8-11-2015

Searching Genome-wide Disease Association Through SNP Data

Xuan Guo

SEARCHING GENOME-WIDE DISEASE ASSOCIATION THROUGH SNP DATA

by

XUAN GUO

Under the Direction of Yi Pan, PhD

ABSTRACT

Taking the advantage of the high-throughput Single Nucleotide Polymorphism (SNP) genotyping technology, Genome-Wide Association Studies (GWASs) are regarded holding promise for unravelling complex relationships between genotype and phenotype. GWASs aim to identify genetic variants associated with disease by assaying and analyzing hundreds of thousands of SNPs. Traditional single-locus-based and two-locus-based methods have been standardized and led to many interesting findings. Recently, a substantial number of GWASs indicate that, for most disorders, joint genetic effects (epistatic interaction) across the whole genome are broadly existing in complex traits. At present, identifying high-order epistatic interactions from GWASs is computationally and methodologically challenging.

My dissertation research focuses on the problem of searching genome-wide association with considering three frequently encountered scenarios, i.e. one case one control, multi-cases multi-controls, and Linkage Disequilibrium (LD) block structure. For the first scenario, we present a simple and fast method, named DCHE, using dynamic clustering. Also, we design two methods, a Bayesian inference based method and a heuristic method, to detect genome-wide multi-locus epistatic interactions on multiple diseases. For the last scenario, we propose a block-based Bayesian approach to model the LD and conditional disease association simultaneously. Experimental results on both synthetic and real GWAS datasets show that the proposed methods improve the detection accuracy of disease-specific associations and lessen the computational cost compared with current popular methods.

INDEX WORDS: Algorithm, GWAS, SNP analysis, epistatic interactions, epistasis, clustering, Bayesian Theory, Markov Chain Monte Carlo

SEARCHING GENOME-WIDE DISEASE ASSOCIATION THROUGH SNP DATA

by

XUAN GUO

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2015

SEARCHING GENOME-WIDE DISEASE ASSOCIATION THROUGH SNP DATA

by

XUAN GUO

Committee Chair:	Yi Pan
Committee:	Alexander Zelikovsky Rajshekhar Sunderraman Jing Maria Zhang

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
August 2015

DEDICATION

To my parents Zhaogong Guo and Ximei Wang.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Yi Pan. Without his constant encouragement and guidance, I would most certainly not have completed much of anything. I could not imagine having selected a different advisor for my Ph.D. program. I am truly indebted to him. I am grateful to the Chair Prof. Rajshekhar Sunderraman for his constant encouragement, and great advises throughout my stay at Georgia State University. I would like to give special thanks to Dr. Alexander Zelikovsky for his invaluable guidance throughout my Ph.D. program. I thank Dr. Jing Zhang for sharing her valuable knowledge and reviewing my dissertation work.

Special thanks to Computer Science Department of Georgia State University faculty, staff and students.

Special thanks to Mrs. Tammie, Mrs. Adrienne, Mrs. Venette, and Mrs. Celena for helping me with all my paperwork.

Special thanks to the GSU Molecular Basis of Disease Initiative (MBD).

Special thanks go out to my dear friends Ning Yu, Xiaojun Ding, Yu Meng, Weinian Rao, and Bing Li.

Finally I am grateful to my parents Zhaogong Guo and Ximei Wang for their support and advice; Special thanks with all my heart to my girlfriend Liya, for her patience and understanding; Special thanks to my sister, Hui Guo, for being the best sibling someone could possibly have. I love you all dearly and would be most certainly not be who I am without all of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- GWAS - Genome-Wide Association Study
- LD - Linkage Disequilibrium
- LE - linkage Equilibrium

PART 1

INTRODUCTION

Most common diseases, such as hypertension, cancer, diabetes and heart disease, are complex traits resulting from the joint effects of various genetic variants, environmental factors or their interactions. It is of great interest to identify the genetic risk factors for complex diseases, to understand disease mechanisms, develop effective treatments and improve public health. The cost of genomic technologies is falling exponentially over time. For instance, the Human Genome Project took 13 years and cost $2.7 billion, whereas the current cost of sequencing a genome is approaching $1000 and takes less than a week. Within the next ten years, it is expected to drop to as low as $10 and take just a few minutes to genotype all 3 billion nucleotides in the human genome. With the availability of large-scale genotyping technologies together with their rapid improvement, the cost of genome-wide analyses has been widely decreased, and a great number of large-scale genetic association studies is initiated.

Complex diseases do not show the “pure” inheritance pattern observed in Mendelian diseases, where alterations in a single gene or a unique locus are causal for a phenotype. In a complex disease, multiple genes are involved, each with low-penetrance, where each gene modestly increases the probability of disease and does not ultimately determine disease status. These factors often render the traditional genetic dissection approaches, such as linkage analysis, ineffective tools to study complex diseases. My dissertation research focuses on three main problems in SNP data analysis, namely, detection of disease-related associations using one case, multiple cases, and modelling the block structure caused by linkage disequilibrium. Giving one case and one control, we are possible to find some associations connecting to a disease phenotype. With multiple disease cases, the same genetic factors with varied effects on different diseases could be detected. Inferring the block-wise structure in the human genome can further diminish the significant false associations due to the LD effects.

1.1 Single Nucleotide Polymorphism

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 [1]. 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 300bp 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 vast majority of them with minor impacts on biological systems, it is not necessary to study all the SNPs [2][3]. 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 [4]. 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.

1.2 Genome-wide Association Studies

Genome-wide Association Studies (GWAS) 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. In the last two decades, extensive computational efforts have been provided to study the functional and structural consequences of the SNPs [5]. 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 particular SNP allele by differential hybridization of the sample DNA. More details of these two competing technologies can be found in [6]. By far, over 600 GWA studies 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 first exciting finding of GWAS is on Age-related Macular Degeneration (AMD), which identified the *Complement Factor H* gene as a major risk factor [7]. The associated *Complement Factor H* gene demonstrates not only the DNA sequence variations but also the biological basis for the effect. Another successful application of GWAS is in the area of pharmacology, which heavily depends on the 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 [8]. Accordingly, GWAS is increasingly used to identify biological pathways and underlying networks of complex diseases [9]. 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.

1.3 Genome-wide Association Search

The concept of epistasis [10] was introduced around 100 years ago. It was referred as an extension of the notion of dominance for alleles within the same allelomorphic pair [10]. In recent literature, epistasis has been defined as the interaction among different genes (SNPs) [11]. 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 ($k \geq 2$) with phenotypes.

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 the binary variables. 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 the interested trait, another has a lower prevalence of such alleles [12]. The primary analysis paradigm for GWAS is dominated by the analysis of the susceptibility of individual SNPs, which can only explain a small part of causal genetic effects for complex diseases [13]. As a matter of fact, single locus-based approaches are insufficient to detect all interacting genes, in particular 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 [14]. Many studies have demonstrated that the epistasis is an important contributor to genetic variation in complex diseases, such as asthma, breast cancer [15], diabetes, coronary heart disease [16], and obesity [17].

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 [18][19][20]. 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 vast number of hypothesis tests are often conducted on limited sample sizes with a high degree of freedom, many false epistases are significantly associated with a disease trait by random chance [21][22].

1.4 Main Contributions

The main contributions are as follows:

- A simple, fast, and cloud-based method, named DCHE, for detection of genome-wide multi-locus epistatic interactions including:
  - develop of the novel method for grouping genotypes displaying similar genetic factors.
  - develop of the evaluation of interaction based on χ^2 statistic tests.
  - compare the proposed method to other popular approaches on simulated data with various main effects and join effects.
  - experimental validation of the findings on two real GWAS data.
- Two novel methods for search of genome-wide multi-locus epistatic interactions from multiple disease cases including:
  - develop a novel method, named DAM, based on Bayesian inference model for describing diverse join effects on distinct disease traits.

  - develop a novel heuristic approach, named SAM, for clustering SNP factors showing almost identical effects on multiple disease outcomes.
  - develop a novel evaluation of interaction accounting for redundant combinations of SNPs.
  - validation of the proposed methods on both simulated and real data.

- A novel Bayesian inference based method, named BAM, is designed to model the block structure in human genome and identify multi-locus epistatic interactions including:
  - develop an LD-block model to describe a block structure pattern.
  - incorporate the Bayesian variable partition model with LD-block model to capture the disease-related SNPs.
  - propose a two-level MCMC scheme to sample the posterior distribution.
  - validation of the block structure on simulated and real data.

1.5 Roadmap of the Rest of the Dissertation

The remainder of the dissertation is devoted to the detection of genome-wide high-order epistatic interaction problem. Chapter 2 introduces the problem and discusses state-of-the-art methods and models for the disease associations mapping. Chapter 3 presents a novel method, DCHE, for high-order interaction searching via dynamic clustering and stepwise evaluation of interaction. Chapter 4 describes a theoretical and algorithmic techniques using Bayesian partition model for identifying disease-specific associations and also presents a heuristic method based on Jensen-Shannon Divergence and a improved version of k-mean clustering. Chapter 5 gives an advanced Bayesian inference model capable of capturing block structure and disease associations at the same time. At last, chapter 6 summarizes our conclusions and future work.

1.6 Publications

Book Chapters

1. Y. Pan, and **X. Guo**, “Chapter: Cloud Computing for NGS Data Analysis,” *Computational Methods for Next Generation Sequencing Data Analysis*, 2016.

Refereed Journal Articles

1. **X. Guo**, J. Zhang, Z. Cai, D. Zhu, and Y. Pan, “A novel Bayesian method for detecting genome-wide associations on multiple diseases,” *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, Submitted.

2. **X. Guo**, N. Yu, X. Ding, J. Wang, and Y. Pan, “DIME: A novel framework for de novo metagenomic sequence assembly,” *Journal of Computational Biology*, vol. 22, no. 2, pp. 159–177, 2015.

3. X. Ding, J. Wang, A. Zelikovsky, **X. Guo**, M. Xie, and Y. Pan, “Searching high-order snp combinations for complex diseases based on energy distribution difference.” *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, no. 99, pp. 1, 2014.

4. **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.

Refereed Survey Article

1. **X. Guo**, N. Yu, F. Gu, X. Ding, J. Wang, and Y. Pan, “Genome-wide interaction-based association of human diseases-a survey,” *Tsinghua Science and Technology*, vol. 19, no. 6, pp. 596–616, 2014.

Refereed Conference Articles

1. **X. Guo**, J. Zhang, Z. Cai, D.-Z. Du, and Y. Pan, “DAM: A Bayesian method for detecting genome-wide associations on multiple diseases,” in *Bioinformatics Research and Applications.* Springer, 2015, pp. 96–107.

2. N. Yu, **X. Guo**, F. Gu, and Y. Pan, “Dna as x: An information-coding-based model to improve the sensitivity in comparative gene analysis,” in *Bioinformatics Research and Applications.* Springer, 2015, pp. 366–377.

3. **X. Guo**, X. Ding, Y. Meng, and Y. Pan, “Cloud computing for de novo metagenomic sequence assembly,” in *Bioinformatics Research and Applications*, ser. Lecture Notes in Computer Science, Z. Cai, O. Eulenstein, D. Janies, and D. Schwartz, Eds. New York: Springer Berlin Heidelberg, 2013, vol. 7875, pp. 185–198.

4. T. Zeng, **X. Guo**, and J. Liu, “Discovering negative correlated gene sets from integrative gene expression data for cancer prognosis,” in *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on.* IEEE, 2010, pp. 489–492.

PART 2

STATE-OF-THE-ART IN GENOME-WIDE ASSOCIATION DETECTION

The goal of detection of epistatic interactions is to identify d-SNP ($d \geq 2$) modules significantly associated with the phenotype. Furthermore, epistatic interactions can be classified into two types: epistasis displaying main effects (eME) and epistasis displaying no main effects (eNME) [23]. There are two challenges in searching epistasis: First, the total number of tests grows exponentially as k increases, leading to the inability of the exhaustive search to examine all the combinations. For example, using epiSNP [24] 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 vast number of hypotheses are tested using limited samples, a high proportion of significant associations are expected to be false positives. Therefore, retaining the statistical power while reducing false positive rate is another important issue.

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. Here, 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 often. 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.

2.1 Model-based VS. Model-free-based

There are two ways to categorize the methods for detection of 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 Figure 2.1, and eight of them with the names in bold are reviewed and discussed in Section 2.2. 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.

We describe two routines in this 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 the case and 0 for the control). For the analysis of two-locus epistatic interactions, we need to collect a contingency table (showed in Table 2.1), where n_{ijy} is the count of individuals with genotype $s_a = i$, $s_b = j$, and phenotype $Y = y$.

Table 2.1 The Genotype Counts in Cases and Controls.

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

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:

$$\log \frac{P\left(Y = 1 | s_a = i, s_b = j\right)}{P\left(Y = 0 | s_a = i, s_b = j\right)} = \beta_0 + \beta_i^{s_a} + \beta_j^{s_b} + \beta_{ij}^{s_a s_b} \tag{2.1}$$

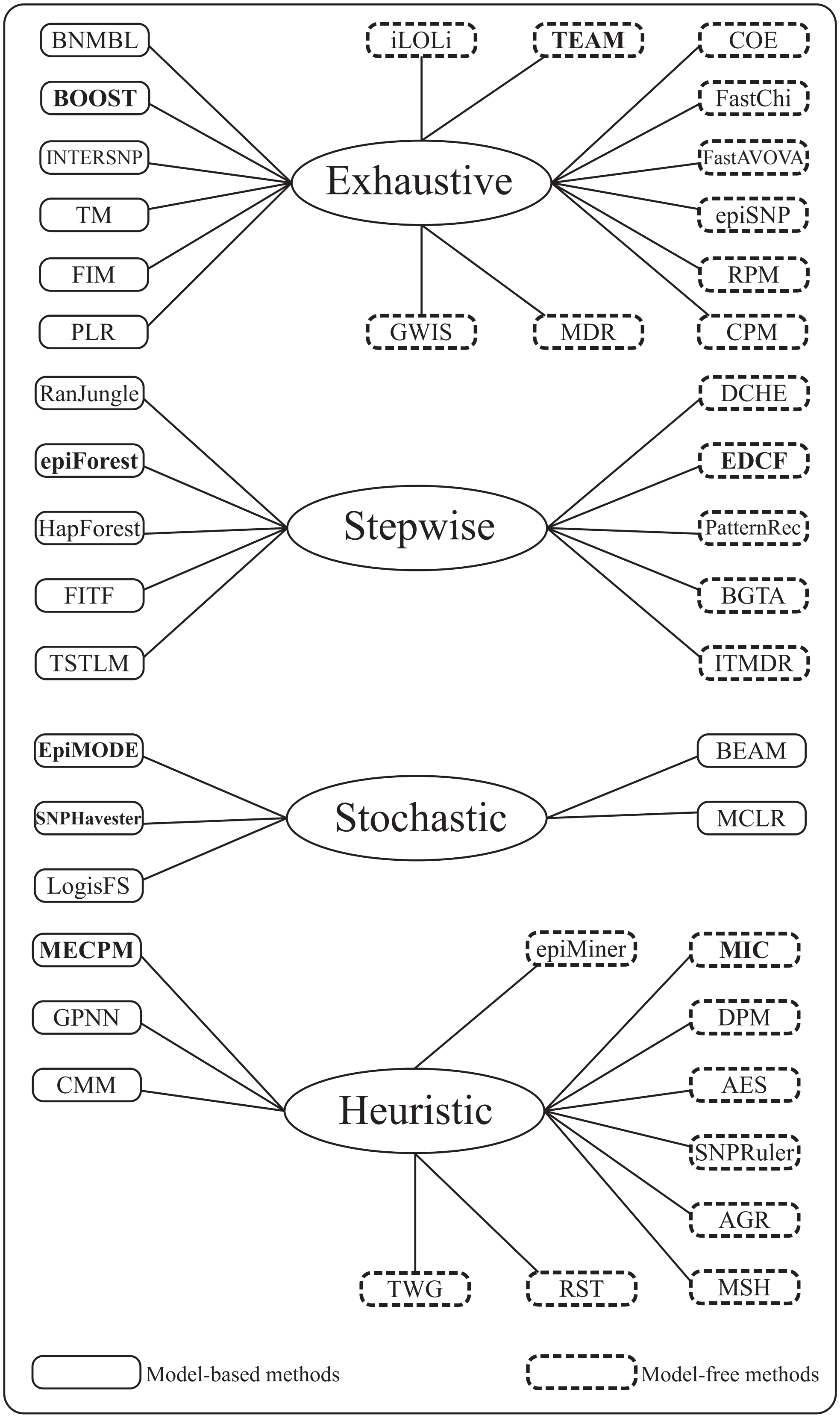

Figure 2.1 Classification of the methods that detect epistasis.

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

$$\log \frac{P\left(Y=1|s_a=i,s_b=j\right)}{P\left(Y=0|s_a=i,s_b=j\right)}=\beta_0 \tag{2.2}$$

There are nine coefficients in Equation 2.1 and one coefficient in Equation 2.2. 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 Equation 2.1 and 2.2. By evaluating the values at their Maximum Likelihood Estimations (MLEs), i.e., $\hat{L_F}-\hat{L_N}$, 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 showed in Table 2.1. (2) obtain the p-value using χ^2 statistic (Equation 3) with 8 degrees of freedom [25].

$$\chi^2=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i} \tag{2.3}$$

n is the count of genotype combinations by giving a set of SNPs. The observed frequency O is corresponding to the count of the individual with 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.

2.2 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 the review of recent studies, 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.

2.2.1 Exhaustive Searching Methods

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.* [14] 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 [15] 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 [26] is a GPU framework based implementation of BOOST, and PIAM [27] is developed by Liu *et al*, which uses 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 are too computationally expensive to be feasible, especially for large GWAS datasets with millions SNPs. In this section, we use BOOST and TEAM as examples of exhaustive searching methods. The overview and resource information of method falling into exhaustive search is showed in Table 2.2.

BOOST BOOST is a model-based, exhaustive search method, which is the abbreviation of “BOolean Operation-based Screening and Testing” [18]. Indicated by the name, there are two features of BOOST: first, a new boolean representation is used to accelerate

Table 2.2 Exhaustive search methods for detecting epistasis.

Model-based	
BNMBL	Bayesian Network Minimum Bit Length score (2010) [28]
BOOST	Boolean Operation-based Screening and Testing (2009) [18] http://bioinformatics.ust.hk/BOOST.html
INTERSNP	INTERSNP (2009) [29] http://intersnp.meb.uni-bonn.de
TM	Tukey's 1 d.f model for interaction (2006) [30]
FIM	Full Interaction Model (2005) [14] http://www.cbil.ece.vt.edu/ResearchOngoingSNP.htm
PLR	Restricted Partitioning Method (2004) [31]
Model-free	
GWIS	Genome Wide Interaction Search (2013) [32] http://bioinformatics.research.nicta.com.au/gwis
iLOLi	Interacting Loci (2012) [33] http://www4a.biotec.or.th/GI/tools/iloci
TEAM	Tree-based Epistasis Association Mapping (2010) [19] http://www.csbio.unc.edu/epistasis/download.php
COE	COE (2009) [34] http://www.csbio.unc.edu/epistasis/download.php
FastChi	FastChi (2009) [35] http://www.csbio.unc.edu/epistasis/download.php
FastAVOVA	FastANOVA (2008) [36] http://www.csbio.unc.edu/epistasis/
epiSNP	epiSNP (2008) [24] http://animalgene.umn.edu/episnp/index.html
RPM	Restricted Partitioning Method (2004) [37]
CPM	Combinatorial Partitioning Method (2001) [16]
MDR	Multifactor Dimensionality Reduction (2001) [15] http://sourceforge.net/projects/mdr/

the collecting of contingency table; second, an upper bound for the likelihood ratio test based on log-linear models and Kirkwood superposition approximation [38] 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 here. 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,

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

BOOST denotes the log likelihood of the full model under MLE as $\hat{L_F}$, the log likelihood of the main effect model under MLE as $\hat{L_M}$, the log likelihood of log-linear saturated model as $\hat{L_S}$, which was equivalent to the full logistic regression model, and the log likelihood of the homogeneous model as $\hat{L_H}$, which is equivalent to the main effect model. According to the likelihood ratio test, interaction effects are measured by the difference between two log likelihood of the main effect model and the full model evaluated at their MLEs, i.e. $(\hat{L_M} - \hat{L_F})$. Directly using $(\hat{L_S} - \hat{L_H})$ to test interactions in GWAS still has some difficulties, because iterative methods are needed in model fitting to compute $\hat{L_H}$, 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, $(\hat{L_{KSA}} \leq \hat{L_H})$ of $(\hat{L_S} - \hat{L_H})$. The reason for the replacement is that the calculation of $\hat{L_{KSA}}$ was straightforward and no iteration is involved. BOOST contains two stages: screening, evaluate all pairwise interactions by using the KSA; testing, for each pair with $2(\hat{L_S} - \hat{L_{KSA}}) > \tau$, test the interaction effect using the likelihood ratio statistic $2(\hat{L_S} - \hat{L_H})$.

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 only used to detect two-locus epistatic interaction, although it runs very fast (it only takes 170 seconds to analyze 10,000 SNPs with 5,000 samples on a 3.0 GHz CPU with 4G memory running the Windows XP Professional system.)

TEAM TEAM (Tree-based Epistasis Association Mapping) [19] 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 4.4. 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 $N = 1,000$, $L = 100,000$ and $H = 1,000$, 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.* [19] 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 2.1). The authors proved that given a 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 a MST, TEAM constructed an approximate MST instead.

The overall time complexity of TEAM is $O(NLH + NL^2 + W_T NK)$, where $O(NLH)$ is for generating all single-locus contingency tables, $O(NL^2)$ is for building the minimum spanning tree, and $O(W_T NK)$ is for updating the value of $n_{111}, n_{121}, n_{211}, n_{222}$ for H permutations. Comparing to the complexity of brute force approach $O(NL^2H)$, the performance of TEAM was faster than the latter by an order of magnitude. As TEAM did 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.

2.2.2 Stepwise Searching Methods

Although the exhaustive search is computationally possible to test all two-locus epistatic interactions for a moderate size of GWAS data, it requires enormous computation time and loses statistic power when searching higher-order interactions. 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 showed in a recent theoretical study [39], 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, the stepwise search may not be able to find interactions involving loci with small or no marginal effects.

epiForest Jiang *et al.* [41] 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

Table 2.3 Stepwise search methods for detecting epistasis.

Model-based	
RanJungle	Random Forests for high-dimensional data (2010) [40] `http://www.randomjungle.org/rjungle/rjunglenews`
epiForest	Random forest for the detection of epistatic interactions (2009) [41] `http://bioinfo.au.tsinghua.edu.cn/epiForest`
HapForest	forest-based approach to identifying gene-gene interactions (2007) [42] `http://c2s2.yale.edu/softwarepackages/HapForest/`
FITF	Focused Interaction Testing Framework (2006) [43] `http://hydra.usc.edu/fitf`
TSTLM	Two-Stage Two-Locus Models (2006) [44]
Model-free	
DCHE	Dynamic Clustering for High-order genome-wide Epistatic interactions detecting (2014) [45] `http://www.cs.gsu.edu/~xguo9/DCHE.html`
EDCF	Epistasis Detector based on the Clustering of relatively Frequent items (2012) [46] `http://www.cs.ucr.edu/~minzhux/EDCF.zip`
PatternRec	Genotype Pattern based on Difference Frequencies (2009) [47] `http://www.genemapping.cn`
BGTA	Backward Genotype-Trait Association (2006) [48] `http://statgene.stat.columbia.edu`
ITMDR	Information Theory and MDR (2006) [49] `http://www.cbil.ece.vt.edu/ResearchOngoingSNP.htm`

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 Leo Breiman [50]. 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. To measure the contribution of a 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 were very consistent [50], and the computation cost of the gini importance is much more economic [41].

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 [51] 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.

EDCF Xie *et al.* [46] proposed a stepwise search algorithm called EDCF (Epistasis Detector based on the Clustering of relatively Frequent items) 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 ($k \geq 3$), it is impractical to exhaustively search the whole space since there are $\begin{pmatrix} L \\ k \end{pmatrix}$ 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 $f_s \geq 1$ 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 with the same genotype followed a Binomial distribution. For the Binomial distribution, the parameter n equals to the total count of individual, 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 [25] 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-base exhaustive search approaches, like BOOST, extensive experiments on simulated data illustrates that EDCF tends to lose certain powers on detecting embedded disease models without main effects (eNME) [46].

2.2.3 Stochastic Searching Methods

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) [51] 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.* [52] 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 essential framework of stochastic search strategy can be generalized as follows:

Given a set of states (or configurations) $X = \{X_1, \ldots, X_M\}$ 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 reaching the maximum iteration count.

Particularly, the states are the assignments of SNPs (jointly associated with the 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. In the following, we use SNPHarvester and epiMODE to illustrate the basic idea in stochastic search for detecting epistasis.

Table 2.4 Stochastic search methods for detecting epistasis.

Model-based	
epiMODE	epistatic MOdule DEtection (2009) [52] `http://bioinfo.au.tsinghua.edu.cn/epiMODE/`
SNPHavester	filtering-based approach for detecting epistatic interactions (2009) [53] `http://bioinformatics.ust.hk/SNPHarvester.html`
LogicFS	logicFS (2008) [54] `http://bioconductor.org/packages/2.4/bioc/html/logicFS.html`
BEAM	Bayesian Epistasis Association Mapping (2007) [51] `http://www.fas.harvard.edu/~junliu/BEAM/`
MCLR	Monte Carlo Logic Regression (2005) [55] `http://cran.r-project.org/web/packages/LogicReg/index.html`

SNPHarvester Yang *et al.* proposed a method, SNPHarvester, using a path selection procedure to sample the searching space [53]. SNPHarvester first identifies disease-associated SNP groups from thousands of SNPs to reduce the number 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 [31] 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 make SNPHarvester possible to directly apply to large GWAS dataset for detecting high-order epistatic interaction.

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 [53]:

- 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. It uses the score function to measure the association between a k-SNP group and the phenotype, and records the SNP group whose score exceeds a fixed threshold. PathSeeker adopts the χ^2 value as the score function, and the threshold is determined by Bonferroni correction. It first removes significant single SNPs according to their χ^2 test values, since SNPHarvester is only interested in epistatic interactions that have weak main effects but significant joint effect. Then it randomly picks k-SNPs to form an active set $S = \{SNP_1, SNP_2, \dots, SNP_k\}$, and leaves the rest of the SNPs to form a candidate set Sc for the next random selection. A swapping operation is applied between S and Sc to switch two SNPs, $SNP_i \in S, SNP_j \in Sc$, if the new k-SNPs group achieves better χ^2 test value. For generating one k-SNPs 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 identify m k-SNP groups are employed for the model-fitting in step 2. 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 penalized logistic regression model.

epiMODE Tang *et al.* [52] developed an extension of BEAM using the Gibbs sampling strategy, named epiMODE (*epi*static *MO*dule *DE*tection), to facilitate the detection of epistatic modules. In epiMODE, the epistatic interaction module is considered as the fundamental 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 Linkage Disequilibrium (LD) between genetic variants. The genetic variants (SNPs) belonging to an epistasis module are simulated in a procedure, called reversible jump Markov chain Monte Carlo (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, S_2 of SNPs can be described as

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

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

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

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 specified modules. Particularly, the assignment for an SNP can be done by first calculating the probability of the observed data given a particular partition pattern using a Bayesian model. EpiMODE assumes that all loci are in linkage equilibrium (LD), also known as independent. It uses first-order Markov model to account for the situation in which a set of SNPs is at 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 [23]. Apparently, 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.

2.2.4 Heuristic Searching Methods

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 showed in Table 2.5. 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 that are treated as model constraints. MIC is a model-free method that defines significant tests based on mutual information, and it also uses k-means clustering to narrow down the candidate groups.

MECPM Miller *et al.* [56] proposed a method, named MECPM (maximum entropy conditional probability modeling), to identify markers/interactions by building the 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

Table 2.5 Heuristic Search methods for detecting epistasis.

Model-based	
MECPM	Maximum Entropy Conditional Probability Modelling (2009) [56] `http://www.cbil.ece.vt.edu/ResearchOngoingSNP.htm`
GPNN	Genetic Programming optimized Neural Network (2006) [57]
CMM	Tree and spline based association analysis (2004) [58] `http://lib.stat.cmu.edu/`
Model-free	
epiMiner	epistasis Miner (2014) [59] `https://sourceforge.net/projects/epiminer/files/`
MIC	Mutual Information (2014) [60]
DPM	Discriminative Pattern Mining (2012) [61]
AES	AntEpiSeeker (2010) [62] `http://nce.ads.uga.edu/~romdhane/AntEpiSeeker/index.html`
SNPRuler	Predictive rule inference for epistatic interaction detection (2010) [63] `http://bioinformatics.ust.hk/SNPRuler.zip`
AGR	Association Graph Reduction (2009) [64]
MSH	MegaSNPHunter (2009) [65]
RST	Rough Set Theory (2009) [66]
TWG	Trimming, Weighting and Grouping (2001) [67] `http://linkage.rockefeller.edu/ott/sumstat.html`

the principle of maximum entropy (ME), a probability model should agree with all known information and remain maximal uncertainty [68]. 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 amount of the ME distribution's entropy which is decreased by applying the constraint. The ME optimization problem is convex with linear constraints, and the standard solutions guarantees to the convergence.

A variant of greedy search based on BIC (Bayesian information criterion) is used for choosing the ME constraints (interactions up to five-way) when model grows, and determining the number of constrains 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 [69] 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 order, 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 1,000 SNPs and 2,000 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).

MIC Leem *et al.* [60] 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 $A = \{A_1, A_2, \ldots, A_n\}$ 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 $\left\{A^{(1)}, A^{(2)}, \ldots, A^{(k)}\right\}$ and a partition Y can be defined as follows:

$$\begin{aligned} I(A^{(1)}, A^{(2)}, \ldots, A^{(k)}; Y) = {} & H(A^{(1)}, A^{(2)}, \ldots, A^{(k)}) \\ & + H(Y) - H(A^{(1)}, A^{(2)}, \ldots, A^{(k)}; Y) \end{aligned}$$

where $H(A^{(1)}, A^{(2)}, \ldots, A^{(k)})$ 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 $I(A^{(1)}, A^{(2)}, \ldots, A^{(k)}; Y)$.

Since the size of current GWAS data can reach up to hundreds of thousands of SNPs, it takes too much time to calculate the $I(A^{(1)}, A^{(2)}, \ldots, A^{(k)}; Y)$ for every k-modules ($k \geq 3$). Therefore, MIC uses k-means clustering to reduce time complexity by placing strongly interacting SNPs into different clusters. Furthermore, 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 the mutual information between two SNPs.

- Step 2. Top d SNPs is 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 the 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 modules.

2.3 Summary

All eight methods reviewed in this chapter have demonstrated respective utilities based on the experimental results conducted in the recent studies [70][23][60]. We summarize their merits and weaknesses in Table 2.6. 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. 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 search only select a subset of SNPs for further tests. A common disadvantage of them is the power lost. As showed in a recent theoretical study [39], the possibility that a high-order (size-k) combination with high 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 EDCF [46]. Stochastic methods use random sampling procedures to search the space of interactions. The key factor influencing the performance of the stochastic method is the selection of sampling procedure. As showed in the experimental results from recent studies [70][23][60], 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

Table 2.6 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 $k \geq 3$; 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, and D. stands for disadvantage.

*Note: All stochastic search methods listed are belonging to model-free category.

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 hard to tell how good they are without control of type I error and to avoid false epistatic interactions.

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 close 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. We identify 43 methods for detecting disease-associated epistatic interactions. Eight of them are discussed in details in the chapter. Finally, we summarize the advantages and disadvantages of popular GWAS tools.

PART 3

SEARCHING HIGH-ORDER GENOME-WIDE EPISTASIS ON ONE DISEASE

3.1 Introduction and Contributions

At present, traditional single-locus-based methods are insufficient to detect interactions consisting of multiple-locus, which are broadly existing in complex traits. In addition, statistic tests for high-order epistatic interactions with more than 2 SNPs propose computational and analytical challenges because the computation increases exponentially as the cardinality of SNPs combinations gets larger. In this chapter, we provide a novel method, named "Dynamic Clustering for High-order genome-wide Epistatic interactions detecting" (DCHE), to address challenges. DCHE adopts an elaborate dynamic clustering procedure to maximize statistic significance for SNP combinations and ranks top ones as results. DCHE conducts statistic tests on merged groups of genotype categories determined by the dynamic clustering. Each grouped genotype category tends to share a similar effect associating with corresponding phenotypes. Truly disease-related joint genetic effects will gain higher ranking values if genotype combinations can be correctly clustered together. Systematic experiments on simulated two- and three-locus disease models datasets show that DCHE is more powerful in finding epistatic interactions than some recently proposed methods including TEAM [19], SNPRuler, BOOST and EDCF [46]. Our experiments on two real genome-wide case/control datasets, Age-related macular degeneration (AMD) and Rheumatoid arthritis (RA) demonstrate that DCHE is feasible for the full-scale analyses of multi-locus associations on large GWAS datasets and it enriches a great deal of novel and significant high-order epistatic interactions that have not been reported in the literature.

3.2 DCHE: *D*ynamic *C*lustering for *H*igh-order Genome-wide *E*pistatic Interactions Detecting

3.2.1 Notation

Suppose a GWAS dataset has M diallelic SNPs and N samples. In general, bi-allelic genetic markers use uppercase letters (e.g. A, B,...) to denote major alleles and lowercase letters (e.g. a, b) to denote minor alleles. For encoding three genotypes, one popular way is to use $\{0, 1, 2\}$ to represent $\{AA, Aa, aa\}$, respectively. In a GWAS case-control dataset, N^U denotes the number of cases (i.e. disease individuals) and N^D denotes the number of controls (i.e. normal individuals). X is utilized to indicate the ordered set of the M SNPs, and X_i represents the i-th SNP in X. $MAF(X_i)$ denotes the minor allele frequency of X_i and g_i^j denotes the genotype of j-th individual at X_i. For d-locus interaction, $(X_{i_1}, \ldots, X_{i_d})$, one genotype combination denotes as $(g_{i_1}, \ldots, g_{i_d})$.

3.2.2 Dynamic Clustering

An intuitive strategy to detect genome-wide epistatic interactions is to test differences of genotypes' frequencies for single SNP or SNPs' combinations in cases and controls. The contingency table in Table 3.1 gives an example for two-locus disease model, where there are 9 genotype combinations and $N^U = N^D = 800$. Numbers within the parentheses are counted from controls. Cells with higher frequencies in cases are coloured by grey background. Some methods, like Multifactor dimensionality reduction (introEpsBreast) [15] and its extensions [71], take the case/control ratio of each genotype combination to test associations between SNP combinations and disease status. However, the frequency cannot be a fair indicator to uncover disease-related associations, because it can be biased by many factors, including effect size, allele frequency, linkage disequilibrium between markers and disease loci as well as sampling errors. Other recent developed strategies use Pearson's χ^2 test, exact likelihood ratio test and entropy-based test to examine the independence of observations. For the example in Table 3.1a, the unadjusted p-value from Pearson's χ^2 test with 8 degrees of freedom is 1.724×10^{-18}. If we use Bonferroni correction to adjust p-value, this pair can still be significant with threshold 0.05 for a large GWAS dataset. However, it will not always be the case, and some limitations, including uneven or insufficient samples, tiny penetrance on single genotype, would dramatically affect the adjusted p-value. Another toy example sampled from a two-locus multiplicative effects model (see Table 3.2) is showed in Table 3.1b, where $N^U = N^D = 400$. Normally, the approximation to the χ^2 distribution breaks down if more than 20% expected frequencies below 5. The unadjusted p-value of Table 3.1b is 1.09×10^{-6} and the adjusted p-value is 0.547 if $M = 1,000$, which is obviously larger than 0.05. Another popular method, BOOST [18] which utilizes the likelihood ratio to test statistic, cannot get a significant result for Table 3.1b by setting the significant level to 0.1. We can observe the same result by applying EDCF, which utilizes the concept of the frequent item to group genotype combinations and adopts the χ^2 test with $df = 2$.

Table 3.1 Examples for contingency tables.

(a)

	BB	Bb	bb
AA	71(108)	97(151)	44(47)
Aa	89(138)	184(184)	93(55)
aa	29(43)	113(57)	80(17)

(b)

	BB	Bb	bb
AA	40(55)	49(76)	13(21)
Aa	43(62)	110(103)	49(22)
aa	16(24)	50(27)	30(10)

To address the preceding issues, we propose a dynamic clustering procedure. Basically, we first merge all 3^{n^d} genotype cells to n^d groups based on certain combination criteria, where n^d ranges from 3 to 6. The criterion to combine two genotype categories is rooted from their similar effects that associate with phenotypes. Secondly, we collect statistic test values on merged groups. For better illustration, we take Table 3.2 for example. Although there are 9 genotype categories, some have same penetrances, so we can partition them into 4 groups, where penetrances are $\alpha, \alpha\left(1+\theta\right), \alpha\left(1+\theta\right)^2$ and $\alpha\left(1+\theta\right)^4$. In reality, it is different to predict the order of complex disease model and its penetrance table. Therefore, we try to find a statistically significant evaluation of interactions in a stepwise manner by merging genotype categories into a range of the number of groups and test levels of significance. We select the most significant one as the evaluation for d SNPs interaction. The full algorithm of dynamic clustering is as follows.

- Step 1. For a set of SNPs, cross-tabulate genotype combinations of SNPs with the categories of the dependent variable (phenotype).
- Step 2. Find a pair of genotype combinations whose 2×2 sub-table is least significantly different. If this significance does not reach a critical value, merge the two combinations and consider this merger as a single compound combination, and repeat this step.
- Step 3. Calculate the significant evaluation for each merged group pattern when categories' number is larger than three and less than six. Select the most significant one as the unadjusted p-value as the evaluation for the current interaction.

Table 3.2 Two-locus interaction multiplicative effects.

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

In Step 2, there are several ways to measure the difference of a 2-by-2 contingency table, like Pearson's χ^2 test with $df = 1$ and phi coefficient. In our algorithm, we adopt Pearson's χ^2 test with $df = 1$ to measure the difference. Following the dynamic clustering procedure, we can get the most significant group pattern as 161, 129, 110 in cases and 238, 59, 103 in controls for Table 3.1. Note that the df varies when the number of clusters changes. According to our setting that n^d ranges from 3 to 6, the df changes from 2 to 5, respectively. Along the clustering, we calculate the $p-value_{unadjusted}$ for each clustering with corresponding df. The trace of merging for Table 3.1b is (0, 1, 2, 3, 6), (5, 7, 8) and (4), if cells are labelled from left to right, from top to bottom and start from 0. It is easy to output this ground-truth interaction by the combination of Bonferroni correction and permutation tests for controlling type-I error. ($p - value_{unadjusted} = 1.15 \times 10^{-9}$ and the significant level is $\alpha = 3.0 \times 10^{-9}$ with false positive rate nearly equals to 0.1).

3.2.3 Evaluation of Interactions

The goal of DCHE is to identify d-SNP ($d \geq 2$) epistatic interactions significantly associated with the phenotype. As stated in [11, 18], epistasis can be interpreted as the statistical interaction or the full association. The evaluation of interactions used by DCHE is similar to detect full associations in model-based methods. In terms of logistic regression, the epistatic interaction we are looking for may contain main effects or interaction effects or both. In order to detect the significant association between genotype and phenotype, we use a model-free method and p-value from Pearson's χ^2 test to indicate the significance. Since DCHE aims to find high-order genome-wide epistatic interaction, the high-order interaction module may consist of one or some redundant SNPs, which do not contribute to increasing the significance. To avoid such cases, we give a definition of the minimal significant epistatic interaction.

Definition 1. *A SNPs module $(X_{i_1}, X_{i_2}, \ldots, X_{i_d})$ is considered as the minimal significant epistatic interaction by giving the significant level α, if it meets the following two conditions:*
(1) the $p - value$ of clusters of $(X_{i_1}, X_{i_2}, \ldots, X_{i_d}) \leq \alpha$;
(2) the $p - value$ of clusters of any subset of $(X_{i_1}, X_{i_2}, \ldots, X_{i_d}) <$ the $p - value$ of clusters of $(X_{i_1}, X_{i_2}, \ldots, X_{i_d})$.

3.2.4 Algorithm

Since testing all high-order SNP combinations is impossible for large GWAS datasets of millions of SNPs, we utilize a stepwise strategy to emulate and run dynamic clustering on all two-locus SNPs combinations. As showed in a recent theoretical study [39], the possibility that a high-order (size-d) combination with strong differentiation shows zero differentiation in all of its subsets decreases dramatically when d increases (generally it becomes impossible for $d \leq 5$). Therefore, we use top-l_d low-order SNPs combinations which demonstrate some significance as candidates. For higher order, we add one SNP X each time to interactions and re-invoke the dynamic clustering procedure until d reaches the defined value. We adopt same bitwise operations and Boolean Representation as introduced in BOOST [18] to collect and compress contingency tables. The details of the sequential algorithm are showed in Algorithm 1.

Algorithm 1: The DCHE Algorithm

```
Input: An $N \times (M+1)$ matrix, where $N = N^D + N^U$ and the first column denotes
        disease statuses; $d$, critical level $\alpha$ and $l = \{l_2, ..., l_d\}$
Output: The top-$l_d$ significant $d$-locus interaction with $p - value_{adjusted} > \alpha$
Read $N \times (M+1)$ matrix file;
Boolean represent $N \times (M+1)$ matrix as $(3 \times M) \times N$ matrix $W$;
Initialize an ascending list $L$ with length as $\max(l)$;
Set $d' = 2$;
for each pair of SNPs, $(X_i, X_j)$, $(1 \leq i < j \leq M)$ do
    Collect the contingency table $C_{i,j}$;
    Set $p - value_{i,j} =$ DynamicClustering$(C_{i,j})$;
    Insert $p - value_{i,j}$ into $L$;
end
$d' = d' + 1$;
Initialize another ascending list $L'$ with length $\max(l)$;
while $d' <= d$ do
    for each interaction $\left(X_{i_1}, \ldots, X_{i_{d'}}\right)$ in top-$l_{d'}$ positions of $L$ do
        for each SNP $X_j$, $(1 \leq j \leq M)$ do
            if $j \notin \{i_1, \ldots, i_{d'}\}$ then
                Collect the contingency table $C_{i_1,\ldots,i_{d'},j}$;
                Set $p - value_{i_1,\ldots,i_{d'},j} =$ DynamicClustering$\left(C_{i_1,\ldots,i_{d'},j}\right)$;
                Insert $p - value_{i_1,\ldots,i_{d'},j}$ into $L'$;
            end
        end
    end
    Clean list $L$, Initialise $L_{d'}$, Copy top-$l_{d'}$ elements in $L'$ to $L, L_{d'}$, Clean list $L'$;
end
for each interaction $(X_{i_1}, \ldots, X_{i_d})$ in the top-$l_d$ position of $\{L_2, \ldots, L_d\}$ do
    if $p - value_{i_1,\ldots,i_d} \leq \alpha$ then
        for subset $Q$ of $(X_{i_1}, \ldots, X_{i_d})$ do
            if $Q \in \{L_2, \ldots, L_{d-1}\}$ and $p - value$ of $Q \geq$ the $p - value$ of $(X_{i_1}, \ldots, X_{i_d})$
            then
                break;
            end
        end
        Write $\langle (X_{i_1}, \ldots, X_{i_d}), p - value_{i_1,\ldots,i_d} \rangle$ into result file;
    end
end
```

Each column in matrix M is converted to 3 rows in matrix W based on Boolean Representation (line 1-2). An ascending list where redundancy is not allowable is initialized with size $\max(l)$. The structure of an element in L consists a pair of key and value that the key is SNPs combinations and value is the unadjusted p-value (line 3). DCHE uses bitwise operations to collect contingency tables for all two-locus interaction and calculates evaluations of significance via DynamicClustering procedure. The p-value will be inserted into L (line 4-9). DCHE only selects top-$l_{d'}$ interactions to extend in DynamicClustering procedure and inserts estimated significance into another candidate list L'. At the end of each iteration for specific d, list L gets cleaned and DCHE transfers top-$l_{d'}$ elements from L' to L and new list $L_{d'}$ (line 10-23). When d reaches the defined value, top-l_d interaction modules with $p - value > \alpha$ will be written into the result file (line 24-33).

The time complexity of dynamic clustering is $O\left(3^d\right)$. According to the theory in [39], we only need to apply dynamic clustering procedure for up to 4 order of SNPs combinations. So when $d = 2$, the time complexity to test all 2-locus interactions is $O\left(NM^2\right)$. Inserting an element into ascending list takes time $O\left(\log\left(\max\left(l\right)\right)\right)$. The total time complexity for 2-locus interaction detection is $O\left(NM^2\right)+O\left(\log\left(\max\left(l\right)\right)\right)$. When $d = 3$, the time complexity to extend all candidate 2-locus interactions to 3-locus modules is $O\left(l_3M\right)$. Hence, the entire time complexity reaches to $O\left(l_3M\right) + O\left(NM^2\right) + O\left(\log\left(\max\left(l\right)\right)\right)$, if the user plans to search 3-locus interaction. Similar time complexity analysis can be applied to higher order interaction detection by using our DCHE.

3.3 Experiment Results

We first give definitions of 6 simulated disease models and the power metric used to evaluate the effectiveness of DCHE in comparison with other 4 popular epistatic interactions detecting methods, i.e. TEAM [19], SNPRuler [63], EDCF [46], BOOST [18]. Three reasons for choosing above 4 approaches are as follows: (1) TEAM, EDCF and BOOST all use the exhaustive search strategy for detecting two-locus interactions, so the comparison of their performance is fair; (2) a recent review tested five available methods and recommended

BOOST and TEAM as a powerful tool for searching epistatic interactions on a genome-wide scale [70]; (3) our goal is to discover high-order epistatic interactions from GWAS data, and among 4 detectors excluding DCHE, only SNPRuler and EDCF are equipped the ability to search interactions with more than 2 SNPs. Before experiments on simulated datasets, a discussion on how to control the false positive rate is illustrated because the Bonferroni correction, the most common method for controlling error rate, can be too conservative to filter significant interactions. We also present results of DCHE on two real GWAS dataset, Age-related macular degeneration (AMD) and Rheumatoid Arthritis (RA). Interactions detected by DCHE from different orders demonstrate a considerable number of novel, potentially disease-related genetic factors.

3.3.1 Experimental Design

Data Simulation To evaluate the effectiveness of DCHE, we perform extensive simulation experiments using six disease models with two- and three-locus associations. The unassociated SNP genotypes are generated by the same procedure used in previous studies [18]. Minor allele frequencies (MAFs) are uniformly sampled from the set $[0.05, 0.5]$. By assuming HardyWeinberg equilibrium, we can sample the genotype g_i^j for individual j. For embedded disease models, 4 two-locus epistasis models and 2 three-locus epistasis models are chosen by given odds tables or penetrance table that can be found in [45] and named these six models from model 1 to 6. In addition, we conduct tests on 50 two-locus epistasis models without marginal effects as BOOST and EDCF did in [18][46]. For models 1 to 4 and 50 models without marginal effect, each simulated dataset contained $M = 1000$ SNPs and $N = 800$ or 1600 with balanced samples in case and control under each parameter setting. For model 5, one dataset has 1000 SNPs and 2000 or 4000 samples with $N^U = N^D$. For model 6, $M = 2000$ and N is reduced to 400 and 800 with balanced cases and controls.

A disease model can be defined either by specifying the penetrance table or the odds table. Relations among penetrance $p(D)$, odds ODD_{g_i} and the probability $p(D|g_i)$ that an individual will be affected with a given genotype combination g_i can be calculated as

Equation 3.1,3.2.

$$ODD_{g_i} = \frac{p\left(D|g_i\right)}{p\left(\overline{D}|g_i\right)} = \frac{p\left(D|g_i\right)}{1 - p\left(D|g_i\right)} \tag{3.1}$$

$$p\left(D|g_i\right) = \frac{ODD_{g_i}}{1 + ODD_{g_i}} \tag{3.2}$$

Following [18], the disease prevalence $p(D)$ and genetic heritability h^2 are given by Equation 3.3,3.4.

$$p\left(D\right) = \sum_i p\left(D|g_i\right) p\left(g_i\right) \tag{3.3}$$

$$h^2 = \frac{\sum_i \left(p\left(D|g_i\right) - p\left(D\right)\right)^2 p\left(g_i\right)}{p\left(D\right)\left(1 - p\left(D\right)\right)} \tag{3.4}$$

For simplicity, we adopt same parameters as used in [18] for model 1 to 4, i.e. $p\left(D\right) = 0.1$, $h^2 = 0.03$ for model 1 and $h^2 = 0.02$ for Models 2, 3 and 4, MAF $\in \{0.1, 0.2, 0.4\}$. For model 5, we adopt similar setting in [46], i.e. $p\left(D\right) = 0.1$, effect size $\lambda = 0.2$, $\beta \in \{4, 1.5, 1, 0.7, 0.5\}$ and MAF $\in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. For model 6, MAFs of disease associated loci are fixed to 0.5. Effect parameters α and θ in odds tables for all six models are determined numerically using same procedures in [51]. Settings for 50 models without marginal effect are similar to [72], i.e. h^2 ranges from 0.05 to 0.4 with five intervals and MAF equals to 0.2 or 0.4.

Statistical power In the comparison of performances on simulated data, 100 datasets are generated for each setting. In one dataset, we embed one ground-truth epistatic interaction. The measure of discrimination power used in [18] is adopted, which is defined as the fraction of 100 datasets on which only top interaction given by the method matches the ground-truth. For all programs, the ground-truth interaction are detected if it is set to the most significant one and its adjusted p-value is larger than the critical value that is setted to 0.1 in following experiments.

3.3.2 False Positive Rate

Since DCHE uses stepwise strategy similar to EDCF, we also adopt two levels of multiple comparisons: (1) test $\binom{M}{d}$ combinations for d loci for a dataset with M SNPs; (2) test dynamic clustering results, which could end with up to 3^d possible genotype combinations with k groups, $d \in \{3, 4, 5, 6\}$. If we use the Bonferroni correction for above two-level multiple tests, the upper bound of possible ways to do combination is $\binom{M}{d} 6^{3^d}$. Hence, it is too conservative to obtain significant interaction modules. In order to reasonably loose the strictness, inspired by EDCF we combine the Bonferroni correction and permutation tests for these two levels, that is Bonferroni corrections for d loci combinations and permutation tests for the dynamic clustering procedure. More specifically, the significant level of an epistatic interaction is calculated from Equation 3.5.

$$\alpha = \alpha_0 / \binom{M}{d} \tag{3.5}$$

In Equation 3.5, α_0 is estimated from permutation tests for different ds on null simulations and $\binom{M}{d}$ represents the Bonferroni correction. To accurately control the false positive rate, we simulated datasets with five different settings for each d, i.e. we either fix $M = 1000$ and set N to 400, 800 and 1600 or fix $N = 800$ and set M to 1000, 2000 and 4000. Note that one thousand datasets are generated under one setting. The false positive rate is defined as $n_{false}/1000$, where n_{false} is the number of datasets where DCHE has found one or more interaction modules. Test results showed in Figure 3.1 illustrate: for a general setting of critical level 0.1, a recommended α_0 is 1.5×10^{-3} for two-locus disease model detection, 1.2×10^{-8} for three-locus disease model detection and 1.0×10^{-21} for four-locus disease model detection. In addition, the false positive rates tend to decrease or remain nearly unchanged as the number of samples and SNPs go up (Figure 3.1B and 3.1C). Therefore, in tests of simulated datasets and two real GWAS datasets, we set $\alpha_0 = 1.5 \times 10^{-3}, 1.2 \times 10^{-8}, 1.0 \times 10^{-21}$ for $d = 2, 3, 4$, respectively, to control the overall false positive rate for DCHE, unless otherwise stated.

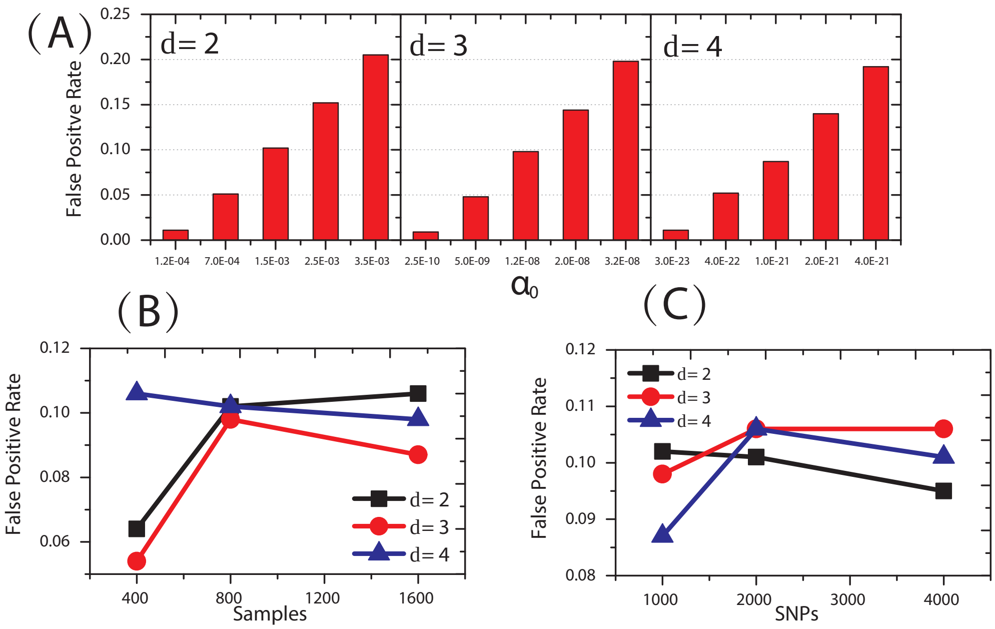

Figure 3.1 **False positive rates under null models.** The plot in **A** shows the false positive rates of DCHE using different α_0s for different ds, and the plots in **B** and **C** show the false positive rates of DCHE for different ds when sample size and the number of SNPs vary.

3.3.3 Two-locus Disease Models

For a fair comparison, interactions reported by all programs are filtered using the critical value 0.1 as the significant threshold. Test results are illustrated in Figure 3.2 for model 1 to 4. A common trend for all programs is that power is increasing as sample size increases from 800 to 1600. Most methods show more power when ground-truth model interactions' MAFs are larger, except that BOOST shows less power on model 1 and 2 when MAFs goes up. We can see that DCHE achieves highest or comparable powers on almost all datasets. More specifically, with 24 parameter settings for four disease models, DCHE outperforms other four methods at 9 settings and obtains full powers at 10 settings and gains comparable results at 5 settings. Taking results from datasets with $N = 1600$ for example, it is obvious that DCHE defeats other approaches with nearly 100% powers. For a more straight comparison, we introduce a new concept, the overall quality $q = n_{correct}/n_{total}$, where $n_{correct}$ is the number of datasets where programs successfully detect the ground-truth interactions and n_{total} is the total number of datasets. When $N = 800, M = 1000$, the overall quality for DCHE, TEAM, SNPRuler, EDCF and BOOST are 0.541, 0.455, 0.087, 0.508 and 0.31, respectively. When $N = 1600, M = 1000$, all five programs achieve higher accuracies than former settings and q are 0.981, 0.912, 0.162, 0.944 and 0.681, respectively. Note that DCHE, TEAM, and EDCF have abilities to achieve more than 90% powers, and powers for DCHE reach to at least 98% on datasets with 1600 samples. Note that BOOST is designed to identify significant statistical interaction without considering the main effects, so it is reasonable that our method DCHE and other two methods, i.e. EDCF and TEAM, outperform BOOST for detecting the Model 1 through 4. The reason we still put the BOOST into the experiments is that the biologists might be more interested in epistatic interaction as long as it shows significant association genotypes with phenotypes. In addition, similar designs of experiments can be found in other studies [46, 70, 18].

Moreover, we conduct tests on 50 disease models with little marginal effects. For convenience, penetrance tables for 50 models are not listed, and they are available in literature [72]. Since most methods gain near full powers, we use box plots to demonstrate overall performances in Figure 3.3. We can see that DCHE, EDCF and BOOST achieve comparable results in two subfigures. Specifically, they can accurately detect embedded associated SNPs interactions under most settings. On the contrary, TEAM and SNPRuler lose significant powers on both datasets with $MAFs = 0.2$ or 0.4. A common trend to previous experimental results is that five methods tend to possess more powers as $MAFs$ increase. After carefully examining results from five techniques, we can find that DCHE apparently outperforms other three methods except BOOST, although the difference is not too much. A possible explanation is that these embedded models with little main effects are more suit-

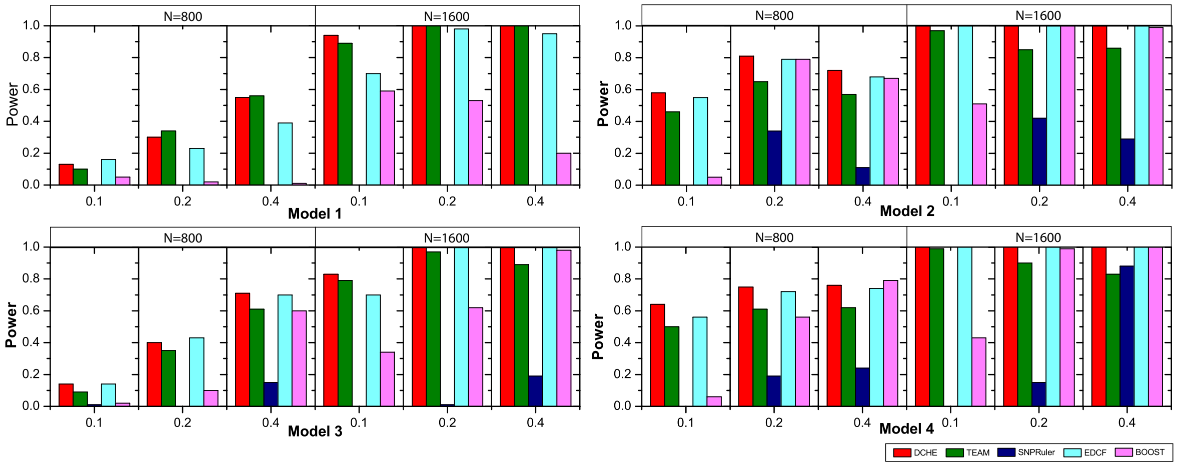

Figure 3.2 **Performance comparison of DCHE, TEAM, SNPRuler, EDCF and BOOST on disease models 1-4.** Performance comparison of DCHE, TEAM, SNPRuler, EDCF and BOOST on four disease models for different allele frequencies and sample sizes. The red, green, blue, cyan and magenta bars show powers of DCHE, TEAM, SNPRuler, EDCF and BOOST, respectively. Models are ordered from top to bottom and from left to right and they are model 1, model 2, model 3 and model 4.

able for model-based detection strategy, and DCHE is a model-free based method. If we adopt the same overall quality defined in previous paragraph to evaluate performances, q are 0.972, 0.656, 0.891, 0.951 and 0.984 for DCHE, TEAM, SNPRuler, EDCF and BOOST, respectively.

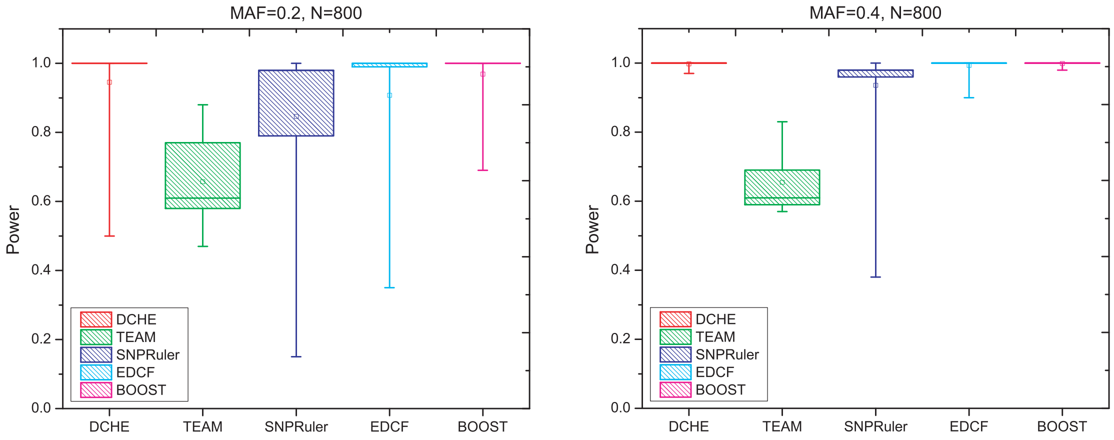

Figure 3.3 **Performance comparison on 50 models without main effects.** For each model, we simulate data using sample size 800 and $MAF \in \{0.2, 0.4\}$. The red, green, blue, cyan and magenta boxes show powers of DCHE, TEAM, SNPRuler, EDCF and BOOST, respectively.

3.3.4 Three-locus Disease Models

For comparisons of three-locus disease models, two methods are dropped, i.e. TEAM and BOOST, because both of them are designed only for detecting two-locus interactions. Based on settings of Model 5 given in previous sections, we get 10 groups of datasets with 100 replicates, which can be further categorized into two families with $N = 2000$ or 4000. Note that when $MAF = 0.5$, there is no marginal effect, otherwise disease models have $\lambda = 0.2$. Experimental results are illustrated in Figure 3.4. Similar to Models 1 to 4, three programs tend to get more powers as MAF or sample size increases. Considering parameter β in (β, MAF), we can see that powers go up when β goes down for all methods. A significant difference can be found is that SNPRuler can only obtain acceptable results when $(\beta = 0.5, MAF = 0.5)$; otherwise SNPRuler hardly gets powers. Although the distinction between DCHE and EDCF is not too much, we can still observe that DCHE hits more ground-truth SNPs interactions than EDCF does at datasets with $N = 4000$. Additionally, overall qualities for DCHE and EDCF are 0.514 and 0.52 with $N = 2000$, and the overall quality for DCHE rises to 0.914 comparing with 0.866 for EDCF.

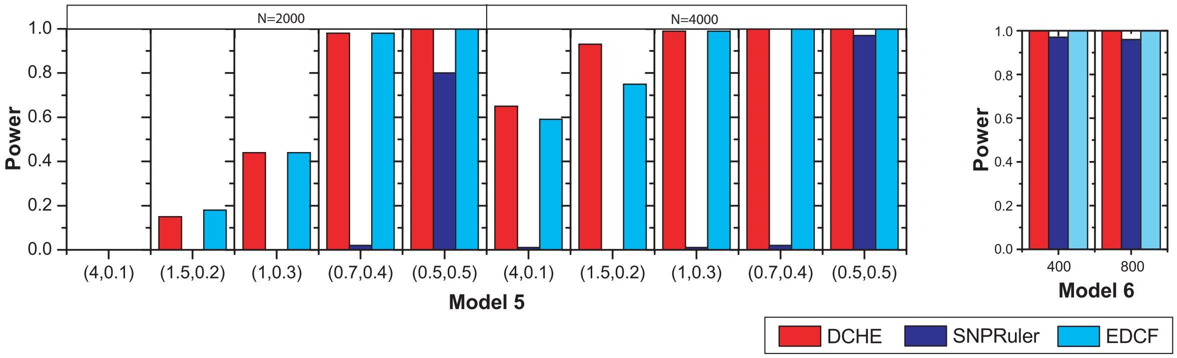

Figure 3.4 **Performance comparison on the three-locus epistasis models.** Performance comparison of DCHE, SNPRuler and EDCF on two three-locus epistasis models, Model 5 and Model 6, for different allele frequencies and sample sizes. The red, blue and cyan bars show powers of DCHE, SNPRuler and EDCF, respectively.

For Model 6, we set $MAF = 0.5$ and population prevalence $p = 0.01$ as EDCF does in [46]. Note that Model 6 gets the maximum h^2 when $p \in (0, \frac{1}{16}]$. Three methods' results are showed in Figure 3.4, from which we can see that all methods can get nearly full powers for Model 6. With considering the overall quality, DCHE and EDCF both reach 100% and SNPRuler hits to 0.965.

3.3.5 Experiments on AMD Data

Age-related macular degeneration (AMD) is an acquired degeneration of the retina that usually affects older adults and results in a loss of vision in the centre of the visual field. Like many other chronic diseases, AMD is caused by a combination of genetic and environmental risk factors, including and not limited to macular degeneration gene, too much exposure to sunlight and smoking. The reported AMD dataset contains genotypes of 103,611 SNPs from 96 affected individuals and 50 controls [73]. Before applying DCHE on AMD dataset, the same quality control used in [18] is applied: SNPs with more than 10% missing data or $MAF < 0.05$ or p-values from Hardy-Weinberg Equilibrium (HWE) tests less than 0.001 are removed. Subsequently, 90,449 SNPs from 50 controls and 96 cases are left in the dataset. The setting of parameters for DCHE is as follows: $l = \{10000, 4000, 2000\}$, $d = 2, 3, 4$ and $\alpha_0 = 1.5 \times 10^{-3}, 1.2 \times 10^{-08}, 1.0 \times 10^{-21}$ for two-, three- and four-locus interactions detections, respectively.

Table 3.3 Centre SNPs identified in top-1000/500 SNPs interactions on AMD dataset.

# SNPs per interaction	Centre SNPs from analyses of AMD dataset	
	Centre SNPs (Genomic position)	#Interacting SNPs
2	**rs380390 (Ch1: 196701051)**	524
	rs1329428 (Ch1: 196702810)	253
	rs1394608 (Ch5: 155783294)	23
	rs1740752 (Ch10: 38538771)	20
	rs1363688 (Ch5: 174609731)	11
	rs10512174 (Ch9: 88886574)	11
3	rs380390 (Ch1: 196701051)	709
	rs1329428 (Ch1: 196702810)	106
	rs1363688 (Ch5: 174609731)	63
	rs618499 (Ch11: 108148839)	47
	rs1926489 (Ch13: 92667989)	35
	rs3781868 (Ch11: 108059569)	34
4	rs380390 (Ch1: 196701051)	459
	rs618499 (Ch11: 108148839)	188
	rs3781868 (Ch11: 108059569)	115
	rs294278 (Ch3: 31127911)	36
	rs300780 (Ch2: 110819)	35
	rs315511 (Ch1: 84849116)	28

When we set $\alpha_0 = 1.5 \times 10^{-3}$ to filter out insignificant interactions for two-locus epistatic interactions detecting, DCHE can hardly report any qualified modules, so we select top-k modules to conduct the analysis. For $d = 3$ and 4, DCHE generates more than 1000 pairs epistatic interactions. In order to give a straight view of results, we introduce two concepts, "centre SNPs" and "centre genes", similar to those in [33]: we arrange those SNPs and genes in descending manner according to their frequencies showing in top-k interactions, and select top-s SNPs or genes as "centre". Based on the previous procedure, Table 3.3 and Table 3.4 give a general view of DCHE's results on AMD dataset with $k = 1000$ for $d = 2$, 3, $k = 500$ for $d = 4$, and $s = 6$. Names of SNPs or genes showed in bold font indicate that they present in tables. As we can see, top-k-s SNPs or genes tend to share some common elements among different settings of the order of interactions, d. For AMD dataset, two SNPs (rs380390 and rs1329428) already have been reported as the disease-associated SNPs with AMD based on results from single allelic association tests with $df = 1$ [73]. In our findings, DCHE also ranks rs380390 and rs1329428 as top 2 centre SNPs both in two- and three-locus epistatic interactions detecting. Both rs380390 and rs1329428 locate inside the gene CFH whose location is 1p32, and their protein products have an essential role in the regulation of complement activation and restricting the innate defence mechanism to microbial infections. In addition to rs380390 and rs1329428, we also find another interesting SNP, rs3781868, in the category $d = 4$. rs3781868 locates in the gene NPAT with location 11q22-q23 that is known to be essential for histone mRNA 3' end processing and recruiting CDK9 to replication-dependent histone genes.

We also analyse results using gene only from the top-1000/500 SNPs subset. Top-1000/500 SNPs are mapped to disease-related genes, which have been annotated in the HuGE Navigator database, and we get 720, 851 and 424 genes for $d = 2, 3, 4$ showed in Table 3.5. It is obvious that the majority of centre genes have not yet been reported by HuGE as associated with the AMD disease. Applying the similar analysis used in [33], we submit centre genes indicated in Table 3.5 to the ToppGene, a candidate gene prioritization tool [74], to evaluate biological significances of these novel genes. From DCHE's

Table 3.4 Centre genes identified in top-1000/500 SNPs interactions on AMD dataset.

#Genes per interaction	Centre genes from gene-only SNP analyses	
	Centre genes	#Interacting genes
2	**CFH: complement factor H**	777
	ZNF25: zinc finger protein 25	23
	SGCD: sarcoglycan, delta (35kDa dystrophin-associated glycoprotein)	23
	LRIG3: leucine-rich repeats and immunoglobulin-like domains 3	14
	DRD1: dopamine receptor D1	11
	ISCA1: iron-sulfur cluster assembly 1	11
3	CFH: complement factor H	815
	DRD1: dopamine receptor D1	63
	ATM: ataxia telangiectasia mutated	47
	GPC5: glypican 5	43
	NPAT: nuclear protein, ataxia-telangiectasia locus	34
	KDM4C: lysine (K)-specific demethylase 4C	25
4	CFH: complement factor H	459
	ATM: ataxia telangiectasia mutated	191
	NPAT: nuclear protein, ataxia-telangiectasia locus	115
	LRIG3: leucine-rich repeats and immunoglobulin-like domains 3	73
	TGFBR2: transforming growth factor, beta receptor II (70/80kDa)	38
	ACP1: acid phosphatase 1, soluble	35

results, ToppGene enriches a cell-cell communication pathway with the name 'REACTOME_ADHERENS_JUNCTIONS_INTERACTIONS'. Reported in Reactome, this pathway contains 14 centre genes and only one gene in this pathway presents in HuGE Navigator database. As gene names are given in HGNC, these genes are PVRL3, CDH18, CDH10, CDH11, CDH12, CDH13, CDH2, CDH4, CDH7, CDH6, CDH9, CDH8, CADM1, CADM3.

Table 3.5 The disease association of DCHE selected genes from gene-only SNP analyses.

# SNPs per interaction	# DCHE genes in top 1000 (500) SNP pairs	Reported in HuGE Navigator database	
		# Analyzed genes	# DCHE genes
2	720	151	20
3	851		28
4	424		13

For the detection of two-locus epistatic interaction on AMD dataset, DCHE successfully identifies rs380390 and rs1329428 reported in the original paper. Comparing with results obtained by other existing methods, we find that there are some overlaps between them. For example, DCHE lists two pairs of SNPs with ranking 246 and 247 (rs1394608 and rs3743175, rs1394608 and rs2828155), which have been identified by epiMODE applied on AMD dataset [52]. In addition, DCHE reports another interaction module (rs1394608 and rs6847164), whose $p - value$ is more significant than the above two ($p - value_{unadjusted} = 6.78 \times 10^{-10}$). rs1394608 resides within the intron of SGCD, a gene located on chromosome 5q33-34, which has been implicated in AMD [52]. rs6847164 resides within PDE5A, a gene located on chromosome 4q27. According to the databases of NCBI and Entrez, PDE5A is involved in the regulation of intracellular concentrations of cyclic nucleotides and is important for smooth muscle relaxation. DCHE has also detected other significant three-locus and four-locus interaction modules: (rs10487833, rs10495593, rs1740752) and (rs9302001, rs10497231, rs380390, rs1940041) whose unadjusted p-values are 3.24×10^{-18} and 8.28×10^{-28}, respectively. rs10487833 locates about 0.3Mb upstream of gene NAMPT on chromosome 10. NAMPT encodes a protein that is thought to be involved in many important biological processes, including metabolism, stress response and aging. rs10497231 resides at about 0.3Mb downstream of gene KCNH7 on chromosome 2. KCNH7 encodes a member of the potassium channel, voltage-gated, subfamily H related to the functions including regulating neurotransmitter release, heart rate, insulin secretion, neuronal excitability, epithelial electrolyte transport, smooth muscle contraction, and cell volume. rs9302001 locates about 0.4Mb upstream of gene ABCC4 on chromosome 13. The protein encoded by ABCC4 is a member of the superfamily of ATP-binding cassette (ABC) transporters, which is thought to play a role in cellular detoxification as a pump for its substrate, organic anions. The clustering details of genotype combinations of the above three interaction modules can be found in [45].

3.4 Conclusion

In this chapter, a novel algorithm DCHE, for detection of high-order genome-wide epistatic interactions is proposed. The critical step of DCHE is the dynamic clustering procedure, which is used to guide on how to merge genotype categories to a limited and variable number of groups. By dynamic clustering, DCHE tries to categorize approximately genotypes with similar genetic effects on phenotypes. Comprehensive and systematic comparisons of simulated datasets shows that DCHE can obtain more or comparable powers for both two- and three-locus interaction modules detecting comparing to other four recently developed algorithms, i.e. TEAM, SNPRuler, EDCF and BOOST. Furthermore, experiments on two real datasets of AMD and RA demonstrate that DCHE discovers many novel high-order associations that are significantly enriched in cases and many centre-SNPs and genes which only appear in detections of high-order epistatic interactions. Therefore, our method provides a promising way to accelerate large genome-wide association studies accurately.

PART 4

SEARCHING HIGH-ORDER GENOME-WIDE EPISTASIS ON MULTIPLE DISEASES

4.1 Introduction and Contributions

To the best of our knowledge, current epistasis detecting tools are only capable of identifying interactions on the data of GWAS with two groups, *i.e.* case-control studies. These tools are incompetent to discover the genetic factors with diverse effects on multiple diseases. Moreover, only use limited case samples, they may lose the benefit of alleviating deficiency of statistical powers by pooling different disease samples together. In this chapter, we first introduce a novel designed and implemented Bayesian inference method for Detecting genome-wide Association on Multiple diseases, named DAM, to address challenges. DAM employs Markov Chain Monte Carlo (MCMC) sampling based on the Bayesian variable partition model.

One of the limitations of DAM is that it takes very long time to analyze fully a real large GWAS data due to the enormous number of iterations needed by the MCMC. Thus, we present a heuristic method, named SAM, by using Jensen-Shannon divergence and a modified k-mean clustering to filter out a candidate set of SNPs showing potentially diverse effects on multiple phenotypic traits. A stepwise evaluation of association is engaged for both DAM and SAM to further determining the genetic effect types of associations. Systematic experiments on both simulated and real GWAS datasets demonstrate that our methods are feasible for identifying multi-locus interaction on GWAS datasets and enriches some novel, significant high-order epistatic interactions with specialties on various diseases.

4.2 DAM: A Bayesian Method for *D*etecting Genome-wide *A*ssociations on *M*ultiple Diseases

4.2.1 Notations

Suppose we have M SNPs, L groups and each group with N^l samples, $l \in \{1, 2, \cdot, L\}$. Let D be the genotypes of all samples at SNP loci. We use B to denote the count of partitions for L groups. The number is also known as Bell numbers. The M SNPs are assigned to $2B$ different label types, where the SNPs of the first B types are independently associated with the phenotype traits, and the SNPs of the last B types are jointly associated with the phenotype traits. Let $\widetilde{T}$ denote the set of the independent B types, and $\overline{T}$ denote the dependent B types. We use T_k to denote the kth type and T can be either $\widetilde{T}$ or $\overline{T}$. An example to show the partitions and types for a three-group dataset is showed in Figure 4.1, where there are 10 possible types in which type 1 through 5 indicate those SNPs are independently associated with phenotypes, and types 6 through 10 indicate the SNPs are dependently associated with the phenotypes. We use $\mathbb{P}$ to represent the total disjoint sets in a partition of L groups, and there are $|\mathbb{P}_{T_k}|$ disjoint sets for Type T_k. Let $I = (I_1, \ldots, I_M)$ be the memberships of M SNPs, in which $I_i \in \widetilde{T} \bigcup \overline{T}$. Let $\mathbb{M}_{T_k}$ denote the number of SNPs in type T_k ($\sum_{i=1}^{|T|} \mathbb{M}_i = M$), $D^{(T_k)}$ denote the genotypes of SNPs of type T_k, and $D_i^{(T_k)}$ denote the genotypes in ith disjoint set of partition for SNPs in type T_k. Please note that the superscript inside the parenthesis is merely a label and does not represent the exponent.

In the experiment of 3 groups GWAS data, we use group 1 and 2 as two case groups and group 3 as the shared control group. Since we only interested in identifying SNPs associated with disease traits, SNPs in types 2 through 5 and types 7 through 10 are the target ones.

4.2.2 Bayesian Variable Partition Model

Consider a categorical variable x, which can be sampled in t different distributions $\left\{\Delta^{(1)}, \Delta^{(2)}, \ldots, \Delta^{(t)}\right\}$ with t different parameter settings $\left\{\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(t)}\right\}$, where Θ_k is the parameter set used for describing x following kth distribution. The model used to

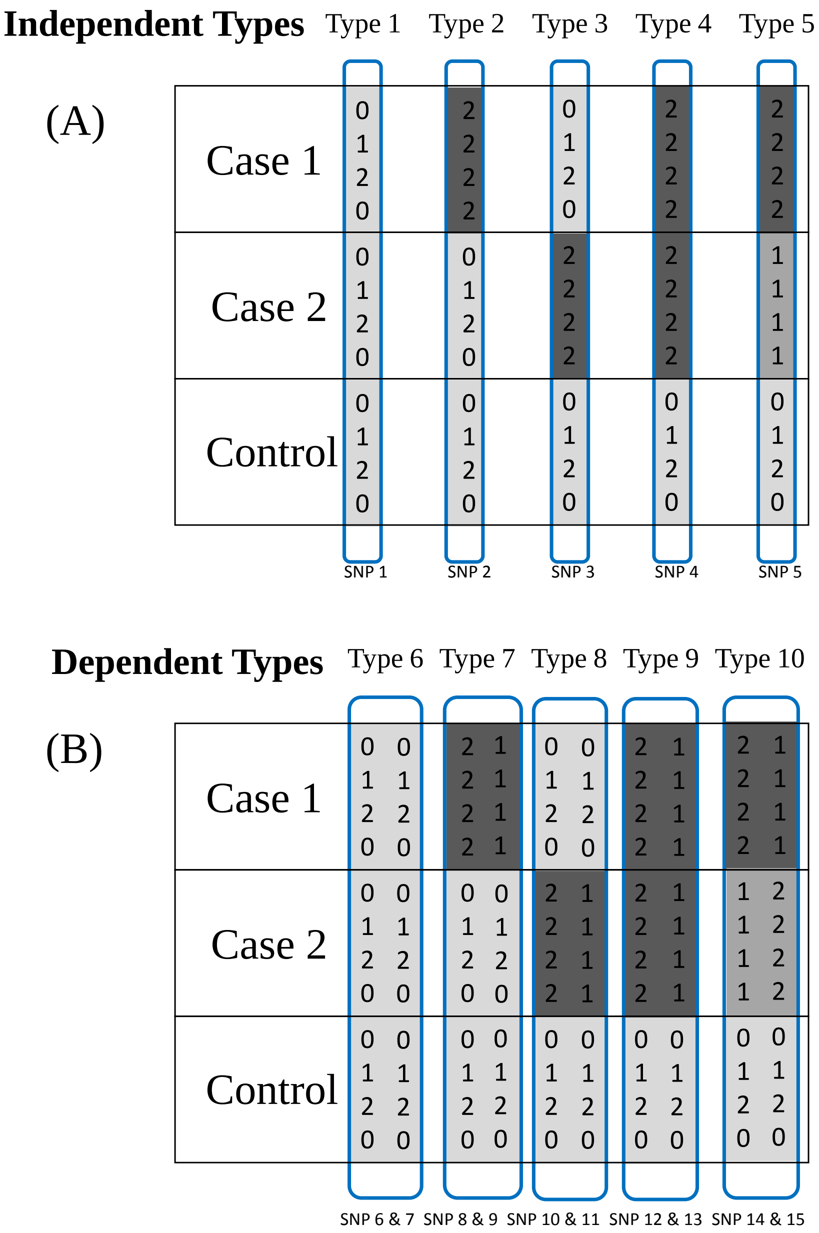

Figure 4.1 Illustration for 10 types on 3 groups.

capture the sums of independently and identically distributed mixture categorical variables in different distributions is referred as a ‘multinomial model’, meaning that it can be partitioned into t inseparable multinomial models. Consider a model for a vector of M categorical variables $X = \{x_1, x_2, \ldots, x_M\}$. If all variables are independent, the model can be simply treated as the union of M univariate multinomial models. If interactions exist among multiple variables, we can replace these multiple variables with a new model in which a single variable is used to collapse the interacting variables. The sample space of the collapsed variable is the product of the sample spaces of the variables before collapsing. Bayesian variable partition model (BVP) is a multinomial model based on Bayesian theorem, and we use BVP to capture the associations in the GWAS SNP data. For simplicity, we first describe the BVP by assuming independence between SNPs. The likelihood for the BVP of a variable of kth type distribution is

$$\begin{aligned} P(D^{(i)}|\Delta_k) &= \int P(D^{(i)}|\Delta_k, \Theta_k) d\Theta_k \\ &= \int_{\theta_1, \theta_2, \ldots, \theta_g} P(D|\theta_1, \theta_2, \ldots, \theta_g) P(\theta_1, \theta_2, \ldots, \theta_g) dp \qquad (4.1) \end{aligned}$$

where $D^{(i)}$ is the observations for the categorical variable x_i, and g is the number of category value for the variable x_i. Assuming a Dirichlet prior $Dir(\alpha_1, \alpha_2, \ldots, \alpha_g)$ for $P(\Theta = (\theta_1, \theta_2, \ldots, \theta_g))$, we obtain a closed form for Equation 4.1:

$$\begin{aligned} P(D^{(i)}|\Delta_k) &= \int_{\theta_1, \theta_2, \ldots, \theta_g} P(D^{(i)}|\theta_1, \theta_2, \ldots, \theta_g) P(\theta_1, \theta_2, \ldots, \theta_g) dp \\ &= \int_{\theta_1, \theta_2, \ldots, \theta_g} \frac{1}{Beta(\alpha_1, \alpha_2, \ldots, \alpha_g)} \prod_{i=1}^{g} p_i^{n_i + a_i - 1} dp \\ &= \left(\prod_{i=1}^{g} \frac{\Gamma(n_i + \alpha_i)}{\Gamma(\alpha_i)} \right) \frac{\Gamma(|\alpha|)}{\Gamma(\mathbb{N} + |\alpha|)} \qquad (4.2) \end{aligned}$$

where p_i is the probability of the ith category, $\mathbb{N}$ is the total number of observations, n_i is the frequency of ith category, and $|\alpha|$ is the sum of $(\alpha_1, \alpha_2, \ldots, \alpha_g)$. Suppose the vector I is a vector of the memberships of distribution types for the categorical variable vector X, we obtain the posterior distribution of I as

$$P(I|D) \;\propto\; \left(\prod_{i=1}^{M} P(D^{(i)}|I)\right) P(I) \tag{4.3}$$

Based on Bayesian theorem, we design the specific BVP for the genome-wide association mapping as follows. For the SNPs independently associated with phenotypes, we use $\Theta_{ij}^{(\widetilde{T}_k)} = ((\theta_{j1}, \theta_{j2}, \theta_{j3}) : I_j \in \widetilde{T}_k)$ to denote the genotype frequencies of SNP in ith disjoint set of the partition of type $\widetilde{T}_k$. Note that the SNP with membership value in $\widetilde{T}$ does not have interaction with other SNPs. The likelihood of $D^{\widetilde{T}}$ based on BVP model is that

$$P(D^{(\widetilde{T})}|\Theta^{(\widetilde{T})}) = \prod_{k=1}^{|\widetilde{T}|} \prod_{I_j=k} \prod_{i=1}^{|\mathbb{P}_k|} (\Theta_{ij}^{(\widetilde{T}_k)})^{n_{ij}^{(\widetilde{T}_k)}}, \tag{4.4}$$

where

$$(\Theta_{ij}^{(\widetilde{T}_k)})^{n_{ij}^{(\widetilde{T}_k)}} = \prod_{s}^{3} \theta_{js}^{n_{js}}$$

and $n_{ij}^{(\widetilde{T}_k)} = (n_{j1}, n_{j2}, n_{j3})$ are the genotype counts of SNP j in ith disjoint set of partition of type $\widetilde{T}_k$. Similarly, assuming a Dirichlet distribution $Dir\,(\alpha)$ with parameter $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ to $\Theta_{ij}^{(\widetilde{T}_k)}$, we can integrate out $\Theta_{ij}^{(\widetilde{T}_k)}$ and obtain the marginal probability:

$$P(D^{(\widetilde{T})}|I) = \prod_{k=1}^{|\widetilde{T}|} \prod_{I_j=k} \prod_{i=1}^{|\mathbb{P}_k|} \left(\left(\prod_{s=1}^{3} \frac{\Gamma(n_{js} + \alpha_s)}{\Gamma(\alpha_s)} \right) \frac{\Gamma(|\alpha|)}{\Gamma(\mathbb{N}_{ij} + |\alpha|)} \right) \tag{4.5}$$

where $\mathbb{N}_{ij}$ is the count of samples in groups belonging to ith disjoint set of partition of type $\widetilde{T}_k$ for the SNP with I_j, and $|\alpha|$ represents the sum of all elements in α. Please note that, for simplicity of the equation, we omit the disjoint set labels and type labels of n_{js}, α_s and α.

For the rest SNP assigned to types $\overline{T}$, they are supposed to contribute to the phenotype traits with interactions. Thus, we concatenate SNPs of type $\overline{T}_k$ into a single categorical variable to resolve the interactions. Note that there are $3^{\mathbb{M}_{\overline{T}_k}}$ possible concatenated genotype combinations. Let $\Theta_{ij}^{(\overline{T}_k)} = ((\phi_1, \phi_2, \dots, \phi_{3^{\mathbb{M}_{\overline{T}_k}}})$ be the concatenated genotype frequencies over $\mathbb{M}_{\overline{T}_k}$ SNPs in the ith disjoint set of the partition of type $\overline{T}_k$. Similarly, we use a Dirichlet prior $Dir\left(\beta\right)$ for $\Theta_{ij}^{(\overline{T}_k)}$, $\beta = (\beta_1, \beta_2, \dots, \beta_{3^{\mathbb{M}_{\overline{T}_k}}})$. According to Equation 4.2, we obtain the marginal probability:

$$P(D^{(\overline{T})}|I) = \prod_{k=1}^{|\overline{T}|} \prod_{i=1}^{\mathbb{P}_k} \left(\left(\prod_{s=1}^{3^{\mathbb{M}_{\overline{T}_k}}} \frac{\Gamma(n_s) + \beta_s}{\Gamma(\beta_s)} \right) \frac{\Gamma(|\beta|)}{\Gamma(\mathbb{N}_i + |\beta|)} \right) \tag{4.6}$$

where $\mathbb{N}_i$ is the count of individuals belonging to ith disjoint set of partition of type $\overline{T}_k$, n_s is the count of sth concatenated genotype combination, and for simplicity, we again omit the type labels for $\mathbb{N}_i$, $n_i^{(\omega)}$, and β.

Combining Equation 4.3, 4.5 and 4.6 , we obtain the posterior distribution of I as

$$P(I|D) \propto P(D^{(\widetilde{T})}|I)P(D^{(\overline{T})}|I)P(I) \tag{4.7}$$

We set $P(I) \propto \prod_{k=1}^{|\widetilde{T}|} p_{\widetilde{T}_k}^{\mathbb{M}_{\widetilde{T}_k}} \prod_{k=1}^{|\overline{T}|} p_{\overline{T}_k}^{\mathbb{M}_{\overline{T}_k}}$ to incorporate the prior knowledge of the phenotype association types of SNPs. In our experiments with three groups, we set $p_k = 0.001, k \in \{2, \dots, 10\}$, and $\alpha_s = \beta_s = 0.5, \forall s$.

4.2.3 MCMC Sampling

We apply MCMC method to sample the indicator I from the distribution defined by Equation 4.7. According to the prior $P(I)$, DAM first initializes I randomly, then use the Metropolis-Hastings (MH) algorithm [75] to construct a MCMC to update I. Three types of updating strategies are employed: (i) randomly change an SNP's type, (ii) randomly exchange two SNPs' types when their types are different, or (iii) randomly shuffle the SNP set with types of $\overline{T}$. At each iteration, the acceptance of newly generated indicator is based on the MH ratio, which is a Gamma function. DAM begins to record the membership of all SNPs from the accepted indicator after the burn-in process. The frequencies of the memberships of types are treated as the posterior distribution SNPs being associated with one or more phenotype traits. The number of iteration in the burn-in process is fixed to $10M$, and the number of sampling iteration is set to M^2 in our experiments, where M is the number of SNPs. We also apply a distance constraint that the physical distance between two SNPs in a multi-locus module is at least 1Mb. This constraint is used to avoid associations that might be attributed to the LD effects [11].

4.3 SAM: A Jensen-*S*hannon Divergence Based Method for Detecting Genome-wide *A*ssociations on *M*ultiple Diseases

One of the limitations of DAM is that it takes very long time to analyze fully a real large GWAS data due to the enormous number of iteration of MCMC. We propose a fast algorithm for detection of high-order epistatic interactions in GWAS with considering multiple disease cases. The basic idea of SAM is to group SNPs displaying similar association effects on one or more diseases together based on JensenShannon divergence, a traditional method of measuring the similarity between two probability distributions, and an improved k-means clustering algorithm. These candidates are examined to find the representative SNPs from groups for further statistic association tests. Before we directly apply k-means clustering algorithm, there are three questions needs to clarify:

1. What distance could be a good measure for two distributions?
2. How to decide the centroids of clustered groups?
3. What SNPs in the clusters that we select for further statistic tests?

From Chapter 2, χ^2 statistic test can be used as the statistical correlation between two loci for the filtering stage by choosing appropriate thresholds. There are two defects of χ^2

test: 1) the approximation of χ^2 distribution is broken when cells' frequencies in contingency tables are less than 5; 2) χ^2 test is not satisfying the conditions to be a distance measure, which make us impossible to apply distance-based clustering algorithms. To capture the correction between SNPs and interest traits, the Jensen-Shannon divergence is employed. Since the genotypes of samples and SNPs are not mapping to Euclidean space, the centroid of clusters cannot be obtained by averaging the coordinates of the SNPs in the cluster. So the SNPs with the sum of distance to the rest SNPs in the cluster is treated as the centroid, and we use a heuristic routine to sample the centroids instead of emulate all the pair-wise distances between all SNPs. At last, a score function also on the basis of Jensen-Shannon divergence is employed to select candidates from each cluster.

4.3.1 Notation

We follow the most common notations used in the previous section. Let L denote the total number of groups including $L-1$ case groups and one control group. Let N be the total count of individuals from L groups, and M be the number of SNP markers in the GWAS data.

4.3.2 Jensen-Shannon Divergence

The Jensen-Shannon divergence (JS) is a popular distance measurement based on Kullback-Leibler divergence [76], which evaluate the similarity between two probability distributions. Given two probability distributions, p and q with g categories, the Kullback-Leibler divergence (KL divergence) is defined as follows:

$$\mathbb{KL}\left(p \parallel q\right) = \sum_{i=1}^{g} p_g log \frac{p_g}{q_g} \tag{4.8}$$

The KL divergence is not a distance since it is not symmetric. One symmetric version of the KL divergence is the Jensen-Shannon divergence, defined as follows:

$$JS\left(p,q\right) = 0.5\mathbb{KL}\left(p \parallel \frac{p+q}{2}\right) + 0.5\mathbb{KL}\left(q \parallel \frac{p+q}{2}\right) \tag{4.9}$$

By given a L groups GWAS data, there are $|H| = \binom{L}{2}$ combinations of two groups out of L groups. Let H denote the set of all possible combinations. For each combination h_i and two SNPs, x_a and x_b, two probability distributions, p^{h_i} for the first group and q^{h_i} for the second group where each category indicates the possibility of individuals having a certain genotypes defined by x_a and x_b. Based on the Jensen-Shannon divergence, we define the distance between two SNPs, x_a and x_b as follows

$$Dist(x_a, x_b) = \max_{h_i \in H} JS\left(p^{h_i}, q^{h_i}\right) \tag{4.10}$$

4.3.3 K-means Clustering

We want to find the set of d SNPs that maximizes the JS dissimilarity between any two groups. But it is computational expensive to examine every d-combinations of nearly million SNPs for $d \geq 3$. In order to diminish time complexity, we make use of k-means clustering to group SNPs into clusters where interacting SNPs tend to be placed into separate clusters and SNPs with similar genotype frequencies distributions between groups tend to be put into the same clusters. Using some ranking technique, we only need to investigate the interactions of those SNPs that are assigned to different clusters. We define the distance of an SNP, x_a, to a cluster, C, as follows

$$Dist(x_a, C) = \frac{1}{|C|} \sum_{x_i \in C} Dist(x_a, x_i) \tag{4.11}$$

where $|C|$ is the number of SNPs in that cluster. There is no geometric coordinate of each SNP in our distance measure, so it is inexplicit to calculate the centroid of each cluster by averaging the sum of coordinates of all SNPs. Instead of doing so, we define the centroid of a cluster as the SNP with the smallest amount of distance to the rest of SNPs in the same cluster. To avoid emulating all pairs of SNPs to obtain the centroid, we design a heuristic routine, *Centralizing*, as follows. Note that, assumption made on the clustering data is that the SNPs are around particular centroid, and the probability of an SNP showing near the centroid is larger than the possibility of an SNP showing far away the centroid. We say this kind of data as a centred data. First, an SNP, x_b, is randomly selected from a cluster, and the distances from x_b to the rest of SNPs, C/x_b, in the same cluster are computed and sorted according to the magnitude. Next, we put those SNPs, C/x_b, into bins with fixed width, like using the histogram to represent graphically the distribution of numerical data. Along the increasing of these distances, the bin in which the number of SNPs starts to decrease is marked and the corresponding distance value is used as threshold τ to filter out SNPs. At the end, only those SNPs with distances larger than τ will be selected to calculate the centroid. The correctness of this heuristic routine to locate the centroid is ensured in Theorem 1.

Theorem 1. *For a centred data in any dimensional space, given an arbitrary point x_a, the distance, $Dist$, between x_a and the rest points is considered as a random variable. The distance between the centroid x_0 and x_a, $Dist(x_a, x_0)$ is indicating the last mode (if exists) for the probability distribution of $Dist$.*

Proof. Without loss of generality, we prove the theorem in 2-dimensional space, and for the other dimensional spaces, similar proof can be applied. As shown in Figure , there are two rings, R_A, R_B, whose area amounts are equal. R_A and R_B have the same centroid, x_a, and x_0 is also inside these two rings. Two lines can be drawn to cut out two pieces, s_a, s_b, as illustrated in the figure, where the areas of s_a and s_b are equal. Obviously, the possibility of points falling in s_a is larger than the possibility of points falling in s_b because of the centred data assumption. We can cut the rest of R_A and R_B in the same way and we can gain that the possibility of points falling in R_a is larger than the possibility of points falling in R_b. Therefore, as the distance getting larger than $Dist(x_a, x_0)$, the possibility to have a point is decreasing. Based on the mode definition, $Dist(x_a, x_0)$ is indicating the last mode for the probability distribution of $Dist$. □

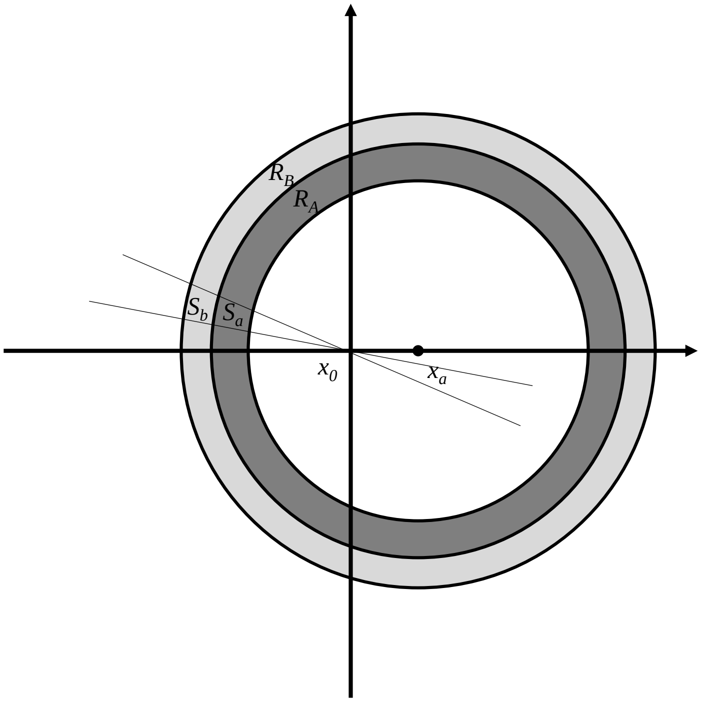

Figure 4.2 Illustration for a centred data in 2-dimensional space.

Based on the above theorem, we can further reduce the computation by choosing multiple SNPs and narrow down the searching space for the centroid.

After clustering, top $d \times k$ candidates from all clusters based on a ranking score. A candidate in a cluster is an SNP, which shows high dissimilarity between any two groups, in other words, far away from the elements in the other clusters. We define the score as follows, where C_i is the centroid SNP of ith cluster.

$$Score(x) = \sum_{x \in C_j, j \neq i} Dist(x, C_i) \tag{4.12}$$

4.3.4 Algorithm

The details of the SAM algorithm is showed in Algorithm 2 consisting two parts: clustering and candidates ranking. The convergence of K-means clustering is very fast on the simulated data. Usually setting the number of iteration to 10 is large enough to get an overview of the SNPs' belongings. Inside the k-means clustering, the distance between each SNP and the centroid is computed according to Equation 4.11 and the *Centralizing* procedure is employed to update the centroids. In the second part, all SNPs are ranked based on Equation 4.12 and inserted into a size-limited descending list to select promising candidates.

Algorithm 2: The SAM Algorithm

```
Input: An $N \times (M+1)$ matrix
Output: The top-$kd$ SNP candidates
Read $N \times (M+1)$ matrix file;
Initialize $n_{iteration}$;
Randomly select $k$ SNPs as centroid;
for $i$ in $n_{iteration}$ do
    for each SNP $x$ do
        Calculate $DIST(x, C)$ ;
        Assign $x$ to a cluster ;
    end
    Apply the procedure Centralizing;
    $i++$;
end
$t' = t' + 1$;
Initialize descending list List with length $kd$;
for each SNP $x$ do
    Calculate $Score(x)$;
    Place $x$ into List if $Score(x)$ is among top $kd$ SNPs ;
end
```

4.4 Stepwise Evaluation of Interaction

With the candidate SNPs generated by MCMC and BVP, we apply the χ^2 statistic and the conditional χ^2 test to measure the significance for a module of SNPs. Let $\mathbb{A} = (x_1, x_2, \ldots, x_d : T_k)$ denote an SNP module $\mathbb{A}$ with d SNPs of type T_k. We use

$\chi^2(x_1, x_2, \ldots, x_d : T_k)$ to denote the χ^2 statistic of $\mathbb{A}$ and $\chi^2(x_1, x_2, \ldots, x_d | x_{c_1}, x_{c_2}, \ldots, x_{c_{d'}} : T_k)$ as the conditional χ^2 statistic by given a subset of $\mathbb{A}'$, $(x_{c_1}, x_{c_2}, \ldots, x_{c_{d'}})$ with d' SNPs. The χ^2 statistic can be calculated as

$$\chi^2(x_1, x_2, \ldots, x_d : T_k) = \sum_{i=1}^{|\mathbb{P}_{T_k}|} \sum_{s=1}^{3^d} \frac{(n_{i,s} - e_{i,s})^2}{e_{i,s}} \tag{4.13}$$

where $n_{i,s}$ is the frequency of sth genotype combination in ith disjoint set of the type T_k, and $e_{i,s}$ is the corresponding expected frequency. The degrees of freedom for Equation 4.13 is $(|\mathbb{P}_{T_k}| - 1) \cdot (3^d - 1)$.

The conditional independent test based on the χ^2 statistic is defined as follows

$$\begin{gathered}\chi^2(x_1, \ldots, x_d | x_{c_1}, \ldots, x_{c_{d'}} : T_k) = \\ \sum_{\iota=1}^{3^{d'}} \sum_{i=1}^{|\mathbb{P}_{T_k}|} \sum_{s=1}^{3^{d-d'}} \frac{(n_{i,s}^{(\iota)} - e_{i,s}^{(\iota)})^2}{e_{i,s}^{(\iota)}}\end{gathered} \tag{4.14}$$

where we calculate χ^2 statistic for $\mathbb{A} - \mathbb{A}'$ separately for the given genotype combinations of $\mathbb{A}'$. The degrees of freedom for Equation 4.14 is $3^{d'} \cdot (|\mathbb{P}_{T_k}| - 1) \cdot (3^{d-d'} - 1)$. We treat those SNPs as redundant SNPs when they are conditional independent by giving a subset of the SNP module. To avoid these redundant SNPs, we define an SNP module ($d \geq 2$) as a compact epistatic interaction by the following Definition.

Definition 2. *an SNPs module $\mathbb{A} = (x_1, x_2, \ldots, x_d : T_k)$ is considered as a significant compact interaction by giving a significant level α_d, if it meets the following three conditions:*
(1) the p-value of $\chi^2(x_1, x_2, \ldots, x_d : T_k) \leq \alpha_d$;
(2) the p-value of $\chi^2(x_1, x_2, \ldots, x_d : T_k) =$ the minimum p-value of $\chi^2(x_1, x_2, \ldots, x_d : T)$;
(3) the p-value of $\chi^2(x_1, x_2, \ldots, x_d | x_{c_1}, x_{c_2}, \ldots, x_{c_{d'}} : T_k) \leq \alpha_d$ for $\forall \mathbb{A}' = (x_{c_1}, x_{c_2}, \ldots, x_{c_{d'}} : T_k)$ whose p-value $\leq \alpha_{d'}$.

Based on the Definition 2, we develop a stepwise algorithm to search for top-f significant

d-locus significant compact interactions, where the searching space only includes the SNP markers generated by MCMC and BVP. Here, f is a user defined number. We assume that one SNP can only participate in one significant interaction with one type. For the SNP markers of type $\widetilde{T}$, we first searches all the modules with just one SNP based on Definition 2, then we recursively tests all the possible combinations by setting the module size with one more SNP. For the SNPs reported as jointly contributing to the disease risk, we calculate the p-value under different types and use the conditional test if part of SNPs already reported as significant. All SNPs with significant marginal associations after a Bonferroni correction are reported in a list $\mathbb{L}$. The algorithm recursively searches the interaction space with larger module size until d reaches a user pre-set value. We add all novel d-way interactions (i.e. none of the SNPs in the module has been reported earlier) that are significant to $\mathbb{L}$ after the Bonferroni correction for $B \cdot \binom{M}{d}$ tests. For the interactions whose subsets have been reported as significant before, we use the conditional independent test, and put the interaction in $\mathbb{L}$ if it is still significant after Bonferroni correction of $B \cdot \binom{M}{d} \cdot \binom{d}{d'}$ tests. We also apply a distance constraint that the physical distance between two SNPs in a multi-locus module should be at least 1Mb. This constraint is used to avoid associations that might be due to the linkage disequilibrium effects [11].

4.5 Experiment Results

We first give definitions of 8 simulated multi-disease models and the metric power measurement, and then evaluate the effectiveness of our two methods. We also apply DAM and SAM on two real GWAS datasets, Rheumatoid Arthritis (RA) and Type 1 Diabetes (T1D), and we find not only the results reported by other literature but also some novel interesting interactions.

4.5.1 Experimental design

Data simulation To evaluate the effectiveness of DAM and SAM, we perform extensive simulation experiments using four disease models with two-locus associations on three groups. The genotypes of unassociated SNP are generated by the same procedure used in previous studies [45] with Minor Allele Frequencies (MAFs) sampled from $[0.05, 0.5]$. The odds tables for four models are identical to the first four models in Section 3.3.1. The settings for four datasets are showed in Table 4.1. In a setting, all models are using the same MAF $\in \{0.1, 0.2, 0.4\}$, we generate 100 replicas per setting. Therefore, by given an MAF, a dataset contains at most 8 associations labeled as Ep 1 to 8, where there are 4 single-locus associations and 4 two-locus associations. Note that there are 3 epistatic interactions in Model 5, because the combination of three 2-locus models does not exist when MAF $= 0.1$. Each simulated replica contained $M = 1000$ SNPs. The sizes of three groups are set to $(1000, 1000, 2000)$ or $(2000, 2000, 4000)$, where the first two groups are the case groups and the third one is the control group.

Table 4.1 Four settings of model combinations for four datasets.

Epistasis (Ep)	1	2	3	4
Type	7	8	9	10
Setting 1	Model 1	Model 1	Model 1	Model 1 & 2
Setting 2	Model 2	Model 2	Model 2	Model 2 & 3
Setting 3	Model 3	Model 3	Model 3	Model 3 & 4
Setting 4	Model 4	Model 4	Model 4	Model 4 & 1

Statistical power In the evaluation of performances on simulated data, 100 datasets are generated for each setting. The measure of discrimination power is defined as the fraction of 100 datasets on which the ground-truth associations are identified as compact and significant by DAM or SAM.

4.5.2 Null Simulation to Test Type I Errors

Here we examine the false positive rate for varied sizes of interaction, *i.e.* $d = 2, 3, 4$. We generate 1000 null data sets for 6 settings, respectively. Specifically, we fix the number of SNP to 1000 and vary the number of samples in each group. The first four settings regarding the numbers of the individuals are $N1 = \{200, 200, 400\}$, $N2 = \{400, 400, 800\}$,

$N3 = \{800, 800, 1600\}$, and $N4 = \{1600, 1600, 3200\}$, where the first two numbers indicate the sizes of two case groups, and the last number is the size of the control group. For the rest two settings, using N4 as the sizes for groups, we increase the number of SNP to 2000 and 4000. All of the SNPs are generated independently, with MAFs uniformly distributed in $[0.05, 0.5]$. Note that we set the significance level to 0.1 and also apply the Bonferroni correction introduced in Section 4.4. The degree of freedom for Pearson's χ^2 test is $df = (|T| - 1)(|G| - 1)$, where T denotes the type and G is the set of genotype given the SNP module. The degree of freedom for conditional χ^2 test is $|G'|(|T|-1)(|G/G'|-1)$, where G' is the set of genotype given the subset of SNP module and G/G' is the set of genotype for the rest SNPs. We use the same measurement defined in the Section of Statistical power to record the false positive rates. The result is showed in Figure 4.3.

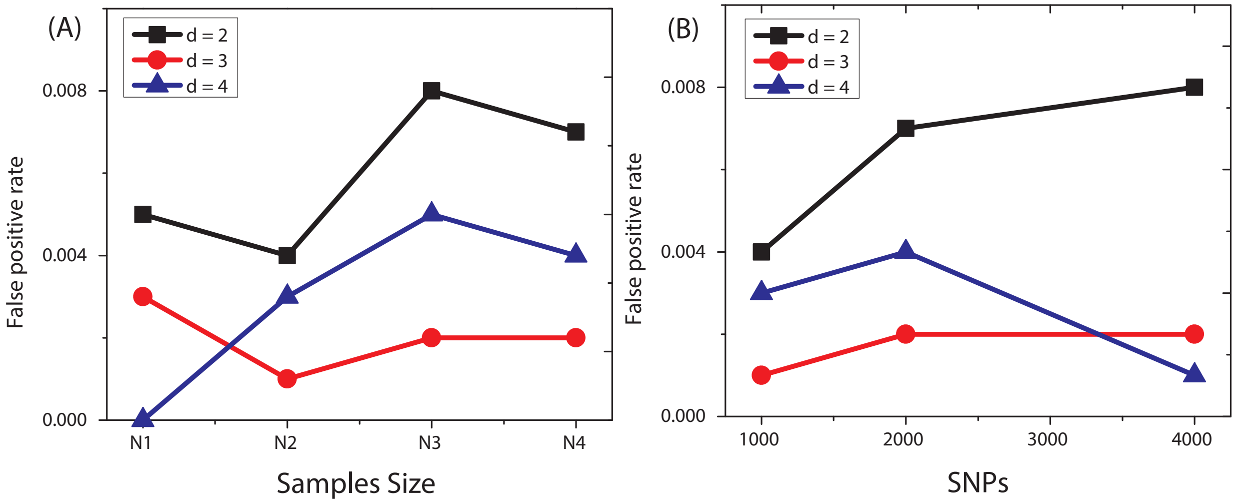

Figure 4.3 False positive rates of DAM under null models. The plots in (A) and (B) show the false positive rates of DAM for different ds when sample sizes and the numbers of SNP vary.

By setting the significance level to 0.1, we observe that the false positive rates of SAM and DAM under different sizes of interactions are all below 0.1, which complies with the fact that Bonferroni correction is too conservative. The overall false positive rate does not fluctuate too much as the increasing of the numbers of samples or SNPs.

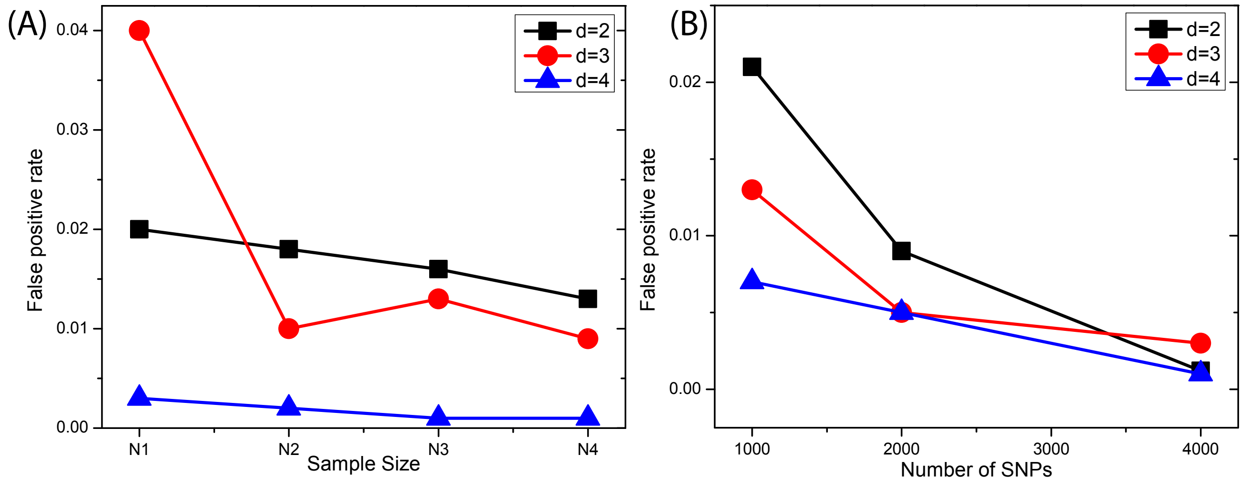

Figure 4.4 False positive rates of SAM under null models. The plots in (A) and (B) show the false positive rates of SAM for different *d*s when sample sizes and the numbers of SNP vary.

4.5.3 Simulation Experiment from Four Epistasis Models

Test results for SNPs contributing jointly to the disease risks are illustrated in Figure 4.5. We can find that both methods can report nearly 100% of embedded interactions for dataset 1 and 2. It also obtained nearly full power when MAF is 0.1 for dataset 1, 2, and 4. Similar to the results on single-locus disease models, after the stepwise procedure, more interactions were assigned to correct types. The overall quality of SAM and DAM are 0.732 and 0.805 for the sample size of $\{1000, 1000, 2000\}$, and 0.9195 and 0.9197 for the sample size of $\{2000, 2000, 4000\}$, respectively. We can find that along the increasing of the sample size, the accuracies of SAM and DAM do not differ very much.

4.5.4 Computation Time

We tested the running time of SAM and DAM on our desktop computer. The results are showed in Table 4.2. We can see that SAM is much faster than DAM under these three cases, and the speed-up of SAM is becoming higher when the number of SNPs is increasing.

Table 4.2 Time Comparison of SAM and DAM

Data size	SAM	DAM
N=6,000, M=1,000	8.6s	131.4s
N=6,000, M=5,000	60.4s	793.2s
N=6,000, M=10,000	359.8s,	2165.1s

4.5.5 Experiments on The WTCCC Data

We applied DAM to analyze data from the WTCCC (3999 cases in total and 3004 shared controls) on two common human diseases: Rheumatoid Arthritis (RA), Type 1 Diabetes (T1D), where RA is treated as group 1, T1D is treated as group 2, and control group is group 3. The procedure of quality control is the same as presented in the [45]. After the SNP filtration, the data set finally has 333,739 high-quality SNPs. DAM ran about 36 hours, for a total of 1×10^{11} iterations. We started the MCMC chains 10 times independently and pool the posterior distribution about the whole human genome together showed in Figure 4.6.

Table 4.3 Some significant interactions obtained by DAM on the WTCCC data. Following each SNP is its the chromosome index.

State Index	DAM p-value	SNP 1	SNP 2	SNP 3
10	1.33E-95	rs3099844 (Chr6)	rs3135376 (Chr6)	
9	3.01E-21	rs1230649 (Chr1)	rs11984645 (Chr8)	
8	1.70E-18	rs9467704 (Chr6)	rs12924729 (Chr16)	
7	5.84E-13	rs2958371 (Chr8)	rs1420247 (Chr16)	
10	1.09E-10	rs9393713 (Chr6)	rs2104286 (Chr10)	
10	2.04E-9	rs9358932 (Chr6)	rs7085631 (Chr10)	
10	5.58E-96	rs2488457 (Chr1)	rs2894254 (Chr6)	rs2958371 (Chr8)
10	3.07E-89	rs3099844 (Chr6)	rs3135376 (Chr6)	rs11984645 (Chr8)
10	1.03E-18	rs1230649 (Chr1)	rs9467704 (Chr6)	rs12924729 (Chr16)
10	9.47E-12	rs9393713 (Chr6)	rs2104286 (Chr10)	rs1420247 (Chr16)

The MHC region in chromosome 6 with respect to infection, inflammation, autoimmunity, and transplant medicine has been heavily reported [77] [18] [78], which is also verified by the results from DAM on Chromosome 6, that most SNPs with high posterior probabilities are located inside chromosome 6. But we still can find some inconsistency between what has been reported by other methods and what DAM has reported. For example, many SNPs contributing to RA are not located within the MHC region, while the SNPs associated with T1D gather in MHC region, where we can see two vertical lines of blue dots in Type 7 with one higher line in the MHC region. We select top 30 SNPs according to their posterior probabilities and analyze them with the stepwise evaluation procedure introduced in Section 4.4. Table 4.3 summarizes some novel findings of the significant interactions with p-values adjusted by 2.78×10^{11} for two loci and 3.10×10^{16} for three loci interactions, respectively. We again find some epistasis with diverse associations, and they are not only located in chromosome 6. For example, rs1230649 is located inside the coding region of the gene, PHTF1 (putative homeodomain transcription factor 1). PHTF1 can recruit FEM1B to the endoplasmic reticulum membrance and FEM1B belonging to the death receptor-associated family of proteins plays an important role in mediating apoptosis. The associated SNP with rs1230649 is rs11984645, which is located near the gene, MRPL15, mitochondrial ribosomal protein L15. By using the LD plot from HapMap and the NCBI dbSNP (Figure 4.7), we found rs1230647 and MRPL15 are both inside a block caused by LD effect. MRPL15 is encoded by nuclear genes and helps in protein synthesis within the mitochondrion, and MRPL15 is also recognized as associated with T1D by other study [79]. For another SNP, rs2488457 is in the coding region of the gene, AP4B1-AS1, which has been inferred to be associated with RA by many studies [80][81]. rs2488457 is also linked to rs2958371 on Chromosome 8. rs2958371 is inside the coding region of gene NCOA2, which encodes nuclear receptor coactivator 2. NCOA2 helps in the function of nuclear hormone receptors that plays important roles in various aspects of cell growth and development. Although the biological verifications of these interactions beyond this work, they still provide some insight relations between RA and T1D.

4.6 Conclusion

The enormous number of SNPs genotyped in genome-wide case-control studies poses a significant computational challenge in the identification of gene-gene interactions. During the last few years, many computational and statistical tools are developed to finding gene-gene interactions for data with only two groups, *i.e.* case and control groups. Here, we present two novel methods, named “DAM” and “SAM”, to address the computation and statistical power issues for multiple diseases GWASs. We have successfully applied our methods to systematic simulation and also analyzed two datasets from WTCCC. Our experimental results on both simulated and real data demonstrate that our methods are capable of detecting high-order epistatic interactions for multiple diseases at the genome-wide scale.

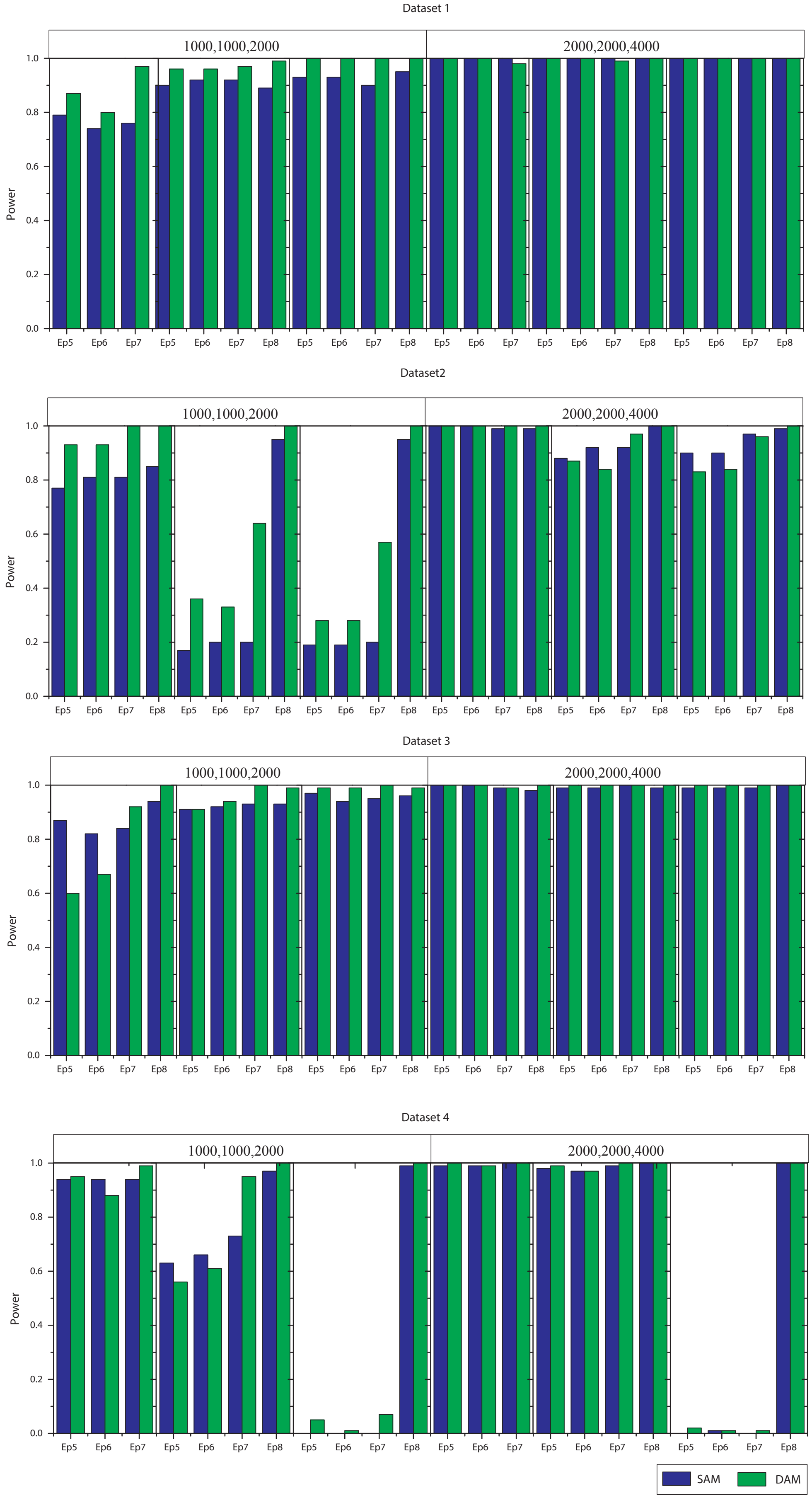

Figure 4.5 Performance comparison between DAM and SAM on simulated disease datasets 1-4 embedded with Joint effect SNPs. Note that the combinations of model 1 with the rest

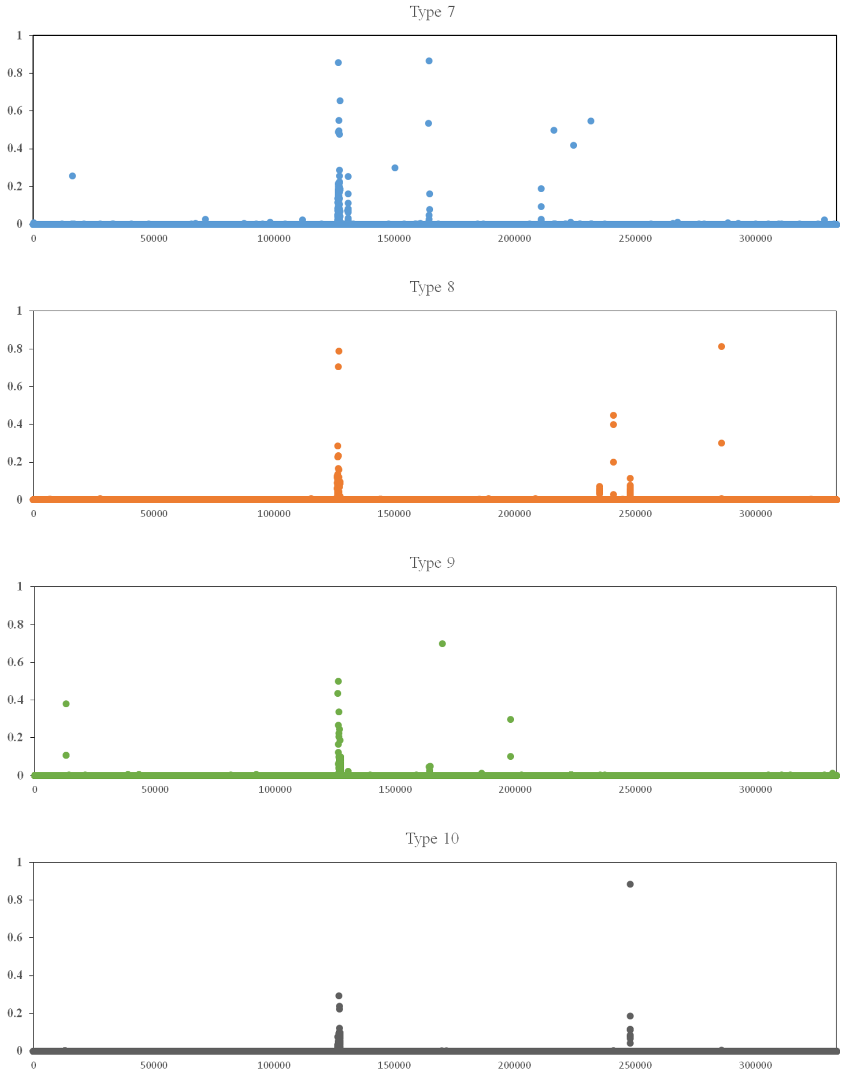

Figure 4.6 Posterior probabilities of SNPs on the chromosome 1 to 22. Type 7 to 10 are the disease association types. The x-axis indicates the position orders of SNPs on the chromosomes, and the y-axis shows the posterior probabilities of SNPs.

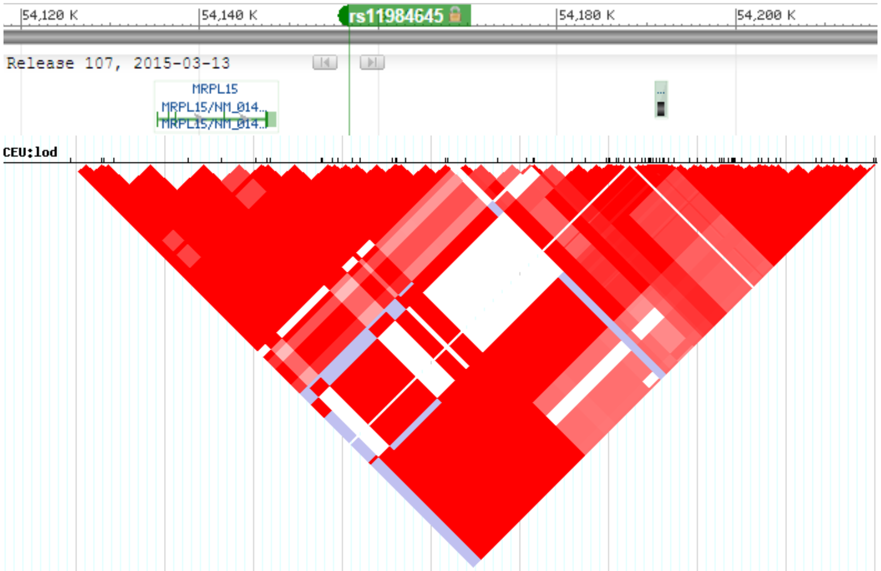

Figure 4.7 The 54120k - 54220k region of Chromosome 8. The LD plot from HapMap shows that a block structure exists from 54120k to 54220k.

PART 5

BLOCK-BASED DETECTION OF HIGH-ORDER GENOME-WIDE EPISTASIS ON MULTIPLE DISEASES

5.1 Introduction and Contributions

Linkage disequilibrium (LD) is the well-known dependence structure between adjacent SNPs in the human genome. We say two loci at LD if they are non-randomly inherited, and we say they at linkage equilibrium (LE) if the inheritances are independent to each other. One of the applications resulting of LD in whole-genome SNP association studies is the SNP tagging. An SNP is selected as a tag SNP to represent a region of the genome if it is in high LD with the group of SNPs in that region. Although the SNP data from large GWASs are already using tag SNPs to sample the genetic variants, it is inevitable to have SNPs at strong LD. If some of genuinely causal SNPs are correlated with nearby SNPs due to the LD effect, most popular association detecting tools are incapable to tell whether the significance is caused by the interaction or the LD effect. In this chapter, we extend the DAM method to model the block structures (referred as LD-block [79]) and capture the significant associations within and between the inferred blocks. We name the novel method as BAM, Block-based detection of genome-wide Association on Multiple diseases. Experimental results on simulated data sets demonstrate that BAM is capable to recover perfectly the LD-blocks and complicated embedded associations. We also applied BAM on two real WTCCC data sets, and some novel and interesting findings were reported. The chapter is organized as follows. In Section 5.2, we present the extended Bayesian variable partition model with the incorporation of the LD-block model. We design a two-level MCMC updating scheme and a stepwise interaction evaluation in Section 5.2.5. In Section 5.3, we demonstrate the superior performance of BAM by simulation studies comparing to two other approaches, DAM and SAM. We also report our results by applying BAM to the RA and

T1D data from WTCCC. We conclude the article in Section 5.4.

5.2 BAM: A *B*lock-Based Method of Genome-wide *A*ssociation on *M*ultiple Diseases

5.2.1 Notations

We follow most of the notations introduced in Section 4.2. Let B denote the block set with $|B|$ blocks, and B_i be the ith block with $|B_i|$ SNPs. Let $D_{(B_i)}$ be the genotypes of SNPs of ith block.

5.2.2 Assumptions

To embed the LD effects in identifying genome-wide associations, we make the following assumptions:

- Each SNP only belongs to one and only one LD-block.
- The epistatic interaction may contain SNPs from more than one LD-blocks.
- At most one SNP in an LD-block for a type $\overline{T}_k \in \overline{T}$.

5.2.3 LD-block Model

We assume that the genotypes, $D_{(\mathbb{B}_i)}$, follows a multinomial distribution with the frequency parameters $\Theta = (\theta_1, \ldots, \theta_g, \ldots, \theta_{3^{|\mathbb{B}_i|}})$, where θ_g is frequency parameter for a particular genotype or genotype combination. There are up to $3^{|\mathbb{B}_i|}$ possible genotypes or genotype combinations. Assuming a Dirichlet prior distribution on Θ, we obtain the likelihood of observations in the block as follows

$$P(D_{(\mathbb{B}_i)}|\Theta) = \prod_{g=1}^{3^{|\mathbb{B}_i|}} \theta_g^{n_g} \tag{5.1}$$

$$\tag{5.2}$$

where n_g denotes the combined count of the genotype combination g observed in the data, and

$$P(\Theta) = \frac{\Gamma(\alpha\circ)}{\prod_{g=1}^{3^{|\mathbb{B}_i|}} \Gamma(\alpha_g)} \prod_{g=1}^{3^{|\mathbb{B}_i|}} \theta_g^{\alpha_g - 1} \tag{5.3}$$

Here, $\alpha\circ$ denotes the sum of values over all elements in α. By integrating out $\{\theta_g\}$, we can obtain the marginal probability of the data in the block as

$$\begin{aligned} P\left(D_{(\mathbb{B}_i)}|\mathbb{B}_i\right) &= \int_\Theta P\left(D_{(\mathbb{B}_i)}|\mathbb{B}_i, \Theta\right) P\left(\Theta\right) d\Theta \\ &= \frac{\Gamma(\alpha\circ)}{\prod_{g=1}^{|\mathbb{B}_i|} \Gamma(\alpha_g)} \int_\Theta \prod_{g=1}^{|\mathbb{B}_i|} \theta^{n_g+\alpha_g-1} d\Theta \\ &= \left(\prod_{g=1}^{|\mathbb{B}_i|} \frac{\Gamma(n_g + \alpha_g)}{\Gamma(\alpha_g)} \right) \frac{\Gamma(\alpha\circ)}{\Gamma(N + \alpha\circ)} \end{aligned} \tag{5.4}$$

By default, we set $\alpha = \frac{\varphi}{3^{|\mathbb{B}_i|}}$ for a genotype combination g, and let $\varphi = 1.5$.

5.2.4 Bayesian Variable Partition Model with LD-block Model

To incorporate epistatic interactions and LD-block, we use two indicator vectors, I^T and $I^{\mathbb{B}}$, where I^T represent the type of the M SNP markers, and $I^{\mathbb{B}}$ denotes the block partition of the M SNPs. In DAM, SNPs of type $\widetilde{T}$ are independent to each other, if they belong to the same block, they become dependent of each other in the joint model of Bayesian partition model and LD-block model. To illustrate our model, supposing there are only two types, $\widetilde{T}_1$ and $\widetilde{T}_2$, of SNPs in a block, we can obtain the joint probability of this block as

$$P(D_{(\mathbb{B}_i)}) = P(D_{(\mathbb{B}_i)}^{(\widetilde{T}_1)}|D_{(\mathbb{B}_i)}^{(\widetilde{T}_2)})P(D_{(\mathbb{B}_i)}^{(\widetilde{T}_2)}) \tag{5.5}$$

Thus, SNPs of type $\widetilde{T}_0$ and $\widetilde{T}_1$ are no longer mutually independent to each other as assumed in DAM, but are related to each other because of the LD-block structure. To include all the types in $\widetilde{T}$, we introduce a chain rule to put conditional independence on the SNPs of $\widetilde{T}$ based on the order of types. We revise formula 5.5 to take all the types of $\widetilde{T}$ into consideration:

$$P(D_{(\mathbb{B}_i)}) \;=\; \left(\prod_{k=1}^{|\widetilde{T}|-1} P\left(D_{(\mathbb{B}_i)}^{(\widetilde{T}_k)} | D_{(\mathbb{B}_i)}^{(\widetilde{T}_{k\star|\widetilde{T}|})}\right)\right) P(D_{(\mathbb{B}_i)}^{(\widetilde{T}_{|\widetilde{T}|})}) \tag{5.6}$$

where $\widetilde{T}_{k\star|\widetilde{T}|}$ denotes a set of types with subscripts in $\left[k+1, |\widetilde{T}|\right]$.

SNPs of types $\overline{T}$ are assumed to be dependently associated with the disease(s), then we assume SNPs of types $\widetilde{T}$ to be conditional on the genotypes or genotype combinations of the SNPs of types $\widetilde{T}$ when they are all in the same block. Similar to the chain rule of SNPs of $\widetilde{T}$, we also assume a conditional order for SNPs of $\overline{T}$ if there are multiple SNPs of types $\overline{T}$ within the block. We further revise formula 5.6 as

$$\begin{aligned} P(D_{(\mathbb{B}_i)}) \;=\; & \left(\prod_{k=1}^{|\widetilde{T}|-1} P\left(D_{(\mathbb{B}_i)}^{(\widetilde{T}_k)} | D_{(\mathbb{B}_i)}^{(\widetilde{T}_{k\star|\widetilde{T}|} \cup \overline{T})}\right)\right) P\left(D_{(\mathbb{B}_i)}^{(\widetilde{T}_{|\widetilde{T}|})} | D_{(\mathbb{B}_i)}^{(\overline{T})}\right) \\ \times\; & \left(\prod_{k=1}^{|\overline{T}|-1} P\left(D_{(\mathbb{B}_i)}^{(\overline{T}_k)} | D_{(\mathbb{B}_i)}^{(\overline{T}_{k\star|\overline{T}|})}\right)\right) P\left(D_{(\mathbb{B}_i)}^{(\overline{T}_{|\overline{T}|})}\right) \end{aligned} \tag{5.7}$$

where $P\left(D_{(\mathbb{B}_i)}^{(T_k)} | D_{(\mathbb{B}_i)}^{(T_{k'})}\right)$ and $P\left(D_{(\mathbb{B}_i)}^{(T_{k''})}\right)$ are specified in formula 5.5.

Note that when there are multiple SNPs of type $\overline{T}_k$ which are not belonging to the same block, the conditional probability only considers those SNPs within the block, so $D_{(\mathbb{B}_i)}^{(\overline{T}_k)}$ indicates the genotypes of SNPs within the ith block and having the type $\overline{T}_k$. Also, for simplicity, we omit the partitions of observations based on the types of SNPs in the above formulas. More details about the partitions introduced by T can be find in Section 4.2.1.

To complete the type assignments and block structures determined by I^T and $I^{\mathbb{B}}$ in formula 5.7, we express the joint probability of the whole data as

$$P\left(D|I^T,I^{\mathbb{B}}\right) \;=\; \prod_{k=1}^{|\overline{T}|} P\left(D^{(\overline{T}_k)}|I^T\right) \times \prod_{i=1}^{|\mathbb{B}|} P\left(D_{(\mathbb{B}_i)}|I^T,I^{\mathbb{B}}\right) \tag{5.8}$$

We set the prior distribution of the block variable $I^{\mathbb{B}}$ as the product of independent Bernoulli probabilities $P(I^{\mathbb{B}}) = p^{|\mathbb{B}|}(1-p)^{M-|\mathbb{B}|}$. We further assume a priori that there are 50,000 blocks in the human genome, and thus we set $p = \min(0.5, 50000R/(3\times 10^9 M))$. Here, R denotes the length of the region spanned by the M SNP markers, and 3×10^9 is the length of the human genome. Finally, the Bayesian variable partition model with LD-model is write as

$$P\left(D,I^T,I^{\mathbb{B}}\right) = P\left(D|I^T,I^{\mathbb{B}}\right)P\left(I^T\right)P\left(I^{\mathbb{B}}\right) \tag{5.9}$$

where the conditional distribution, $P\left(D|I^T,I^{\mathbb{B}}\right)P\left(I^T\right)$ is showed in formula 5.8.

5.2.5 MCMC Updates and Follow-up Tests

The parameters of interest in our model are the block partition $I^{\mathbb{B}}$ and the assignments of types of SNPs I^T. We develop an MH algorithm to update $I^{\mathbb{B}}$, and, simultaneously, we develop a mix of a Gibbs sampler and an MH algorithm to update I^T.

We propose the three MH updates: (1) randomly select a block and split it into two new blocks at a random position, (2) randomly select two adjacent blocks to merge into a new block, and (3) randomly select two adjacent blocks and shift their common boundary by random positions. The update is accepted based on the MH ratio, which is a Gamma function.

To update the type assignments, we use the same MCMC sampling scheme in Section 4.2.3. In each iteration, a Gibbs sampler is used to update the types of SNPs inside each block first, and then the MCMC sampling scheme is used to update the whole SNP set. We set a minimum threshold on the posterior probabilities of I^T) and feed those SNPs larger than the threshold to the stepwise interaction evaluation process introduced in Section 4.4.

5.3 Experiment Results

To show the performance of BAM with consideration of LD effects, we first introduce the simulation using HapMap data [82]. Following that we show the null simulation to test type I errors of BAM. With the simulated data in which we embed four types of disease-specific associations, we compared BAM to DAM and SAM. We also applied BAM to two real WTCCC data sets, RA and T1D, in which BAM produced accurate block partitions that match well to the visual blocks displayed by Haploview [83].

5.3.1 Experimental Design

Data simulation To evaluate the effectiveness of BAM, we perform extensive simulation experiments using four disease models with two-locus associations on three groups. The genotypes of unassociated SNP are generated by the same procedure used in previous studies [45] with Minor Allele Frequencies (MAFs) sampled from $[0.05, 0.5]$. The odds tables for four models are identical to the first four models in Section 3.3.1. In a setting of an epistatic interaction, we use the MAF ranged in $\{0.1, 0.2, 0.4\}$, we generate 100 replicas per setting per epistatic interaction. Therefore, by given an MAF, a main and a secondary model, we can have four associations labelled as Ep 5 to 8. Note that there are only three epistatic interactions in Model 5, because the combinations of Model 5 with other models do not have a mathematical solution when MAF = 0.1. Each simulated replica contained $M = 1000$ SNPs. The sizes of three groups are set to $(800, 800, 800)$ or $(1600, 1600, 1600)$, where we consider the first two groups as the case groups and the third one as the control group. There are totally 92 settings.

To mimic real genetic data in human populations, we select an arbitrary region in Chromosome 22. Given a disease model setting, we first sample the genotype for the disease SNPs, and then randomly sample 2400 or 4800 individuals from a pool of controls generate by HAPGEN [84] using CBH sample with odds ratio = 1. We keep the positions of these

SNPs in Chromosome 22. For these data, we also applied a quality control procedure that we remove those SNPs with $MAF < 0.05$ or p-values from Hardy-Weinberg Equilibrium (HWE) tests less than 0.001.

Statistical power In the evaluation of performances on simulated data, 100 datasets are generated for each setting. We define the measurement of discrimination power as the fraction of 100 datasets on which the ground-truth associations are identified as compact and significant by the methods.

5.3.2 Null Simulation to Test Type I Errors

To display type I errors of our method, we conduct the null simulation experiments. We simulate 1000 replica datasets without disease association embedded. The false positive rate of BAM is showed in Figure 5.1, in which we can see that we obtained much less false positive rate under all different group sizes and numbers of SNPs by setting the significance level to 0.1. It shows that the Bonferroni correction is conservative to adjust the p-value. Under most of the situations, BAM did not report any significant interactions. By looking at the output candidate SNP lists of MCMC sampling, we found that there is no significant SNP with the posterior probability larger than the posterior probability threshold, 0.5.

5.3.3 Simulation Experiment from Four Epistasis Models with LD

To verify the effectiveness of BAM, we use the simulation strategy introduced in Section 5.3.1, where the block structure due to LD is from a real human genome. Although the detailed boundary information is unavailable, the block structure widely exists across the entire human genome. We took the SNPs from Chromosome 22 to generate the simulated data for this section. The performance comparison between BAM, DAM, and SAM is illustrated in Figure 5.2. We use the same concept of overall quality to evaluate the overall performance of these three methods on the simulated datasets. The overall qualities for BAM, SAM, and DAM are 0.564, 0.283, and 0.119 for the group size of 2400 individuals, and 0.742, 0.663, and 0.321 for the group sizes of 4800 individuals, respectively. The improvement shows that

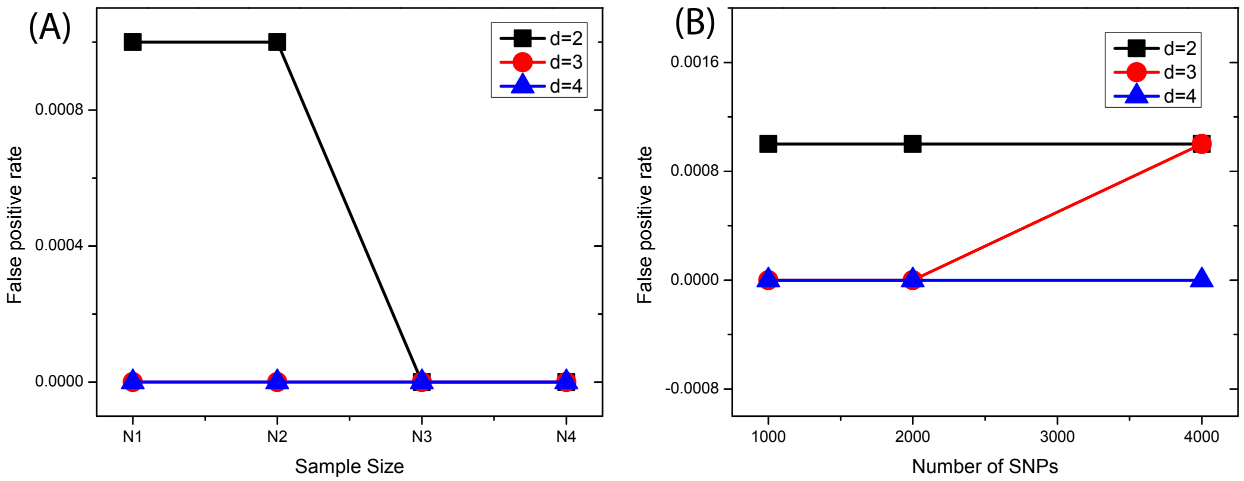

Figure 5.1 False positive rates of BAM under null models. The plots in (A) and (B) show the false positive rates of BAM for different *d*s when sample sizes and the numbers of SNPs vary.

by incorporating the LD-Model into DAM, BAM is greatly improving the discrimination power of those false interaction SNPs caused by LD effects. Here, we also found that SAM performed better than DAM when block structures existed. We examined the candidate SNP lists generated by SAM and DAM and found that DAM cannot pick up correctly the disease-associated SNPs in the block. In the contrast, SAM uses the clustering algorithm to group the SNPs when they show limited dissimilarities between data groups, which cause causal SNPs and their nearby SNPs placed into different clusters. With the ranking of SNPs, the disease-associated SNPs gain higher ranking scores than their adjacent SNPs, which give those causal SNP factors changes to go to the stepwise interaction evaluation.

5.3.4 Experiments on The WTCCC Data

We also applied BAM to two real GWAS data, RA and T1D, especially in the region of Chromosome 6. The block structures due to LD effects have many different definitions. Therefore, to give an intuitive idea of the performance of the block structure yielded by BAM, we randomly select four regions with lengths about 200kb and use HapMap to draw the corresponding Haploviews of these areas. Figure 5.3 shows the recovered block structures by BAM. We set the number of iterations to 300 times with extra 300 times in the burn-in process. Comparing to the visual blocks displayed by Haploview, BAM produced accurate block partitions for these four regions.

5.4 Conclusion

In this chapter, we proposed a block-based Bayesian method for LD-block inference and detection of high-order epistatic interaction on multiple diseases. Extensive experimental results on simulated datasets indicate that BAM can identify most embedded disease associations even with block structures mimicked from real human genomes. By comparing to the methods, SAM and DAM introduced in Chapter 4, BAM substantially improve the statistic power of DAM by incorporating the LD model to Bayesian variable partition model. The results from experiments in Chromosome 6 show that BAM is capable to recover the block structures accurately.

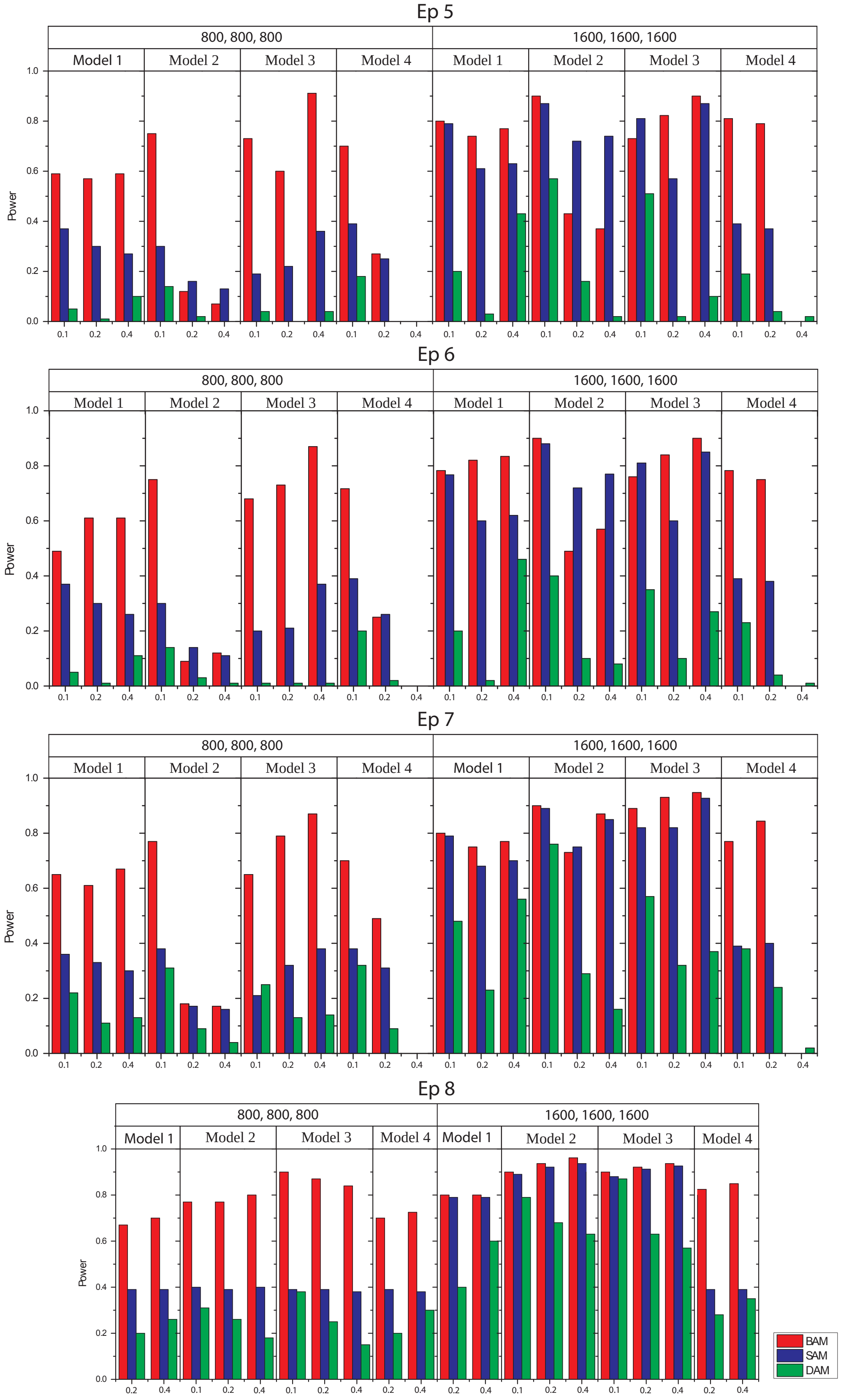

Figure 5.2 Performance comparison between BAM, SAM, and DAM after the same stepwise interaction evaluation test on the dataset with 4 simulated disease associations. The X-axis is the MAF value.

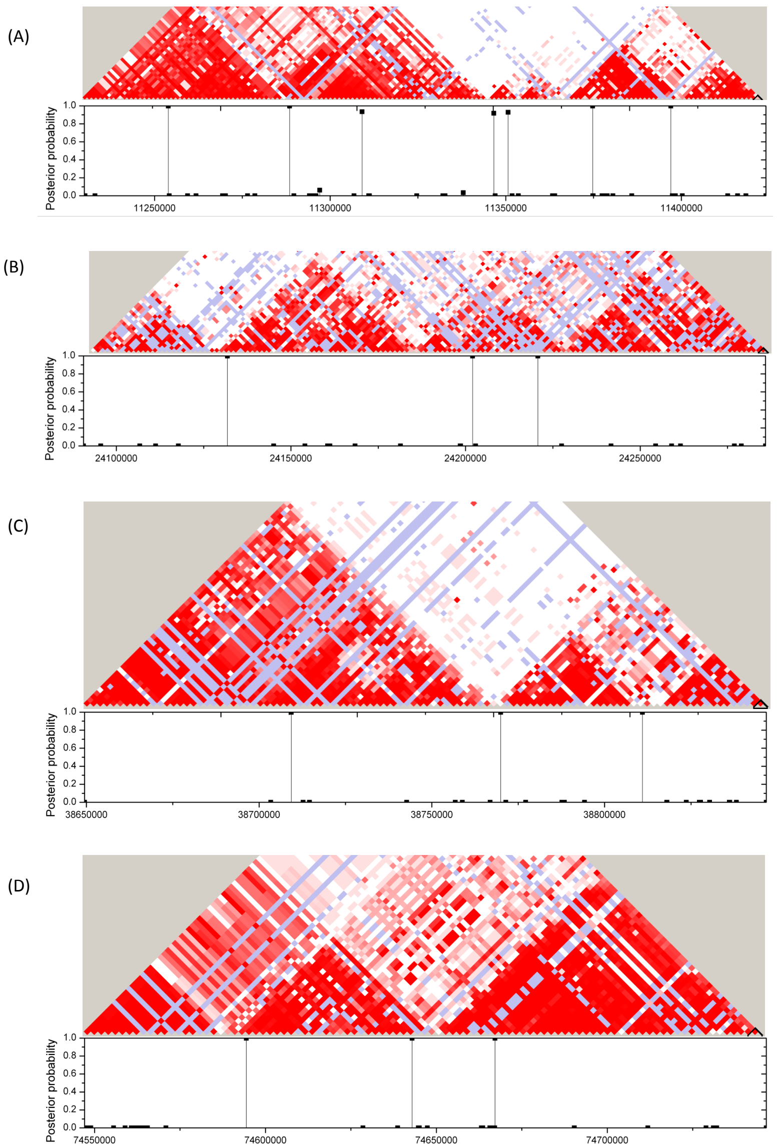

Figure 5.3 Four block structures recovered by BAM in Chromosome 6. The X-axis in the bottom half figure is the physical locations of the SNPs, and the Y-axis is the posterior probabilities of SNPs. The top half figure is the Haploview of the corresponding areas spanned by the SNPs.

PART 6

DISCUSSION AND FUTURE WORK

The enormous amount of SNPs genotyped in genome-wide association studies poses a significant computational challenge in the identification of gene-gene interactions. During the last few years, there have been fast-growing interests in developing and applying computational and statistical approaches to finding gene-gene interactions. In this work, we presented four methods to address a series of problems in genome-wide association studies in three scenarios. We first proposed a simple and fast method, named DCHE, for detection of genome-wide multi-locus epistatic interactions using one disease status. With the observations of diverse effects of epistasis on multiple diseases, we designed two novel methods, named DAM and SAM, for the search of genome-wide multi-locus epistatic interactions using multiple disease cases. DAM uses a Bayesian variable partition model to capture the disease-specific associations. SAM uses a Jenson-Shannon Divergence and K-mean clustering to reduce the searching space of SNPs. A stepwise interaction evaluation process based on χ^2 test and conditional χ^2 test was applied to the candidate list of SNPs generate by both SAM and DAM. For the last scenario, combining the LD-model and Bayesian variable partition model, we developed a novel Bayesian inference method, named BAM, with the capability of modelling the block structures in the human genome and identify multi-locus epistatic interactions simultaneously. Systematic simulated data were generated to evaluate the performance of all these methods. By comparing to other popular tools, our methods show significant improvement in terms of the discrimination power. Experimental results on real data show that our four methods can find some novel biologically meaningful associations and block structures.

In our future work, we will concentrate on the following five research directions:

- Parallelize our Jensen-Shannon Divergence based method, SAM and make it be able to be deployed on some popular cloud platforms, for example, Amazon EC2, Windows Azure.

- Add block structure detection ability to SAM. One possible way is to use clustering algorithms find the optimal number of blocks merely consisting with adjacent SNPs.

- Parallelize our block-based Bayesian method, BAM, to make it fast for handling entire human genome-scale data.

- Apply BAM to the GWASs with the whole human genomes, to verify its performance and try to find some statistically and biologically significant disease associations.

- Use well-defined block structure data by other block detecting tools to verify the recovery accuracy of BAM.

In addition to the above future topics, 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 3D conformation of protein structure or their corresponding active domains. Consequently, they can potentially disrupt the recruitment scaffold of the similar protein. Besides, the epigenome consists of a record of histone modifications and DNA methylation of an organism. The existing evidence shows that there are some correlations between SNPs and the quantitative traits of DNA methylation. 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.

REFERENCES

[1] G. P. Consortium *et al.*, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

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

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

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] W. Bateson, *Mendel’s principles of heredity.* Cosimo, Inc., 2007.

[11] 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.

[12] 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.

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

[14] J. Marchini1, P. Donnelly1, and L. R. Cardon, “Genome-wide strategies for detecting multiple loci that influence complex diseases,” *Nature Genetics*, vol. 37, pp. 413 – 417, 2005.

[15] 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 estrogen-metabolism genes in sporadic breast cancer,” *The American Journal of Human Genetics*, vol. 69, pp. 138–147, July 2001.

[16] 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.

[17] H. J. Cordell, “Detecting gene-gene interactions that underlie human diseases,” *Nat Rev Genet*, vol. 10, pp. 392–404, 06 2009.

[18] 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.

[19] X. Zhang, S. Huang, F. Zou, and W. Wang, "Team: efficient two-locus epistasis tests in human genome-wide association study," *Bioinformatics*.

[20] 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.

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

[22] D. Anastassiou, "Computational analysis of the synergy among multiple interacting genes," *Molecular systems biology*, vol. 3, no. 1, 2007.

[23] 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.

[24] 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.

[25] R. A. Fisher, "On the interpretation of $\chi 2$ from contingency tables, and the calculation of p," *Journal of the Royal Statistical Society*, pp. 87–94, 1922.

[26] L. S. Yung, C. Yang, X. Wan, and W. Yu, "Gboost: a gpu-based tool for detecting genegene interactions in genomewide case control studies," *Bioinformatics*, vol. 27, no. 9, pp. 1309–1310, 2011.

[27] Y. Liu, H. Xu, S. Chen, X. Chen, Z. Zhang, Z. Zhu, X. Qin, L. Hu, J. Zhu, G.-P. Zhao, and X. Kong, "Genome-wide interaction-based association analysis identified multiple

new susceptibility loci for common diseases," *PLoS Genet*, vol. 7, no. 3, p. e1001338, 03 2011.

[28] X. Jiang, M. M. Barmada, and S. Visweswaran, "Identifying genetic interactions in genome-wide data using bayesian networks," *Genetic epidemiology*, vol. 34, no. 6, pp. 575–581, 2010.

[29] C. Herold, M. Steffens, F. F. Brockschmidt, M. P. Baur, and T. Becker, "Inter-snp: genome-wide interaction analysis guided by a priori information," *Bioinformatics*, vol. 25, no. 24, pp. 3275–3281, 2009.

[30] 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.

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

[32] B. Goudey, D. Rawlinson, Q. Wang, F. Shi, H. Ferra, R. M. Campbell, L. Stern, M. T. 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.

[33] J. Piriyapongsa, C. Ngamphiw, A. Intarapanich, S. Kulawonganunchai, A. Assawamakin, C. Bootchai, P. J. 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.

[34] X. Zhang, F. Pan, Y. Xie, F. Zou, and W. Wang, "Coe: a general approach for efficient genome-wide two-locus epistasis test in disease association study," in *Research in Computational Molecular Biology.* Springer, 2009, pp. 253–269.

[35] X. Zhang, F. Zou, and W. Wang, "Fastchi: an efficient algorithm for analyzing gene-gene interactions," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing.* NIH Public Access, 2009, p. 528.

[36] X. Zhan, 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.* ACM, 2008, pp. 821–829.

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

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

[39] 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*, J. Huang, L. Cao, and J. Srivastava, Eds. New York: Springer Berlin Heidelberg, 2011, vol. 6634, pp. 338–350.

[40] D. F. Schwarz, I. R. König, and A. Ziegler, "On safari to random jungle: a fast implementation of random forests for high-dimensional data," *Bioinformatics*, vol. 26, no. 14, pp. 1752–1758, 2010.

[41] 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.

[42] 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. 19 199–19 203, 2007.

[43] 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.

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

[45] 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.

[46] 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.

[47] Q. Long, Q. Zhang, and J. Ott, "Detecting disease-associated genotype patterns," *BMc bioinformatics*, vol. 10, no. Suppl 1, p. S75, 2009.

[48] 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, p. 196, 2006.

[49] 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.

[50] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

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

[52] 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, 05 2009.

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

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

[55] C. Kooperberg and I. Ruczinski, "Identifying interacting snps using monte carlo logic regression," *Genetic epidemiology*, vol. 28, no. 2, pp. 157–170, 2005.

[56] 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.

[57] A. A. Motsinger, S. L. Lee, G. Mellick, and M. D. 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.

[58] N. R. Cook, R. Y. 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.

[59] 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.

[60] 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.

[61] 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, 04 2012.

[62] 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.

[63] 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.

[64] J. R. Kilpatrick, "Methods for detecting multi-locus genotype-phenotype association," Ph.D. dissertation, RICE UNIVERSITY, 2009.

[65] X. Wan, C. Yang, Q. Yang, H. Xue, N. L. 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.

[66] C. Aporntewan, D. H. Ballard, J. Y. Lee, J. S. Lee, Z. Wu, and H. Zhao, "Gene hunting of the genetic analysis workshop 16 rheumatoid arthritis data using rough set theory," in *BMC proceedings*, vol. 3, no. Suppl 7. BioMed Central Ltd, 2009, p. S126.

[67] J. Hoh, A. Wille, and J. Ott, "Trimming, weighting, and grouping snps in human case-control association studies," *Genome research*, vol. 11, no. 12, pp. 2115–2119, 2001.

[68] X. Guo, X. Ding, Y. Meng, and Y. Pan, "Cloud computing for de novo metagenomic sequence assembly," in *Bioinformatics Research and Applications*, ser. Lecture Notes in Computer Science, Z. Cai, O. Eulenstein, D. Janies, and D. Schwartz, Eds. New York: Springer Berlin Heidelberg, 2013, vol. 7875, pp. 185–198.

[69] S. Kullback, *Information theory and statistics.* Courier Corporation, 1997.

[70] 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.

[71] S. Oh, J. Lee, M.-S. Kwon, B. Weir, K. Ha, and T. Park, “A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based mdr,” *BMC Bioinformatics*, vol. 13, no. Suppl 9, p. S5, 2012.

[72] 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.

[73] 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, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh, “Complement factor h polymorphism in age-related macular degeneration,” *Science*, vol. 308, no. 5720, pp. 385–389, 2005.

[74] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, “Toppgene suite for gene list enrichment analysis and candidate gene prioritization,” *Nucleic Acids Research*, vol. 37, no. suppl 2, pp. W305–W311, 2009.

[75] J. S. Liu, *Monte Carlo strategies in scientific computing.* springer, 2008.

[76] J. Lin, “Divergence measures based on the shannon entropy,” *Information Theory, IEEE Transactions on*, vol. 37, no. 1, pp. 145–151, 1991.

[77] R. Lechler and A. N. Warrens, *HLA in Health and Disease.* Academic Press, 2000.

[78] J. Zhang, Z. Wu, C. Gao, and M. Q. Zhang, “High-order interactions in rheumatoid arthritis detected by bayesian method using genome-wide association studies data,” *American Medical Journal*, vol. 3, no. 1, pp. 56–66, 2012.

[79] Y. Zhang, J. Zhang, and J. S. Liu, “Block-based bayesian epistasis association mapping with application to wtccc type 1 diabetes data,” *The annals of applied statistics*, vol. 5, no. 3, p. 2052, 2011.

[80] H.-S. Lee, B. D. Korman, J. M. Le, D. L. Kastner, E. F. Remmers, P. K. Gregersen, and S.-C. Bae, “Genetic risk factors for rheumatoid arthritis differ in caucasian and korean populations,” *Arthritis & Rheumatism*, vol. 60, no. 2, pp. 364–371, 2009.

[81] Q. Wang, C. Yang, J. Gelernter, and H. Zhao, “Mediated pleiotropy between psychiatric disorders and autoimmune disorders revealed by integrative analysis of multiple gwas,” *bioRxiv*, p. 014530, 2015.

[82] I. H. Consortium *et al.*, “A haplotype map of the human genome,” *Nature*, vol. 437, no. 7063, pp. 1299–1320, 2005.

[83] P. I. de Bakker, R. Yelensky, I. Pe’er, S. B. Gabriel, M. J. Daly, and D. Altshuler, “Efficiency and power in genetic association studies,” *Nature genetics*, vol. 37, no. 11, pp. 1217–1223, 2005.

[84] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, “A new multipoint method for genome-wide association studies by imputation of genotypes,” *Nature genetics*, vol. 39, no. 7, pp. 906–913, 2007.