and the performance relies on random chances to select phenotype-associated SNPs. With the number of SNPs going higher, the chances of correct hit drop accordingly. Heuristic methods use some heuristics to avoid exhaustive searches to obtain locally optimal solutions based on available information. Another way to categorize the epistatic interaction methods depends on the usage of models, i.e., model-based methods and model-free methods. By assuming a statistical model between phenotypes and genotypes, model-based methods use the case/control data to fit models and select the best model to rank the SNP modules. In contrast, model-free methods only examine the statistics of each possible epistasis associated with phenotypes rather than put prior assumption on the observed data. More details of the categorized methods are provided in Section 4.

Considering the wide variety of algorithms and techniques used in the detection of epistatic interactions, this paper is dedicated to offering a clear and complete picture of the epistasis detection by discussing and summarizing different tools according to their features. The organization of this article is as follows. We first give a statement of the problem in Section 2. Then we give two basic methods for searching epistatic interactions in view of model-based and model-free in Section 3. Total eight methods are elaborated with critial comments at the end of each discussion in Section 4 in respect of searching strategies. We describe the evaluation methodologies in Section 5 including the calculation of statistical power, the control for false discovery rate, resources of synthetic and real GWAS data, and a summary of current tools for detecting epistatic interactions. Finally, we discuss the future research directions of high-order epistasis detection in Section 6.

2 Problem Statement

Two types of data are collected in GWAS: genotype data, which encodes the genetic variants of individuals, and phenotype data that indicates the affected statuses of individuals. Like the methods reviewed in this paper, we only consider bi-allelic SNPs, which means that an SNP has only two alleles. The SNP is termed as a minor allele if the allele occurs less frequently, otherwise it is termed as a major allele if the allele occurs more frequently. Usually, we use lowercase letter to denote the minor allele and uppercase letter to denote the major allele, like a and A; so the two alleles form three genotypes, AA, Aa, and aa, and they can be encoded as 0, 1, and 2 in raw data. For phenotype data, the binary variable is used that 0 indicates unaffected and 1 indicates affected.

The goal of detection of epistatic interactions is to identify k-SNP ($k \geqslant 2$) modules significantly associated with the phenotype. Furthermore, epistatic interactions can be classified into two types: epistasis displaying Marginal Effects (eME) and epistasis displaying No Marginal Effects (eNME)[25]. More details can be found in Section 5.2. Basically, there are two challenges in searching epistasis: First, the total number of tests grows exponentially as k increases, leading to the inability of exhaustive search to examine all the combinations. For example, using epiSNP[26] to emulate and calculate all the two-locus epistatic interactions on a GWAS dataset with 1 000 000 SNPs will take 5 years on a 2.66 GHz single processor. Second, because a huge number of hypotheses are tested using limited samples, a large proportion of significant associations are expected to be false positives. Therefore, retaining the statistical power while reducing false positive rate is another important issue.

3 Methods in View of Model-Based and Model-Free

There are two ways to categorize the methods for detecting epistatic interaction: One is according to the assumption on the observed data; another is according to the searching strategy. We will cover the latter one in next section. An overview of recently developed 43 methods is depicted in Fig. 1, and eight of them with name in bold are reviewed and discussed in Section 4. If a predefined statistical model is set up between phenotypes and genotypes, we say that it is a model-based approach in which some parameters require to be estimated; otherwise it is a model-free method that no prior assumption is made on the data or the model.

In order to give an overview of model-based and model-free methods on detecting epsitasis, we describe two routines in the following section: (1) model fitting using logistic regression models and (2) Pearson's χ^2 test of goodness of fit. Obviously, the former is model-based, and the latter is model-free. Assuming that there are L SNPs and N samples, we use S to denote the ordered set of the L SNPs, s_i to denote the i-th SNP in S ($i \in [1, L]$), and Y to denote the class label (1 for case and 0 for control). For the analysis of two-locus epistatic interactions, we need to collect a contingency table (shown in Table 1), where n_{ijy} is the count of individuals with genotype $s_a = i, s_b = j$, and phenotype $Y = y$.

For the model-based methods, the model defining the epistasis via logistic regression models must be established at first. The logistic regression model with both main effect (marginal effect) terms and interaction terms, i.e., the full model, has the following form:

$$\log \frac{P\left(Y=1|s_a=i, s_b=j\right)}{P\left(Y=0|s_a=i, s_b=j\right)} = \beta_0 + \beta_i^{s_a} + \beta_j^{s_b} + \beta_{ij}^{s_a s_b} \quad (1)$$

The null logistic regression model without main effect term or interaction terms has the following form:

$$\log \frac{P\left(Y=1|s_a=i, s_b=j\right)}{P\left(Y=0|s_a=i, s_b=j\right)} = \beta_0 \quad (2)$$

There are nine coefficients in Eq. (1) and one coefficient in Eq. (2). We denote the log-likelihood of the full model as L_{F} and the log-likelihood of the null model as L_{N}. According to the likelihood ratio test, the effect of epistasis in this paper is defined as the difference between two log-likelihoods of models in Eqs. (1) and (2). By evaluating the values at their Maximum Likelihood Estimations (MLEs), i.e., $\hat{L_{\mathrm{F}}} - \hat{L_{\mathrm{N}}}$, we are able to estimate epistasis effect based on the departure of observed data from the null model naturally.

For the model-free method using Pearson's χ^2 test of goodness of fit, the following steps are conducted: (1) Collect the contingency table as shown in Table 1; (2) Obtain the p-value using χ^2 statistic (Eq. (3)) with 8 degrees of freedom[27].

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

where n is the count of genotype combinations by giving a set of SNPs. The observed frequency O is

Table 1 The genotype counts in cases and controls.

		s_a=0	s_a=1	s_a=2
Y=0	s_b=0	n_{000}	n_{100}	n_{200}
	s_b=1	n_{010}	n_{110}	n_{210}
	s_b=2	n_{020}	n_{120}	n_{220}
Y=1	s_b=0	n_{001}	n_{101}	n_{201}
	s_b=1	n_{011}	n_{111}	n_{211}
	s_b=2	n_{021}	n_{121}	n_{222}

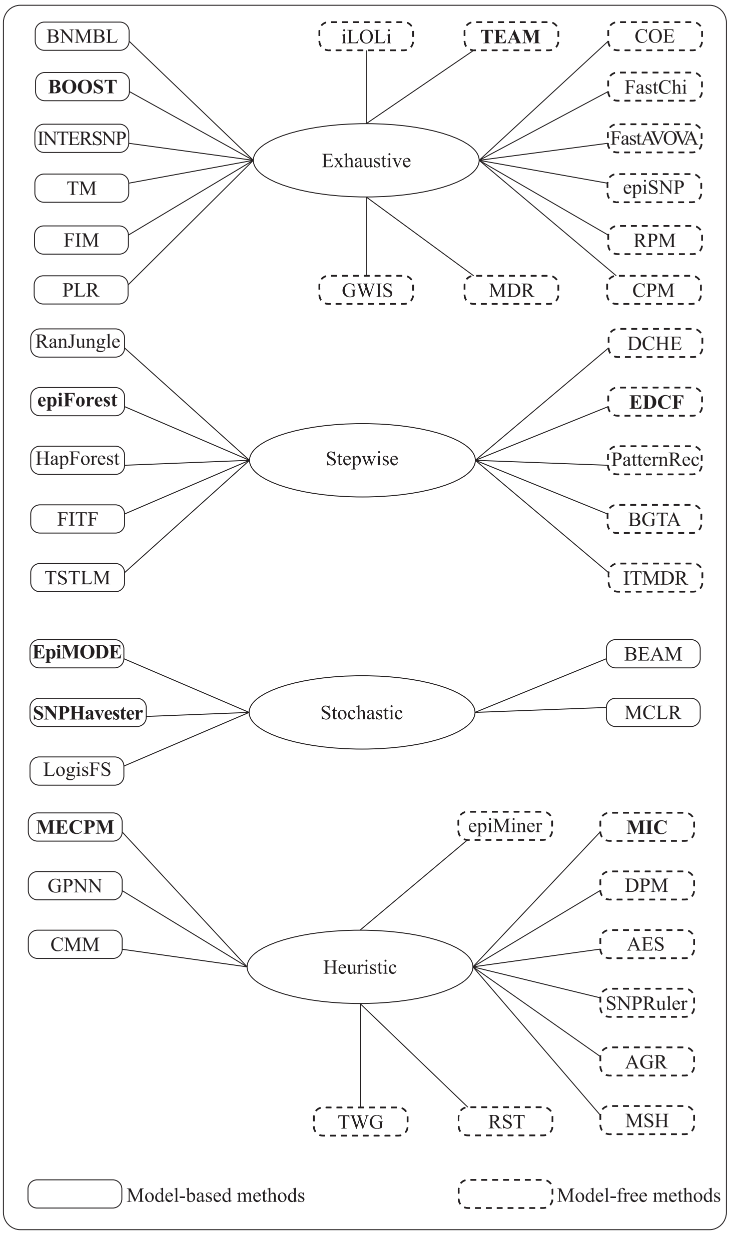


Fig. 1 Classification of the methods that detect epistasis.

corresponding to the count of the individual with a certain genotype combinations and a class label. SNP modules with p-value larger than a predefined threshold are reported as significant epistasis. Note that not all model-free methods are using the above described approaches, but they share the same feature that similar tests are applied for detection without any estimation of parameters of models.

4 Methods in View of Searching Strategies

According to the search strategy, existing approaches for searching epistatic interactions can be grouped into four broad categories, exhaustive search, stepwise search, stochastic search, and heuristic approaches. In review of recent literatures, we identify 43 methods used to detect epistasis, excluding specializations, tweaks, and simply paralleled methods. In the following sections, we scrutinize eight methods in these four categories, and point out their advantages and disadvantages.

4.1 Exhaustive searching methods for detecting epistatic interactions

The naive solution to tackle the problem of detecting epistatic interaction is exhaustive search using χ^2 test, exact likelihood ratio test or entropy-based test for all modules of multiple-locus. Marchini et al.[13] show that it is computationally possible to test two-locus associations allowing for interactions in GWAS based on current computing capability. Examples in exhaustive search, like MDR[17] and its extensions, utilize repeated cross-validations and permutation tests to evaluate accuracy and significance of classification. A major barrier for exhaustive search is the intensive computation, and thus parallel computing was adopted to further speed up the analysis of gene-gene interactions. For example, GBOOST[28] is a GPU framework based implementation of BOOST, and PIAM is developed by Liu et al.[29], which used the multi-thread to perform Genome-Wide Interaction-Based Association (GWIBA) analysis for exhaustive two-locus searches. However, finding higher order (more than 2 loci) disease-related associations is too computationally expensive to be feasible, especially for large GWAS datasets with millions SNPs. In this section, we use BOOST and Tree-based Epistasis Association Mapping (TEAM) as examples of exhaustive searching methods. The overview and resource information of method falling into exhaustive search is shown in Table 2.

Table 2 Exhaustive search methods for detecting epistasis.

	Model-based
BNMBL	Bayesian Network Minimum Bit Length score (2010)[30]
BOOST	Boolean Operation-based Screening and Testing (2009)[20], http://bioinformatics.ust.hk/BOOST.html
INTERSNP	INTERSNP (2009)[31], http://intersnp.meb.uni-bonn.de
TM	Tukey's 1 d.f model for interaction (2006)[32]
FIM	Full Interaction Model (2005)[13], http://www.cbil.ece.vt.edu/ResearchOngoingSNP.htm
PLR	Restricted Partitioning Method (2004)[33]
	Model-free
GWIS	Genome Wide Interaction Search (2013)[34], http://bioinformatics.research.nicta.com.au/gwis
iLOLi	Interacting Loci (2012)[35], http://www4a.biotec.or.th/GI/tools/iloci
TEAM	Tree-based Epistasis Association Mapping (2010)[21], http://www.csbio.unc.edu/epistasis/download.php
COE	COE (2009)[36], http://www.csbio.unc.edu/epistasis/download.php
FastChi	FastChi (2009)[37], http://www.csbio.unc.edu/epistasis/download.php
FastAVOVA	FastANOVA (2008)[38], http://www.csbio.unc.edu/epistasis/
epiSNP	epiSNP (2008)[26], http://animalgene.umn.edu/episnp/index.html
RPM	Restricted Partitioning Method (2004)[39]
CPM	Combinatorial Partitioning Method (2001)[18]
MDR	Multifactor Dimensionality Reduction (2001)[17], http://sourceforge.net/projects/mdr/

4.1.1 BOOST

BOOST is a model-based, exhaustive search method, which is the abbreviation of "BOolean Operation-based Screening and Testing"[20]. Indicated by the name, there are two features of BOOST: First, a new boolean representation is used to accelerate the collecting of contingency table; Second, an upper bound for the likelihood ratio test based on log-linear models and Kirkwood superposition approximation[40] is used to prune insignificant epistatic interactions. Instead of using one row for each SNP, the boolean representation uses three rows, with each row for one specific genotype from 0 through 2. Each row consisted of two strings of boolean values (0 or 1), one for control samples and another for case samples. Each bit in the string represented one individual, and its value was set to 1 if the individual had the corresponding genotype, otherwise 0. By transforming the data representation to the boolean type, the collecting of contingency table information can be efficiently accomplished by performing 64-bit **AND** operation in one instruction, and the counting of "1" bits in a bit string can be treated as hamming weight.

The interested interactions focused by BOOST is not totally equivalent to the epistasis defined in this review. In the terms of the logistic regression model, the likelihood ratio test used in BOOST is based on the deviance of difference between the full model and the main effect model,

$$\log \frac{P\,(Y=1|s_a=i, s_b=j)}{P\,(Y=0|s_a=i, s_b=j)} = \beta_0 + \beta_i^{s_a} + \beta_j^{s_b}.$$

BOOST denotes the log likelihood of the full model under MLE as $\hat{L}_{\mathrm{F}}$, the log likelihood of the main effect model under MLE as $\hat{L}_{\mathrm{M}}$, the log likelihood of log-linear saturated model as $\hat{L}_{\mathrm{S}}$, which was equivalent to the full logistic regression model, and the log likelihood of the homogeneous model as $\hat{L}_{\mathrm{H}}$, which is equivalent to the main effect model. According to the likelihood ratio test, interaction effects are measured by the difference between two log likelihoods of the main effect model and the full model evaluated at their MLEs, i.e., $(\hat{L}_{\mathrm{M}} - \hat{L}_{\mathrm{F}})$. Directly using $(\hat{L}_{\mathrm{S}} - \hat{L}_{\mathrm{H}})$ to test interactions in GWAS still has some difficulties, because iterative methods are needed in model fitting to compute $\hat{L}_{\mathrm{H}}$, which is computationally intensive when hundreds of billions of SNP pairs were required to test. The Kirkwood Superposition Approximation (KSA) is used to approximate the homogenous association model to get a lower bound, $(\hat{L}_{\mathrm{KSA}} \leqslant \hat{L}_{\mathrm{H}})$ of $(\hat{L}_{\mathrm{S}} - \hat{L}_{\mathrm{H}})$. The reason for the replacement is that the calculation of $\hat{L}_{\mathrm{KSA}}$ is straightforward and no iteration is involved. BOOST contains two stages: screening, evaluates all pairwise interactions by using the KSA; testing, for each pair with $2(\hat{L}_{\mathrm{S}} - \hat{L}_{\mathrm{KSA}}) > \tau$, tests the interaction effect using the likelihood ratio statistic $2(\hat{L}_{\mathrm{S}} - \hat{L}_{\mathrm{H}})$.

BOOST only focuses on detecting the eNME, i.e., epistasis displaying no marginal effects, so it achieves high power when applied to simulated dataset with only eNME. One weakness of BOOST is that it can be only used to detect two-locus epistatic interaction, although it runs very fast (it only takes 170 seconds to analyze 10 000 SNPs with 5000 samples on a 3.0 GHz CPU with 4 GB memory running the Windows XP Professional system).

4.1.2 TEAM

TEAM[21] is a model-free, exhaustive search method to detect two-locus epistatic interactions in GWAS. TEAM is dedicated to address the heavy computation aroused by the permutation test. Because many SNPs are correlated, and their correlation structures among genotype profiles can be preserved across enumeration, permutation test is preferred over simple Bonferroni correction. In permutation test, we perform significance test each time when class labels were shuffled. More details about permutation test are covered in Section 5. Following the above notations, the entire search space of two-locus interaction is $HLN(L-1)/2$ with H different permutations. Considering a moderate GWAS setting that $N = 1000$, $L = 100\,000$, and $H = 1000$, we need to conduct 5×10^{15} pairwise tests. Obviously, it is expensive to compute the contingency table for every combination of SNPs on all permutations for calculating the p-values.

Zhang et al.[21] stated that many statistics, such as χ^2 test and likelihood ratio test, were defined as the functions of the counts collected in contingency table. In particular, calculating the two-locus test value needed all 18 observed frequencies in two-way contingency table (Table 1). The authors proved that given an SNP pair and two single-locus contingency tables of each SNP, once the value of $(n_{111}, n_{121}, n_{211}, n_{222})$ fixed, the two-locus test value can be calculated for any permutations. In addition, these four values can be determined incrementally utilizing a Minimum Spanning Tree (MST) built on SNPs. In the MST, the nodes were SNPs, and the edges were the SNP pairs with weights indicating the number of individuals having different genotypes. In other words, the computation of a contingency table in other permutations can be achieved by considering only the individuals with different values, and they had been represented as weights in MST. As it is costly to construct an MST, TEAM constructed an approximate MST instead.

The overall time complexity of TEAM is $O(NLH + NL^2 + W_T NH)$, where $O(NLH)$ is for generating all single-locus contingency tables, $O(NL^2)$ is for building the minimum spanning tree, and $O(W_T NH)$ is for updating the value of $n_{111}, n_{121}, n_{211}$, and n_{222} for H permutations. Comparing to the complexity of brute force approach $O(NL^2 H)$, the performance of TEAM was faster than the latter 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, and to detect both eME and eNME. However, if there is no close-form solution for calculating the statistic test value, using the same framework in TEAM is still computationally intensive when we deal with tons of SNPs.

4.2 Stepwise searching methods for detecting epistatic interactions

Although exhaustive search is computationally possible to test all two-locus epistatic interactions for a moderate size of GWAS data, it requires huge computation time, and loses statistic power when searching higher-order interactions as discussed in Section 2. Instead of explicitly enumerating all possible combinations of k-locus, stepwise search approaches first select a subset of SNPs based on single-locus tests or model-free measures, then conduct tests for multi-locus interactions on the selected subset of SNPs. Compared to exhaustive approaches, stepwise algorithms usually are much faster, and may perform reasonably well for disease associated interactions when the marginal effects exist. As shown in a recent theoretical study[41], the possibility that a high-order (size-k) combination with strong differentiation between case and control groups displaying zero differentiation in all of its subsets decreases dramatically when k increases (generally become impossible for k greater than 5). However, since it removes a considerable portion of SNPs, stepwise search may not be able to find interactions involving loci with small or no marginal effects. The methods are shown in Table 3.

4.2.1 epiForest

Jiang et al.[43] proposed a stepwise approach, called epiForest (detection of *epi*static interactions using random *Forest*), for detecting multi-locus epistasis. epiForest uses SWSFS (Sliding Window Sequential forward Feature Selection) algorithm to select a small set of SNPs as candidates, and then statistically tests up to three-way interactions on the candidates.

In epiForest, the GWAS can be treated as a binary classification problem where cases are positive samples and controls are negative samples, and it utilizes the random forest for the classification. The SNP markers are used as categorical features with three possible values in the classification formulation. The random forest is an ensemble learning methodology originated by Breiman[52]. The basic idea of ensemble learning is to boost the performance of a number of weak learners via a voting scheme, where a weak learner can be an individual decision tree, a single perceptron/sigmoid function, or other simple and fast classifiers. In order to measure the contribution of an SNP to the classification performance, epiForest uses the *gini importance*, which is defined as the summation of all gini decrease of a centain feature over all trees in the forest. It shows that the *gini importance* and the raw importance are very consistent[52], and the computation cost of the *gini importance* is much more economic[43].

Table 3 Stepwise search methods for detecting epistasis.

	Model-based
RanJungle	Random Forests for high-dimensional data (2010)[42], http://www.randomjungle.org/rjungle/rjunglenews
epiForest	Random forest for the detection of epistatic interactions (2009)[43], http://bioinfo.au.tsinghua.edu.cn/epiForest
HapForest	Forest-based approach to identifying gene-gene interactions (2007)[44], http://c2s2.yale.edu/software packages/HapForest/
FITF	Focused Interaction Testing Framework (2006)[45], http://hydra.usc.edu/fitf
TSTLM	Two-Stage Two-Locus Models (2006)[46]
	Model-free
DCHE	Dynamic Clustering for High-order genome-wide Epistatic interactions detecting (2014)[47], http://www.cs.gsu.edu/∼xguo9/DCHE.html
EDCF	Epistasis Detector based on the Clustering of relatively Frequent items (2012)[48], http://www.cs.ucr.edu/∼minzhux/EDCF.zip
PatternRec	Genotype Pattern based on Difference Frequencies (2009)[49], http://www.genemapping.cn
BGTA	Backward Genotype-Trait Association (2006)[50], http://statgene.stat.columbia.edu
ITMDR	Information Theory and MDR (2006)[51], http://www.cbil.ece.vt.edu/ResearchOngoingSNP.htm

There are two stages of epiForest. On the first stage, the random forest is built on all SNPs to classify the GWAS data, and the objective is to obtain the contribution for every SNP measured by *gini importance*. SWSFS algorithm greedily searches for a small subset of SNPs that could minimize the classification error. It adds one SNP at a time by the order from the most significant SNP to the least significant one. SWSFS selects a small set of l ($<< L$, the total number of SNP markers) candidate SNPs that have the most significant contribution to the discrimination of cases against controls. On the second stage, a hierarchical procedure is adopted for one-, two-, and three-way statistical tests to declare the statistical significance that the candidate SNPs are associated with the disease. In the one-way tests, the B statistic proposed by Zhang and Liu[53] is applied to every candidate SNP. Given a predefined significance level α (e.g., 0.05), all SNPs whose p-values are more significant than α after Bonferroni corrections for L tests, are reported. In the two-way tests, the B statistic was also used, and interactions whose p-values were less than α after Bonferroni corrections for $L(L-1)/2$ tests, are reported. If both two SNPs in two-locus interaction have already been detected in the one-way tests, further interaction tests will be skipped, otherwise they are considered for two- and three-way interaction tests. Similarly, in the three-way tests, the B or conditional B statistics were applied to all three-way interactions of the candidates, and those with p-values less than α after Bonferroni corrections for $L(L-1)(L-2)/6$ tests are reported.

With a limit on the size of subset of most important SNPs, the random forest is constructed very fast. epiForest is capable to detect up to three-way epistatic interactions including eME and eNME. However, the detection of epistasis is not as the same as the feature selection process by tradition classification. Phenotype associated combinations of SNPs may not be the only factor leading to the disease effects. Therefore, merely relying on the decision tree, epiForest is insufficient to capture all the SNPs linked to the disease status.

4.2.2 EDCF

Xie et al.[48] proposed a stepwise search algorithm called EDCF to detect multi-locus epistatic interactions in genome-wide case/control studies. The number of SNPs in current GWAS ranges from several hundreds of thousands to a few millions. For interactions involving k loci ($k \geqslant 3$), it is impractical to exhaustively search the whole space since there are $\binom{L}{k}$ possible combinations. Therefore, based on the assumption that the subsets of significant interaction modules were possibly significant, EDCF selects top-df_{s} significant $(k-1)$-locus modules for k-locus interaction test. EDCF starts with searching for the top-df_{s} significant two-locus interaction (where $f_{\mathrm{s}} \geqslant 1$ is a scale factor), and EDCF evaluates all 2-locus combinations ($k = 2$). It recursively searched the interaction space with the top selected SNPs until k reaches user defined value. Due to the large number of reported significant interactions, biologists may only be interested in the d most significant ones, so only top-d interactions are generated by EDCF.

To measure the statistic significance of epistasis, the test utilized by EDCF is Pearson's χ^2 test. In order to give a reasonable elevation of SNP combinations, EDCF partitions all 3^k genotype combinations for k-locus into three groups, defined as G_0, G_1, and G_2. G_0 contains all combinations that occur significantly more frequently in cases than in controls (presumably high-risk combinations); G_2 contains those who occur significantly more frequently in controls than in cases (presumably low-risk combinations); and G_1 contains the remaining genotypes. To group the genotype combinations, EDCF assumes the population po with the same genotype followed a Binomial distribution. For the Binomial distribution, the parameter n equals to the total count of po, and another parameter p equals to the ratio of case or control count over n. To obtain high-risk combinations, EDCF uses case count over n, while it used control count over n to obtain low-risk combinations. Given a significance level α_{s}, let T_{a} and T_{u} denote the critical value corresponding to α_{s} for cases and controls, respectively. The genotype is treated as high-risk if the count of cases in this genotype is larger than T_{a}. Similarly, the genotype is treated as low-risk if the number of controls in this genotype was larger than T_{u}. Once all genotype combinations for k SNPs have been grouped into G_0, G_1, and G_2, EDCF collects a 3×2 contingency table, where the rows represented three groups and the columns represented two class labels for case and control. The χ^2 statistic with 2 degrees of freedom[54] is used to measure the significance of the interactions.

By combining the advantages of the χ^2 test and high/low-risk genotype combinations, EDCF is an effective and efficient algorithm for detecting epistatic interactions for GWAS, especially when interactions contained strong main effects. Comparing to the model-based exhaustive search approaches, like BOOST, extensive experiments on simulated data illustrate that EDCF tends to lose certain powers on detecting embedded disease models without main effects (eNME)[48].

4.3 Stochastic searching methods for detecting epistatic interactions

Instead of explicitly enumerating all possible combinations of k-locus, stochastic methods use random sampling procedures to search the space of interactions. Among them, Bayesian Epistasis Association Mapping (BEAM)[53] is one representative. BEAM takes case-control genotypes as input, and iteratively uses the Markov Chain Monte Carlo (MCMC) approach to calculate the posterior probability of a locus or multiple loci associated with the disease. Tang et al.[55] extended BEAM in their epistatic MOdule DEtection (epiMODE) method. epiMODE uses Gibbs sampling and a reversible jump MCMC procedure to search for significant epistatic modules. A basic framework of stochastic search strategy can be generalized as follows.

Given a set of states (or configurations) $X = \{X_1, \cdots, X_M\}$ and a function, Eval$(\cdot)$, that evaluates each configuration, four basic steps are employed in stochastic search to find X^* such that Eval(X^*) is greater than all Eval(X_i) for all other possible values of X_i:

- **Step 1** Initialize the configuration X.
- **Step 2** Calculate the function value, Eval(X).
- **Step 3** Randomly select X' in the neighbors of X.
- **Step 4** Obtain the new function value, Eval(X'); if Eval(X') is better than Eval(X), set X to X'; go to Step 2 until reach the maximum iteration count.

Particularly, the states are the assignments of SNPs (jointly associated with disease or not) and the function can be the posterior probability from predefined models. It is necessary to note that existing methods falling to stochastic search are all model-based as shown in Table 4. In the following, we use SNPHarvester and epiMODE to illustrate the basic idea in stochastic search for detecting epistasis.

4.3.1 SNPHarvester

Yang et al.[56] proposed a method, SNPHarvester, using a path selection procedure to sample the searching space. SNPHarvester first identifies disease-associated SNP groups from thousands of SNPs. It assumes that multiple epistatic interactions rather than a single one are expected to be found due to the sophisticated regulatory mechanism encoded in the human genome. SNPHarvester then generates multiple paths with a generic score function to identify multiple significant SNP groups. After that, L_2 penalized logistic regression model[59] is used as a post-processing step to extract epistasis from selected SNP groups. The screening process based on path selection greatly reduces the number of SNPs for further statistic measure, and it makes SNPHarvester possible to directly apply to large GWAS dataset for detecting high-order epistatic interaction.

Table 4 Stochastic search methods for (model-based) detecting epistasis.

epiMODE	epistatic MOdule DEtection (2009)[55], http://bioinfo.au.tsinghua.edu.cn/epiMODE/
SNPHavester	Filtering-based approach for detecting epistatic interactions (2009)[56], http://bioinformatics.ust.hk/SNPHarvester.html
LogicFS	LogicFS (2008)[57], http://bioconductor.org/packages/2.4/bioc/html/logicFS.html
BEAM	Bayesian Epistasis Association Mapping (2007)[53], http://www.fas.harvard.edu/~junliu/BEAM/
MCLR	Monte Carlo Logic Regression (2005)[58], http://cran.r-project.org/web/packages/LogicReg/index.html

Before giving the details of SNPHarvester, we need to introduce its assumption used by it. SNPHarvester partitions the L SNP markers into three classes as follows[56]:

- Class 0: SNPs are unassociated to the disease.
- Class 1: SNPs influence the disease risk independently, i.e., they show marginal effects.
- Class 2: SNPs contribute little effects to the disease risk individually but influence the disease risk jointly.

SNPHarvester consists of two steps: the filtering and the model-fitting steps. In the filtering, it randomly initializes the starting point of each path, and generates the path by a local search algorithm called PathSeeker. PathSeeker uses the score function to measure the association between a k-SNP group and the phenotype, then records the SNP group whose score exceeds a fixed threshold. It adopts the χ^2 value as the score function, and the threshold is determined by Bonferroni correction. PathSeeker first removes significant single SNPs according to their χ^2 test values, because SNPHarvester is only interested in epistatic interactions that have weak main effects but significant joint effect. Then it randomly picks k-SNP to form an active set $S = \{\text{SNP}_1, \text{SNP}_2, \cdots, \text{SNP}_k\}$, and leaves the rest of the SNPs to form a candidate set S_c for the next random selection. A swapping operation is applied between S and S_c to switch two SNPs, $\text{SNP}_i \in S, \text{SNP}_j \in S_\text{c}$, if the new k-SNP group achieves better χ^2 test value. For generating one k-SNP group, PathSeeker needs to try a total of $k(n-k)$ combinations. The identified group with the local optimum χ^2 test value is removed from the L SNPs. In next iteration, PathSeeker continues to select k SNPs to form another active set, and the rest $n-2k$ SNPs form a candidate set. The time complexity to generate m groups is $O(kLm)$, which is affordable even when there were $> 100\,000$ SNPs. The identified m k-SNP groups are employed for model-fitting in second step. The model fitting is used to distinguish SNPs that have joint effects from those SNPs that only have marginal effects.

Due to the feature of randomization technique, SNPHarvester is expected to perform no better than exhaustive search. Since there are numerous local optimal paths, the performance of the filtering step is poor, which leads to little power to detect the ground-truth interactions. Comparing to brute-force approaches, SNPHarvester is suitable to detect three-way or higher-order epistatic interactions. SNPHarvester focuses only on eNME, because it utilizes the L_2 penalizes logistic regression model.

4.3.2 epiMODE

Tang et al.[55] developed an extension of BEAM using the Gibbs sampling strategy, named epiMODE, to facilitate the detection of epistatic modules. In epiMODE, the epistatic interaction module is considered as the basic units of disease susceptibility loci that independently influence the phenotype. On the basis of this notion, it adopts a Bayesian marker partition model to explain the observed case-control data, and further generalizes this model to account for the existence of LD between genetic variants. The genetic variants (SNPs) belonging to a epistasis module are simulated in a procedure, called Reversible Jump MCMC (RJ-MCMC), based on Gibbs sampling strategy. Further hypothesis testing is applied to screen out statistically significant modules.

In epiMODE, the penetrance of the combinatory genotypes of two subsets, S_1 and S_2, of SNPs can be described as

$$p(D|G_{S_1}, G_{S_2}) = f(G_{S_1}, G_{S_2}),$$

where G represents a combinatory genotype of the multiple loci, and $f(\cdot)$ is the function denoting how combinatory genotypes determines the disease penetrance. If

$$p(D|G_{S_1}, G_{S_2}) = f(G_{S_1}, G_{S_2}) = f_1(G_{S_1}) f_2(G_{S_2}),$$

is always true, the relationship between the two subsets of loci S_1 and S_2 is defined as “independently contributing” to the disease. Otherwise, the relationship between them is defined as “epistasis”. With these concepts, the problem of finding epistatic interactions was equivalent to a problem of assigning the SNP markers to the defined modules. Particularly, the assignment for an SNP can be done by first calculating the probability of the observed data given a certain partition pattern using a Bayesian model. epiMODE assumes that all loci are in LD, also known as independent. It uses first-order Markov model to account for the situation in which a set of SNPs are in LD with a disease susceptibility. Finally, epiMODE resorts to hypothesis testing to screen out significant epistatic interactions.

As discussed in recent reviews, a weakness of epiMODE is that it can not deal with datasets with more than 10 000 SNPs in affordable time[25]. Obviously, epiMODE spends too much time on the iteration of the reversible jump Markov chain Monte Carlo procedure. Another drawback of epiMODE is that it has no ability on datasets with 5% genotyping error, which is common in real GWAS datasets.

4.4 Heuristic searching methods for detecting epistatic interactions

Heuristics approaches adopt machine learning techniques, such as neural networks and predictive rules, to search the space of epistatic interactions rather than explicitly enumerating and testing all the combinations of k-locus. The overview and resource information of method falling into heuristic search is shown in Table 5. Two examples falling into this field are MECPM and MIC. MECPM proposes a phenotype posterior under a maximum entropy principle, and uses greedy searching to find epistatic interactions which are treated as model constraints. MIC is a model-free method which defines significant tests based on mutual information, and uses k-means clustering to narrow down the candidate groups.

4.4.1 MECPM

Miller et al.[60] proposed a method, named MECPM, to identify markers/interactions by building the

Table 5 Heuristic search methods for detecting epistasis.

Model-based	
MECPM	Maximum Entropy Conditional Probability Modelling (2009)[60], http://www.cbil.ece.vt.edu/ResearchOngoingSNP.htm
GPNN	Genetic Programming optimized Neural Network (2006)[61]
CMM	Tree and spline based association analysis (2004)[62], http://lib.stat.cmu.edu/
Model-free	
[1mm] epiMiner	epistasis Miner (2014)[63], https://sourceforge.net/projects/epiminer/files/
MIC	Mutual Information (2014)[64]
DPM	Discriminative Pattern Mining (2012)[65]
AES	AntEpiSeeker (2010)[66], http://nce.ads.uga.edu/~romdhane/AntEpiSeeker/index.html
SNPRuler	Predictive rule inference for epistatic interaction detection (2010)[67], http://bioinformatics.ust.hk/SNPRuler.zip
AGR	Association Graph Reduction (2009)[68]
MSH	MegaSNPHunter (2009)[69]
RST	Rough Set Theory (2009)[70]
TWG	Trimming, Weighting and Grouping (2001)[71], http://linkage.rockefeller.edu/ott/sumstat.html

phenotype-predictive models. MECPM treats the problem as a supervised feature selection in statistical classification where "cases" and "controls" are two classes. The goal is to select the feature subset which leads to the best classification performance. According to the principle of Maximum Entropy (ME), a probability model should agree with all known information and remain maximal uncertainty[72]. In the ME framework, a posterior model (classifier) is defined by the interactions to satisfy the specified constraints and maximize the conditional entropy at the same time. Without any constraint, the ME posterior is a uniform probability mass function over all classes, and the accuracy of the resulting model is compromised. Each encoded constraint reduces the (maximum) entropy, which yields a more predictive posterior. The importance of constraint is measured by the decreased amount of the ME distribution's entropy by applying the constraint. The ME optimization problem is convex with linear constraints, and the standard solutions guarantee to the convergence.

A variant of greedy search based on Bayesian Information Criterion (BIC) is used for choosing the ME constraints (interactions up to five-way) when model grows, and determining the number of constraints to terminate the model growing. BIC is a model selection criterion that it captures a trade-off between data likelihood and model complexity among a finite set of models. In the seed selection and accretion of constraints, MECPM measures the Kullback-Leibler divergence[73] between the probability mass functions for all possible one- and two-way SNP constraints. It shows that at a given order, the constraints, which are furthest from the existing model, will decrease BIC cost the most if they are added to the model. Therefore, an alternative to accretion of all possible constraints is to use seed pool, which comprises the constraints with largest Kullback-Leibler divergence at first and second orders, and MECPM adds one constraint at a time.

MECPM builds the phenotype posterior under a maximum entropy principle, and encodes constraints into the model with a 1-to-1 correspondence to the epistatic interactions. From the experimental results, the time complexity is considerably high that it took 750 hours and 7.5 hours to detect five- and two-way interaction for a dataset with only 1000 SNPs and 2000 balanced samples. Another flaw of MECPM is that it does not give any significance assessment, and thus it is hard to tell the types of reported epistatic interactions (eME or eNME).

4.4.2 MIC

Leem et al.[64] proposed an algorithm (MIC) based on mutual information for detecting high order epistatic interactions in GWAS. As claimed by the authors, mutual information does not suffer the approximation issue as other statistic tests. Take Pearson's χ^2 test for example, the approximation to the χ^2 distribution breaks down, if the expected frequencies are too low. The issue can be even worse when we detect higher order epistatic interactions. Mutual Information $I(S;Y)$ is defined to capture the amount of the information shared by two random variables, S and Y. Let S denote a subset of L SNPs, and Y denote the class labels. MIC uses the value of $I(S;Y)$ to imply the significance of the association between the SNP combination and the disease. Let $A = \{A_1, A_2, \cdots, A_n\}$ be a partition of S. The entropy $H(A)$ of A is defined as follows:

$$H(A) = -\sum_{i=1}^{n} \frac{|A_i|}{L} \log \frac{|A_i|}{L}.$$

Then the mutual information between the joined partition $\{A^{(1)}, A^{(2)}, \cdots, A^{(k)}\}$ and a partition Y can be defined as follows:

$$I(A^{(1)}, A^{(2)}, \cdots, A^{(k)}; Y) = H(A^{(1)}, A^{(2)}, \cdots, A^{(k)}) + H(Y) - H(A^{(1)}, A^{(2)}, \cdots, A^{(k)}; Y),$$

where $H(A^{(1)}, A^{(2)}, \cdots, A^{(k)})$ is the extension of $H(A)$ to multiple partition. Based on the definition of mutual information, MIC tries to find the set of k SNPs that maximized the value $I(A^{(1)}, A^{(2)}, \cdots, A^{(k)}; Y)$.

Since the size of current GWAS data can reach up to hundreds of thousands of SNPs, it takes too much time to calculate the $I(A^{(1)}, A^{(2)}, \cdots, A^{(k)}; Y)$ for every k-modules ($k \geqslant 3$). Therefore, MIC uses k-means clustering to reduce time complexity by placing strongly interacting SNPs into different clusters. MIC can be separated into three steps: clustering, candidates selection, and finding the k-SNP module with high mutual information value. The summary of three steps is as follows:

- **Step 1** k-means clustering is used on the set of SNPs with the distance measured by mutual information between two SNPs.
- **Step 2** Top d SNPs are selected in each cluster according to their scores, which is calculated as the sum of all mutual information values measured between the selected SNP to the rest SNPs in the same cluster.

- **Step 3** MIC exhaustively searches k SNP modules with the highest value of mutual information among kd candidates.

An explicit advantage of MIC is that the speed of algorithm is very fast since k-means takes linear time to do the clustering, and the exhaustive search can be done fast when f is small. However, MIC does not display its false control rate under the null hypothesis, so the significance of reported interaction is unclear. Based on the experimental results of MIC, it shows power to detect eME and eNME, although it cannot distinguish the types of reported modules.

5 Evaluation Methodology

The methods for searching epistatic interactions in GWAS should be evaluated not only with small simulated datasets, but also with large and complicated real datasets. In the experiments of the simulated data, different disease models with or without marginal effects should be embedded into the datasets. In the experiments of real data, well-studied GWAS projects with established epistatic interactions can be employed as standard measurement. In order to evaluate and compare the statistic power of existing methods, more than one metrics should be considered. In the following sections, definitions of five popular power metrics are given. Then we list multiple disease models with or without marginal effects for simulation. Several real GWAS projects with links to the resource are also provided. Finally, we discuss the control of false discovery rate, and summarize the general advantages and disadvantages of existing methods in seven categories.

5.1 Detection powers

Detection power can be defined in several ways, depending on what we desire to measure. Before giving the definitions of detection power, several terms and notations should be introduced. Since a complex disease may be caused by multiple epistatic interactions, each of which consists of one or more SNPs, it is necessary to simulate multiple epistasis models in a dataset. Suppose that we generate W datasets with the same parameter settings, i.e., the embedded epistatsis models have the same number of SNPs and the same values of parameters. For the i-th dataset, we use c_i to denote the count of epistasis models, and x_{ij} to denote the number of SNPs involved in model j. The total number of ground-truth SNPs for i-th dataset is $C_i = \sum_{j=1}^{c_i} x_{ij}$. Usually, the method returns a ranked list of SNPs or SNP combinations, which implies their descending importance. Since the statistical significance increases as the number of SNPs in one model increases, it is difficult to compare the importance of modules with uneven number of SNPs. Therefore, we set c_i to a fixed value $k \geqslant 2$, and we assume that it is able to get top ι_i SNPs from the methods for the i-th dataset.

- Precise Power (PP) is defined as the proportion of datasets in which all ground-truth SNPs are ranked at highest by a method. This power definition evaluates the sensitivity to detect the interaction as a whole. It is written as
$$\text{PP} = \frac{1}{W}\sum_{i=1}^{W} z_{\text{PP},i},$$
where $z_{\text{PP},i} \in \{0, 1\}$ is the detection indicator, i.e., if the detected set of top ι_i ($\iota_i = C_i$) SNPs only consists of ground-truth SNPs in the i-th dataset, $z_{\text{PP},i} = 1$; otherwise, $z_{\text{PP},i} = 0$.
- Average Power (AP) is defined as an average proportion of true positives in the top ι_i ($\iota_i = C_i$) SNPs. It is written as
$$\text{AP} = \frac{1}{W}\sum_{i=1}^{W}\sum_{j=1}^{c_i}\frac{|x_{ij} \cap \iota_i|}{x_{ij}},$$
where $|x_{ij} \cap \iota_i|$ is the number of ground-truth SNPs in the top ι_i SNPs identified in the i-th dataset.
- Extended Power (EP) is defined as the ratio of the number of ground-truth SNPs appearing in the top ι_i ($\iota_i > C_i$) SNPs for the i-th dataset by a method. It can be written as
$$\text{EP} = \frac{1}{W}\sum_{i=1}^{W}\frac{1}{C_i}\sum_{j=1}^{c_i}|\iota_i \cap x_{ij}|,$$
where $|\iota_i \cap x_{ij}|$ is the number of ground-truth SNPs in the top ι_i ($\iota_i > C_i$) SNPs detected in dataset i for j-th epistatsis model.
- General Power (GP) is defined as the percentage of W datasets in which at least 1 ground-truth SNPs is identified in the top ι_i SNPs for all embedded models in a dataset. It is written as
$$\text{GP} = \frac{1}{W}\sum_{i=1}^{W} z_{\text{GP},i},$$
where $z_{\text{GP},i} \in \{0, 1\}$ is the detection indicator, i.e., if the detected set of top ι_i SNP consists at least on ground-truth SNPs for all embedded models in dataset i, $z_{\text{GP},i} = 1$; otherwise, $z_{\text{GP},i} = 0$.

- Receiver Operating Characteristic (ROC) curve is a graphical plot showing how many ground-truth SNPs detected for a given false positive SNP count. It is generated by plotting the fraction of true positives out of the total actual positives vs. the fraction of false positives out of the total actual negatives at various threshold settings[74].

5.2 Simulated disease models

There is a wide spectrum of epistases. Some show both marginal (main) effects and interactive effects, and others show no marginal effects but interactive effects. We refer to the former as eME and the latter as eNME[25]. A disease model can be defined either by specifying the penetrance table or the odds table. Let $p\,(D|g_i)$ denote the probability that an individual will be affected with a given genotype combination g_i. Relations among penetrance $p\,(D)$, odds ODD_{g_i}, and $p\,(D|g_i)$ can be calculated as Eqs. (4) and (5).

$$\text{ODD}_{g_i} = \frac{p\,(D|g_i)}{p(\overline{D}|g_i)} = \frac{p\,(D|g_i)}{1 - p\,(D|g_i)} \tag{4}$$

$$p\,(D|g_i) = \frac{\text{ODD}_{g_i}}{1 + \text{ODD}_{g_i}} \tag{5}$$

In Ref. [20], the disease prevalence $p(D)$ and genetic heritability h^2 are given by Eqs. (6) and (7).

$$p\,(D) = \sum_i p\,(D|g_i)\,p\,(g_i) \tag{6}$$

$$h^2 = \frac{\sum_i \left(p\,(D|g_i) - p\,(D)\right)^2 p\,(g_i)}{p\,(D)\,(1 - p\,(D))} \tag{7}$$

As introduced in Ref. [13], the marginal odds at locus A for two-locus epistasis model containing locus A and B are defined by

$$\frac{p(D|g_A)}{p(\overline{D}|g_A)} = \frac{\sum_{g_B} p(D|g_A, g_B)p(g_B)}{\sum_{g_B} p(\overline{D}|g_A, g_B)p(g_B)},$$

where $p(D|g_A, g_B)$ is the probability that an individual has the disease given that they have a combination of genotype g_A at locus A and genotype g_B at locus B. For each model, the parameter λ is used to measure the marginal effect size, which is written as follows.

$$\lambda = \frac{p(D|s_a = 0)}{p(\overline{D}|s_a = 0)} \Big/ \frac{p(D|s_a = 1)}{p(\overline{D}|s_a = 1)} - 1,$$

where 0 and 1 are the genotypes at locus A used in the Section 4. Based on the marginal effect size, if $\lambda = 0$, the model displays no marginal effects; otherwise, the model has marginal effects. We describe several examples of epistasis models with or without marginal effects in the following paragraphs.

- Models with marginal effects. We consider four two-loucs epistasis models whose odds tables are given in Table 6. Model 1 is a multiplicative model. Model 2 is an epistasis model that has been used to describe handedness and the color of swine. Model 3 is a classical epistasis model. Model 4 is the well known XOR (exclusive OR) model. Two three-locus epistasis models are shown in Tables 7 and 8. In Model 5, increased disease risk is assigned to certain genotype combinations, and marginal effect of each disease locus ranges from very small to zero. Model 6 is a three-locus model, and it achieves the maximum heritability at the low end of disease prevalence ($\theta \in \left(0, \dfrac{1}{16}\right]$).

Table 6 Odds tables of epistasis models 1, 2, 3, and 4.

Model 1	BB	Bb	bb
AA	α	α	α
Aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$
aa	α	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^4$
Model 2	BB	Bb	bb
AA	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
Aa	$\alpha(1+\theta)$	α	α
aa	$\alpha(1+\theta)$	α	α
Model 3	BB	Bb	bb
AA	α	α	$\alpha(1+\theta)$
Aa	α	$\alpha(1+\theta)$	α
aa	$\alpha(1+\theta)$	α	α
Model 4	BB	Bb	bb
AA	α	$\alpha(1+\theta)$	α
Aa	$\alpha(1+\theta)$	α	$\alpha(1+\theta)$
aa	α	$\alpha(1+\theta)$	α

Table 7 Odds table of epistasis model 5.

	$BBCC$	$BbCC$	$bbCC$
AA	α	α	α
Aa	α	α	$\alpha(1+\theta)$
aa	α	$\alpha(1+\theta)$	α
	$BBCc$	$BbCc$	$bbCc$
AA	α	α	$\alpha(1+\theta)$
Aa	α	$\alpha(1+\beta\theta)$	α
aa	$\alpha(1+\theta)$	α	α
	$BBcc$	$Bbcc$	$bbcc$
AA	α	$\alpha(1+\theta)$	α
Aa	$\alpha(1+\theta)$	α	α
aa	α	α	α

Table 8 Penetrance table of epistasis model 6.

	CC			Cc			cc		
	BB	Bb	bb	BB	Bb	bb	BB	Bb	bb
AA	0	0	16θ	0	0	0	0	0	0
Aa	0	0	0	0	4θ	0	0	0	0
aa	0	0	0	0	0	0	16θ	0	0

- Models without marginal effects. Disease models displaying no main effects have been carefully discussed, and a wide spectrum of these models have been provided[75]. Velez et al.[75] generated a total of 70 different penetrance functions that define a probabilistic relationship between genotype and phenotype. The susceptibility to disease is dependent on genotypes from two loci in the absence of any marginal effects. A total of five models for each of the 14 heritability-allele frequency combinations were generated for a total of 70 models. The details of the 70 penetrance functions are available online (http://discovery.dartmouth.edu/epistatic_data/).

In the simulation, we have two ways to numerically solve the parameters (α and θ): either (1) specify the disease prevalence $p(D)$, the genetic heritability h^2, and Minor Allele Frequency (MAF), or (2) fix the MAFs and marginal effect size λ. A detail depiction of the simulation process is available in Ref. [53].

5.3 Real GWAS data

In order to evaluate the performance of the method for identifying epistatic interactions truly involving biological processes, we need to test the methods on well studied GWAS datasets. In Table 9, we list 11 GWAS datasets with their references. The Wellcome Trust Case Control Consortium (WTCCC) is a collaboration of many British research groups. The genetic signals of seven common human diseases have been examined in the first phase of WTCCC. Take the disease AMD for example, it is the leading cause of blindness for people over 50, and it is a common eye disease that is associated with aging and gradually destroys sharp, central vision. For the AMD dataset, Klein et al.[76] reported two SNPs (rs380390 and rs1329428) that were associated with AMD. Therefore, it is reasonable to tell the performance of methods whether these two SNPs have been reported by proposed methods. In addition, it is interesting to compare the results generated by new methods with the results from existing tools to check novel findings.

Table 9 Real GWAS datasets.

Name	Number			Ref.
	SNP	Case	Con.	
AMD (Age-related Macular Degeneration)	116 204	96	50	[76]
CD (Crohn's Disease)	317 503	513	515	[77]
LOAD (Late-Onset AD)	502 627	861	550	[78]
RA (Rheumatoid Arthritis)	545 080	868	1194	[79]
WTCCC				
RA (Reumatoid Arthritis)	459 012	1860	2983	
HT (HyperTension)	459 012	1952	2983	
CD (Crohn's Disease)	459 012	1748	2983	
CAD (Coronary Artery Disease)	459 012	1926	2983	[80]
BD (Bipolar Disorder)	459 012	1868	2983	
T1D (Type 1 Diabetes)	459 012	1963	2983	
T2D (Type 2 Diabetes)	459 012	1924	2983	

5.4 False discovery rate control

False Discovery Rate (FDR) control is used in multiple hypothesis testing to correct for multiple comparisons. In a list of findings (i.e., studies where the null-hypotheses are rejected), FDR procedures are designed to control the expected proportion of incorrectly rejected null hypotheses ("false discoveries")[81]. In the detection of epistatic interactions, there are two popular procedures for FDR control: permutation test and Bonferroni correction, of which the brief introductions are given in the following paragraphs.

5.4.1 Permutation test

Without a normal assumption for the distribution of data, like genome-wide case-control data, permutation test is an effective non-parametric approach to establish the null distribution of a test statistic. In GWAS, two hypotheses need to be tested: H_0, the module is not associated with the disease; H_1, the module is associated with the disease. The following procedures can be applied for permutation test in detecting gene-gene interactions:

- **Step 1** Use the proposed method to the original case-control data with the appropriate parameter setting, list the candidate epistatic modules in order of their significance indicators, like p-value.
- **Step 2** Generate new case-control data by shuffling individuals' labels between case and control, apply the method with the same parameter setting to the newly permuted data, and record the value of indicator which is most significant from the reported epistasis modules.

- **Step 3** Repeat Step 2 for ω times, and make a list of ω significance indicators.
- **Step 4** For each candidate epistasis module generated using the original case-control data, record how many significance indicators from permuted data are more significant than the current candidate module. The quotient using this count divided by ω is the p-value of permutation test for the candidate module by employing proposed method.

Usually, we set $\omega = 1000$. By giving a significant level $\alpha = 0.05$, we remove the non-significant epistatic interactions if their p-values from permutation test are larger than α.

5.4.2 Bonferroni correction

In statistics, the Bonferroni correction is a method used to counteract the problem of multiple comparisons. The Bonferroni correction is based on the idea that if an experimenter is testing n dependent or independent hypotheses on a set of data, the probability of type I error is offset by testing each hypothesis at a statistical significance level $\dfrac{1}{n}$ times what it would be if only one hypothesis is tested. For example, given a GWAS dataset with L SNPs, the adjusted p-value of reported epistatic interaction is calculated by multiplying the number of tests. For detecting two-locus interactions, the number of tests is $L(L-1)/2$, and for detecting three-locus interactions, the number of tests is $L(L-1)(L-2)/6$.

5.5 Advantages and disadvantages

All eight methods reviewed in this paper have demonstrated respective utilities based on the experimental results conducted in the recent literatures[25, 64, 82]. We summarize their merits and weaknesses in Table 10. For the exhaustive methods, since they enumerate and test all possible combinations of k-locus, it can report all significant epistasis without losing any power. An explicit drawback of exhaustive search is the intensive computation. In order to accelerate the process, finding an approximation with less intensive computation for calculating the significance value is a possible solution, but it will lose power. Instead of testing all the k-locus combinations of SNPs, stepwise, stochastic, and heuristic searches only select a subset of SNPs for further tests. A common disadvantage of them is the power lost. As shown in a recent theoretical study[41], the possibility that a high-order (size-k) combination with strong differentiation displaying zero differentiation in all of its subsets decreases dramatically when k increases (generally become impossible for k greater than 5). Based on the above theory, stepwise methods usually exhaustively test all two-locus interaction, and select top-d SNPs for higher-order tests. The scalability of stepwise methods is good if one single test can be done very fast, like

Table 10 The advantages and disadvantages of existing method for searching epistatic interactions.

	Model-based	Model-free
	Exhaustive search	
A	High power; Discrimination of eME and eNME	Low complexity for single significant test; Enumerating two-locus tests is computational possible.
D	Time consuming in model fitting if the size of SNP module $k \geqslant 3$; Power lost if using approximation test	No discrimination on the model types
	Stepwise search	
A	Discrimination of eME and eNME	Low complexity for single significant test; The size of tested SNP module can reach up to 5.
D	Power lost if the higher-order modules displaying insignicant marginal effects; Time consuming if the screening process is complicated.	
	Stochastic search*	
A	Performance is good if SNP models display strong marginal effects.	
D	Power lost if model cannot capture the relationship; Time consuming if the number of iteration for sampling is huge.	
	Heuristic search	
A	High power for GWAS data with moderate size	Low complexity for single significant test; Screening process is efficient.
D	Power lost if model cannot capture the relationship; Time consuming if model building is compuatationally intensive.	Power lost if the high-order modules displaying insignicant marginal effects; Lack of controlling of false discovery rate.

Note: A, stands for advantage; D, stands for disadvantage; *, all stochastic search methods in this review are belonging to model-free category.

EDCF[48]. Stochastic methods use random sampling procedures to search the space of interactions. The key factor influencing the performance of stochastic method is the selection of sampling procedure. As shown in the experimental results from recent literatures[25, 64, 82], stochastic methods lose more power than the other three strategies, and the execution of stochastic methods is time consuming if the number of iteration of sampling is large for huge GWAS data. Heuristics approaches utilize machine learning techniques, such as neural networks and predictive rule, to search the space of epistatic interactions. Most heuristic methods are running very fast compared to the proceeding three strategies. However, most of them lack the control of FDR, and thus it is difficult to tell how good they are without control of type I error, and to avoid false epistatic interactions.

6 Summary and Outlook

As we have seen, there are numerous methods and an even larger number of software implementations allowing investigators to examine disease-associated epistatic interaction based on available GWAS data generated from large-scale genotyping projects. Although the precise details of the methods are different, in many cases there are closely conceptual links between the approaches. Existing methods for searching epistasis can be grouped into two types, model-based and model-free by considering the assumption on the data. In addition, based on the searching strategies, they can also be categorized into four types, i.e., exhaustive, stepwise, stochastic, and heuristic approaches. In this review, we identify 43 methods for detecting disease-associated epistatic interactions. Eight of them are discussed in details in the paper. The evaluation methodologies for searching epistasis are elaborated, including five definitions of detecting power, disease models for simulating dataset, resources of well studied real GWAS datasets, and the control for FDR. Finally, we generally summarize the advantages and disadvantages of popular GWAS tools.

Current approaches merely focus on the relationship between genotypes and the phenotype traits. However, more information, like SNP positions on the chromosomes and suspicious epigenome patterns, can be helpful to construct a systematically causal relation behind the diseases. SNPs are found in coding areas as well as in the non-coding regions of genes. In general, SNPs in coding regions, termed as non-synonymous SNPs, may have a greater impact on the gene function than those in non-coding regions. The non-synonymous SNPs may cause pathological consequences either by affecting the 3-D conformation of protein structure or their corresponding active domains. Consequently they can potentially disrupt the recruitment scaffold of corresponding protein. The epigenome consists of a record of histone modifications and DNA methylation of an organism. There are some evidences of correlation between SNPs and the quantitative traits of DNA methylation[83-85]. Therefore, how to integrate the region information of SNPs and the changes on epigenome into the detection of disease-associated epistatic interactions will be a promising and challenging direction in genome-wide association studies.

Acknowledgements

This study was supported by the Molecular Basis of Disease (MBD) program at Georgia State University. This work was also supported in part by the National Natural Science Foundation of China (Nos. 61379108 and 61232001).

References

[1] W. S. Bush and J. H. Moore, Genome-wide association studies, *PLoS Computational Biology*, vol. 8, no. 12, p. e1002822, 2012.

[2] J. L. Haines, M. A. Hauser, S. Schmidt, W. K. Scott, L. M. Olson, P. Gallins, K. L. Spencer, S. Y. Kwan, M. Noureddine, J. R. Gilbert, et al., Complement factor h variant increases the risk of age-related macular degeneration, *Science*, vol. 308, no. 5720, pp. 419-421, 2005.

[3] G. M. Cooper, J. A. Johnson, T. Y. Langaee, H. Feng, I. B. Stanaway, U. I. Schwarz, M. D. Ritchie, C. M. Stein, D. M. Roden, J. D. Smith, et al., A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose, *Blood*, vol. 112, no. 4, pp. 1022-1027, 2008.

[4] M. Fareed and M. Afzal, Single nucleotide polymorphism in genome-wide association of human population: A tool for broad spectrum service, *Egyptian Journal of Medical Human Genetics*, vol. 14, no. 2, pp. 123-134, 2013.

[5] The 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing, *Nature*, vol. 467, no. 7319, pp. 1061-1073, 2010.

[6] K. Christensen and J. C. Murray, What genomewide association studies can do for medicine, *N. Engl. J. Med.*, vol. 356, no. 11, pp. 1094-1097, 2007.

[7] A. K. Daly, Genome-wide association studies in pharmacogenomics, *Nature Reviews Genetics*, vol. 11, no. 4, pp. 241-246, 2010.

[8] O. L. Griffith, S. B. Montgomery, B. Bernier, B. Chu, K. Kasaian, S. Aerts, S. Mahony, M. C. Sleumer, M. Bilenky, M. Haeussler, et al., Oreganno: An open-access community-driven resource for regulatory annotation, *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D107-D113, 2008.

[9] S. Mooney, Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis, *Briefings in Bioinformatics*, vol. 6, no. 1, pp. 44-56, 2005.

[10] J. K. DiStefano and D. M. Taverna, Technological issues and experimental design of gene association studies, in D*isease Gene Identification*. Springer, 2011, pp. 3-16.

[11] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn, Genome-wide association studies for complex traits: Consensus, uncertainty and challenges, *Nature Reviews Genetics*, vol. 9, no. 5, pp. 356-369, 2008.

[12] Q. He and D.-Y. Lin, A variable selection method for genome-wide association studies, *Bioinformatics*, vol. 27, no. 1, pp. 1-8, 2011.

[13] J. Marchini, P. Donnelly, and L. R. Cardon, Genome-wide strategies for detecting multiple loci that influence complex diseases, *Nat. Genet.*, vol. 37, no. 4, pp. 413-417, 2005.

[14] W. Bateson and G. Mendel, *Mendel's Principles of Heredity*. Putnam's, 1909.

[15] W. Bateson, *Mendel's Principles of Heredity*. Cambridge University Press, 1909.

[16] H. J. Cordell, Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans, *Human Molecular Genetics*, vol. 11, no. 20, pp. 2463-2468, 2002.

[17] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, Multifactor-dimensionality reduction reveals high-order interactions among estrogenmetabolism genes in sporadic breast cancer, *The American Journal of Human Genetics*, vol. 69, no. 1, pp. 138-147, 2001.

[18] M. Nelson, S. Kardia, R. Ferrell, and C. Sing, A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation, *Genome Research*, vol. 11, no. 3, pp. 458-470, 2001.

[19] H. J. Cordell, Detecting gene-gene interactions that underlie human diseases, *Nat. Rev. Genet.*, vol. 10, pp. 392-404, 2009.

[20] X. Wan, C. Yang, Q. Yang, H. Xue, X. Fan, N. L. Tang, and W. Yu, Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies, *The American Journal of Human Genetics*, vol. 87, no. 3, pp. 325-340, 2010.

[21] X. Zhang, S. Huang, F. Zou, and W. Wang, Team: Efficient two-locus epistasis tests in human genome-wide association study, *Bioinformatics*, vol. 26, no. 12, pp. i217-i227, 2010.

[22] D. Brinza, M. Schultz, G. Tesler, and V. Bafna, Rapid detection of gene-gene interactions in genome-wide association studies, *Bioinformatics*, vol. 26, no. 22, pp. 2856-2862, 2010.

[23] J. Lehár, A. Krueger, G. Zimmermann, and A. Borisy, High-order combination effects and biological robustness, *Molecular Systems Biology*, vol. 4, no. 1, pp. 415-425, 2008.

[24] D. Anastassiou, Computational analysis of the synergy among multiple interacting genes, *Molecular Systems Biology*, vol. 3, no. 1, p. 83, 2007.

[25] J. Shang, J. Zhang, Y. Sun, D. Liu, D. Ye, and Y. Yin, Performance analysis of novel methods for detecting epistasis, *BMC Bioinformatics*, vol. 12, no. 1, p. 475, 2011.

[26] L. Ma, H. B. Runesha, D. Dvorkin, J. R. Garbe, and Y. Da, Parallel and serial computing tools for testing single-locus and epistatic snp effects of quantitative traits in genome-wide association studies, *BMC Bioinformatics*, vol. 9, no. 1, p. 315, 2008.

[27] R. A. Fisher, On the interpretation of _2 from contingency tables, and the calculation of p, *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87-94, 1922.

[28] L. S. Yung, C. Yang, X. Wan, and W. Yu, Gboost: A gpu-based tool for detecting gene-gene interactions in genome–wide case control studies, *Bioinformatics*, vol. 27, no. 9, pp. 1309-1310, 2011.

[29] Y. Liu, H. Xu, S. Chen, X. Chen, Z. Zhang, Z. Zhu, X. Qin, L. Hu, J. Zhu, G.-P. Zhao, et al., Genome-wide interaction-based association analysis identified multiple new susceptibility loci for common diseases, *PLoS Genetics*, vol. 7, no. 3, p. e1001338, 2011.

[30] X. Jiang, M. M. Barmada, and S. Visweswaran, Identifying genetic interactions in genomewide data using Bayesian networks, *Genetic Epidemiology*, vol. 34, no. 6, pp. 575-581, 2010.

[31] C. Herold, M. Steffens, F. F. Brockschmidt, M. P. Baur, and T. Becker, Intersnp: Genomewide interaction analysis guided by a priori information, *Bioinformatics*, vol. 25, no. 24, pp. 3275-3281, 2009.

[32] N. Chatterjee, Z. Kalaylioglu, R. Moslehi, U. Peters, and S. Wacholder, Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions, *The American Journal of Human Genetics*, vol. 79, no. 6, pp. 1002-1016, 2006.

[33] M. Y. Park and T. Hastie, Penalized logistic regression for detecting gene interactions, *Biostatistics*, vol. 9, no. 1, pp. 30-50, 2008.

[34] B. Goudey, D. Rawlinson, Q. Wang, F. Shi, H. Ferra, R. Campbell, L. Stern, M. Inouye, C. S. Ong, and A. Kowalczyk, Gwis-model-free, fast and exhaustive search for epistatic interactions in case-control gwas, *BMC Genomics*, vol. 14, no. Suppl 3, p. S10, 2013.

[35] J. Piriyapongsa, C. Ngamphiw, A. Intarapanich, S. Kulawonganunchai, A. Assawamakin, C. Bootchai, P. Shaw, and S. Tongsima, iloci: A snp interaction prioritization technique for detecting epistasis in genome-wide association studies, *BMC Genomics*, vol. 13, no. Suppl 7, p. S2, 2012.

[36] X. Zhang, F. Pan, Y. Xie, F. Zou, and W. Wang, Coe: A general approach for efficient genomewide two-locus epistasis test in disease association study, *Lecture Notes in Computer Science*, vol. 5541, pp. 253-269, 2009.

[37] X. Zhang, F. Zou, and W. Wang, Fastchi: An efficient algorithm for analyzing genegene interactions, in *Pac. Symp. Biocomput.*, 2009, pp. 528-539.

[38] X. Zhang, F. Zou, and W. Wang, Fastanova: An efficient algorithm for genome-wide association study, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '08*, New York, NY, USA, 2008, pp. 821-829.

[39] R. Culverhouse, T. Klein, and W. Shannon, Detecting epistatic interactions contributing to quantitative traits, *Genetic Epidemiology*, vol. 27, no. 2, pp. 141-152, 2004.

[40] H. Matsuda, Physical nature of higher-order mutual information: Intrinsic correlations and frustration, *Physical Review E*, vol. 62, no. 3, p. 3096, 2000.

[41] M. Steinbach, H. Yu, G. Fang, and V. Kumar, Using constraints to generate and explore higher order discriminative patterns, in *Advances in Knowledge Discovery and Data Mining*. Springer, 2011, pp. 338-350.

[42] D. Schwarz, I. Knig, and A. Ziegler, On safari to random jungle: A fast implementation of random forests for high-dimensional data, *Bioinformatics*, vol. 27, no. 3, pp. 439, 2010.

[43] R. Jiang, W. Tang, X. Wu, and W. Fu, A random forest approach to the detection of epistatic interactions in case-control studies, *BMC Bioinformatics*, vol. 10, no. Suppl 1, p. S65, 2009.

[44] X. Chen, C.-T. Liu, M. Zhang, and H. Zhang, A forest-based approach to identifying gene and gene-gene interactions, *Proceedings of the National Academy of Sciences*, vol. 104, no. 49, pp. 19199-19203, 2007.

[45] J. Millstein, D. V. Conti, F. D. Gilliland, and W. J. Gauderman, A testing framework for identifying susceptibility genes in the presence of epistasis, *The American Journal of Human Genetics*, vol. 78, no. 1, pp. 15-27, 2006.

[46] D. M. Evans, J. Marchini, A. P. Morris, and L. R. Cardon, Two-stage two-locus models in genomewide association, *PLoS Genet.*, vol. 2, no. 9, p. e157, 2006.

[47] X. Guo, Y. Meng, N. Yu, and Y. Pan, Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering, *BMC Bioinformatics*, vol. 15, no. 1, p. 102, 2014.

[48] M. Xie, J. Li, and T. Jiang, Detecting genome wide epistases based on the clustering of relatively frequent items, *Bioinformatics*, vol. 28, no. 1, pp. 5-12, 2012.

[49] Q. Long, Q. Zhang, and J. Ott, Detecting disease-associated genotype patterns, *BMC Bioinformatics*, vol. 10, no. Suppl 1, p. S75, 2009.

[50] T. Zheng, H. Wang, and S.-H. Lo, Backward genotype-trait association (bgta)-based dissection of complex traits in case-control designs, *Human Heredity*, vol. 62, no. 4, pp. 196-212, 2006.

[51] J. H. Moore, J. C. Gilbert, C.-T. Tsai, F.-T. Chiang, T. Holden, N. Barney, and B. C. White, A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility, *Journal of Theoretical Biology*, vol. 241, no. 2, pp. 252-261, 2006.

[52] L. Breiman, Random forests, *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[53] Y. Zhang and J. S. Liu, Bayesian inference of epistatic interactions in case-control studies, *Nat. Genet.*, vol. 39, no. 9, pp. 1167-1173, 2007.

[54] R. A. Fisher, On the interpretation of _2 from contingency tables, and the calculation of p, *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87-94, 1922.

[55] W. Tang, X. Wu, R. Jiang, and Y. Li, Epistatic module detection for case-control studies: A Bayesian model with a gibbs sampling strategy, *PLoS Genet.*, vol. 5, no. 5, p. e1000464, 2009.

[56] C. Yang, Z. He, X. Wan, Q. Yang, H. Xue, and W. Yu, Snpharvester: A filtering-based approach for detecting epistatic interactions in genomewide association studies, *Bioinformatics*, vol. 25, no. 4, pp. 504-511, 2009.

[57] H. Schwender and K. Ickstadt, Identification of snp interactions using logic regression, *Biostatistics*, vol. 9, no. 1, pp. 187-198, 2008.

[58] C. Kooperberg and I. Ruczinski, Identifying interacting snps using monte carlo logic regression, *Genetic Epidemiology*, vol. 28, no. 2, pp. 157-170, 2005.

[59] M. Y. Park and T. Hastie, Penalized logistic regression for detecting gene interactions, *Biostatistics*, vol. 9, no. 1, pp. 30-50, 2008.

[60] D. J. Miller, Y. Zhang, G. Yu, Y. Liu, L. Chen, C. D. Langefeld, D. Herrington, and Y. Wang, An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions, *Bioinformatics*, vol. 25, no. 19, pp. 2478-2485, 2009.

[61] A. Motsinger, S. Lee, G. Mellick, and M. Ritchie, Gpnn: Power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease, *BMC Bioinformatics*, vol. 7, no. 1, p. 39, 2006.

[62] N. R. Cook, R. Y. L. Zee, and P. M. Ridker, Tree and spline based association analysis of gene-gene interaction models for ischemic stroke, *Statistics in Medicine*, vol. 23, no. 9, pp. 1439-1453, 2004.

[63] J. Shang, J. Zhang, Y. Sun, and Y. Zhang, Epiminer: A three-stage co-information based method for detecting and visualizing epistatic interactions, *Digital Signal Processing*, vol. 24, pp. 1-13, 2014.

[64] S. Leem, H. H. Jeong, J. Lee, K. Wee, and K.-A. Sohn, Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure, *Computational Biology and Chemistry*, vol. 50, pp. 19-28, 2014.

[65] G. Fang, M. Haznadar, W. Wang, H. Yu, M. Steinbach, T. R. Church, W. S. Oetting, B. Van Ness, and V. Kumar, High-order snp combinations associated with complex diseases: Efficient discovery, statistical power and functional interactions, *PLoS ONE*, vol. 7, no. 4, p. e33531, 2012.

[66] Y. Wang, X. Liu, K. Robbins, and R. Rekaya, Antepiseeker: Detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm, *BMC Research Notes*, vol. 3, no. 1, p. 117, 2010.

[67] X. Wan, C. Yang, Q. Yang, H. Xue, N. L. Tang, and W. Yu, Predictive rule inference for epistatic interaction detection in genome-wide association studies, *Bioinformatics*, vol. 26, no. 1, pp. 30-37, 2010.

[68] J. R. Kilpatrick, Methods for detecting multilocus genotype-phenotype association, PhD dissertation, Rice University, 2009.

[69] X. Wan, C. Yang, Q. Yang, H. Xue, N. Tang, and W. Yu, Megasnphunter: A learning approach to detect disease predisposition snps and high level interactions in genome wide association study, *BMC Bioinformatics*, vol. 10, no. 1, p. 13, 2009.

[70] C. Aporntewan, D. Ballard, J. Lee, J. Lee, Z. Wu, and H. Zhao, Gene hunting of the genetic analysis workshop 16 rheumatoid arthritis data using rough set theory, *BMC Proceedings*, vol. 3, no. Suppl 7, p. S126, 2009.

[71] J. Hoh, A. Wille, and J. Ott, Trimming, weighting, and grouping snps in human casecontrol association studies, *Genome Research*, vol. 11, no. 12, pp. 2115-2119, 2001.

[72] E. T. Jaynes, *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*. Springer, 1989.

[73] S. Kullback, *Information Theory and Statistics*. Courier Dover Publications, 2012.

[74] J. A. Hanley and B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, vol. 143, no. 1, pp. 29-36, 1982.

[75] D. R. Velez, B. C. White, A. A. Motsinger, W. S. Bush, M. D. Ritchie, S. M. Williams, and J. H. Moore, A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction, *Genetic Epidemiology*, vol. 31, no. 4, pp. 306-315, 2007.

[76] R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, et al., Complement factor h polymorphism in agerelated macular degeneration, *Science*, vol. 308, no. 5720, pp. 385-389, 2005.

[77] R. H. Duerr, K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg, M. J. Daly, A. H. Steinhart, C. Abraham, M. Regueiro, A. Griffiths, et al., A genome-wide association study identifies il23r as an inflammatory bowel disease gene, *Science*, vol. 314, no. 5804, pp. 1461-1463, 2006.

[78] E. M. Reiman, J. A. Webster, A. J. Myers, J. Hardy, T. Dunckley, V. L. Zismann, K. D. Joshipura, J. V. Pearson, D. Hu-Lince, M. J. Huentelman, et al., GAB2 alleles modify alzheimer's risk in APOE "4 carriers, *Neuron*, vol. 54, no. 5, pp. 713-720, 2007.

[79] NARAC, Genetic analysis workshop 16, http://www.gaworkshop.org/index.html, 2010.

[80] The wellcome trust case control consortium, http://www.wtccc.org.uk/, 2007.

[81] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57. no. 1, pp. 289-300, 1995.

[82] Y. Wang, G. Liu, M. Feng, and L. Wong, An empirical comparison of several recent epistatic interaction detection methods, *Bioinformatics*, vol. 27, no. 21, pp. 2936-2943, 2011.

[83] J. R. Gibbs, M. P. van der Brug, D. G. Hernandez, B. J. Traynor, M. A. Nalls, S.-L. Lai, S. Arepalli, A. Dillman, I. P. Rafferty, J. Troncoso, et al., Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain, *PLoS Genetics*, vol. 6, no. 5, p. e1000952, 2010.

[84] J. T. Bell, A. A. Pai, J. K. Pickrell, D. J. Gaffney, R. Pique-Regi, J. F. Degner, Y. Gilad, J. K. Pritchard, et al., DNA methylation patterns associate with genetic and gene expression variation in hapmap cell lines, *Genome Biol.*, vol. 12, no. 1, p. R10, 2011.

[85] R. Shoemaker, J. Deng, W. Wang, and K. Zhang, Allele-specific methylation is prevalent and is contributed by cpg-snps in the human genome, *Genome Research*, vol. 20, no. 7, pp. 883-889, 2010.

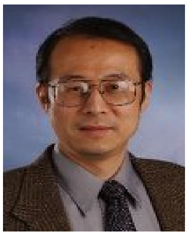
Xuan Guo received his BS degree and MS degree in computer science and technology from Southwest University and Wuhan University in 2009 and 2011, respectively. He is currently a PhD candidate in computer science at Georgia State University. His research interests include genome mapping, genome assembly, genome-wide association studies, and cloud computing.

Ning Yu currently is a PhD student in computer science in Georgia State University. He received MS degree in computer science from Southern Illinois University Carbondale in 2009. His BS degree in computer science and MEng degree in signal processing were received from Communication University of China in 2001 and 2004, respectively. His research area includes data mining, bioinformatics, and parallel and cloud computing.

Feng Gu received his BS degree in mechanical engineering from China University of Mining and Technology in 1998 and MS degree in information systems from Beijing Institute of Machinery in 2003. He received his MS and PhD degrees in computer science from Georgia State University in 2009 and 2011, respectively. He was an assistant professor of computer science at Voorhees College from 2010 to 2013. He is currently an assistant professor of computer science at College of Staten Island, The City University of New York, and the doctoral faculty member of Graduate Center of The City University of New York. He is the recipient of Natural Science Foundation Research Initiation Award. His research interests include modeling and simulation, complex systems, high performance computing, and bioinformatics.

Xiaojun Ding received his BS degree and MS degree in computer science and technology from Central South University, China. in 2001 and 2008. He is currently a PhD student in Central South University. He was a visiting PhD student in the Computer Science Department at Georgia State University for 2 years during his PhD study. He is interested in bioinformatics and machine learning.

Jianxin Wang received his BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the PhD degree in computer science from Central South University, China, in 2001. He is the vice dean and a professor in School of Information Science and Engineering, Central South University, China. His current research interests include algorithm analysis and optimization, parameterized algorithm, bioinformatics, and computer network. He has published more than 150 papers in various international journals and refereed conferences. He is a senior member of the IEEE.



Yi Pan is a distinguished university professor of the Department of Computer Science and an Interim Associate Dean at Georgia State University, USA. Dr. Pan received his BEng and MEng degrees in computer engineering from Tsinghua University, China, in 1982 and 1984, respectively, and his PhD degree in computer science from the University of Pittsburgh, USA, in 1991. His profile has been featured as a distinguished alumnus in both Tsinghua Alumni Newsletter and University of Pittsburgh CS Alumni Newsletter. Dr. Pan's research interests include parallel and cloud computing, wireless networks, and bioinformatics. Dr. Pan has published more than 150 journal papers with over 50 papers published in various IEEE journals. In addition, he has published over 150 papers in refereed conferences. He has also co-authored/co-edited 37 books. His work has been cited more than 4000 times. Dr. Pan has served as an editor-in-chief or editorial board member for 15 journals including 7 IEEE Transactions. He is the recipient of many awards including IEEE Transactions Best Paper Award, IBM Faculty Award, JSPS Senior Invitation Fellowship, IEEE BIBE Outstanding Achievement Award, NSF Research Opportunity Award, and AFOSR Summer Faculty Research Fellowship. He has organized many international conferences and delivered over 40 keynote speeches at various international conferences around the world.