IDENTIFYING GENE-GENE INTERACTIONS ASSOCIATED WITH COMPLEX DISEASES AND COMPLEX TRAITS

ZHOU XIANGDONG

PhD

The Hong Kong Polytechnic University

2018

The Hong Kong Polytechnic University

Department of Computing

Identifying Gene-gene Interactions Associated with Complex Diseases and Complex Traits

Zhou Xiangdong

A thesis submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

March 2018

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

Zhou Xiangdong (Name of student)

Abstract

Diseases are usually associated with genetic variants, mainly single nucleotide polymorphisms (SNPs) or Single Sequence Repeat Polymorphisms (SSRPs). Therefore it's an important task for researchers in human genetics to search for genetic factors having influence on diseases as it can be used in many medical case-control studies. In recent years, this research has been greatly improved by using genome-wide association studies (GWASs) which use a single-locus approach, where each variant is tested individually for association with a specific disease. However most complex diseases are considered to be the results of gene-gene and gene-environment interactions. many computational methods have been proposed to detect if a particular set of genes has epistatic interaction with a particular complex disease.

However, even though many such methods have been shown to be useful, they can be made more effective if the properties of gene-gene interactions can be better understood. Towards this goal, we have attempted to uncover patterns in gene-gene interaction and the patterns reveal an interesting property that can be reflected in an inequality that describes the relationship between two genotype variables that takes on the genotypes of two different genes as values, and a disease-status variable that takes on binary values representing the presence or absence of a complex disease. We show that this inequality can be derived for generalization to *n* genotype variables. Based on this inequality, we establish a conditional independence and redundancy (CIR) based definition of gene-gene interaction and the concept of an interaction group. We discuss the properties of these concepts and explain how they can be used in a novel algorithm that can be used to detect gene-gene interaction with an order of two and above greater than two. Experimental results using both simulated and real datasets show that the proposed algorithm can be very promising. Possible ways to further improve the effectiveness of the new algorithm are also provided.

Like complex diseases, complex quantitative traits (QTs) are also usually

associated with genetic variants. The majority of innate and acquired body and behavioral characteristics. Many physiological characteristics are also reflected by complex traits. In addition, most diseases exhibit various symptoms through complex traits.

The Multifactor Dimensionality Reduction (MDR) method was originally proposed as a nonparametric and model-free data reduction approach for identifying interactions without significant main effects and has been successfully applied to identify gene-gene interactions in many common complex diseases. Some efforts have been made to extend MDR to QTs.

However these methods are still not computationally efficient or effective. Therefore we propose Extended Fuzzy Quantitative trait MDR (EFQMDR) to strengthen identification of gene-gene interactions associated with a quantitative trait by first transforming it to an ordinal trait and then using a balanced accuracy measure based on extended member functions of fuzzy sets to select multiple best sets of genetic markers as having strongest associations with the trait. Experimental results on simulated datasets and real datasets show that our algorithm has better performance in terms of test accuracy and consistency in identifying gene-gene interactions associated with QTs.

Multiple correlated phenotypes often appear in complex traits or complex di seases. These correlated phenotypes are useful in identifying gene-gene interact ions associated with complex traits or complex disease more effectively. Some approaches have been proposed to use correlation among multiple phenotypes to identify gene-gene interactions that are common to multiple phenotypes. H owever these approaches either didn’t find truly gene-gene interactions or the results are hard to explain, especially using all correlated phenotypes to identi fy gene-gene interactions make identified interactions unreliable. Multivariate Quantitative trait based Ordinal MDR (MQOMDR) algorithm is therefore prop osed to effectively identify gene-gene interactions associated with multiple corr elated phenotypes by selecting the best classifier according to not only the tra

ining accuracy of the phenotype under consideration but also other phenotypes with weights determined mainly by their pair correlation with the phenotype under consideration and also by repeated selection process to make use of tru ly useful correlated phenotypes . Experimental results on two real datasets sho w that our algorithm has better performance in identifying gene-gene interacti ons associated with multiple correlated phenotypes.

List of Publications

1. Zhou, X, and Keith C. C. Chan. "A new information-theoretic approac h to detect gene-gene interactions in case-control studies." *IEEE, Inter national Conference on Bioinformatics and Bioengineering* IEEE, 2015: 1-5.
2. Zhou, X, and Keith C. C. Chan. "An effective approach to identify ge ne-gene interactions for complex quantitative traits using generalized fu zzy accuracy." *Computational Intelligence in Bioinformatics and Compu tational Biology* IEEE, 2017:1-6.
3. Zhou, X, K. C. C. Chan, and Zhu, D.. "A multi-stage approach to det ect gene-gene interactions associated with multiple correlated phenotype s." *IEEE Conference on Computational Intelligence in Bioinformatics a nd Computational Biology* IEEE, 2017:1-8.
4. Zhou, X, and Keith C. C. Chan. “Detecting gene-gene interactions for complex quantitative traits using generalized fuzzy classification.” BM C *Bioinformatics*. (Second revision)
5. Zhou, X, and Keith C. C. Chan. “An extended fuzzy classification me thod for identifying gene-gene interactions associated with complex qu antitative traits.” *Fuzzy Sets and Systems*. (First revision)

Acknowledgements

I would like to express my deepest appreciation to my supervisor Professor Keith C. C. Chan, who has broad knowledge, deep thinking and exploring spirit in his research field and other related fields. His earnest guidance, helpful advice, continuous support and enthusiastic encouragement when I encounter difficulty give me great help in my research. Without his supervision and help this thesis would not have been completed.

I would like to thank to my friends Zhuhong You, Lun Hu, Bing Li, Peiyuan Zhou, Tiantian He and Penwei Hu. Their advice and discussion with me give me much help and many enlightments in my research and I also enjoy a good time with them in my campus life.

Thanks will also be given to my parents who greatly encourage and support me to do my PhD study after I had worked for so many years.

Table of contents

Abstract……I

List of Publications……..…V

Acknowledgements……VII

Table of Contents……..IX

List of Figures…….……...XIII

List of Tables…….……...XV

Chapter 1 Introduction……1

1.1 Background……..1

1.2 Problem Statements…….3

1.3 Overview of Solutions……6

Chapter 2 Background Knowledge and Related Work……...9

2.1 Background Knowledge…….9

2.2 Related Works……11

2.2.1 Traditional analytical methods used in studies of gene-gene interactions……..……..…. …… 11

2.2.2 Data-mining methods…….……..15

2.2.2.1 Multifactor dimensionality reduction method……...16

2.2.2.2 Ordinal MDR…….17

2.2.2.3 Random forest method……..…18

2.2.2.4 Pathway-based gene-set analyses……..19

2.2.3 Biological interpretation……21

Chapter 3 A Novel Approach for Identifying Epistatic Interactions for Complex Disease Prediction…….……..23

3.1 Introduction……23

3.2 Related works……24

3.3 Methodology……..29

3.3.1 New definitions of gene-gene interaction and an interaction group……29

3.3.2 Some Properties Related to the New Definitions…….33

3.3.3 Measurement of Gene-Gene Interaction……34

3.3.4 An Algorithm to Detect High Order Gene-Gene Interaction…37

3.4 Experimental results and analysis……..40

3.4.1. Experiments on Simulated Datasets…… …...40

3.4.1.1.Program gs 2.0…….40

3.4.1.2. Evaluation of the New Definition of Gene-Gene Interaction……42

3.4.1.3. Type I error…….43

3.4.1.4. Evaluation of the Proposed Algorithm……...44

3.4.2. Experiments on a Real Datasets……..48

3.5. Conclusion……..50

Chapter 4 An Extended Fuzzy Classification Method for Identifying Gene-Gene Interactions Associated with Complex

Quantitative Traits……52

4.1 Introduction…….52

4.2 Related works……...53

4.3 Methods…….62

4.3.1 Extended fuzzy classification using extended member functions

……62

4.3.2 EFQMDR Algorithm……..65

4.4 Experimental results and analysis……68

4.4.1 Experiments and analysis of results on simulated data……..68

4.4.1.1 Experimental setup……..68

4.4.1.2 Experimental Results……70

4.4.2 Experiments and analysis of results on real data……..74

4.4.2.1 Experimental setup…….74

4.4.2.2 Experimental results……75

4.5 Conclusion……...84

Chapter 5 A Multi-stage Approach to Detect Gene-gene Interactions Associated with Multiple Correlated Phenotypes……..87

5.1 Introduction……..87

5.2 Related works……...87

5.3 Methods…….….101

5.3.1. Deciding Weight of Correlated Phenotypes……..,.101

5.3.2. Filtering of Correlated Phenotypes……102

5.3.3. The MQOMDR Algorithm……..102

5.4 Experimental Results and Analysis……..……...104

5.4.1. Experimental setup……...104

5.4.2. Experimental results……106

5.5 Conclusion……...111

Chapter 6 Conclusions and Suggestions for Future Research……113

References…….116

List of Figures

Figure 2.1. Eight examples of two-locus models (Dong *et al.*, 2008).....…..13

Figure2.2. An example of a classification tree (Cordell, 2009).……..…..…….19

Figure 2.3. Sample pathway (Ogata *et al*., 2000).…….…….…….21

Figure 3.1. $|S_1 \cup S_2 \cup S_3|=|S_1|+|S_2|+|S_3|-|S_3 \cap (S_1 \cup S_2)|-|S_1 \cap S_2)|$..…….……31

Figure. 3.2**.** The graphical model representation of property 3.5……...37

Figure. 3.3**.** The graphical model representation of Theorem 3.1……..38

Figure 3.4. Hit ratios for the first three locus epistasis model. …….…….46

Figure 3.5. Hit ratios for the second three loci epistasis model. ……46

Figure 3.6. Hit ratios for different orders of epistasis models. ……..48

Figure 4.1. The linear membership functions of high(H), average(A) and low(L) levels of a QT. …….…….。…..63

Figure 4.2. The extended linear membership functions of high(H), average(A) and low(L) levels of a QT. ……..……65

Figure 4.3. Models of two way interactions for ordinal traits. White, light grey, dark grey represent normal, low risk, high risk of an ordinal trait respectively. ….68

Figure 4.4. Comparison of AMTSBCA1(average maximum testing balanced classification accuracy on 2-way, 3-way and 4-way interactions) and AMTSBCA2 (average AMTSBCA1 on all four QTs) among EFQOMDR, FQMDR, OMDR and MDR when k=1. …….…….……..77

Figure 4.5. Comparison of AMTSBCA1(average maximum testing balanced classification accuracy on 2-way, 3-way and 4-way interactions) and AMTSBCA2 (average AMTSBCA1 on all four QTs) among EFQOMDR, FQMDR, OMDR and MDR when k=5. …….83

Figure 5.1. Absolute values of correlation coefficient between 6 phenotypes in experiment 1 presented as graylevel values…….106

Figure 5.2. Comparison of average CVCs for three groups of phenotypes among MQOMDR, MDR, QMDR and Multi-QMDR for the DBA2╳NMRI8 dataset. …..108

Figure 5.3. Absolute values correlation coefficient between phenotypes in experiment 2 presented as gray level values……..……109
Figure 5.4. Comparison of average CVCs for three groups of phenotypes among MQOMDR, MDR, QMDR and Multi-QMDR for the DBA2╳DU6i dataset. Group 2 represents phenotypes: afw, mw and kidney for 2-way interactions. ……..…..111

List of Tables

Table 2.1. The Expected Distribution of Cases expressed by Gene Frequencies in the Population and Risks Associated with Gene for Gene- Gene Interaction Analysis in a Case-Only Design. (Yang *et al.*, 1999) …….15

Table 2.2. The distribution of cases by genotype combinations in a Case-Only Design. (Yang *et al.*, 1999)……..……..…….……..15

Table 3.1. CIR based algorithm. ……..……..……..…….39

Table 3.2. Penetrance table for two two-locus interaction models. ……....42

Table 3.3. Hit ratios for threshold model. ……43

Table 3.4. Hit ratios for exclusive or model. ……..…..……43

Table 3.5. Type I error rate with the significance level α of 0.05 from datasets with 1000 replicates. ……..……44

Table 3.6. Penetrance table for the first three-locus interaction model. ……....45

Table 3.7. Penetrance table for the second three-locus interaction model. ……..….45

Table 3.8. Average ratios for the first three locus epistasis model. ……...…47

Table 3.9. Average ratios for the second three locus epistasis model. …….……47

Table 3.10. Hospital admissions with malaria and severe malaria by hemoglobin type and a+- thalassemia genotype. ……..49

Table 4.1. Hit ratios (%) for model 1. ……..……..….70

Table 4.2. Hit ratios (%) for model 2. …….…....70

Table 4.3. Hit ratios (%) for model 3. …….……71

Table 4.4. Hit ratios (%) for model 4. ..……..……..72

Table 4.5. Hit ratios (%) for model 5. ……..…...72

Table 4.6. Type I Error Rate with the Significance Level α of 0.05 from Datasets with 1000 1000 Replicates. ……..74

Table 4.7. Comparison of MTSBCA and GCVC among EFQMDR, FQMDR, OMDR and MDR when k=1. ……76

Table 4.8. Comparison of MTSBCA and GCVC of PI classifiers among EFQMDR,

FQMDR, OMDR and MDR when k=5. (a) For two loci. (b) For three loci. (c) For four loci. ……...77
Table 4.9. Comparison of MTSBCA and GCVC of bw classifiers among EFQMDR, FQMDR, OMDR and MDR when k=5. (a) For two loci. (b) For three loci. (c) For four loci. ……79
Table 4.10. Comparison of MTSBCA and GCVC of afw classifiers among EFQMDR, FQMDR, OMDR and MDR when k=5. (a) For two loci. (b) For three loci. (c) For four loci. ……80
Table 4.11. Comparison of MTSBCA and GCVC of mw classifiers among EFQMDR, FQMDR, OMDR and MDR when k=5. (a) For two loci. (b) For three loci. (c) For four loci. ……..……..…….82
Table 5.1. Best set of snps and corresponding cvc of the dba2╳nmri8 dataset for mqomdr, mdr, qmdr and multi-qmdr. ……..……..……107
Table 5.2. Best set of snps and corresponding cvc of the dba2╳du6i dataset for mqomdr, mdr, qmdr and multi-qmdr. ……..110

Chapter 1 Introduction

1.1 Background

Diseases are usually associated with genetic variants, mainly single nucleotide polymorphisms (SNPs).The appearance of high-throughput genotyping technology made it possible and easy to scan whole-genome single-nucleotide polymorphisms (SNPs) for genes associated with diseases. As a result, doctors can utilize genetic data to analyze the mechanisms of diseases and customize medical treatment. Searching for genetic factors having influence on complex diseases and complex traits becomes an important and challenging for modern geneticists.

In recent years, this research has been greatly improved by using genome-wide association studies (GWASs) to detect associations of SNPs with many diseases (WTCC Consortium, 2007). A single-locus approach, where each variant is tested individually for association with a specific phenotype is used by most of these studies. However confirmed associations account for a small part of the heredity of complex traits and complex diseases (Franke . *et al.*, 2009). Most complex diseases are considered to be influenced by gene-gene and gene-environment interactions (Manolio *et al.*, 2009). For example, two conditions of the hemoglobinopathies were previously found to be protective against malaria. One is structural variant hemoglobin S: heterozygote HbAS (homozygote HbSS is not considered since it can lead to premature death) and the other is lack of the normal α-globin component of hemoglobin，α^{+}-thalassemia, which is caused by two variants: heterozygote -α/αα and homozygote -α/-α. However, in a malaria cohort study performed by Williams *et al* in Kenya (Williams *et al*, 2005), it was found that when two conditions were inherited together, the protection provided by each condition inherited alone disappeared. If a gene influences a disease mainly through interaction with other genes or environmental factors, the association might be missed if the gene is assessed individually without considering its interactions with other genes.

Except for a few cases such as replicating a previous study or testing a specific biological hypothesis, researchers are not satisfied with testing known interactions. They would more often search for possible unknown interactions at potentially many sites with given genotype data from a GWA study or from a local candidate gene study.

However when the methods for GWAS are extended to multiple loci for identifying gene-gene interactions associated with complex diseases and complex traits, they will have decreased statistical power and be computationally costful due to high dimensionality and small sample size. For example, if we want to consider all combinations of two SNPs to identify second order gene-gene interactions for 10 thousand SNPs in a genome, we will examine nearly 50 million possibilities.

A variety of methods including principal components analysis, information gain and multifactor dimensionality reduction have been proposed to make complexity algorithms tractable by reducing dimensionality.

Pathway and gene set methods use a set of genes having functional relation to jointly identify their association with a disease or a trait. These methods also have the advantage to identify genetic variants that individually have little association with a disease but collectively have a significant association and would be missed if they are individually tested in GWAS.

Systems biology and network approaches make use of external biological knowledge from genome, transcriptome, metabolome, proteome or functional and regulatory networks (Kohl *et al.*, 2010) to decide which genes or combinations of genes are more likely to have association, therefore greatly reduce the number of SNPs to be searched. A disease or a phenotype having response to a drug can be viewed as a perturbation of networks from their stable state (Auffray *et al.*, 2009). For example, in one research, chemical similarity metrics, pharmaco- genomic interactions and PPI were integrated to predict pharmacogenes (Hansen *et al.*, 2009), in another research, similarity of drug ligand sets were used to predict ‘off-target’ interactions (Keiser *et al.*, 2007).

The research achievements of association study can be applied in clinical practice

to improve medical care. Traditionally, drugs are developed to be applied in medical care without considering individual situation. Genetic variation is an important factor to consider in drug selection, dosing and adverse events (Giacomini *et al.*, 2007) which was showed evidence for by many examples of drugs such as thiopurines for cancer (Weinshilboum, 2001) and the anticoagulant clopiogrel (Shuldiner *et al.*, 2009). Drug development would benefit in therapy from a genetically tailored approach. For example, a hypothetical clinical application of the anticoagulant warfarin driven pharmacogenetically could reduce 40% of the cost and risk of adverse events (Ohashi and Tanaka, 2010).

More measurements need to be taken to make personalized medicine a routine approach for many physicians in their clinical practice. These measurements include popularizing personalized medicine to physicians, further proving the efficacy of drugs developed and prescribed pharmacogeneticly, making discoveries available to the clinic by storing them in public databases, integrating bioinformatics with the electronic medical record (EMR) (Busis, 2010).

1.2 Problem Statements

A variety of diseases and quantitative traits have shown there association with multiple genetic variants. The combined effect of these genetic variants could be additive, therefore we could evaluate individual effect of each genetic variant first, then their overall effect could be accumulated from their individual effects. However in many cases, the combined effect could also be non-additive. The example in 1.1 shows such a non-additive effect. Either heterozygote HbAS or α^{+}-thalassemia is protective against malaria. However when two conditions are inherited together, the combined effect is not stronger protection against malaria, but results in the loss of protection.

In order to detect these non-additive combined effects or gene-gene interaction, a set of genes should be examined as a whole. A number of different computational

methods have been proposed to detect gene-gene interactions existing in complex diseases and complex traits. As a first step, there should be an accurate definition of gene-gene interaction and a measure to detect such interaction. Bateson gave the first definition of gene-gene interaction (Bateson, 1909) which is actually a qualitative definition. Statistical definitions (Armitage et al., 2002; McCullagh and Nelder, 1989) and definitions based on information theory (Jakulin and Bratko, 2003; Jakulin *et al*., 2003; Chanda *et al*., 2007; Shang *et al*., 2016; Dong *et al*., 2008; Yee *et al*., 2013) were later proposed.

However these definitions of gene-gene interaction are not quite reasonable. Statistical definitions depend very much on the specific models. Definitions based on information theory are mainly based on interaction gain. The value of interaction gain can be positive, zero or negative and the explanation of its sign is difficult and confusing.

In a reasonable measure of gene-gene interaction, the measure should be computed without depending on any specific model and has its minimum value when there is no interaction.

With the increase of the number of interaction genes, the exponential increase of the number of possible combinations of interaction genes will greatly increase the computational cost, on the other hand, the increase of the number of sparse cells will decrease the statistical power. If we could find the relation between high order interaction and low order interaction under some conditions, we would identify some high order interactions by identifying low order interactions, thus greatly decrease computational cost and increase statistical power.

Complex traits are reflected in many aspects of human body. A variety of physiological parameters and body characteristics such as blood pressure, body temperature, height and weight are complex traits. Many innate and acquired behavioral characteristics such as memory, motivation, intelligence, emotion and learning are also complex traits. In addition symptoms of many diseases such as hypertension, obesity, cardiovascular diseases and neuropsychiatric disorders are reflected by complex traits. Therefore complex traits are closely related to our health

and diseases. To better understand complex diseases, we also need to understand complex traits.

Complex traits not only have association with genetic factors, but also are related to many other factors. Genetic factors related to a specific complex trait could not determine alone a value of a complex trait precisely. Therefore an appropriate way to predict the value of a complex trait with genetic factors associated with it is to classify it into several categories and predict its category. For example, in Ordinal Multifactor Dimensionality Reduction (OMDR) (Kim *et al*., 2012), complex traits are classified into several ordinal levels and an extended MDR method is proposed. However quantitative information is lost in these methods. In Quantitative MDR (QMDR) (Gui *et al*., 2013), to better utilize quantitative information contained in complex traits, a t-distribution test statistic is employed to select the best interaction classifier. However this method only classified the trait into two levels, which results in the loss of the large variability of the quantitative outcome.

Therefore we need a method which can not only fully utilize quantitative information, but also can classify the trait into any number of levels according to practical situations.

Multiple correlated phenotypes often appear in complex traits or complex diseases. For example, hypertension is diagnosed by systolic and diastolic blood pressure, cognitive ability is usually measured by memory, intelligence, language, executive function and visual-spatial function. Since GWAS analyzed each phenotype separately, it has low power to detect genetic variants with small effects which are very common in genetic association studies. If these genetic variants have small effects across multiple phenotypes or pleiotropy effects which result in strong correlation among them, these correlated phenotypes could be analyzed jointly so that these genetic variants would be detected with better power.

A variety of methods have been proposed to combine multi-locus analysis with multi-phenotype analysis in genetic association studies. These methods use all correlated phenotypes to identify gene-gene interactions. If a set of SNPs have interaction on different phenotypes, then these phenotypes would have correlation

among each other. Conversely if some phenotypes have correlation among them, there may be not the result of a common set of SNPs having interaction on these phenotypes. Therefore there should be a method to filter out those correlated phenotypes whose correlation with the phenotype under studying is not caused by pleiotropy effects.

1.3 Overview of Solutions

In this section, we give an outline of the solutions to the problems stated in the last section.

To identify gene-gene interaction associated with complex diseases, we derive a reasonable definition and measure of gene-gene interaction. Based on this new definition, we find an efficient way to identify high order gene-gene interactions. Since complex traits are common and important to understand complex diseases, we propose an extended MDR method to identify gene-gene interactions associated with complex traits by better utilizing quantitative information contained in complex traits. In addition, we propose an appropriate method to filter out those correlated phenotypes whose correlation with the phenotype under studying is not caused by pleiotropy effects for detecting gene-gene interactions associated with multiple correlated phenotypes

To give a reasonable definition of gene-gene interaction, we first give and prove an inequality which describes the relationship between two genotype variables that takes on the genotypes of two different genes as values, and a disease-status variable that takes on binary values representing the presence or absence of a complex disease. This inequality can be further generalized to *n* genotype variables. Based on this inequality, we establish a conditional independence and redundancy (CIR) based definition of gene-gene interaction and the concept of an interaction group. CIR is not only intuitive but is also non-confusing as its properties can be proven mathematically. CIR can also be computed without depending on any specific model and reaches its

minimum value when there is no interaction. We also derive a kai square statistic to measure gene-gene interactions. According to some properties of the new definition, we find the relation between high order interaction and low order interaction under some conditions, and apply it to a novel algorithm to detect high order gene-gene interactions. This algorithm can greatly decrease computational cost and increase statistical power. Possible ways to further improve the effectiveness of the novel algorithm are also provided.

To better utilize quantitative information contained in complex traits, we use extended member functions of fuzzy sets which extend the outcome range of traditional member functions of fuzzy sets from [0,1] to [-1,1] as a new measure to evaluate classification accuracy of a candidate interaction model. We then propose Extended Fuzzy Quantitative trait MDR (EFQMDR) to strengthen identification of gene-gene interactions associated with a quantitative trait by first transforming it to an ordinal trait and then using a balanced accuracy measure based on extended member functions of fuzzy sets to select multiple best sets of genetic markers as having strongest associations with the trait. Extended member functions is not only applied to the computation of training and testing accuracies, but also applied to the classification of each cell or genotype combination. EFQMDR can not only better utilize quantitative information contained in complex traits, but also can classify the trait into any number of levels according to practical situations.

Multivariate Quantitative trait based Ordinal MDR (MQOMDR) algorithm is proposed to effectively identify gene-gene interactions associated with multiple correlated phenotypes by selecting the best classifier according to not only the training accuracy of the phenotype under consideration but also other phenotypes with weights determined mainly by their pair correlation. At first, all correlated phenotypes are used to identify interactive genetic loci. Then in order to filter out those correlated phenotypes whose correlation with the phenotype under studying is not caused by pleiotropy effects, phenotypes which have the same set of SNPs that has the largest cross validation consistency (CVC) for a fixed order of interaction are grouped together. In the next stage, for each phenotype in each group, all phenotypes in the

same group only are used to identify interactive genetic loci and the average CVC of the same set of SNPs for each phenotype in each group is calculated. Remove a phenotype in each group which has the smallest CVC and calculate the average CVC again. This process is repeated until the average CVC is equal to or smaller than that in the last repetition or there are only two phenotypes left in the group. Then the average CVC of each of the remaining groups is compared with that of MDR to decide whether it is retained or abandoned. Through such a filtering process, those correlated phenotypes whose correlation with the phenotype under studying is not caused by pleiotropy effects can be filtered out and genetic variants have small effects across multiple phenotypes could be detected with better power.

Chapter 2
Background Knowledge and Related Work

2.1 Background Knowledge

Each of two complementary strands of DNA consists of a chain of nucleotides. There are four types of nucleotides based on four kinds of bases: adenine, thymine, guanine and cytosine (abbreviated as A, T, G and C respectively). DNAs from different people are almost the same except variations in a small part of these nucleotides, 90% of which are single nucleotide differences between the pairs of homologous chromosomes.

A SNP can strictly be defined as a single nucleotide variant with the allele frequency higher than 1%. It can also be used in a broader sense to include variants with smaller allele frequencies (Fernald *et al.*, 2011). Although theoretically a nucleotide can have four different forms, practically, only two of the four possible DNA bases (A,T,G,C) are seen in most SNPs; multiple base variations at a single SNP site are usually rare.

The ratio of SNPs in the human genome which includes about 3×10^{9} base pairs is estimated to fall between 3.7×10^{-4} and 8.3×10^{-4} (Carlson, 2004). SNPs may occur within both coding and noncoding regions of genes, or in the regions between genes. Since the genetic coding is redundant, some SNPs do not change the amino acid sequence of the protein produced. Such SNPs are termed synonymous, otherwise they are nonsynonymous. SNPs in noncoding regions may influence gene splicing, transcription factor binding, or the sequence of noncoding RNA. Although most SNPs in humans are neutral, some can affect their susceptibility to diseases and responses to drugs, chemicals and treatments.

Sequencing of human genome greatly promotes the increase in the amount of SNP data and the number of SNP databases. These SNP data are very useful for disease

research and drug development. The dbSNP database (Sherry *et al.*, 2001) is the most famous SNP database. In the dbSNP database Build 132, there are over 20 million validated human SNPs (Build 132, September 2010). Another important SNP database is the Online Mendelian Inheritance in Man (OMIM) database (Amberger *et al.*, 2009) containing human SNPs associated with Mendelian disorders. The Human Gene Mutation Database (HGMD) includes germline mutations in genes associated with human inherited diseases. There are more than 76 000 mutations from ~2900 genes which are free for academic and nonprofit use. The SwissVar database contains 56 000 manually annotated missense SNPs (mSNPs) from more than 11 000 genes. The PharmGKB database collects genetic variations having known drug response, including over 40 very important pharmacogenes (VIPs) and more than 3400 variants with annotated drug–response.

Conventionally, a complex trait, also called a complex phenotype, is defined as a phenotype whose features are regulated by multiple genetic and environmental factors. This contrasts to monogenetic traits, which are directly controlled by variations in a single gene. Complex diseases are disease-associated complex traits. Complex traits do not follow the rules of Mendelian inheritance, the relationships between their genetic variants and phenotypes are not linear, which means their associated genes do not interact additively.

The above definition can be further explained from clinical phenomenology and molecular backgrounds。

Clinically, rather than single qualities , complex traits are usually described in terms of combinations of different heterogeneous phenotypes or symptoms which can involve multiple organs and/or tissues types. Some of the phenotypes or symptoms can be partially or entirely shared between two or more other complex traits.

As to their molecular backgrounds, complex traits are not only regulated by genetic and environmental factors, but also products of genes (e.g., RNA, proteins or metabolites). These factors interact in different combinations and at different levels to form series of complex networks.

2.2 Related Works

The effect of gene-gene interactions is considered to play a more important role than the main effect of any individual gene in the susceptibility to common human diseases. Traditional statistical methods are not appropriate for the task. Therefore detecting and characterizing gene-gene interactions are important and challenging problems that need to be solved to diagnose and treat complex diseases. New approaches need to be developed to address this problem.

In the following sections, we will begin by introducing traditional analytical methods in gene-gene interaction studies of, and then focus on some prospective new approaches.

2.2.1 Traditional analytical methods used in gene-gene interaction studies

Statistically, a gene-gene interaction is defined as a departure from a given model on a particular scale. For simplicity, the following discussion is concentrated on interaction related to two genetic factors (two-locus interactions).

An interaction associated with a quantitative trait Y is usually represented as a deviation from a model with an additive genetic effect on the phenotype, and is tested by adding a product term in the model:

$$Y=\beta_0+\beta_{G1}\times G1+\beta_{G2}\times G2+\beta_{G1G2}\times G1\times G2 \qquad (2.1)$$

where G1 and G2 are the genotypes (usually assuming values 0,1,2 to represent the number of copies of the minor allele) of gene 1 and gene 2 respectively. The parameter β_{G1G2} represents the effect of gene-gene interaction; if $\beta_{G1G2} = 0$, there is no interaction, otherwise there exists an interaction between G1 and G2 on the trait Y.

For a binary trait or disease D (the presence/absence of a disease), the logarithm of the odds of the disease state D can be estimated using logistic regression. A gene-gene interaction is represented as a deviation from an additive genetic effect on the

log-odds of disease, and examined by testing whether $\beta_{G1G2} = 0$ in the following model:

$$\log(P(D)/1-P(D))= \beta_0+\beta_{G1}\times G1+\beta_{G2}\times G2+\beta_{G1G2}\times G1\times G2 \quad (2.2)$$

The parameters including β_{G1G2} in (2.1) and (2.2) are estimated and the null hypothesis that $\beta_{G1G2}=0$ is tested.

If we let $R=\log(P(D)/1-P(D))$, then $P(D)=1/(1+e^{-R})$. So P(D) is actually a sigmoid function of R, which tend to saturate or converge to 1 or 0 when R approaches infinity.

The parameters in these two models are scale dependent. The absence of interaction indicated by $\beta_{G1G2}=0$ in model (2.2) may be replaced by the presence of interaction when the model is transformed from the logit to the penetrance or probit scale. Conversely, an interaction shown by $\beta_{G1G2}\neq 0$ may disappear after a specific transformation (Cordell, 2002). Such transformations may result in more parsimonious models that have better fit and greater power to detect association with the contributing factors for some types of interactions (Satagopan & Elston, 2013).

In addition to linear or logistic regression models, other approaches have also been proposed. For example additive allelic effects can be replaced by dominant or recessive ones. There is also a trend to use information theory to model genetic interactions (Moore *et al.*, 2006; Chanda *et al.*, 2007; Kang *et al.*, 2008; Dong *et al.*, 2008).

In (Dong *et al.*, 2008), an entropy-based method was developed to identify two locus gene-gene interaction and, furthermore the best-fit model from all two locus interaction models. Examples (Figure 2.1) of two locus interaction models include the threshold model, jointly recessive–recessive model, jointly dominant-dominant model, and so on. Li and Reich have enumerated all possible two-locus models, some of which have significant biological meaning (Li & Reich, 2000). The existing approaches usually identify gene-gene interaction without identifying interaction models, resulting in lack of biological or genetic meaning of identified interaction.

Interaction effects of two SNPs in two genes are measured by gain ratio $\Delta R_{1,2}$:

$$\text{Gain}(D|S_{1,2})=H_0-H_{1,2}-\max\{(H_0-H_1), (H_0-H_2)\}=\min\{H_1,H_2\}-H_{1,2} \quad (2.3)$$

jointly recessive-recessive

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	0
Aa	0	0	0
Aa	0	0	1

jointly recessive-dominant

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	0
Aa	0	0	0
Aa	0	1	1

single-gene recessive

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	0
Aa	0	0	0
Aa	1	1	1

threshold

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	0
Aa	0	0	1
Aa	0	1	1

modifying-effect

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	0
Aa	0	0	1
Aa	1	1	1

jointly dominant-dominant

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	0
Aa	0	1	1
Aa	0	1	1

exclusive OR

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	1
Aa	0	0	1
Aa	1	1	0

diagonal

	SNP 2		
SNP 1	BB	Bb	bb
AA	0	0	1
Aa	0	1	0
Aa	1	0	0

Figure 2.1 Eight examples of two-locus models. 1 and 0 represent high-risk and low-risk genotype combinations respectively. (Dong *et al.*, 2008)

$$\Delta R_{1,2}=\frac{H_0-H_{1,2}-\max\{(H_0-H_1),(H_0-H_2)\}}{\min\{H_1,H_2\}}=\frac{\min\{H_1,H_2\}-H_{1,2}}{\min\{H_1,H_2\}} \qquad (2.4)$$

where H_0 is the entropy of disease status, H_1, H_2 are conditional entropies of disease status given SNP1 and SNP2 respectively, $H_{1,2}$ is the conditional entropy of disease status given SNP1 and SNP2 simultaneously.

To further identify the best-fit model from all interaction models, the case and control dataset D is divided into high- and low-risk subsets for each model: $S'_{1,2}(D)= S'_{1,2}\{D_{high}, D_{low}\}$, where D_{high} consists of data from all high-risk genotype combinations and D_{low} consists of data from all low-risk genotype combinations of a

specific model (a genotype combination is considered as a high-risk combination if it has a large case-control ratio than the total case-control ratio of the data set).

$$H'_{1,2}=P(D_{high})H(D_{high})+ P(D_{low})H(D_{low}) \quad (2.5)$$

$$\Delta R'_{1,2}=\frac{\min\{H_1,H_2\}-H'_{1,2}}{\min\{H_1,H_2\}} \quad (2.6)$$

where $H'_{1,2}$ is the entropy and $\Delta R'_{1,2}$ is the gain ratio.

New gain ratio $\Delta R'_{1,2}$ is evaluated for each candidate model and the model with maximal $\Delta R'_{1,2}$ is chosen as the best-fit model.

A more powerful approach for measuring gene-gene interaction is to use case-only analysis under the assumption that the frequencies of genes are independent in the population (Yang *et al.*, 1999).

For simplicity, suppose each of two disease susceptibility genes (gene 1 and gene 2) has two allelic variants (susceptible and nonsusceptible) that follow an autosomal dominant inheritance pattern and they are not in linkage disequilibrium, therefore their frequencies are independent in the population. For the two diallelic genes, let the first subscript i and the second subscript j indicate that the variant of gene 1 and gene 2 are present (1) or absent (0) respectively. Let P_{ij} denote the proportion of the population who have the variant of gene 1 at level i and the variant of gene 2 at level j. Let R_{ij} indicate the risk associated with the combinations of present and absent of the variants of gene 1 and gene 2. Table 2.1 shows the distribution of the number of cases expected to arise during follow-up of a "fixed" population in terms of gene frequencies in the population and risks associated with the combination of present and absent of the gene variants.

Using cases only by the presence and absence of the gene 1 and gene 2 variants, we can construct a 2×2 table (Table 2.2), from which the case-only cross-product Ψ_{co} can be calculated as follows:

$$\Psi_{co}=\frac{ad}{bc}=\frac{(P_{11}\times N\times R_{11})(P_{00}\times N\times R_{00})}{(P_{10}\times N\times R_{10})(P_{01}\times N\times R_{01})}=\frac{(P_{11}\times R_{11})(P_{00}\times R_{00})}{(P_{10}\times R_{10})(P_{01}\times R_{01})} \quad (2.7)$$

Let a period in the subscript refer to the marginal frequency of the genes in the population, then $P_{ij}=P_{i.}\times P_{.j}$. Therefore

$$\Psi_{co}=\frac{(P_{1.}\times P_{.1}\times R_{11})(P_{0.}\times P_{.0}\times R_{00})}{(P_{1.}\times P_{.0}\times R_{10})(P_{0.}\times P_{.1}\times R_{01})}=\frac{R_{11}\times R_{00}}{R_{10}\times R_{01}} \quad (2.8)$$

TABLE 2.1. The Expected Distribution of Cases expressed by Gene Frequencies in the Population and Risks Associated with Gene for Gene- Gene Interaction Analysis in a Case-Only Design. (Yang *et al.*, 1999)

Gene 1 Variant	Gene 2 Variant	Gene Frequencies in Population	Risk Associated with Genes	No. of Expected Cases
+	+	P_{11}	R_{11}	$P_{11} \cdot R_{11} \cdot N$
+	−	P_{10}	R_{10}	$P_{10} \cdot R_{10} \cdot N$
−	+	P_{01}	R_{01}	$P_{01} \cdot R_{01} \cdot N$
−	−	P_{00}	R_{00}	$P_{00} \cdot R_{00} \cdot N$

"+" and "-" represent minority allele and majority allele respectively, subscripts "1" and "0" of P and R represent minority allele and majority allele respectively.

TABLE 2.2. The distribution of cases by genotype combinations in a Case-Only Design. (Yang *et al.*, 1999)

Gene 1 Variant	Gene 2 Variant: +	Gene 2 Variant: −
+	*a*	*b*
−	*c*	*d*

a, $P_{11} \cdot R_{11} \cdot N$; *b*, $P_{10} \cdot R_{10} \cdot N$; *c*, $P_{01} \cdot R_{01} \cdot N$; *d*, $P_{00} \cdot R_{00} \cdot N$. ψ_{co}, *ad*/*bc*.

If we define the risk ratios as $RR_{ij}=R_{ij}/R_0$, Ψ_{co} can be represented in terms of risk ratios as $\Psi_{co}= RR_{11}/RR \times R_{10}R_{01}$. If the effects for the two genes conform to a multiplicative relation, then the case-only Ψ_{co} should equal unity, that is, $RR_{11}/RR \times R_{10}R_{01} =1$. Therefore departure of case only Ψ_{co} from unity provides a measure of gene-gene interaction under the assumption of independent gene frequencies in the population.

2.2.2 Data-mining methods

Since traditional regression-based methods measure interaction by departure from a linear model, they are not appropriate for nonlinear models and high order interaction corresponding to sparse contingency tables with a lot of empty cells.

To solve this problem, a variety of data-mining methods have been proposed recently. These methods search for patterns in high-dimensional data in a computationally efficient way. However, these methods test for association with a specific genetic factor and interaction with other genetic factors combined, rather than testing for interaction separately. Therefore, gene-gene interaction should be identified using additional statistical modeling. In addition, to avoid overfitting problems, cross validation is used.

A variety of data-mining approaches have been proposed to detect interactions in genetic association studies, such as logic regression (Kooperberg *et al.*, 2001; Kooperberg *et al.*, 2005), genetic programming (Nunkesser *et al.*, 2007), neural networks (Motsinger, *et al.*, 2006; Motsinger-Reif, *et al.*, 2008) and pattern mining (Li *et al.*, 2007; Long *et al.*, 2009). In the remainder of this section, several popular and promising methods for detecting gene-gene interactions are discussed.

2.2.2.1 Multifactor dimensionality reduction method

One particularly popular data-mining method is Multifactor dimensionality reduction (MDR) (Ritchie *et al.*, 2001; Moore *et al.*, 2004; Chung *et al.*, 2007). In order to detect high- dimensional gene-gene interaction, MDR groups genotype combinations of multiple genetic factors into two categories: high risk or low risk categories, then tests association between a binary trait or disease with this new one dimensional variable. Rather than testing for interaction separately, MDR tests main effects and interactions of multiple genetic factors combined.

The MDR method is proceeded as follows: the 10-fold cross validation is used. A set of *n* genetic loci is specified and all of their combinations or cells form an *n* dimensional space. Each locus has three genotypes and *n* loci can form *n* square of three genotype combinations. Then the ratio of the number of cases to the number of controls is estimated for each combination, which is then labeled either as “high-risk”, if the cases:controls ratio is equal or greater than some threshold, or otherwise as

"low-risk". Thus all cells are allocated to either of high risk group or low risk group, which reduces the n-dimensional space into a one dimensional space. The processe is repeated for all possible *n*-loci combinations. The combination having maximal case–control ratio of the high-risk group of the training data is selected and its prediction error can be estimated using the testing data. The model having minimal prediction error is selected as the final best *n*-locus model. The cross-validation consistency is defined as the number of cross-validation replicates in which that same *n*-locus model was chosen as the best model. A best multifactor model is selected for each of the two up to a certain maximum number of loci. The combination of loci having minimal prediction error is selected from these best multifactor models. Hypothesis testing for this final model can be performed by evaluating its cross-validation consistency.

Like other exhaustive search techniques, the main problem with MDR is its prohibitive computational cost to search all *n*-locus combinations which increase exponentially with the increase of the number of locus (Ritchie *et al.*, 2001).

2.2.2.2 Ordinal MDR

An extension of MDR, ordinal MDR (OMDR), is proposed to extend the application of MDR from binary traits to ordinal traits which often appear in trait description (e.g., obesity can be classified as normal, pre-obese, mild obese and severe obese) (Kim *et al.*, 2012).

Suppose a given ordinal phenotype has J classes labeled as 1, 2,..., J. For an m-locus combination, n_{ij} denotes the number of individuals with the ith m-locus genotype and n_{+j} denotes the overall number of individuals in class j, where i = {1, 2,...,3^m} and j = 1, 2,..., J.

The dataset is analyzed using L-fold cross-validation (CV). The ith m-locus genotype is labeled as class c(i) by OMDR as follows:

$$c(i)= \left(\frac{n_{ij}}{n_{+j}}\right) \qquad (2.9)$$

Then classification accuracy is used to select K best classifiers for each CV set. The

general cross-validation consistency based on top-K selection ($GCVC^K$) which is the number of times a classifier is selected as one of the K best classifiers for all CV sets is calculated for each of the K best classifiers for all CV sets. Classifiers having the maximum predictability and maximum $GCVC^K$ for all CV sets are selected as best classifiers. Finally, classifiers having best predictability and $GCVC^K$ among those best classifiers for different number of locus are selected as the overall best classifiers.

2.2.2.3 Random forest method

Another data-mining approach extensively used to study of gene-gene interactions is random forest (RF). RF is an ensemble or 'forest' of some kind of classification or regression trees.

A classification or regression tree maps each genotype combination to a disease status (Figure 2.2). Each node in the tree represents a genetic factor and is connected by arcs or edges to 'child' nodes. Each edge represents some possible values the parent node could take. A path through the tree forms a specific combination of values taken by the genetic factors in that path. The tree stops to grow at a node when no better classification accuracy can be obtained (for example, the node includes only one status: cases or controls, the path ending with the node contains or when all possible SNPs) or some stopping conditions are satisfied. Therefore the trees test for main effect and interaction combined, rather than testing for interaction separately..

One problem for recursive partitioning is that, since it selects a variable at the root node according to its main effect, i.e., its ability to partition the data into more homogeneous sub-groups and at the internal nodes according to its main effect conditional on variables selected beforehand, pure interactions without main effects are missed.

To address this problem, an ensemble of trees can be used. The random forest approach is a popular one (Breiman, 2001) employed in several association studies (Lunetta *et al.*, 2004; Bureau *et al.*, 2005). The trees in a random forest are grown on

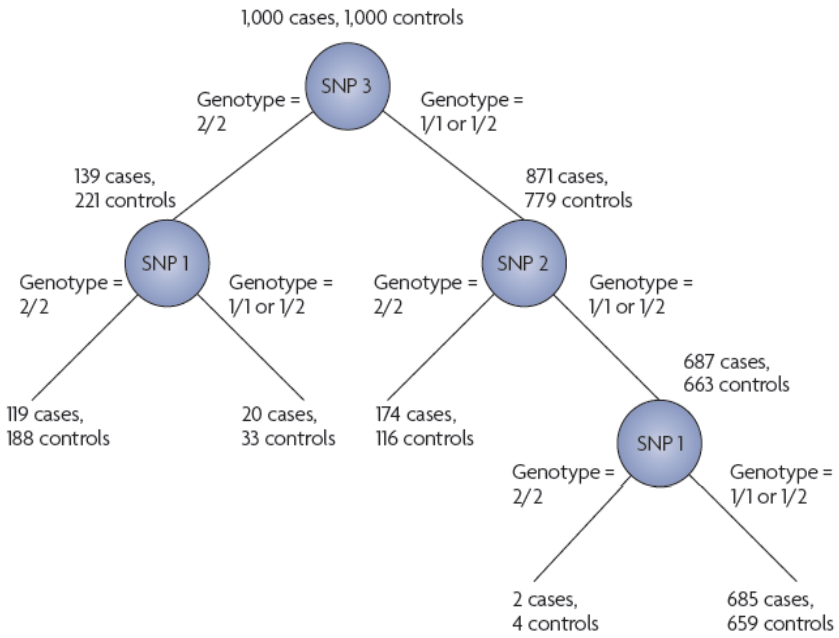

Figure2.2 An example of a classification tree. (Cordell, 2009)

bootstrap samples of the original data. At each node, only a subset of rather than all possible genetic factors is randomly selected to determine the best split, therefore partially solving the problem of missing pure interactions. Each tree is trained on a different bootstrap sample, and prediction error is estimated using remaining sample.

For each individual, its class is predicted across all trees where it was not in the corresponding bootstrap sample, and its final predicted class is the class predicted for the most times in all trees. Each genetic factor is assigned an importance score that measures its importance and therefore its priority by random forests. By using ensemble of trees in such a manner, it's more likely to identify interactions among genetic factors with weak main effects.

2.2.3 Pathway-based gene-set analyses

Prior biological or functional knowledge has also been used to increase possibility to detect genetic association. An important approach is to use pathway-based gene-set analyses which use a set of genes having functional relation to collectively evaluate their association with a trait.

Although GWAS increase power to detect genetic association and have found novel

genes for several complex diseases, but still many associations are lost. First, genetic factors that individually have a small impact but collectively have a significant impact on a disease may be missed by GWAS which only detect significant SNPs/genes. Second, even those factors that bring a significant effect may be missed due to the small sample size. To address these two limitations, the effects of biologic network context, especially metabolic pathways, can be considered and become feasible due to fast expansion of databases for metabolic pathways. The KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway reflects reaction and interaction in a complex network (Ogata *et al*., 2000) and could be used in pathway-based gene-set analyses of complex diseases.

In (Chen *et al*., 2008), the method of prioritizing risk pathways (PRP) is used. According to matrix D which depicts the frequency of cases and controls for each allele of a SNP locus, a *P*-value with the following statistics χ^2 is calculated for each SNP:

$$\chi^2=\frac{(ad-bc)^2(a+b+c+d)}{(a+b)(b+d)(d+c)(c+a)} \qquad (2.10)$$

SNPs with a *P*-value beyond the significance level of 0.05 are filtered out for further study and their *risk* statistics calculated. *Risk* values for other SNPs are directly set to be 0.

In KEGG, a pathway is a network whose node is the metabolite and edge is an enzyme or a gene cluster. First, each KEGG pathway is transformed into a graph K in which an edge represents a metabolite and a node represents an enzyme or a gene cluster. Figure 2.3 gives an example of such a pathway .

The degree attribution of the node which is the number of edges connecting to it in the graph K, also referred as the biologic network context reflects the diversity of the metabolites related to an enzyme or a gene cluster in the original pathway.

All the screened SNPs are mapped to the corresponding g_t (t =1,...,T, where T is the total number of genes involved in the pathway) that are located <500 kb away from g_t. The maximum *risk* value *Risk*(g_t) is selected as the genetic statistic value for g_t .

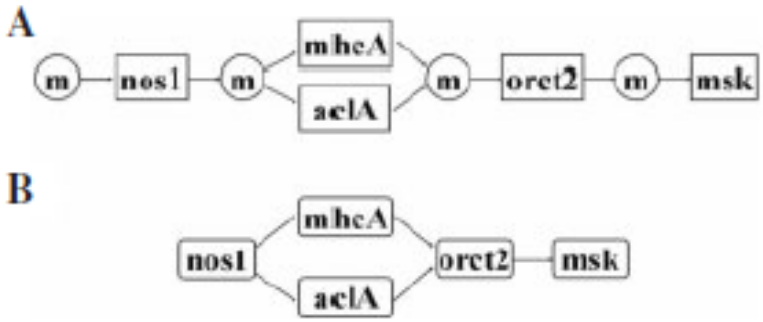

Figure 2.3 Sample pathway. (**A**) Sample pathway of KEGG with rectangles representing reactions named by genes encoding for their catalyzing enzymes and circles representing metabolites labeled with 'm'. (**B**) The graph *K* transformed from the pathway in (A). (Ogata *et al*., 2000)

For each gene cluster s_j (j =1,...,N_i , where N_i is the count of gene clusters in pathway k_i) in pathway k_i (i=1,...,U, where U is the number of human pathways in KEGG database), the biologic network context $E(s_j , k_i)$ of the gene cluster s_j is measured by the number of its edges connecting to it in the transformed pathway k_i and the genetic factor $G(P, s_j)$ of the gene cluster s_j is calculated as:

$$G(P, s_j)=\frac{1}{M_j}\sum_1^{M_j} Risk(g_t) \tag{2.11}$$

The RS (risk score) value $preRS(P, k_i)$ between phenotype P and pathway k_i is:

$$preRS(P, k_i)=\sum_{j=1}^{N_i}\{G(P, s_j) \times E(s_j, k_i)\} \tag{2.12}$$

The standardized value $RS(P, k_i)$ is quantified as:

$$RS(P, k_i)=\frac{preRS(P,k_i)}{\max_{1\ll i\ll U}\{preRS(P,k_i)\}} \tag{2.13}$$

Then risk pathways are prioritized according to $RS(P, k_i)$. This provides a new channel to study the pathogenesis of complex diseases.

2.2.2.4 Biological interpretation

The relation between statistical interaction and biological or functional interaction has been extensively discussed. In a recent review (Phillips, 2008), three different forms of epistasis or interaction are defined: compositional epistasis, statistical epistasis and functional epistasis. Compositional epistasis is defined as the effect of an genetic variant is masked by another genetic variant; statistical epistasis is defined as

the average effect of substitution of alleles at combinations of loci, with respect to the average genetic background of the population and functional epistasis is defined as the interactions between biological molecules such as proteins and other genetic elements. The important thing in interaction modeling is how the main effect of a variable, the independence of the main effects are defined and, therefore, how deviation from the independence of effects is measured.

Although there seems no obvious connection between biological interaction and statistical interaction (Greenland, *et al.*, 2009), some work has been done to evaluate the fit of some biological models to given genetic or genomic data (Sepulveda *et al.*, 2007; Sepulveda *et al.*, 2009; Aylor, *et al.*, 2008). This work is more practical since it tries to use known biological knowledge to explain given genetic or genomic data.

Chapter 3

A Novel Approach for Identifying Epistatic Interactions for Complex Disease Prediction

3.1 Introduction

Identifying genetic variants, in terms of single nucleotide polymorphisms (SNPs) or Single Sequence Repeat Polymorphisms (SSRPs), that are associated with complex diseases is important for the understanding of complex diseases. In most genome-wide association studies (GWASs), a single-locus approach, where each variant is tested individually for association with a disease, is usually used to find statistical associations of SNPs with important common diseases (the Wellcome Trust Case Control Consortium, 2007). However, the associations that are so identified account only for a small part of the heredity of complex diseases (Franke *et al*., 2009) which are considered to be mostly associated with gene-gene or gene-environment interactions (Manolio *et al*., 2009) and this has been confirmed with evidence by several studies (Bateson, 1909; Moore *et al*., 2005; Malmberg *et al*., 2005; Segre *et al*., 2005).

To characterize and detect gene-gene interactions existing in complex diseases, a variety of different computational methods have been proposed which include such methods as logistic regression (Kooperberg *et al*., 2001; Kooperberg and Ruczinski, 2005), recursive partitioning (Zhang and Bonney, 2000; Nelson *et al*., 2001; Culverhouse *et al*., 2004), Multifactor Dimensionality Reduction (MDR) (Ritchie *et al*., 2001; Hahn *et al*., 2003; Moore 2004), genetic programming (Nunkesser *et al*., 2007), artificial neural networks (Motsinger *et al*., 2006; Motsinger *et al*., 2008) and pattern mining (Li *et al*., 2007; Long *et al*., 2009). The effectiveness of these methods

depends very much on how gene-gene interaction is defined but, unfortunately, such a definition is still not quite adequate.

3.2 Related works

The first definition of gene-gene interactions, which is also referred to as *epistasis*, was given in (Bateson, 1909) as a phenomenon where the effects of a given gene on a biological trait are either masked or enhanced by one or more of the other genes. This definition is actually a qualitative rather than a quantitative description.

Currently, the most common statistical definition of interaction is that interaction represents departure from a linear model that predicts a phenotypic outcome with respect to several predictors. This kind of definition of interaction depends very much on the specific models.

In addition to statistical definitions, gene-gene interactions are also defined based on information theory.

According to (Jakulin and Bratko, 2003; Jakulin *et al*., 2003; Moore *et al*., 2006), if the genotypes of two genes are represented as two genotype variables, G_1 and G_2, then their dependency with respect to a disease-status variable, D, can be measured by interaction gain (IG) which is defined as follows. Let $H(X)$ be the entropy of X, then the IG of G_1, G_2, and D can be expressed as:

$$\mathrm{IG}(G_1\ G_2\ D) = I(G_1\ G_2;\ D) - I(G_1;\ D) - I(G_2;\ D) = I(G_1;G_2|\ D) - I(G_1;G_2) \quad (3.1)$$

where I denotes the mutual information measure,

$$I(G_1\ G_2;\ D) = H(G_1\ G_2) + H(D) - H(G_1\ G_2,\ D) \quad (3.2)$$

$$I(G_1;G_2|\ D) = H(G_1|\ D) + H(G_2|\ D) - H(G_1,\ G_2|\ D) \quad (3.3)$$

$$I(G_1;G_2) = H(G_1) + H(G_2) - H(G_1,\ G_2) \quad (3.4)$$

The variables G_1 and G_2 are joined into their Cartisian product $G_1\ G_2$. According to this formula, interaction gain is regarded as the difference between the actual decrease in entropy achieved by the joint variables $G_1\ G_2$ and the expected decrease in entropy, which is $I(G_1;\ D)+I(G_2;\ D)$, with the assumption of independence

between G_1 and G_2. A positive difference indicates interaction between G_1 and G_2 that cannot be linearly decomposed, while a negative difference indicates information redundancy between G_1 and G_2 and a zero difference indicates conditional independence or a mixture of synergy and redundancy.

$I(G_1;G_2)$ is a measure of dependence or "correlation" between two genes G_1 and G_2 regardless of the context D, whereas $I(G_1;G_2|\ D)$ is conditional mutual information, a measure of dependence of G_1 and G_2 given the context of D. Therefore IG is the change in the dependence of two genes by introducing context D. When IG is positive, context increased the amount of dependence between two genes; when IG is zero, context did not change the amount of dependence; when IG is negative, context decreased the amount of dependence.

In (Chanda *et al*., 2007), the k-way interaction information (KWII) or co-information, which is a generalization of the mutual information and includes IG as a special case when k=3, and the total correlation information (TCI) are introduced to identify and visualize gene-gene and gene-environment interactions.

Let X={ X_1,X_2, … ,X_k} is set of k variables. The KWII on X is an alternating sum over all possible subsets T of X.

$$\text{KWII}(X)=-\sum_{T\subseteq X}(-1)^{|X|-|T|}H(T) \tag{3.5}$$

For k=3,

$$\text{KWII}(X_1,X_2,X_3)=-H(X_1)-H(X_2)-H(X_3)+H(X_1\ X_2)+H(X_1\ X_3)+H(X_2\ X_3)-H(X_1X_2X_3)$$

The TCI on X is the difference of entropies of the individual variables $H(X_1)$, $H(X_2)$, … , and $H(X_k)$ and the entropy of their combination $H(X_1,X_2,\ \ldots\ ,X_k)$:

$$\text{TCI}(X_1,X_2,\ \ldots\ ,X_k)=[\sum_{i=1}^{k}H(X_i)\ ]-H(X_1,X_2,\ \ldots\ ,X_k) \tag{3.6}$$

The KWII represents the change of information when all k variables are observed as a whole. The KWII is always positive for two variables, but it can be positive or negative for multiple variables. A positive value indicates synergy among variables, a negative value indicates information redundancy among variables, and a zero value indicates there is no K-way interactions. However an even number of

completely redundant variables will correspond to a positive value of KWII, rather than a negative value.

The TCI is the amount of information shared among k variables or a measure of redundancy or dependency. A zero value shows independence among variables. The TCI reaches its maximal value when one variable is completely redundant with the others, which means one variable brings all the information that the others can provide.

To lower the high computational cost of KWII which requires the entropies of all subsets, a novel information theoretic metric called phenotype associated information (PAI) is used to detect genetic factors involved in gene–gene and gene–environment interactions (Chanda *et al*., 2008).

The PAI is the amount of information shared or dependency between genotype variables and the disease status variable. Therefore the PAI doesn't include interdependencies among genotype variables, making it robust to correlations and redundancies among genotype variables such as LD which are interference factors to identification of gene-gene interaction. It can be obtained by the difference between the TCI representing the overall dependency among the genotype variables and the disease status variable and the TCI representing the interdependencies among the genotype variables:

$$\mathrm{PAI}(G_1,G_2,\ldots,G_k,D)=\mathrm{TCI}(G_1,G_2,\ldots,G_k,D)-\mathrm{TCI}(G_1,G_2,\ldots,G_k) \qquad (3.7)$$

The KWII provides a more valuable and parsimonious interaction measure since it only measures interaction among a set of variables of interest as a whole and does not contain interactions from its subsets. However its computation requires the entropies of all subsets, making it computationally intractable. In addition, since its value could be either positive or negative determined by the nature of the interaction, it could not use hill climbing algorithms to reduce the search space.

In contrast, the PAI don't take negative values and increases monotonically with the increase of interaction order, making it appropriate for hill climbing algorithms. In addition, the computation of the PAI requires only individual and joint entropies, making it far more feasible than the KWII in computation.

The PAI is also closely related to the KWII since it can be derived that

$$\mathrm{PAI}(G_1,G_2,\ldots,G_k,D)=\sum_{T\subseteq G}\mathrm{KWII}(T,D) \tag{3.8}$$

This equation shows that the PAI is the sum of the KWIIs of all subset combinations of the genotype variables and the disease status variable. Therefore a greedy search algorithm which avoids combinatorial explosion can be employed to search for gene-gene interactions associated with complex diseases. It uses the PAI to reduce the search space from the combinatorial space to the interesting regions and then calculate the KWII for the reduced search space.

The PAI is further used to analyze the gene-gene and gene-environment interactions associated with quantitative (Chanda *et al*., 2009).

The entropy $H(Z)$ of a normally distributed variable Z is employed for the entropy of a QT, P:

$$H(P)=\ln(\sigma\sqrt{2\pi e}), \tag{3.9}$$

where σ is the standard deviation.

Accordingly the entropy of the joint distribution of a continuous QT, P, and a set of discrete genotype variables can be computed as:

$$H(G,P)=-\sum_{g}\int_{p}p(P,G=g)\ln p(P,G=g)dP=H(G)+\sum_{g}p(G=g)\ln(\sigma\sqrt{2\pi e}), \tag{3.10}$$

where $H(G)$ contains only discrete variables.

These equations make it feasible to calculate the KWII and the PAI to identify gene-gene interactions associated with a QT.

In (Shang *et al*., 2016), two co-information based measure: NCI, normalized *n*-order interaction effect and CCI which not only measures the impact of a set of SNPs itself but also measures the impact of its subsets are proposed to measure the impact of a set of SNPs to the phenotype.

Co-information is actually the same as the KWII. It has two confusing properties which prevent it from widely adopted as an interaction measure. One is its value. Except for 2-order interaction whose co-information is actually always positive mutual information, its value can be positive, negative or zero. The explanation of its

sign is only given intuitively and confusing. Another is its sensitivity to the SNP combination order which makes it difficult to rank sets of SNPs with different combination orders. To tackle the second problem, NCI, n-order interaction effect which is the averages of co-information values fixed for different orders is proposed to normalize co-information. NCI only measures the impact of a set of SNPs as a whole. The total impact of a set of SNPs to the phenotype should include not only the interaction effect of itself as a whole, but also interaction effects of all its subsets and main effects of all individual SNPs in the SNP combination. Therefore another association measure based on co-information is proposed to measure the total impact of a set of SNPs to the phenotype which includes its impact, and impacts of its subsets with their NCI values greater than or equal to the user-specified thresholds.

The methods as proposed in (Chanda *et al*., 2007; Chanda *et al*., 2008; Chanda *et al*., 2009; Shang *et al*., 2016) are all based on KWII or the co-information measure. However, the explanation of the signs of the co-information measure (including that of the IG) can only be understood intuitively rather than mathematically and is considered a confusing property of the measures (Shang *et al*., 2016).

In [33], the standardized relative information gain (RIG) was proposed to measure the interactions of a set of SNPs. Suppose Y is the disease status and X is the set of SNPs, then RIG U_0 is defined as follows:

$$U_0 = \frac{H(Y) - H(Y \mid X)}{H(Y)} \tag{3.11}$$

where $H(Y)$ and $H(Y|X)$ are the entropy of Y and the entropy of Y given X respectively. It quantifies the proportion of uncertainty of Y that is reduced after X is introduced. The value of U_0 reflects the strength of impact a specific set of SNPs have on the disease.

Since the value of U_0 tends to increase with the order of interactions when a higher order interaction includes SNPs in a lower order interaction as a subset regardless of the true additional contribution, direct comparison of RIGs among different orders of interaction is not appropriate. Therefore RIGs need to be standardized with the mean and standard deviation from the permuted datasets

generated by repeatedly shuffling the phenotypes in original data with all genotypes unchanged. Standardized relative information gain obtained from initial relative information gain of the original data, U_o, is defined as follows:

$$U_r = \frac{U_o - \bar{U}_P}{S_P} \tag{3.12}$$

where U_o and $\bar{U}_P$ are the average and the standard deviation of the maximum RIG of the permuted datasets respectively.

The RIG essentially measures the mutual information of X and Y relative to $H(Y)$.

As discussed above, neither statistical approaches nor information theoretical approaches to the definition of epistatic gene-gene interactions are quite resonable. We propose here new definitions of gene-gene interaction, called CIR which can better allow such interactions to be discovered. CIR is not only intuitive but is also non-confusing as its properties can be proven mathematically. CIR can also be computed without depending on any specific model.

In the following sections, we introduce the details of the derivation of CIR and provide proofs to some properties and a theorem related to CIR. Three cases where there is no interaction among genes on the disease status are also identified. A new algorithm to detect gene-gene interaction with order greater than two is also proposed based on these new definitions and corresponding properties and theorem. Experiments on simulated and real datasets show the effectiveness of these new definitions and the effectiveness and efficiency of this new algorithm.

3.3 Methodology

3.3.1 New definitions of gene-gene interaction and an interaction group

From previous analysis, existing definitions of gene-gene interactions are not quite

reasonable. To introduce our new definition, we first prove the following inequality:

$$I(G_1, G_2; D) \geq I(G_1; D) + I(G_2; D) - I(G_1; G_2). \quad (3.13)$$

where G_1 and G_2 represent two genotype variables; D represents the disease status variable, I denote mutual information.

Proof:

$$I(A;C) = \sum_{A,C} p(A, C) \log \frac{p(A, C)}{p(A)p(C)},$$

$$I(B;C) = \sum_{B,C} p(B, C) \log \frac{p(B, C)}{p(B)p(C)},$$

$$I(A;B) = \sum_{A,B} p(A, B) \log \frac{p(A, B)}{p(A)p(B)},$$

$$I(A,B;C) = \sum_{A,B,C} p(A, B, C) \log \frac{p(A, B, C)}{p(A, B)p(C)},$$

$$I(G_1;D) + I(G_2;D) - I(G_1,G_2;D) = \sum_{G_1,G_2,D} p(G_1, G_2, D) \log \frac{p(G_1, D)p(G_2, D)p(G_1, G_2)p(D)}{p(G_1)p(G_2)p^2(D)p(G_1, G_2, D)} =$$

$$\sum_{G_1,G_2,D} p(G_1, G_2, D) \log \frac{p(G_1, D)p(G_2, D)}{p(D)p(G_1, G_2, D)} + I(G_1;G_2) \leq \log \sum_{G_1,G_2,D} \frac{p(G_1, D)p(G_2, D)}{p(D)} + I(G_1;G_2) =$$

$$\log \sum_{D} \frac{1}{p(D)} [\sum_{G_1} p(G_1, D) \sum_{G_2} p(G_2, D)] + I(G_1; G_2) = I(G_1; G_2).$$

∴ $I(G_1,G_2;D) \geq I(G_1;D) + I(G_2;D) - I(G_1;G_2)$, with equality iff $p(G_1,D)p(G_2,D) = p(D)p(G_1,G_2,D)$, i.e., $p(G_1|D)p(G_2|D) = p(G_1 G_2|D)$. □

If A,B are independent, then we have $I(A,B;C) \geq I(A;C) + I(B;C)$.

This inequality suggests that if $I(G_1,G_2;D) = I(G_1;D) + I(G_2;D) - I(G_1; G_2)$ or $p(G_1|D)p(G_2|D) = p(G_1G_2|D)$, i.e., G_1 and G_2 are conditionally independent of D, then G_1 and G_2 have no interaction on D. This inequality can be further generalized to n genotype variables $G_1, G_2, \ldots, G_n$ and a disease-status D:

$$I(G_1, G_2, \ldots, G_n;D) \geq I(G_1;D) + I(G_2;D) + \ldots + I(G_n;D) - I(G_n; G_1, , \ldots, G_{n-1}) - I(G_{n-1}; G_1, , \ldots, G_{n-2}) - \ldots - I(G_2; G_1). \quad (3.14)$$

with equality iff $p(G_1|D)\, p(G_2|D) \ldots p(G_n|D) = p(G_1, G_2, \ldots, G_n | D)$.

In order to better understand the inequality in (10), we let S_i, where $i \in \{1, \ldots, n\}$,

be a set of elements in any universe X If we map $I(G_1;D)$, $I(G_2;D)$,…, $I(G_n; D)$ to the cardinality of n sets S_1, S_2,…, S_n, respectively, denoted as $|S_1|$, $|S_2|$,…, $|S_n|$, $I(G_n; G_1, , \ldots, G_{n-1})$,…, $I(G_2;G_1)$ to $|S_n \cap (S_1 \cup \ldots \cup S_{n-1})|$,…, $|S_2 \cap S_1|$ respectively and $I(G_1,G_2, \ldots,G_n;D)$ to $|S_1 \cup \ldots \cup S_n|$, then the right hand side of the inequality is equivalent to $|S_1|+\ldots+|S_n|-|S_n \cap (S_1 \cup \ldots \cup S_{n-1})|-\ldots -|S_2 \cap S_1| = |S_1 \cup \ldots \cup S_n|$, which is equivalent to the left hand side of the inequality. This can be illustrated in Figure 3.1 below when the equality holds for three genotype variables and one disease status variable and the proof is also given as follows.

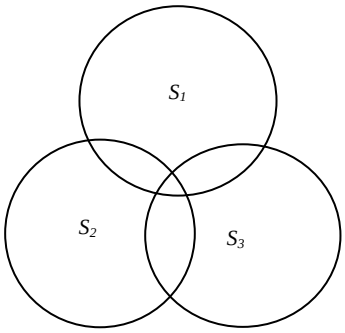

Figure. 3.1 $|S_1 \cup S_2 \cup S_3| = |S_1| + |S_2| + |S_3| - |S_3 \cap (S_1 \cup S_2)| - |S_1 \cap S_2)|$

Proof:

$I(G_1, G_2, \ldots, G_n;D) \geq I(G_1, G_2, \ldots, G_{n-1};D) + I(G_n;D) - I(G_n; G_1, , \ldots, G_{n-1}) \geq$

$I(G_1,G_2, \ldots,G_{n-2};D) + I(G_{n-1};D) - I(G_{n-1}; G_1, , \ldots,G_{n-2}) + I(G_n;D) - I(G_n; G_1, , \ldots, G_{n-1})$

$\geq \ldots \geq I(G_1;D) + I(G_2; D) + \ldots + I(G_n;D) - I(G_n; G_1, , \ldots, G_{n-1}) - I(G_{n-1}; G_1, , \ldots, G_{n-2})$

$-\ldots - I(G_2; G_1)$

with equality iff $p(G_1, G_2, \ldots, G_{n-1} |D)\, p(G_n|D) = p(G_1, G_2, \ldots, G_n |D)$, $p(G_1, G_2, \ldots, G_{n-2} | D)\, p(G_{n-1}| D) = p(G_1, G_2, \ldots, G_{n-1}| D), \ldots, p(G_1| D)\, p(G_2| D) = p(G_1, G_2 | D)$, i.e., $p(G_1| D)\, p(G_2| D) \ldots p(G_n| D) = p(G_1, G_2, \ldots, G_n | D)$ □.

In the case that we have two genotype variables G_1 and G_2, in addition to $I(G_1 ,G_2; D) \geq I(G_1; D) + I(G_2; D) - I(G_1;G_2)$, we also have $I(G_1 ,G_2; D) \geq \max\{I(G_1; D), I(G_2; D)\}$. $I(G_1,G_2; D) = I(G_1; D)$ iff $p(D|G_1) = p(D|G_1,G_2)$, i.e. G_2, G_1, D form a Markov chain (Mceliece, 2002). Likewise, $I(G_1 ,G_2; D) = I(G_2; D)$ iff G_1 ,G_2, D form a Markov chain. In these two cases, G_1 and G_2 are also considered to have no interaction on D.

For n (n>2) genotype variables G_1, G_2, …, G_n, if, for a subset G_{i1}, G_{i2}, …, G_{im} of them, $p(D|G_{i1,} G_{i2,} \ldots, G_{im}) = p(D| G_1, G_2, \ldots, G_n)$, $p(G_{i1,}G_{i2,} \ldots, G_{im} |D) = p(G_{i1}|D)$

$p(G_{i2}|D)\ldots p(G_{im}|D)$, $1\le m<n$, $1\le i1,i2,\ldots, im\le n$, they should also be considered to have no interaction on D. Therefore, we have the following definition:

Definition 3.1 n genotype variables $G_1, G_2, \ldots, G_n$ have no interaction with a disease state variable D if one of the following two conditions is satisfied:

(1) $p(G_1,G_2,\ldots,G_n|D)=p(G_1|D)p(G_2|D)\ldots p(G_n|D)$; or equivalently $I(G_1,G_2,\ldots,G_n;D)=I(G_1;D)+I(G_2;D)+\ldots+I(G_n;D)-I(G_{n;}G_1,,\ldots,G_{n-1})-I(G_{n-1;}G_1,\ldots,G_{n-2})-\ldots-I(G_2;G_1)$.

(2) $\exists$ a subset $G_{i1},G_{i2},\ldots,G_{im}$ of $G_1,G_2,\ldots,G_n$, $p(D|\ G_1, G_2, \ldots, G_n)=p(D|G_{i1,}\ G_{i2,}\ \ldots, G_{im})$, and $p(G_{i1,}\ G_{i2,}\ \ldots, G_{im}|D)= p(G_{i1}\ |D)\quad p(G_{i2}|\ D)\ \ldots\ p(G_{im}|D)$;

or equivalently $I(G_1,\ G_2,\ \ldots,\ G_n;D)= I(G_{i1};D)+\ I(G_{i2};D)+\ldots+\ I(G_{im};D)-\ I(G_{im;}\ G_{i1},\ ,\ \ldots,\ G_{im-1})-\ I(G_{im-1;}G_{i1},\ ,\ \ldots,\ G_{im-2})\ -\ldots-\ I(G_{i2};\ G_{i1})$. $(1\le m<n,\ 1\le i1,i2,\ldots,\ im\le n)$.

We call this new definition of interaction: *Conditional Independence and Redundancy* (CIR) based definition of interaction. Since $I(G;null)=0$, when $m=1$ in condition (2), $I(G_1,\ G_2,\ \ldots,\ G_n;D)= I(G_{i1};D)$. If we replace $1\le m<n$ in (2) with $1\le m\le n$, then (1) and (2) can be merged as "$\exists$ a subset $G_{i1},\ G_{i2},\ \ldots,\ G_{im}$ of $G_1,\ G_2,\ \ldots,\ G_n$, $p(D|\ G_1,\ G_2,\ \ldots,\ G_n)=p(D|G_{i1,}\ G_{i2,}\ \ldots,\ G_{im})$, and $p(G_{i1,}\ G_{i2,}\ \ldots,\ G_{im}|D)= p(G_{i1}\ |D)\ p(G_{i2}\ |D)\ldots p(G_{in}|D)$; or equivalently $I(G_1,\ G_2,\ \ldots,\ G_n;D)=I(G_{i1};D)+\ I(G_{i2};D)+\ldots+\ I(G_{im};D)-I(G_{im;}\ G_{i1},\ ,\ \ldots,\ G_{im-1})-\ I(G_{im-1;}\ G_{i1},\ ,\ \ldots,\ G_{im-2})\ -\ldots-\ I(G_{i2};\ G_{i1})$. $(1\le m\le n,\ 1\le i1,i2,\ldots,\ im\le n)$."

If $m=n$, $G_1,\ G_2,\ \ldots,\ G_n$ are called (completely) conditionally independent of D; if $m=1$, $G_1,\ G_2,\ \ldots,\ G_n$ are called completely redundant on D, in this case, $I(G_1,\ G_2,\ \ldots,\ G_n;D)=I(G_{i1};D)$; otherwise $G_1,\ G_2,\ \ldots,\ G_n$ are called partially redundant on and partially conditionally independent of D.

Definition 3.1 shows that the naïve assumption of class conditional independence in naïve Bayesian classification is not naïve, it actually assumes that there is no interaction among attribute variables on the class variable.

Definition 3.2 n genotype variables $G_1,\ G_2,\ \ldots,\ G_n$ form an interaction group on a disease status variable D if any two groups of variables derived from a partition of them are not conditionally independent of D and for any G_i of them, the following equality is violated:

$$p(D|G_1,\ldots,G_{i-1},G_{i+1}\ldots,G_n)=p(D|G_1,G_2,\ldots,G_n).$$

3.3.2 Some Properties Related to the New Definitions

Relating to the above definitions, we have the following results:

Property 3.1 If for a subset G_{i1}, G_{i2}, …, G_{im} of n genotype variables G_1, G_2, …, G_n and a disease status variable D, $p(D|G_{i1}, G_{i2}, \ldots, G_{im})=p(D| G_1, G_2, \ldots, G_n)$, ($1\le m<n$-1, $1\le i1,i2,\ldots, im \le n$), then $p(D| G_1,\ldots G_{j-1}, G_{j+1}\ldots, G_n)=p(D| G_1, G_2, \ldots, G_n)$, $\{G_{i1}, G_{i2}, \ldots, G_{im}\} \subset \{ G_1,\ldots G_{j-1}, G_{j+1}\ldots, G_n \}$.

Proof: $p(D| G_1,\ldots G_{j-1}, G_{j+1}\ldots, G_n)=\sum_{G_j} p(D, G_j \mid G_1, \ldots, G_{j-1}, G_{j+1}, \cdots, G_n) =$

$\sum_{G_j} p(G_j \mid G_1, \ldots, G_{j-1}, G_{j+1}, \cdots, G_n)p(D \mid G_1, G_2, \cdots, G_n) =$

$\sum_{G_j} p(G_j \mid G_1, \ldots, G_{j-1}, G_{j+1}, \cdots, G_n)p(D \mid G_{i1}, G_{i2}, \cdots, G_{im})$

$= p(D|G_{i1}, G_{i2}, \ldots, G_{im})=p(D| G_1, G_2, \ldots, G_n)$.

Property 3.2 If any two groups of genotype variables derived from a partition of n genotype variables G_1, G_2, …, G_n are not conditionally independent on a disease status variable D, then any k ($3\le k\le n$) groups of variables derived from a partition of them are also not conditionally independent on D.

Proof: Assume that there are k groups of variables S_1, …, S_k derived from a partition of G_1, G_2, …, G_n which are conditionally independent on D ,where $S_1 \cup \ldots \cup S_k = \{G_1, G_2, \ldots, G_n\}$, then

$$p(g_1, g_2, \ldots, g_n |D)= p(s_1 |D)\, p(s_2 |D)\ldots p(s_k|D),$$

where s_i is any combination of values of S_i ($1\le i\le k$),

$$\sum_{s_1} p(g_1, g_2, \ldots, g_n \mid c) = \sum_{s_1} p(s_1 \mid c)\ p(s_2 \mid c)\cdots p(s_k \mid c),$$

$$p(s_2, \ldots, s_k |D)= p(s_2 |D)\ldots p(s_k|D),$$

so $p(g_1, g_2, \ldots, g_n |D)= p(s_1 |D)\, p(s_2, \ldots, s_k |D)$.

This contradicts with the assumption that any two groups of variables derived from a partition of G_1, G_2, …, G_n are not conditionally independent on D.

According to property 3.1 and property 3.2, we have the following equivalent definition of definition 3.2.

Definition 3.2′ n genotype variables G_1, G_2, …, G_n form an interaction group on a

disease status variable D if any two groups of variables derived from a partition of them are not conditionally independent on D and for any G_i of them, the following equality is violated:

$$p(d| g_1,\ldots g_{i-1}, g_{i+1}\ldots, g_n)=p(d| g_1, g_2, \ldots, g_n).$$

Property 3.3 If n random variables $X_1, X_2, \ldots, X_n$ are conditionally independent on a random variable C, then any subset $X_{i1}, X_{i2}, \ldots, X_{is}$ of them have no interaction on C.

Proof: Since $p(x_1, x_2, \ldots, x_n |c)= p(x_1 |c)\, p(x_2 |c)\ldots p(x_n|c)$, we have

$p(x_{i1}, x_{i2}, \ldots, x_{is} |c)= \sum_{x_{j1}, x_{j2}, \ldots, x_{jt}} p(x_1, x_2, \cdots, x_n \mid c) =$

$\sum_{x_{j1}, x_{j2}, \ldots, x_{jt}} p(x_1 \mid c)p(x_2 \mid c)\ldots p(x_n \mid c)\, p(x_{i1},x_{i2},\ldots,x_{is}|c)= p(x_{i1}|c)\, p(x_{i2}|c)\ldots p(x_{is}|c),$

where $\{ X_{j1}, X_{j2}, \ldots, X_{jt}\}=\{ X_1, X_2, \ldots, X_n\}-\{ X_{i1}, X_{i2}, \ldots, X_{is}\}$.

Therefore $X_{i1}, X_{i2}, \ldots, X_{is}$ have no interaction on C.

3.3.3 Measurement of Gene-Gene Interaction

The interaction among n genotype variables $G_1, G_2, \ldots, G_n$ given a disease status variable D can be tested using a χ^2 test. The degree of freedom of the χ^2 test can be determined using the limit theorem of χ^2 statistic due to K. Pearson and its generalization due to R. A. Fisher [34]. To do so, we first need to prove the following property.

Property 3.4 For a subset $G_{i1}, G_{i\,2}, \ldots, G_{i\,m}$ $(1\le m\le n,\ 1\le i1,i2,\ldots,\ im\le n)$ of n genotype variables $G_1, G_2, \ldots, G_n$, $p(D| G_1, G_2, \ldots, G_n)=p(D|G_{i1}, G_{i2}, \ldots, G_{im})$ and $p(G_{i1}, G_{i2}, \ldots, G_{im}|D)=p(G_{i1} |D)\ p(G_{i2} |D)\ldots p(G_{im}|D)$ iff $p(G_1, G_2, \ldots, G_n|D)= \frac{p(G_1,\ldots,G_n)}{p(G_{i1},\ldots,G_{im})}\, p(G_{i1} \mid D)\ldots p(G_{im} \mid D)$.

Proof:

If $p(D| G_1, G_2, \ldots, G_n)=p(D|G_{i1}, G_{i2}, \ldots, G_{im})$ and $p(G_{i1}, G_{i2}, \ldots, G_{im}|D)= p(G_{i1} |D)\ p(G_{i2} |D)\ldots p(G_{im}|D)$, then

$p(G_1, G_2, \ldots, G_n|D)= \frac{p(G_1,\ldots,G_n)p(D \mid G_1,\ldots,G_n)}{p(D)} = \frac{p(G_1,\ldots,G_n)p(D \mid G_{i1},\ldots,G_{im})}{p(D)} =$

$$\frac{p(G_1,\ldots,G_n)\dfrac{p(D)p(G_{i1},\ldots,G_{im}\mid D)}{p(G_{i1},\ldots,G_{im})}}{p(D)} = \frac{p(G_1,\ldots,G_n)}{p(G_{i1},\ldots,G_{im})}\,p(G_{i1}\mid D)\ldots p(G_{im}\mid D).$$

Conversely, if $p(G_1, G_2, \ldots, G_n|D)= \frac{p(G_1,\ldots,G_n)}{p(G_{i1},\ldots,G_{im})}\,p(G_{i1}\mid D)\ldots p(G_{im}\mid D)$, then

$$\sum_{G_{j1},\ldots,G_{jk}} p(G_1,\ldots,G_n\mid D) = \sum_{G_{j1},\ldots,G_{jk}} \frac{p(G_1,\ldots,G_n)}{p(G_{i1},\ldots,G_{im})}\,p(G_{i1}\mid D)\ldots p(G_{im}\mid D)$$

where $\{G_{j1}, G_{j2}, \ldots, G_{jk}\}=\{ G_1, G_2, \ldots, G_n \}-\{G_{i1}, G_{i2}, \ldots, G_{im}\}$, so

$p(G_{i1}, G_{i2}, \ldots, G_{im}|D)= p(G_{i1} |D)\, p(G_{i2} |D)\ldots p(G_{im}|D)$

In addition，by using the equation above, we have

$$p(D|\ G_1,\ G_2,\ \ldots,\ G_n)= \frac{p(D)p(G_1,\ldots,G_n\mid D)}{p(G_1,\ldots,G_n)} = \frac{p(D)}{p(G_1,\ldots,G_n)}\,\frac{p(G_1,\ldots,G_n)}{p(G_{i1},\ldots,G_{im})}$$

$$p(G_{i1}D)\ldots p(G_{im}D)=\frac{p(D)p(G_{i1},\ldots,G_{im}\mid D)}{p(G_{i1},\ldots,G_{im})}=\frac{p(D,G_{i1},\ldots,G_{im})}{p(G_{i1},\ldots,G_{im})}=p(D|G_{i1}, G_{i2}, \ldots,G_{im}).\ \square$$

If $\max\limits_{1\le m\le n}\{I(G_{i1};D)+I(G_{i2};D)+\ldots+ I(G_{im};D)-I(G_{im;}\ G_{i1}\ , \ldots, G_{im\text{-}1})- I(G_{im\text{-}1;}\ G_{i1},\ldots, G_{im\text{-}2})\ \text{-}\ldots\text{-}\ I(G_{i2;}\ G_{i1})\}= I(G_1;D)+ I(G_2;D)+\ldots+ I(G_n;D)-I(G_{n;}\ G_1, , \ldots, G_{n\text{-}1})- I(G_{n\text{-}1;}\ G_1, , \ldots, G_{n\text{-}2})\ \text{-}\ldots\text{-}\ I(G_2;\ G_1)$, $(1\le\ m\le n,\ 1\le i1,i2,\ldots,\ im\le n)$, we need to test $p(G_1, G_2, \ldots, G_n |D)= p(G_1 |D)\, p(G_2 |D)\ldots p(G_n|D)$. In order to use χ^2 test, we transform it to $p(G_1, G_2, \ldots, G_n ,D)= p(D)p(G_1 |D)\, p(G_2 |D)\ldots p(G_n|D)$. The test statistic is:

$$\chi^2=\sum_{i_0=1}^{m_0}\sum_{i_1=1}^{m_1}\cdots\sum_{i_n=1}^{m_n}\frac{(k_{i_0i_1\ldots i_n}\ -\ k\,\hat{p}_{i_0}\,\hat{p}_{i_1|i_0}\ldots\hat{p}_{i_n|i_0})^2}{k\,\hat{p}_{i_0}\,\hat{p}_{i_1|i_0}\ldots\hat{p}_{i_n|i_0}} \qquad (3.15)$$

where $m_0,m_1,\ldots,m_n$ are the numbers of values of $D,G_1, G_2, \ldots, G_n$ respectively, k is the total number of samples, $k_{i_0i_1\ldots i_n}$ is the number of samples when $(D, G_1, \ldots, G_n)=(i_0, i_1,\ldots, i_n)$, $\hat{p}_{i_0}$, $\hat{p}_{i_j|i_0}$ are the maximum likelihood estimation of p_{i_0}, the probability of $D=i_0$, and $p_{i_j.|i_0}$, the probability of $G_j=i_j$ given $D= i_0$, respectively $(1\le j\le n)$.

Since there are $m_0(m_1+\ldots+ m_n\text{-}n)+m_0\text{-}1$ parameters in the test statistic, so in the limit, it obeys a χ^2 distribution with $m_0m_1\ldots m_n\text{-}1\text{-}[m_0(m_1+\ldots+ m_n\text{-}n)+m_0\text{-}1]= m_0m_1\ldots m_n\text{-}\ m_0(m_1+\ldots+ m_n)+(n\text{-}1)\ m_0$ degrees of freedom.

If $\max\limits_{1\le m\le n}\{ I(G_{i1};D)+ I(G_{i2};D)+\ldots+ I(G_{im};D)-I(G_{im;}\ G_{i1}, , \ldots, G_{im\text{-}1})- I(G_{im\text{-}1;}\ G_{i1}, , \ldots, G_{im\text{-}2})\ \text{-}\ldots\text{-}\ I(G_{i2;}\ G_{i1})\}= I(G_j;D)$, $1\le j\le n$, we need to test $p(D|G_1, G_2, \ldots, G_n)= p(D$

$|G_j$). In order to use χ^2 test, we transform it to $p(D, G_1, G_2, \ldots, G_n) = p(G_1, G_2, \ldots, G_n)\, p(D|G_j)$.The test statistic is:

$$\chi^2 = \sum_{i_0=1}^{m_0} \sum_{i_1=1}^{m_1} \cdots \sum_{i_n=1}^{m_n} \frac{(k_{i_0 i_1 \ldots i_n} - k \hat{p}_{i_1 \ldots i_n} \hat{p}_{i_0|i_j})^2}{k \hat{p}_{i_1 \ldots i_n} \hat{p}_{i_0|i_j}} \tag{3.16}$$

Since there are $m_1 \ldots m_n$-1+ $m_j(m_0$-1) parameters in the test statistic, so in the limit, it obeys a χ^2 distribution with $m_0 m_1 \ldots m_n$-1-[$m_1 \ldots m_n$-1+ $m_j(m_0$-1)]=$m_0 m_1 \ldots m_n$-$m_1 \ldots m_n$ - $m_j m_0 + m_j$ degrees of freedom.

If $\max_{1 \le m \le n} \{I(G_{i1};D) + I(G_{i2};D) + \ldots + I(G_{im};D) - I(G_{im;}\ G_{i1}, , \ldots, G_{im\text{-}1}) - I(G_{im\text{-}1;}\ G_{i1}, , \ldots, G_{im\text{-}2}) - \ldots - I(G_{i2;}\ G_{i1})\} = I(G_{i1};D) + I(G_{i2};D) + \ldots + I(G_{im};D) - I(G_{im;}\ G_{i1}, , \ldots, G_{im\text{-}1}) - I(G_{im\text{-}1;}\ G_{i1}, , \ldots, G_{im\text{-}2}) - \ldots - I(G_{i2;}\ G_{i1})$, (1<$m$<n), we need to test $p(D|\ G_1, G_2, \ldots, G_n) = p(D|G_{i1,}\ G_{i2,}\ \ldots,\ G_{im})$ and $p(G_{i1,}\ G_{i2,}\ \ldots,\ G_{im}|D)\ p(G_{i1}\ |D)\ p(G_{i2}\ |D) \ldots p(G_{in}|D)$. In order to use χ^2 test, according property 4, we transform it to $p(G_{1,}\ G_{2,}\ \ldots,\ G_n|D) = \frac{p(G_1, \ldots, G_n)}{p(G_{i1}, \ldots, G_{im})}\, p(G_{i1} \mid D) \ldots p(G_{im} \mid D)$, then to $p(D\ , G_{1,}\ G_{2,}\ \ldots,\ G_n) = p(G_{j1}, \ldots, G_{j(n-m)} \mid G_{i1}, \ldots, G_{im}) p(G_{i1} \mid D) \ldots p(G_{im} \mid c) p(D)$, $\{G_{j1,}\ G_{j2,}\ \ldots,\ G_{j(n\text{-}m)}\} = \{\ G_{1,}\ G_{2,}\ \ldots, G_n\ \} - \{G_{i1,}\ G_{i2,}\ \ldots, G_{im}\}$. The test statistic is:

$$\chi^2 = \sum_{i_0=1}^{m_0} \sum_{i_1=1}^{m_1} \cdots \sum_{i_n=1}^{m_n} \frac{(k_{i_0 i_1 \ldots i_n} - k \hat{p}_{i_0} \frac{\hat{p}_{i_1, \ldots, i_n}}{\hat{p}_{i_{i1}, \ldots, i_{im}}} \hat{p}_{i_{i1}|i_0} \cdots \hat{p}_{i_{im}|i_0})^2}{k \hat{p}_{i_0} \frac{\hat{p}_{i_1, \ldots, i_n}}{\hat{p}_{i_{i1}, \ldots, i_{im}}} \hat{p}_{i_1|i_0} \cdots \hat{p}_{i_n|i_0}} \tag{3.17}$$

Since there are $m_{i1} \ldots m_{im}$ ($m_{j1} \ldots m_{j(n\text{-}m)}$-1)+$m_0$ (m_{i1}+…+ m_{im}-m)+m_0-1 parameters in the test statistic, so in the limit, it obeys a χ^2 distribution with $m_0 m_1 \ldots m_n$-1-[$m_{i1} \ldots m_{im}$ ($m_{j1} \ldots m_{j(n\text{-}m)}$-1)+$m_0$ (m_{i1}+…+ m_{im}-m)+m_0-1] degrees of freedom. Actually the formula of degrees of freedom for the second case is a special case of this formula.

3.3.4 An Algorithm to Detect High Order Gene-Gene Interaction

Before we present the algorithm, We first prove the following property and theorem.

Property 3.5 If two groups of genotype variables $\{G_{11}, G_{12}, \ldots, G_{1m}\}$ and $\{G_{21}, G_{22}, \ldots, G_{2n}\}$ are conditionally independent of a disease status variable D, then any two variables from different groups have no interaction on D.

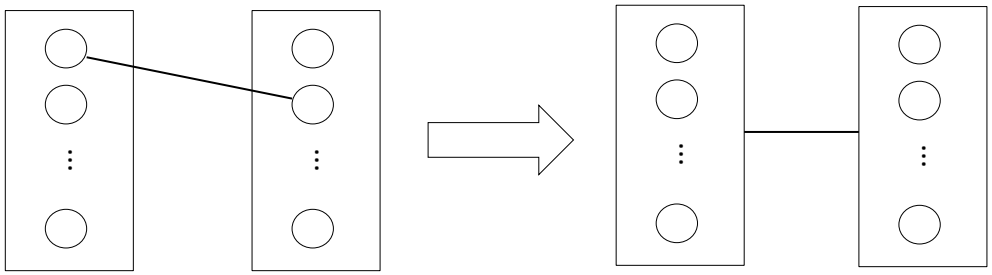

Figure. 3.2 The graphical model representation of property 3.5. (A node represents a genotype variable and an edge represents conditional independency.)

Proof:

Let G_1' denote "$G_{11},\ldots,G_{1,i-1},G_{1,i+1},\ldots,G_{1m}$", G_2' denote "$G_{21},\ldots,G_{2,j-1},G_{2,j+1},\ldots,G_{2n}$". Since $\{G_{11},G_{12},\ldots,G_{1m}\}$ and $\{G_{21},G_{22},\ldots,G_{2n}\}$ have no interaction on D, we have $p(G_{11},G_{12},\ldots,G_{1m}, G_{21},G_{22},\ldots,G_{2n}|D) = p(G_{11},G_{12},\ldots,G_{1m}|D)\, p(G_{21},G_{22},\ldots,G_{2n}|D)$. So for any $G_{1i} \in \{G_{11}, G_{12}, \ldots, G_{1m}\}$, $G_{2j} \in \{G_{21}, G_{22}, \ldots, G_{2n}\}$,

$$\sum_{G_1', G_2'} p(G_{11}, G_{12}, \ldots, G_{1m}, G_{21}, G_{22}, \ldots, G_{2n} \mid D) =$$

$$\sum_{G_1', G_2'} p(G_{11}, G_{12}, \ldots, G_{1m} \mid D)\, p(G_{21}, G_{22}, \ldots, G_{2n} \mid D) =$$

$$\sum_{G_1'} p(G_{11}, G_{12}, \ldots, G_{1m} \mid D) \sum_{G_2'} p(G_{21}, G_{22}, \ldots, G_{2n} \mid D).$$

Therefore $p(G_{1i}, G_{2j}|D) = p(G_{1i}|D)p(G_{2j}|D)$. So G_{1i} and G_{2j} have no interaction on D. □

Theorem 3.1 For n genotype variables $G_1, G_2, \ldots, G_n$, if G_1 and G_2 are not conditionally independent of a disease status variable D, G_3 and one of $\{G_1, G_2\}$ are not conditionally independent of $D,\ldots,G_n$ and one of $\{G_1,G_2,\ldots,G_{n-1}\}$ are not conditionally independent of D, for any G_i of $G_1, G_2, \ldots, G_n$, the following equality is violated:

$$p(D|G_1,\ldots G_{i-1},G_{i+1}\ldots,G_n)=p(D|G_1,G_2,\ldots,G_n),$$

then G_1, G_2, …, G_n form an interaction group.

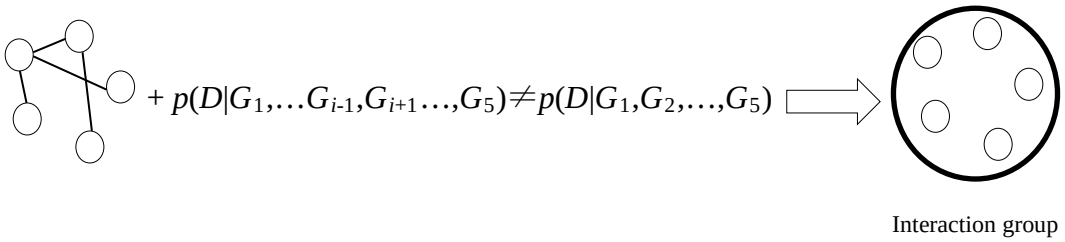

Figure. 3.3 The graphical model representation of Theorem 3.1. (A node represents a genotype variable and an edge represents conditional dependency.)

Proof: If we use a node to represent a genotype variable from G_1, G_2, …, G_n , an edge to mean that the two variables represented by the two nodes connected by the edge are not conditionally independent of D, then nodes representing G_1, G_2, …, G_n respectively and corresponding edges form a connected graph.

If two groups of variables derived from a partition of G_1, G_2, …, G_n are conditionally independent on D, then according to property 3.5, any two random variables from different groups are also conditionally independent on D, therefore these two groups are not connected, a contradiction. □

To search for interaction of more than two genotypes variables on a disease status variable, many methods, such as famous MDR (Ritchie *et al*., 2001; Hahn *et al*., 2003; Nunkesser *et al*., 2007), search all combinations of a fixed number of genotype variables. This is quite time consuming. Theorem 3.1 sheds light to an easy way to find interaction of more than two genotype variables on a disease status variable.

The algorithm which we call CIR based algorithm first finds two genotype variables which have the biggest interaction on a disease status variable D among all pairs of genotype variables in a set of genotype variables A in a training dataset S, i.e.,

$$\arg\max_{i\neq j}\{I(G_i,G_j;D)-\max\{I(G_i;D)+I(G_j;D)-I(G_i;G_j), I(G_i;D), I(G_j;D)\}\} \quad (3.18)$$

and add them to a set B which is initialized as empty. Then find the next variable G_k in A and not in B according to the following criteria and add it to B:

$$\max_k\{\max_i\{I(G_i,G_k;D)-\max\{I(G_i;D)+I(G_k;D)-I(G_i;G_k), I(G_i;D), I(G_k;D)\}\min_j$$

$$\{I(G_j,G_k;D)-I(G_j;D)\}\} \quad (3.19)$$

where G_i, $G_j \in B$, $G_k \in A$, $G_k \notin B$.

The above process is repeated until the required number of attributes is selected or the value of the formula (3.19) is smaller than a threshold.

A complete description of the CIR based algorithm is shown in Table 3.1.

Since directly reducing redundancy among n variables on D is difficult and unreliable, we approximate this goal by reducing redundancy between two variables on D in (3.19), i.e. $\min_j \{ I(G_j,G_k;D) - I(G_j;D)\}$ is used.

Using this method, only all combinations of two genotype variables need to be searched when gene-gen interactions are being identified.

Table 3.1 CIR based algorithm

Algorithm 1: CIR based algorithm

Input: *attributenumber*, *interactionorder*,*threshold*,
a training dataset S

Output: B

1: Initialize B as empty.

2: $(optimalsnp1, optimalsnp2) \leftarrow \arg\max_{1\le i\le attributenumber, i+1\le j\le attributenumber}$
$\{I(G_i,G_j;D) - \max\{I(G_i;D)+I(G_j;D)-I(G_i;G_j), I(G_i;D), I(G_j;D)\}\}$

3: add *optimalsnp1*, *optimalsnp2* to B

4: $n=2$

5: repeat

6: $maxinteraction \leftarrow \max_{1\le k\le attributenumber, k\notin B} \{\max_{i\in B} \{I(G_i,G_k;D) - \max\{I(G_i;D)+I(G_k;D)-I(G_i;G_k)$
$I(G_i;D), I(G_k;D)\}\} \min_{j\in B} \{ I(G_j,G_k;D) - I(G_j;D)\}\}$

7: $optimalsnp \leftarrow \arg\max_{1\le k\le attributenumber, k\notin B} \{\max_{i\in B} \{I(G_i,G_k;D) - \max\{I(G_i;D)+I(G_k;D)-I(G_i;G_k)$
$I(G_i;D), I(G_k;D)\}\}\min_{j\in B} \{ I(G_j,G_k;D) - I(G_j;D)\}\}$

8: add *optimalsnp* to B

9: $n=n+1$

10: until $n=$ *interactionorder* or *maxinteraction*< *threshold*

11: return B

For computational efficiency of the algorithm，if we want to find a k-order interaction group, interaction between any two genotype variables needs to be computed when searching the first two interactive genotype variables. The complexity is $O(mn^2)$, where m is the number of genotype variables and n is the sample size. Then

these computed interactions can be used to find the third until the kth genotype variable in the remaining variables. Since *k* is much smaller than *m*, the complexity is O(*m*). So the overall complexity is O(mn^2), not relevant to the order *k*. Therefore the time costs for finding interaction groups of different orders are close.

3.4 Experimental results and analysis

3.4.1. Experiments on Simulated Datasets

3.4.1.1. Program gs 2.0

Since the true risk of SNPs for most complex diseases are unknown, real world data is not especially useful for assessing performance. For this reason, most approaches are evaluated based on experiments using realistically simulated data for performance evaluation (Assareh *et al*., 2012). Here we use the program gs 2.0 (Li and Chen, 2008; Chen and Li, 2012) to generate simulated data to test the usefulness of the new definition of gene-gene interaction and the performance of the proposed CIR based algorithm.

The program gs 2.0, can quickly generate a large number of samples based on real data that share similar local linkage disequilibrium (LD) patterns as those that can be found in human populations. It is aimed at providing a public available program to compare results from different research groups.. It can be used to implement various interaction models.

Two heuristic methods have been used to generate samples with haplotype/ genotype data. One generate samples from haplotype pairs and the other from patterns of haplotype block structures.

In the first approach, a disease model is first created by using the disease allele frequency (DAF) and the penetrance of each genotype or alternatively the population

prevalence and genotype relative risks. There is a simple relationship between these two sets of parameters. Then a SNP t with the frequency of one allele approximately equal to the specified DAF is selected from the input data such as the haplotype results from the HapMap project or alternatively a SNP at a particular locus can be specified as the disease susceptibility locus. Its genotype g is generated based on the conditional probability of each genotype given that the disease is present:

$$\Pr(g_i|\text{case})=\Pr(g_i)\Pr(\text{case}|\, g_i)/\ \Pr(g_j)\ \Pr(\text{case}|\, g_j),$$

where $\Pr(g_i)$ is the frequency of genotype g_i obtained from allele frequencies under the assumption of Hardy-Weinberg equilibrium and $\Pr(\text{case}|\, g_i)$, the probability of a case given a specific genotype is a user specified penetrance parameter. The haplotype pairs h_1 and h_2 for this case are generated by randomly selecting two haplotypes h_3 and h_4 from the inputs having genotype equal to g at the disease locus t. The haplotype h_1 has the same alleles as h_3 from locus t-l_l to t+l_r, where $l_{min} \le l_l$,$l_r \le l_{max}$ with l_{min} and l_{max} being specified by users. The values of l_l and l_r are specified based on the strength of local LD.The value of l_r is set to l_{min} at first, then it will be increased by 1 continuously until the LD measure D' between locus t+l_r and locus t+l_r+1 is smaller than a random number which follows a uniformly distribution between 0 and 1 or when l_r=l_{max}. The values of l_l can be obtained in a same way on the opposite direction. The haplotype h_2 can be determined similarly as h_1. The process can be repeated to generate the specified number of cases. Normal individuals can be generated in the same way based on the genotype's conditional probability given that the disease is absent. Two parameters l_{min} and l_{max} are adopted to make it possible to consider both long-range LD and short-range LD.

For dense SNPs, a block-like structure of LD patterns is common. Therefore in the second approach, the haplotype block structures rather than haplotype pairs are used as inputs. Each block is a Markov chain state consisting of several common haplotypes with their population frequencies. A transition probability matrix describes the connection patterns between haplotypes in adjacent blocks. A pair of common haplotypes with the genotype at the disease locus generated based on the conditional probability will be selected based on their frequency distribution. Then they will be

extended independently to both directions according the transition probabilities. The process will be repeated to generate a required number of samples having similar LD patterns with real data but different haplotypes and genotypes. SNPs that are not in any blocks and rare haplotypes not appearing the input block file are also considered to maintain a proper level of variety.

These two approaches can be extended to multi-locus disease model in a similar manner.

3.4.1.2. Evaluation of the New Definition of Gene-Gene Interaction

In the first experiment, two different two locus models, the threshold model and the exclusive OR model (Table 3.2), were simulated. For each model, one pair of SNPs was simulated as a causal factor among all possible combinations. Minor allele frequencies (MAF) and effect size (θ) varied with fixed sample size (200 cases and 200 controls), SNP number (882 SNPs) and baseline (α=0.01). Hit ratio which is defined as the proportion of replicated datasets with which the true causal SNPs are detected as the best SNPs among all possible same number of SNPs is used to measure the effectiveness of the new definition of gene-gene interaction. Hit ratios of CIR based definition are compared with that of IG based definition of interaction.

Table 3.2 Penetrance table for two two-locus interaction models

Threshold	BB	Bb	Bb
AA	α	α	α
Aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
Aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
Exclusive OR	**BB**	**Bb**	**Bb**
AA	α	α	$\alpha(1+\theta)$
Aa	α	α	$\alpha(1+\theta)$
Aa	$\alpha(1+\theta)$	$\alpha(1+\theta)$	α

Table 3.3 and Table 3.4 are the results of the first experiment.

From these two tables, we can see that except for MAF=0.1, in most cases, these two models can be detected with high hit ratios with CIR based definition and hit ratios of CIR based definition are generally higher than that of IG based definition.

Table 3.3 Hit ratios for threshold model

		MAF		
θ	Method	0.1	0.3	0.5
9	CIR	0	0.41	0.94
	IG	0	0.1	0.47
19	CIR	0	0.48	0.93
	IG	0	0.13	0.61
49	CIR	0.05	0.7	0.93
	IG	0	0.1	0.79

Table 3.4 Hit ratios for exclusive or model

		MAF		
θ	Method	0.1	0.3	0.5
9	CIR	0	0.59	0.99
	IG	0	0.4	0.99
19	CIR	0	0.59	1
	IG	0	0.53	1
49	CIR	0.01	0.71	1
	IG	0	0.75	1

3.4.1.3. Type I error

To determine type I error rates, the null datasets with no causal pair of SNPs were simulated for different sample sizes (n=200, 400 and 800) and different SNP numbers (m=10, 20, 30). Permutation P values of the identified strongest interaction pair of

SNPs were calculated by permuting disease status of each dataset 1000 times. The ratio of the permutation *P* values smaller than the significance level α=0.05 in 1000 replicates is calculated as the type I error rate. The number of the permutation ensured its accuracy to one decimal place when expressed in percent.

Results given in Table 3.5 show that CIR based definition has type I error rates tightly gathering around 5% with a range from 4.2% to 5.7%, better than that of IG based definition (from 3.5% to 5.4%). Therefore CIR controls type I error better.

TABLE 3.5 Type I error rate with the significance level α of 0.05 from datasets with 1000 replicates

		n		
m	Method	200	400	800
10	CIR	5%	4.2%	5.1%
	IG	4.2%	3.5%	5.4%
20	CIR	5.3%	5.7%	5%
	IG	4.8%	5.3%	4.8%
30	CIR	5.4%	5.6%	4.2%
	IG	5.4%	5.3%	3.9%

3.4.1.4. Evaluation of the Proposed Algorithm

In the third experiment, two three-locus epistasis models (Table 3.6 and Table 3.7) were simulated. Three SNPs were simulated as causal SNPs among all possible combinations. MAF and effect size (θ) varied with fixed sample size (100 cases and 100 controls), SNP number (441 SNPs) and baseline (α=0.01). The performance of CIR based algorithm is compared with that of IG based algorithm (using IG to measure interaction) and MDR method. Figure 3.4 and Fig. 3.5 are the results of the third experiment.

Also, from these two figures, we can see that except for MAF=0.1, in most cases, these two three-locus epistasis models can be detected with relatively high hit ratios in general and the performance of CIR based algorithm is better than that of IG based

Table 3.6 Penetrance table for the first three-locus interaction model

Genotype		BB		Bb		bb
AA	CC	α	CC	α	CC	α
	Cc	α	Cc	α	Cc	α
	cc	α	Cc	α	cc	α(1+θ)
Aa	CC	α	CC	α	CC	α
	Cc	α	Cc	α(1+θ)	Cc	α
	cc	α	Cc	α	cc	α
aa	CC	α(1+θ)	CC	α	CC	α
	Cc	α	Cc	α	Cc	α
	cc	α	Cc	α	cc	α

Table 3.7 Penetrance table for the second three-locus interaction model

Genotype		BB		Bb		bb
AA	CC	α	CC	α	CC	α(1+θ)
	Cc	α	Cc	α	Cc	α
	cc	α(1+θ)	cc	α	cc	α
Aa	CC	α	CC	α	CC	α(1+θ)
	Cc	α	Cc	α(1+θ)	Cc	α
	cc	α	cc	α	cc	α
aa	CC	α	CC	α	CC	α
	Cc	α	Cc	α(1+θ)	Cc	α
	cc	α(1+θ)	cc	α	cc	α

algorithm and MDR in most cases (in the figures the value is 0 where there is no rectangle). Actually for Figure 3.4, besides for MAF=0.1, θ=9 and 19 where all three

methods have 0 hit ratios, MDR also has 0 hit ratios in other five scenarios, therefore the proposed algorithm is more powerful than MDR in nine scenarios in the remaining thirteen scenarios. Similarly for Figure 3.5, the proposed algorithm is more powerful than MDR in a majority of scenarios. In addition the ratio of execution time of MDR versus CIR based algorithm is about 150:1, so CIR based algorithm is much more efficient than MDR.

The proposed algorithm is based on a more reasonable definition of gene-gene interaction which would increase hit ratio and utilizes the relation between high order interaction and low order interaction under some conditions and search high order interactions by searching low order interactions which would improve efficiency but may decrease effectiveness, whereas MDR searches high order interactions directly. Therefore MDR may occasionally attain a higher hit ratio than the proposed algorithm.

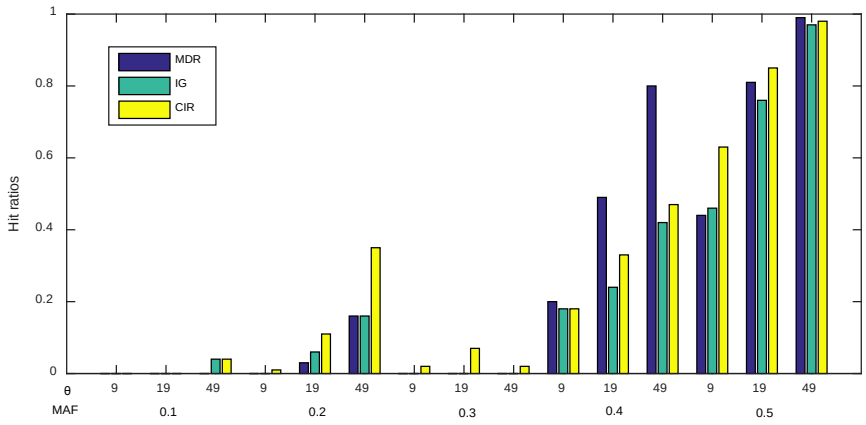

Figure 3.4 Hit ratios for the first three locus epistasis model

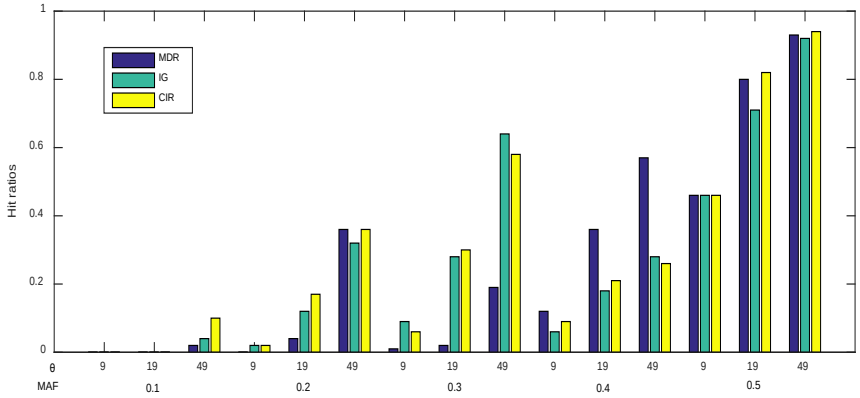

Figure 3.5 Hit ratios for the second three loci epistasis model

The reason for the peak to appear in Figure 3.2 at MAF=0.2 is that for the first three locus epistasis model, when MAF=0.3, two of the three causal markers have neighboring loci, leading to strong linkage disequilibrium, i.e., strong redundancy and therefore resulting in low hit ratios for all algorithms.

Table 3.8 and Table 3.9 show the average ratios of the values in (3.19) of the fourth selected gene to the third selected gene for 100 replicates when the true causal SNPs are selected for these two models. These ratios are much smaller than one in most cases. So in these cases three SNPs rather than other number of SNPs are selected as causal SNPs.

Table 3.8 Average ratios for the first three locus epistasis model

	MAF				
θ	0.1	0.2	0.3	0.4	0.5
9	-*	0.71	0.81	0.28	0.28
19	-	0.99	1.49	0.13	0.07
49	0.63	0.14	0.87	0.09	0.05

"-" represents that hit ratios are 0s in these cases, so there are no such average ratios.

Table 3.9 Average ratios for the second three locus epistasis model

	MAF				
θ	0.1	0.2	0.3	0.4	0.5
9	-*	0.89	0.67	0.57	0.83
19	-	0.30	0.82	0.18	0.73
49	0.551	0.25	0.34	0.26	0.24

"-" represents that hit ratios are 0s in these cases, so there are no such average ratios.

To test the sensitivity of the algorithm scale with the order of interaction, we have done experiments to search for interaction groups with orders ranging from 3 to 6, MAF=0.2 and 0.4, θ=19 and 49, sample size=400 (200 cases and 200 controls). The high risk genotype combinations are 133, 222 and 311 for the 3-order interaction model, 1133, 2222 and 3311 for the 4-order interaction model, 11333, 22222 and

33111 for the 5-order interaction model, 111333, 222222 and 333111 for the 6-order interaction model respectively, where “1” represents the common homogeneous genotype, “2”, the heterogeneous genotype and “3”, the minor homogeneous genotype. Figure 3.6 shows the hit-ratios for these models.

Generally the hit ratios decrease with the increase of the order of interaction and the decrease of MAF or θ. The computational costs are close for different orders.

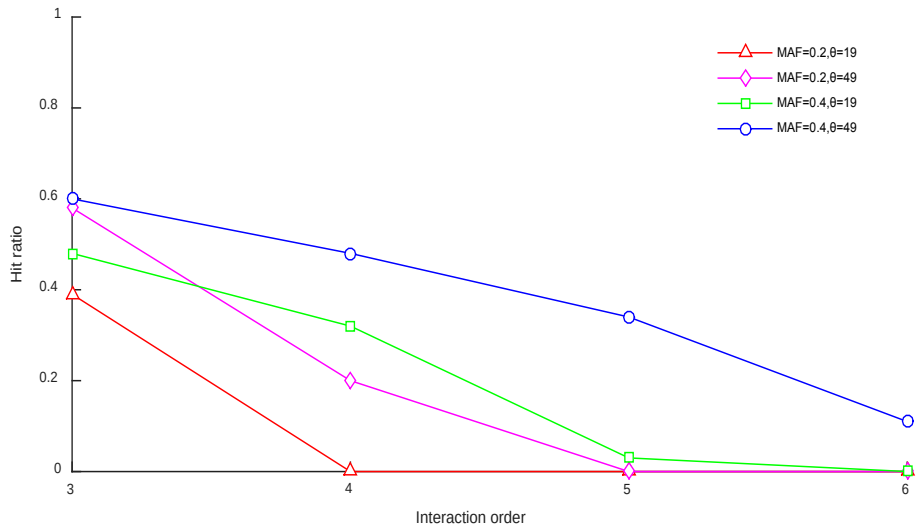

Figure 3.6 Hit ratios for different orders of epistasis models

3.4.2. Experiments on a Real Datasets

A real data set of malaria cohort study, conducted by Williams *et al* in Kenya (Williams *et al*, 2005) was used to further show the effectiveness of the proposed new definition. Two conditions of the hemoglobinopathies were previously found to protect against severe and fatal P. falciparum malaria. One is structural variant hemoglobin S: heterozygote HbAS (homozygote HbSS is not considered since it can lead to premature death) and the other is reduced production of the normal α-globin component of hemoglobin， α^{+}-thalassemia, which is caused by two variants: heterozygote -α/αα and homozygote -α/-α. However a negative epistatic interaction was found between HbAS and a+-thalassemia on malaria infection (Table 3.10).

From Table 3.10, we can get MAF for α-globin component is 0.407, while for hemoglobin, since homozygote HbSS is not considered, the percentage of HbAS can be computed, which is 0.147. Both are above 0.1, so it's appropriate to test interaction here.

Let G_1 denote hemoglobin type, G_2 denote α^+-thalassemia genotype, D denote malaria infection status.

For malaria admission, $I(G_1;D)=0.009979$, $I(G_2;D)=0.001299$, $I(G_1;D)+I(G_2;D)-I(G_1;G_2)=0.01108$, $\max\{I(G_1;D), I(G_2;D), I(G_1;D)+I(G_2;D)-I(G_1; G_2)\}=I(G_1;D)+I(G_2;D)-I(G_1; G_2)$, therefore $p(G_1\ G_2)|D)=p(G_1|D)p(G_2)|D)$ should be tested. The corresponding χ^2 value is 18.08, the degree of freedom is 2×2×3-2×(2+3)+2=4, so P=0.00119.

Table 3.10 Malaria admission and severe malaria by hemoglobin type and a+-thalassemia genotype. (Williams *et al*, 2005).

Hb	α-globin component	*n*	Malaria admission	Severe Malaria
HbAA	αα/αα	626	168	67
	-α/αα	867	187	53
	-α/-α	302	56	17
HbAS	αα/αα	113	6	0
	-α/αα	150	9	2
	-α/-α	46	10	5

For severe malaria, $I(G_1;D)=0.2643$, $I(G_2;D)=0.2628$, $I(G_1;D)+I(G_2;D)-I(G_1;G_2)=0.5269$, $\max\{I(G_1;D), I(G_2;D), I(G_1;D)+ I(G_2;D)-I(G_1; G_2)\}= I(G_1;D)+ I(G_2;D)-I(G_1;G_2)$, therefore again $p(G_1\ G_2|D)=p(G_1|D)\ p(G_2|D)$ should be tested. The corresponding χ^2 value is 21.81, the degree of freedom is also 4, so P=0.000219.

Both P values computed by our proposed measure of gene-gene interaction are much smaller than those of Wald test for interaction given in Williams *et al*, 2005 (P values are 0.026 and 0.0012 respectively), providing more confidence to rejecting the

null hypothesis that there is no interaction between hemoglobin type and α^{+}-thalassemia genotype on malaria infection.

For MDR and IG there are no statistics by now to test the null hypothesis that there is no interaction among a set of markers. This is also an advantage of the proposed approach over MDR and IG.

3.5. Conclusion

In this chapter, we present a new definition of gene-gene interaction according to an inequality and the corresponding definition of an interaction group. We identify three cases where there is no interaction among genes. Based on these new definitions, we also derive a statistic to measure gene-gene interaction. Experimental results using the proposed definition of gene-gene interaction with simulated data show that our new definition of gene-gene interaction can effectively identify two locus gene-gene interaction models among a large number of SNPs. The experiments with the real data sets also show the effectiveness of our proposed measure of gene-gene interaction. With the increase of the number of interaction genes, the number of possible combinations of interaction genes increases exponentially, and the number of sparse cells also increases. A new algorithm is therefore proposed based on the new definition of gene-gene interaction to detect high order gene-gene interactions. Experimental results show that this algorithm can effectively detect many high order gene-gene interactions with high efficiency.

If n genotype variables $G_1, G_2, \ldots, G_n$ are independent of a disease status variable D, i.e. $p(D| G_1, G_2, \ldots, G_n)=p(D)$, then we can prove that any subset $G_{i1}, G_{i2}, \ldots, G_{is}$ of them are also independent of D. Therefore, it is appropriate to define $G_1, G_2, \ldots, G_n$ to have no interaction on D when they are independent of D. However in this case, we have $p(D| G_1)=p(D)=p(D| G_1, G_2, \ldots, G_n)$, the condition (2) in definition 1 is satisfied, so it is not listed out separately in definition 1.

In the proposed CIR based algorithm, we use an approximate method to reduce

redundancy among genes by reducing redundancy between two genes. Experiment results show that it greatly increases hit ratios. Unlike (Ding and Peng, 2005) where $I(G_i,G_j)$ is used to measure redundancy between two genes G_i and G_j, we use $I(G_i,G_j;D)$-$I(G_i;D)$ instead, where G_i is a gene already selected, C is the disease status variable, and the performance is much better.

One limitation of the proposed algorithm is that it could not detect high-order pure epistasis. However, it can detect many other epistases with very high efficiency, as demonstrated in our experiments. Also high-order pure epistasis has not been identified by now and “lower-order effects” are considered by many existing work, including the ones described in (Shang *et al*., 2016), as effective ways to approximate effects of higher-order. In addition, we can detect high order epistasis with our proposed algorithm first, if no satisfactory result can be obtained, we can continue to detect pure epistasis with the definition directly, although it may require more computational effort

Chapter 4
An Extended Fuzzy Classification Method for Identifying Gene-Gene Interactions Associated with Complex Quantitative Traits

4.1 Introduction

Like complex diseases, complex quantitative traits (QTs) are also usually associated with genetic variants, mainly single nucleotide polymorphisms (SNPs) or simple sequence length polymorphic markers (SSLPs). The majority of innate and acquired body and behavioral characteristics such as height, weight, learning, memory and emotions, are complex traits. Many physiological characteristics such as blood pressure and body temperature are also reflected by complex traits. In addition, most diseases such as hypertension, diabetes, obesity, cancer and neuropsychiatric disorders exhibit various symptoms through complex traits.

In many cases, complex QTs with continuous outcomes can provide more accurate analysis.

The Multifactor Dimensionality Reduction (MDR) method was originally proposed as a nonparametric and model-free data reduction approach for identifying interactions without significant main effects and has been successfully applied to identify gene-gene interactions in many common complex diseases (Ritchie et al., 2001; Moore, 2004; Moore et al., 2006).

Some efforts have been made to extend MDR to QTs.

4.2 Related works

The combinatorial partitioning method (CPM) (Nelson et al., 2001) was proposed to identify sets partitions of multi-locus genotypes for predicting variation in quantitative trait levels.

Let M be a subset of L loci that are measured for a sample and corresponding lower letters denote their sizes, G_M denotes the set of *m*-locus genotypes with size g_M. Let *K* be a set of *k* genotypic partitions which is a partition of all the possible *m*-locus genotypes, $2 \leq k \leq g_M$. CPM searches over the state space made up of all possible sets of genotypic partitions of the G_M genotypes obtained from each subset *M* of *L* total loci to identify m loci that divide g_M genotypes into *k* partitions with the mean of a quantitative trait having most similar values within and most dissimilar values between partitions.

The process is composed of three steps.

The first step is to search all possible k sets of genotypic partitions that partition *m*-locus genotypes for all subsets of *l* loci with k ranging from 2 to g_M. The number of k sets of genotypic partitions is a Stirling number of the second kind:

$$S(g_M,k)=\frac{1}{k!}\sum_{i=0}^{k-1}(-1)^i\binom{k}{i}(k-i)^{g_M} \tag{4.1}$$

The sum of the squared deviations of the trait means of the partitions from the trait mean(SS_k) of overall sample is used as a statistical measure of phenotypic characteristics of each set of genotypic partitions to evaluate the state space. The value of this measure increases when the similarities of trait values within genotypic partitions increases and the differences between partitions increase. The partition sum of squares will increase with the increase of *k*, giving advantage to a greater number of partitions of genotypes. To compensate for this bias, a bias-corrected estimate of genotypic variance (Boerwinkle and Sing 1986) is used:

$$s_k^2=\sum_{i=1}^{k}\frac{n_i(\bar{Y}_i-\bar{Y})^2}{n}-\frac{(k-1)}{n}\sum_{i=1}^{k}\sum_{j=i}^{n_i}\frac{(Y_{ij}-\bar{Y}_i)^2}{n-k}=\frac{SS_k}{n}-\frac{(k-1)}{n}MS_W \tag{4.2}$$

where n is the sample size, $\bar{Y}$ is the sample mean, n_i and $\bar{Y}_i$ are the sample size and mean of partition I respectively, Y_{ij} is the phenotype of the jth individual in the ith partition, and MS_W is the mean squared estimate of the phenotypic variability among individuals within genotypic partitions. In this formula, the partition sum of squares increasing with k is penalizes by a term also increasing with k if the estimate of MS_W does not decrease as additional partitions are considered. The ratio of variability explained by a set of genotypic partitions can be computed as:

$$p\gamma_K = \frac{s_K^2}{s_P^2} = \frac{s_K^2}{s_K^2 + MS_W} \tag{4.3}$$

Some criterion (such the significance level of an F-test, biological significance or some proportion of all of the sets considered) is used as a filter to select sets of genotypic partitions for further consideration.

To control estimate bias of partition means and deviations caused by sparse partitions with a few individuals due to low frequency alleles, a partition with the number of individuals below a lower bound is filtered out.

The second step is to validate those retained sets of genotypic partitions. Multifold cross-validation is employed for validation. The predictive ability of a set of genotypic partitions is evaluated by the cross-validated proportion ($pv_{K,CV}$) of the trait variability it explains. Larger $pv_{K,CV}$ implicates more predictability of the set.

The third step is to select a subset of the validated sets of genotypic partitions as classifiers according to some criteria.

The Restricted Partition Method (RPM) is proposed to improve the CPM (Culverhouse et al., 2010). It detects multi-locus genotypes as predictors of a quantitative trait by a partitioning of genotypes into subgroups.

The CPM has two drawbacks. One is its prohibitive computational burden due to huge number of partitions possible with multiple loci when 3-way or above interactions are to be analyzed. Another one is its permutation testing method to evaluate the statistical significance of the models. Since this method needs many (usually 1000) permutations of the data set to generate a null distribution, it makes

the computational burden increase by orders of magnitude.

Realizing that most of the computational burden associated with the CPM can be avoided, the RPM tries to partition the genotypes in the most reasonable way for evaluation that makes a tradeoff between maximization of the between group variation with minal number of groups and the within group variation.

The RPM employs an iterative search procedure to search the best way to partition the genotypes by merging most similar groups of genotypes in each iteration, rather than exhaustively search all possible partition of multi-locus genotypes. The similarity of different groups of genotypes is based on a multiple comparisons test of the mean values of their quantitative trait. The algorithm includes the following steps:

Step 1. Initially, each multi-locus genotype forms a group.

Step 2. A multiple comparisons test is performed to determine the similarity of mean quantitative trait values between any two groups of genotypes. The algorithm ends if all groups have significantly different means

Step 3. Merge the pair of groups which have most similar mean values to form a new group.

Step 4. Return to step 1.

The importance of the final partition can be measured by estimating the R^2 value for the model of the quantitative trait value regressed on the final genotype groups.

For each iteration, before the algorithm ends, the number of groups is reduced by one by merging two groups. Thus the algorithm will end after at most n-1 iterations if there are initially n genotypes. Therefore the RPM is much more efficient than the CPM.

To measure statistical significance of the final partition, P values for the R^2 values are estimated using a permutation test. The trait values in the original data are permutated and then the RPM is executed. Significance is estimated by the frequency with which the R^2 value from the original data exceeds the permuted R^2 values.

The generalized MDR (Lou et al., 2007) extends MDR to continuous phenotypes and includes covariate adjustment.

A phenotype which is dichotomous for a disease and continuous for a quantitative

trait can usually be represented by a generalized linear model with respect to genes and covariates in the exponential family of distributions including the normal, Poisson, and Bernoulli distributions.

Suppose $y_i$is the phenotype of individual i, $\mu_i=E(y_i)$, the expectation of y_i, then we have

$$\theta(\mu_i)=\alpha+x_i^T\beta+z_i^T\gamma \tag{4.4}$$

where $\theta(\mu_i)$ is an appropriate link function, α is the intercept, x_i is the vector that represents gene-by-gene and/or gene-by-environment interactions, z_i is the vector representing for covariates, and β and γ are the parameter vectors. For dichotomous phenotypes having a Bernoulli distribution, the link function is the logit,

$$\theta(\mu_i)=\log\left[\frac{\mu_i}{1-\mu_i}\right] \tag{4.5}$$

For continuous phenotypes having a normal distribution, the link is the identity.

The probability functions of exponential family models coud be expressed as

$$p(y;\theta,\Phi)=\exp\left[\frac{y\theta-f(\theta)}{g(\phi)}+h(y,\phi)\right] \tag{4.6}$$

where *f(·), g(·),* and *h(·)* are known functions, θ is the link function, a function of the expectation μ of *Y*. Therefore the log likelihood for independent observations y_i, $i=1,2,\ldots,n$, can be written as

$$\log L(y|\Omega)=\sum_{i=1}^{n}\{y_i\theta(\mu_i)-f[\theta(\mu_i)]\} \tag{4.7}$$

where **y** is the vector of observations, Ω is the vector of parameters, Ω=(α, β, γ), and $f[\theta(\mu_i)]$ is a function of $\theta(\mu_i)$ with the property that $\partial f[\theta(\mu_i)]/\partial\theta(\mu_i)=\mu_i$. a(ϕ) and c(y,ϕ) don't appear because they make no difference for score defined as the following first partial derivative of the log-likelihood,

$$\frac{\partial\log(y\mid\Omega)}{\partial p}=\sum_{i=1}^{n}\left[\frac{y_i\partial\theta(\mu_i)}{\partial\theta}-\frac{\mu_i\partial\theta(\mu_i)}{\partial\theta}\right] \tag{4.8}$$

where $p\in\Omega$. The residual score vector can be obtained by setting β=0 in model (1),

$$S_\beta(\hat{\alpha}_0,\beta=0,\hat{\gamma}_0)=[S_{\beta_j}(\hat{\alpha}_0,\beta=0,\hat{\gamma}_0)] \tag{4.9}$$

where $S_{\beta_j}(\hat{\alpha}_0, \beta = 0, \hat{\gamma}_0) = \sum_{i=1}^{n} x_{ij}(y_i - \hat{\mu}_i)$, $\hat{\mu}_i = l^{-1}(\hat{\alpha}_0 + z_i^T \hat{\gamma}_0)$ is the estimated expectation, $\hat{\alpha}_0$ and $\hat{\gamma}_0$ are the maximum-likelihood estimates under the null hypothesis $H_0(\beta=0)$, and $x_{ij}(y_i - \hat{\mu}_i)$ is the contribution of individual i to the score for β_j.

The score-based statistics for individual i is defined as normalized contributions:

$$S_i^T = \sum_j \frac{x_{ij}(y_i - \hat{\mu}_i)}{\sqrt{\hat{Var}(y_i)}} \tag{4.10}$$

where $\hat{Var}(y_i)$ is the estimated variance of y_i.

In the GMDR method, the ratio of cases to controls is replaced by the score in each genotype combination or cell to classify it into a high-risk group or low-risk group and calculate classification accuracy and prediction error. First each individual has its score computed under the null hypothesis, H_0: $\beta=0$. Then the scores of individual are summed within each genotype combination and each genotype combination is assigned to either a “high-risk” group if the average score is greater than or equal to a preassigned threshold T (e.g., 0) or a “low-risk” group otherwise. The scores are also used to identify the best model.

This method is flexible in the use of covariates, different study designs, and various kinds of phenotypes including continuous, dichotomous and other phenotypes. It can also be employed for unbalanced case-control, random, and selected samples. In addition to score functions, other statistics, can also be used.

In Model based MDR (MB-MDR) (Calle et al., 2008), MDR is extended to continuous outcomes by using parametric regression.

Although MDR increase the power to identify significant gene-gene interactions by partitioning genotype combinations into only two groups, high-risk and low-risk groups, it has some limitations.

First, some important interactions could be missed. Any cell having a cases/controls ratio above the global threshold will be assigned to high risk group no matter its size of the ratio is. This will lead to missing some cells with significant

association with the disease when they are combined with not significant ones. When combining together, the ratio of the number of cases and controls is similar to that in the overall sample. False positive will also appear if there are a few individuals for both cases and controls in a genotype combination. Therefore more specific alternative hypotheses needs to be considered to increase the power for high-dimensional data.

The second limitation is its lack of adjustment for main effects. For a detected interaction by MDR, it is difficult to decide whether it is because of the main effects, or because a real epistatic interaction.

The third one is its lack of adjustment for confounding factors. MDR can not identify confounding factors without conducting a stratified analysis. However it is important to adjust for confounding factors when two populations not perfectly matched are compared.

The fourth one is that it can only be applied to binary outcomes while gene-gene interactions associated with other kind of outcomes, such as time-to-event variables, which also often appear in practical applications.

The fifth one is its computational burden to evaluate significance. The cross-validation consistency or the average balanced predictive accuracy used in the permutation test to evaluate significance is not invariant reference statistics and therefore the construction of the specific permutation null distribution for any particular case is required.

Finally, MDR has low power when there are genotyping error, missing data, phenocopy and genetic heterogeneity.

To overcome these limitations, Model-Based Multifactor Dimensionality Reduction (MB-MDR) only assigns cells showing significant different cases/controls ratio from the global threshold to the high or low risk group. Those cells which have a cases/controls ratio close the global threshold or have small sample size are assigned to an additional category, that of no evidence of risk. The procedure of MB-MDR is as follows:

Let each multifactor cell be denoted by c_j , where $1 \leq j \leq N$, N is the total number of

multifactor cells.

Step 1:

Each genotype cell, c_j , is labeled as High risk (H), Low risk (L) or no evidence (0) according to its Odd Ratio (OR_j). The null hypothesis is OR_j =1. This association test can either be nonparametric (chi-squared test) or parametric (logistic regression) and, adjustment for main effects and confounder factors can be performed for the latter one.

Cells with an OR of individuals in that cell versus the rest of individuals significantly greater than 1 and smaller than 1 (a p-value smaller than 0.10) are labeled as High risk and Low risk respectively. A conservative threshold of 0.10 is used because the power to detect association using individual cells is very limited. Cells with a p-value larger than 0.1, are labeled as zero.

A new variable X taking values H, L, or 0 is therefore created.

Step 2:

This new predictive variable X on the outcome variable Y leads a new association test. This can also be a nonparametric test (chi-squared test) or a parametric one (logistic regression). Odds ratios for risk categories can be obtained to test the significance of association.

Step 3:

Since after combining cells the statistic don't follow chi-squared distribution, the Wald statistic is employed to test the association instead. The raw p-value should be adjusted for the number of cells combined in each risk category. Permutation null distributions are invariant distributions for interaction of different orders conditional on the number of combined cells. Therefore they could be tabulated and used in future applications.

There are also methods based on information theory. In (Chanda et al., 2009), a method based on two information-theoretic metrics, the k-way interaction information (KWII) and phenotype-associated information (PAI) is developed for identifying gene-gene and gene-environmental interactions associated with quantitative traits. In (Yee et al., 2015), as an extension of information gain, a nonparametric evaluation

method of conditional entropy of a quantitative phenotype associated with a given genotype is proposed.

However none of the above methods is computationally efficient.

In Quantitative MDR (QMDR) (Gui et al., 2013), to exploit continuous outcomes to make the analysis more accurate, a test statistic, rather than the balanced accuracy, is used to determine the best interaction model.

To examine a k-order interaction, K SNPs are selected from a dataset which has m SNPs. The mean value of the phenotype for each genotype combination of the K SNPs is calculate and compared with the overall mean of the phenotype. The genotype combination is labeled high-level if its mean value is larger than the overall mean or low-level otherwise. Therefore all genotype combinations are reduced to an attribute which has two categories: high-level and low-level. Rather than balanced accuracy, a T-test is used to compare phenotype mean values between high and low level groups. The value of the T-test is then used as a training score to select the best interaction model and the best overall model is selected using the maximum testing score. The permutation method can be further used to determine whether the selected best overall model has a significant level to be considered to have association with the phenotype.

This is a computationally efficient algorithm. However this method still classified the outcome into two groups: high and low level groups, which results in the loss of the large variability of the quantitative outcome.

Also there are few methods applied to ordinal categorical traits. Ordinal categorical traits such as the obesity classification based on body mass index (e.g., normal, pre-obese, mild obese and severe obese), the diabetes diagnosis based on glucose level (e.g., normal, impaired glucose tolerance and diabetes) are common in many genetic association studies. These traits are also derived from quantitative traits. In Ordinal MDR (OMDR) (Kim et al., 2013), MDR is extended to analyze gene-gene interaction for ordinal traits and tau-b (Agresti and Kateri, 2011), a common ordinal association measure, is used to replace balanced accuracy to evaluate interactions. However the tau-b measure only measures the degree of tendency of positive

association between true categories of an ordinal trait and predicted categories and doesn't consider the difference between true categories and predicted categories.

In order to better use the information contained in QTs, we first classify the quantitative outcome into several (greater than two) ordinal levels. Then an extended MDR is used to identify gene-gene interactions on this converted ordinal categorical trait. Rather than using balanced accuracy or common ordinal association measures, such as tau-b, we use an extended fuzzy classification method to select the set of genetic markers as having strongest associations with the trait. Usually for each prediction of a category, its accuracy value is either 1, if the prediction is right, or 0, if the prediction is wrong. However for quantitative or ordinal traits, when the prediction is wrong, the closeness of different quantitative values to the true category is different. To reflect such difference, member functions of fuzzy sets could be employed to compute accuracy in classification. Since the range of a member function is between 0 and 1, to better describe the difference of quantitative values to a category, we extend its range to [-1, 1] when it is used in fuzzy classification.

In this paper, a new kind of member functions which have an extended output range from -1 to 1 are proposed to be used in fuzzy classification first. Then Extended Fuzzy Quantitative MDR (EFQMDR) algorithm is given to strengthen identification of gene-gene interactions associated with QTs. This algorithm first transforms a quantitative trait into an ordinal trait and then selects multiple best sets of SNPs as having strongest association with the trait using such kind of member functions in the extended MDR. To test the performance of the proposed algorithm, we use it to identify five different interaction models in simulated data and compare success rates with three other methods. We also use it in two real data sets to select multiple SNPs having strong association with the trait and compare balanced test accuracy and consistency with the same three other methods.

4.3 Methods

4.3.1 Extended fuzzy classification using extended member functions

Real world is complicated, we usually couldn't get or handle simultaneously abundant information to make prediction. Therefore fuzzy set theory proposed by Zadeh (Zadeh, 1965) founds its application in many areas where information is imprecise, such as control theory (Tanaka and Sugeno, 1992; Tanaka and Wang, 2004; Procyk and Mamdani, 1979), data mining (Gustafson and Kessel, 1978; de Oliveira and Pedrycz, 2007; Timm et al., 2004), medicine and bioinformatics (Barro and Marín, 2002; Phuong and Kreinovich, 2001; Angela and Nieto,2006; Dembélé and Kastner, 2003). As an extension of classical set theory where an element has a dichotomous relation with a set:it can either belong to it or not, fuzzy set theory allows an element to partially belong to a set to reflect imprecise situations. Such a relation can be described using a membership function with its values between 0 and 1. Let A be a fuzzy set in the universal space X, its elements can be described using an ordered pairs (x, μ_A), where $x \in X$, μ_A is a membership function taking values on [0,1] and representing the degree of membership of x belonging to X. The classical set can be considered as a special case of the fuzzy set where its membership function can take on only the value 1, if x belongs to A, or 0 if x does not belong to A. Therefore the membership function of a classical set reduces to the indicator function $I_A(x)$ of a set A.

In addition to genetic factors, QTs are also related to many other factors. Genetic factors related to a specific QT could not determine alone a value of a QT precisely. Therefore an appropriate way to predict the value of a QT with genetic factors relating to it is to classify it into several categories and predict its category. To fully utilize the information contained in a QT, each category can be represented by a fuzzy set and

the relation of a QT with a category can be described using a membership function rather than a binary status.

By introducing these fuzzy sets, training balanced accuracy used to select the best classifier can be replaced by training balanced accuracy based on member functions to reflect partial membership of a sample with a particular genotype combination to the category labeled to this genotype combination, while the original MDR assigns either 0 or 1 to a sample to reflect its accuracy to be classified as the category labeled to its genotype combination. Training balanced accuracy used in MDR is a special case of training balanced accuracy based on member functions when the indicator function is used as the member function.

A variety of member functions have been proposed for fuzzy sets. Two popular types are linear and sigmoid. In this paper, linear member functions are used. Here as an example, we divide a QT into three categories or levels: high(H), average(A) and low(L) associated with three fuzzy sets using equal length intervals, as shown in Figure 4.1.

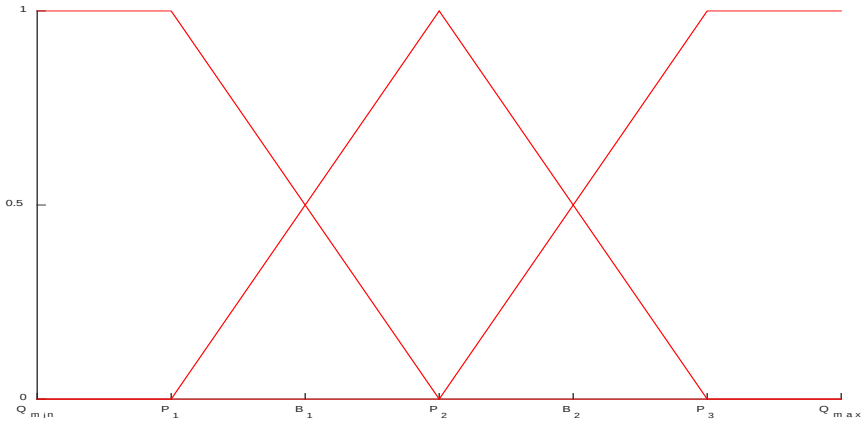

Figure 4.1 The linear membership functions of high(H), average(A) and low(L) levels of a QT.

Let $Q_{\min}$ and $Q_{\max}$ denote the maximum and minimum values that a QT takes on in all samples in a dataset. B_1 and B_2 are upper borders of the low level and the average level respectively. P_1 ,P_2 and P_3 are the middle positions of the low level, average level and high level respectively can be derived as follows:

$$P_1 = \frac{Q_{\min} + B_1}{2} \tag{4.11}$$

$$P_2 = \frac{B_1 + B_2}{2} \tag{4.12}$$

$$P_3 = \frac{B_2 + Q_{\max}}{2}. \tag{4.13}$$

Then member functions for *L*,*A* and *H* levels in Fig.1 can be expressed as:

$$\mu_{L1}(x) = \begin{cases} 1, & \text{if } x <= P_1 \\ \dfrac{P_2 - x}{P_2 - P_1}, & \text{if } P_1 < x <= P_2 \\ 0, & \text{otherwise} \end{cases} \tag{4.14}$$

$$\mu_{A1}(x) = \begin{cases} \dfrac{x - P_1}{P_2 - P_1}, & \text{if } P_1 <= x <= P_2 \\ \dfrac{P_3 - x}{P_3 - P_2}, & \text{if } P_2 < x <= P_3 \\ 0, & \text{otherwise} \end{cases} \tag{4.15}$$

$$\mu_{H1}(x) = \begin{cases} 0, & \text{if } x <= P_2 \\ \dfrac{x - P_2}{P_3 - P_2}, & \text{if } P_2 < x <= P_3 \\ 1, & \text{otherwise} \end{cases} \tag{4.16}$$

Membership functions of fuzzy sets can also be used as an accuracy measure in fuzzy classification. For example, when different values are classified to the high level, we can get different accuracies between 0 and 1 from $\mu_{H1}(x)$. However when selecting a best classifier composed of a set of SNPs to classify a QT, such a range could not fully show differences among different classifiers. For example, if there are both 500 samples in genotypes that are classified as the high level for two classifiers, for classifier 1 there are 300 samples located at P_3, 200 samples located at P_2 and 100 samples located at P_1 in genotypes that are classified as the high level, for classifier 2 there are 300 sample located at P_3, 100 samples located at P_2 and 200 samples located at P_1 in genotypes that are classified as high levels, then the accuracies of the high level for these two classifiers would be the same: 0.6. However classifier 1 is obviously a better classifier to classify the high level. To reflect such difference, we extend the range of member functions from [0,1] to [-1,1] when they are used in fuzzy classification to select the best classifier.

Such an extended linear member function is illustrated in Figure 4.2 and can be

expressed as:

$$\mu_{L2}(x)=\begin{cases}1, & \text{if } x <= P_1\\ \dfrac{P_2 - x}{P_2 - P_1}, & \text{if } P_1 < x <= P_3\\ -1, & \text{otherwise}\end{cases} \tag{4.17}$$

$$\mu_{A2}(x)=\begin{cases}\dfrac{x - P_1}{P_2 - P_1}, & \text{if } x <= P_2\\ \dfrac{P_3 - x}{P_3 - P_2}, & \text{otherwise}\end{cases} \tag{4.18}$$

$$\mu_{H2}(x)=\begin{cases}-1, & \text{if } x <= P_1\\ \dfrac{x - P_2}{P_3 - P_2}, & \text{if } P_1 < x <= P_3\\ 1, & \text{otherwise}\end{cases} \tag{4.19}$$

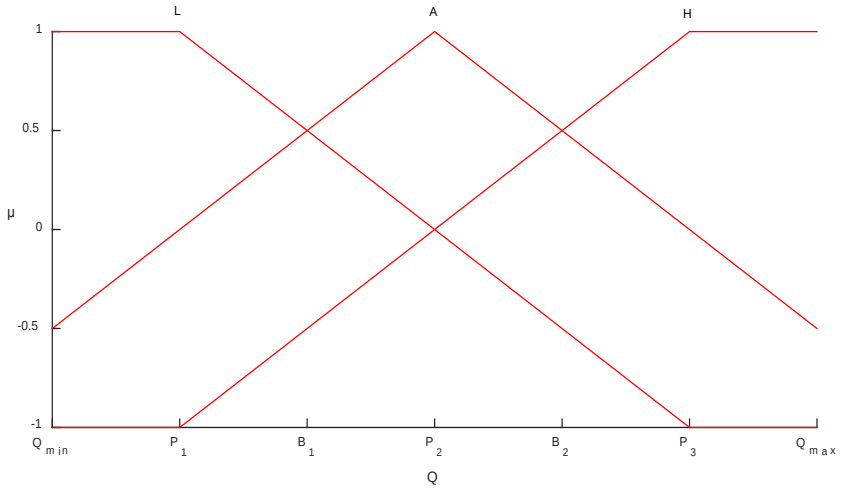

Figure 4.2 The extended linear membership functions of high(H), average(A) and low(L) levels of a QT

4.3.2 EFQMDR Algorithm

In order to detect high- dimensional gene-gene interaction, MDR reduces genotype combinations at multiple loci into a single class variable taking values of either high risk or low risk categories, then tests association between a binary trait or disease with this new one dimensional variable. The training balanced accuracy of the two categories is used to select the best classifier. Balanced accuracy is defined as the

arithmetic mean of sensitivity and specificity:

$$(\text{sensitivity}+\text{specif icity})/2=(TP/(TP+FN)+TN/(TN+FP))/2$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives. The *m*-locus classifier that has the maximum testing balanced accuracy and highest cross-validation consistency is selected as the final best *m*-locus classifier, where cross-validation consistency is used as a tie-break.

For an ordinal categorical trait with *J* levels, an *m* dimensional cell is labeled as one of *J* groups as follows. Let 1, 2,..., *J* be *J* levels or categories for an ordinal trait . For any combination of *m* SNPs, let n_{+j} be the number of individuals in class *j*, n_{ij} be the number of individuals with the *i*th multi-locus genotype in category *j*, where $i = \{1, 2,...,3^m\}$ and $j = 1, 2,..., J$. Then the *i*th m-locus genotype will be labeled as category $c(i)$ as follows:

$$c(i)= \underset{j\in\{1,\ldots,J\}}{\arg\max}\left(\frac{n_{ij}}{n_{+j}}\right)$$

EFQMDR extends MDR to analyze quantitative traits by first converting them to ordinal traits. Then Instead of evaluating each classifier using balanced accuracy or common ordinal association measures, it uses generalized fuzzy classification based on extended member functions to evaluate each classifier and select the best one as having the strongest association with the trait. The procedure of EFQMDR is as follows:

1. Divide the range of a quantitative trait into *J* intervals and label them as categories 1,2,…,*J* respectively.
2. Partition the dataset into *L* subsets for *L*-fold *cross-validation* (CV). Use one of the *L* subsets as a testing set and the rest as a training set.
3. For each *m*-way interaction derived from *m* SNPs or SSLPs, let n_{ij} be the number of individuals belonging to category *j* with the *i*th multi-locus genotype in the training set, n_{+j} be the total number of individuals belonging to category *j* in the training set, where $i = \{1, 2,...,3^m\}$ and $j = 1, 2,..., J$. Then all individuals with the *i*th multi-locus genotype will be assigned into the category $c(i)$ by the

classifier corresponding to the *m* given SNPs as follows:

$$c(i)= \arg\max_{j\in\{1,\ldots,J\}}\left(\frac{n_{ij}}{n_{+j}}\right) \tag{4.20}$$

where n_{ij} and n_{+j} are real numbers, n_{ij} is computed using the extended linear member function, n_{+j}, the size of class *j*, is computed using the traditional linear member function.

4. Compute the training balanced accuracy for each *m*-way interaction:

$$\frac{1}{J}\sum_{i=1}^{3^m}\frac{n_{i,c(i)}}{n_{+c(i)}} \tag{4.21}$$

where $n_{i,c(i)}$, the number of individuals with the *i*th multi-locus genotype which really belong to the class they are classified to, is computed using the extended linear member function.

5. Since multiple gene-gene interactions associated with a QT is common in complex traits, multiple classifiers that have best training balanced accuracies are selected and their testing balanced accuracies based on the extended linear member function are computed.
6. Repeat steps 3-5 on all *L* CV dataset.
7. Multiple candidates of *m*-way gene-gene interactions are selected as having the maximal testing balanced accuracy and highest generalized cross-validation consistency based on top-*K* selection ($GCVC^K$ or simplified as GCVC) [33], where general cross-validation consistency is used as a tie-break.. The $GCVC^K$ is calculated as follows:

$$GCVC^K= \sum_{I=1}^{L} I_I \quad \text{where} \quad I_I= \begin{cases} 1, & \text{if the MDR classifier is identified as one} \\ & \text{of top-}K\text{ classifiers at }I^{\text{th}}\text{ CV dataset} \\ 0, & \text{otherwise} \end{cases} \tag{4.22}$$

4.4 Experimental results and analysis

4.4.1 Experiments and analysis of results on simulated data

4.4.1.1 Experimental setup

The simulation experiment is designed to study the success rate of the proposed method and compare it with that of MDR, OMDR and Fuzzy Quantitative MDR (FQMDR) which uses fuzzy classification based on traditional member functions.

Five different interaction models were used for the ordinal trait transferred from a quantitative trait (Figure 4.3)[33]. For each model, one pair of SNPs were simulated as a causal factor among all possible combinations.

We use gs 2.0 to generate simulated genotype data.

Since the outcome is binary status (case or control), we derive continuous outcome from the penetrance functions (the penetrance function denotes the probability of being a case for each genotype combination.) of the five models as follows:

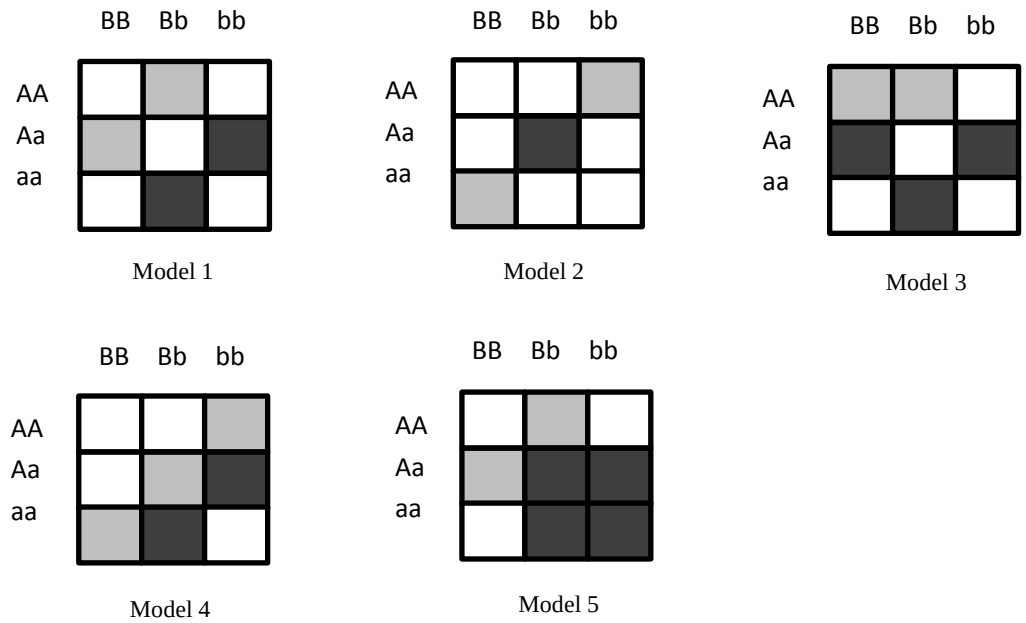

Figure 4.3 Models of two way interactions for ordinal traits. White, light grey and

dark grey represent normal, low risk and high risk of an ordinal trait respectively. (Kim et al., 2013)

Let f_{ij} be the element from the ith row and jth column of a penetrance function for two interacting SNPs, two interacting SNPs, the quantitative trait is generated from the following normal distribution:

$$y|SNP1=i, SNP2=j \sim N(f_{ij}, \sigma^*) \quad (4.23)$$

where f_{ij} and σ^* are the mean and variance of the normal distribution respectively. Then the quantitative trait is transferred to an ordinary trait with three categories. Let μ, σ be the mean value and variance of the quantitative trait, any quantitative trait value smaller than μ-σ/2 is classified as low category; any value between μ-σ/2 and μ+σ/2 is classified as middle category; any value larger than μ+σ/2 is classified as high category.

We use two different minor allele frequencies (MAF=0.2 and 0.4), five different variances (σ^*=0.1, 0.2, 0.3, 0.4 and 0.5) and two different sample size (n=200 and 800) with fixed SNP number (441 SNPs) and penetrance functions (0.01, 0.25, 0.5 for white, light grey, dark grey in Figure 3. respectively) to create simulated datasets. For each interaction model, 100 replicated datasets were generated. Varying variances with fixed penetrance functions is equivalent to varying penetrance functions with fixed variances.

Hit ratio which is defined as the proportion of replicates with which the true causal SNPs are detected as the best SNPs among all possible same number of SNPs is used to measure the success rate.

To test the type I error rate, the null datasets with no causal pair of SNPs were simulated for different sample sizes (n=200, 400 and 600) and different SNP numbers (m=10, 15, 20). Permutation P values of the identified strongest interaction pair of SNPs were calculated by permuting trait values of each dataset 1000 times. The ratio of the permutation P values smaller than the significance level α=0.05 in 1000 replicates is calculated as the type I error rate. The number of the permutation ensured its accuracy to one decimal place when expressed in percent.

4.4.1.2 Experimental Results

Experiment results of five models are shown in Table 4.1 through Table 4.5.

The performance of EFQMDR is better than other three methods in general. The performance of EFQMDR is better than that of MDR except in a few cases for model 4 when the sample size is not small, MAF is low and variance is not small. It is better than that of OMDR except in a few cases for model 1 when the sample size is small,

Table 4.1 Hit ratios (%) for model 1

Sample size	MAF	Method	Variance				
			0.1	0.2	0.3	0.4	0.5
200	0.2	EFQMDR	82	56	18	4	2
		FQMDR	81	51	18	3	2
		OMDR	64	55	25	6	3
		MDR	78	45	9	4	1
	0.4	EFQMDR	99	79	53	27	13
		FQMDR	99	66	38	15	8
		OMDR	97	71	43	17	7
		MDR	94	66	30	11	6
400	0.2	EFQMDR	98	75	46	18	8
		FQMDR	98	76	53	23	9
		OMDR	90	73	44	22	13
		MDR	96	68	40	13	4
	0.4	GFQMDR	100	89	75	54	39
		FQMDR	99	83	64	39	23
		OMDR	100	81	59	41	35
		MDR	99	75	56	35	18
800	0.2	EFQMDR	100	90	70	51	21
		FQMDR	100	92	67	49	36
		OMDR	89	86	63	50	34
		MDR	99	87	60	47	24
	0.4	EFQMDR	100	99	96	89	76
		FQMDR	100	95	91	73	62
		OMDR	100	98	83	71	59
		MDR	100	95	82	66	55

Table 4.2 Hit ratios (%) for model 2

Sample	MAF	Method	Variance

size			0.1	0.2	0.3	0.4	0.5
200	0.2	EFQMDR	90	66	43	19	7
		FQMDR	89	58	38	21	6
		OMDR	89	62	33	19	9
		MDR	82	59	28	6	3
	0.4	EFQMDR	97	82	61	40	23
		FQMDR	96	77	52	36	20
		OMDR	93	80	53	37	26
		MDR	90	69	51	28	11
400	0.2	EFQMDR	98	84	71	54	32
		FQMDR	97	82	66	52	32
		OMDR	99	78	63	48	34
		MDR	92	80	56	40	27
	0.4	EFQMDR	99	95	81	67	48
		FQMDR	98	92	78	63	50
		OMDR	98	92	78	71	50
		MDR	97	91	73	62	42
800	0.2	EFQMDR	100	96	90	75	54
		FQMDR	100	96	88	70	56
		OMDR	100	95	85	68	53
		MDR	99	94	84	63	51
	0.4	EFQMDR	100	100	94	82	70
		FQMDR	100	100	93	83	74
		OMDR	100	100	90	83	73
		MDR	100	98	91	76	66

Table 4.3 Hit ratios (%) for model 3

Sample size	MAF	Method	Variance				
			0.1	0.2	0.3	0.4	0.5
200	0.2	EFQMDR	93	65	41	19	3
		FQMDR	90	51	20	6	1
		OMDR	87	50	20	7	3
		MDR	87	50	17	3	0
	0.4	EFQMDR	83	73	54	34	17
		FQMDR	83	69	51	31	17
		OMDR	80	65	50	36	17
		MDR	80	59	39	17	4
400	0.2	EFQMDR	99	79	61	39	19
		FQMDR	95	66	43	15	2
		OMDR	98	64	34	16	12
		MDR	96	61	28	6	1
	0.4	EFQMDR	100	92	81	69	53
		FQMDR	99	91	75	56	44

		OMDR	100	91	74	55	39
		MDR	96	89	72	51	30
800	0.2	EFQMDR	100	99	84	64	43
		FQMDR	100	95	76	51	30
		OMDR	100	89	72	37	31
		MDR	99	91	71	37	16
	0.4	EFQMDR	100	100	97	93	82
		FQMDR	100	100	93	86	73
		OMDR	100	100	93	82	70
		MDR	100	99	94	82	72

Table 4.4 Hit ratios (%) for model 4

Sample size	MAF	Method	Variance				
			0.1	0.2	0.3	0.4	0.5
200	0.2	EFQMDR	76	35	12	3	1
		FQMDR	76	40	18	4	1
		OMDR	69	41	19	6	3
		MDR	65	36	10	0	1
	0.4	EFQMDR	86	65	46	23	10
		FQMDR	83	59	26	12	5
		OMDR	85	56	36	11	5
		MDR	76	47	19	6	2
400	0.2	EFQMDR	88	52	19	5	1
		FQMDR	85	61	33	9	4
		OMDR	69	47	33	16	8
		MDR	80	59	22	8	3
	0.4	EFQMDR	95	77	52	29	19
		FQMDR	95	66	41	22	10
		OMDR	96	71	44	30	19
		MDR	90	57	32	19	8
800	0.2	EFQMDR	98	75	39	22	9
		FQMDR	98	77	46	27	17
		OMDR	88	61	33	26	13
		MDR	95	71	45	23	13
	0.4	EFQMDR	100	91	74	57	48
		FQMDR	100	87	65	43	32
		OMDR	100	91	61	44	29
		MDR	100	74	55	42	30

Table 4.5 Hit ratios (%) for model 5

Sample size	MAF	Method	Variance				
			0.1	0.2	0.3	0.4	0.5

200	0.2	EFQMDR	83	47	25	6	1
		FQMDR	79	38	9	1	1
		OMDR	85	38	12	2	0
		MDR	71	29	5	1	0
	0.4	EFQMDR	81	51	32	11	5
		FQMDR	75	49	24	7	2
		OMDR	76	47	27	10	4
		MDR	72	37	12	3	1
400	0.2	EFQMDR	94	78	54	26	12
		FQMDR	93	59	21	8	5
		OMDR	97	62	32	15	3
		MDR	90	52	20	6	2
	0.4	EFQMDR	94	78	54	33	18
		FQMDR	93	68	42	21	11
		OMDR	96	64	43	28	13
		MDR	90	63	30	15	5
800	0.2	EFQMDR	99	90	73	58	43
		FQMDR	99	75	55	34	18
		OMDR	99	86	50	36	28
		MDR	98	60	46	22	13
	0.4	EFQMDR	100	94	76	57	40
		FQMDR	100	91	64	41	33
		OMDR	98	86	60	47	31
		MDR	100	81	47	44	28

MAF is low and variance is not small; for model 2 when the sample size is large, MAF is high and variance is large; for model 4 when the sample size is not large, MAF is low and variance is not small. It is also better than that of FQMDR except in a few cases for model 1 when the sample size is not small, MAF is low and variance is not small; for model 2 when the sample size is large, MAF is high and variance is large; for model 4 when MAF is low and variance is not small. It is also observed that the performance of FQOMDR is better than that of MDR in all 30 cases, the performance of OMDR is better than that of MDR in general, and the performance of FQOMDR is slightly better than that of OMDR.

For the type I error rate, results given in Table 4.6 show that EFQMDR has type I error rate tightly gathering around 5% with a range from 4.3% to 5.8%, better than three other methods. Therefore EFQMDR controls type I error rate better.

Table 4.6 Type I Error Rate with the Significance Level α of 0.05 from Datasets with 1000 1000 Replicates

		n		
m	Method	200	400	600
10	EFQMDR	4.6%	4.6%	4.3%
	FQMDR	4.5%	5%	4.9%
	OMDR	5%	5.5%	6.1%
	MDR	3.8%	5.3%	5.7%
15	EFQMDR	5.8%	4.6%	4.9%
	FQMDR	5.2%	4.3%	6.5%
	OMDR	5%	4.4%	5.6%
	MDR	4.2%	3.5%	6.1%
20	EFQMDR	4.9%	5.2%	4.8%
	FQMDR	5.3%	4.7%	5.3%
	OMDR	4.1%	5.7%	3.8%
	MDR	4.2%	5.6%	4.9%

4.4.2

4.4.3

4.4.4 Experiments and analysis of results on real data

4.4.4.1 Experimental setup

We use two real datasets to show applications and performance of the proposed method.

One is *Ultra-violet* (UV) Light-Induced Immunosuppression Data. F1 backcross mice are derived from a backcross between low susceptibility BALB/c female mice and high susceptibility (BALB/c ×C57BL/6) F1 male mice. This dataset contains 64 markers, sex and UV light-induced *percent immunosuppression* (PI) of a contact hypersensitivity response of 134 F1 backcross mice (Clemens et al., 2000). The data were acquired from the Center for Genome Dynamics at the Jackson Laboratory http://cgd.jax.org/nav/qtlarchive1.htm. UV light-induced percent immuno-suppression is the quantitative trait of interest.

Another is intercross mouse population from intercross of DBA2 and NMRI8.

NMRI8 is a long-term high body weight-selected mouse line and analyzed at the age of 6 weeks. It is extremely different in body composition from the control mouse line DBA/2. There are 275 mice (142 females, 133 males), 98 markers and 18 phenotypes. Since genetic factors contributing to obesity and body weight are considered to act through mechanisms affecting muscle weight, fat weight, or both, we use three phenotypes in the population in our experiments which are body weight (bw), abdominal fat (afw) and muscle weight (mw) (Brockmann et al., 2009) The data were downloaded from the QTL Archive curated by the Jackson Laboratories http://phenome.jax.org/db/q?rtn =projects/projdet&reqprojid=213.

For missing values of SNP or SSLP, we set them to the majority value of that SNP or SSLP; for missing values of QTs, we set them to the mean value of that quantitative trait.

All four QTs are divided into three categories. For UV light-induced percent immunosuppression three categories are defined as high percent immunosuppression, medium percent immunosuppression and low percent immunosuppression states respectively; for bw, afw and mw, three categories are defined as heavy weight, medium weight and light weight respectively.

4.4.4.2 Experimental results

The EFQMDR method is used to select the best 2-way, 3-way and 4-way interactions in the above real datasets associated with bw, afw, mw and PI respectively.

The performance of the EFQMDR method is evaluated in maximum testing balanced classification accuracy (MTSBCA) on ten CVs and corresponding GCVC, where GCVC is used as a tie break, and compared with that of FQMDR, OMDR and MDR methods. Balanced accuracy using the extended linear member function, balanced accuracy using the traditional linear member function, tau-b and balanced accuracy are used to select the best interaction SNPs in each CV in EFQMDR, FQMDR, OMDR and MDR methods respectively. We choose k best set of SNPs for

each of 2-way, 3-way and 4-way interactions.

We first set k to 1, i.e. for each CV of a specific QT, we choose one best set of SNPs of a fixed order. In this case GCVC is equivalent to CVC.

From Table 4.7, we can see that the performance of EFQMDR evaluated by MTSBCA and GCVC are better than that of FQMDR, OMDR and MDR in most cases. Figure 4.4 shows that the average maximum testing balanced classification accuracy on 2-way, 3-way and 4-way interactions (AMTSBCA1) with EFQMDR is higher than that with three other methods for each of the four QTs except for afw with FQMDR and average AMTSBCA1 on all four QTs (AMTSBCA2) with EFQMDR is higher than that of three other methods.

To reflect multiple gene-gene interactions associated with a trait in complex traits, we also set k to 5, i.e. for each CV of a specific QT, we choose five best sets of SNPs of a fixed order. MTSBCA1 through MTSBCA5 are used to represent five sets of SNPs which have largest MTSBCAs in the descending order and GCVC1 through GCVC5 are corresponding GCVCs which are used as a tie break.

Again, from Table 4.8 to Table 4.11, we can see that the performance of EFQMDR is better than that of FQMDR, OMDR and MDR in most cases. Figure 4.5 shows that AMTSBCA1 with EFQMDR is higher than that with three other methods for each of the four QTs except for PI with OMDR and bw with FQMDR, AMTSBCA2 with EFQMDR is higher than that of three other methods.

Table 4.7 Comparison of MTSBCA and GCVC among EFQMDR, FQMDR, OMDR and MDR when k=1

		Two-locus classifier		Three-locus classifier		Four-locus classifier	
QT	Method	*MTSBCA*	*GCVC*	*MTSBCA*	*GCVC*	*MTSBCA*	*GCVC*
PI	EFQMDR	0.563	3	0.488	2	0.590	4
	FQMDR	0.597	4	0.333	1	0.375	1
	OMDR	0.347	1	0.625	2	0.587	1
	MDR	0.417	2	0.389	1	0.583	1
bw	EFQMDR	0.6	7	0.583	3	0.5	1
	FQMDR	0.625	2	0.5	1	0.628	1
	OMDR	0.517	1	0.472	2	0.544	2
	MDR	0.533	2	0.522	2	0.5	3

afw	EFQMDR	0.739	1	0.636	1	0.592	1
	FQMDR	0.537	3	0.476	3	0.736	6
	OMDR	0.481	1	0.616	2	0.642	3
	MDR	0.506	4	0.470	1	0.576	1
mw	EFQMDR	0.592	6	0.699	3	0.594	3
	FQMDR	0.630	3	0.490	1	0.520	1
	OMDR	0.505	5	0.520	1	0.556	1
	MDR	0.544	1	0.520	2	0.45	1

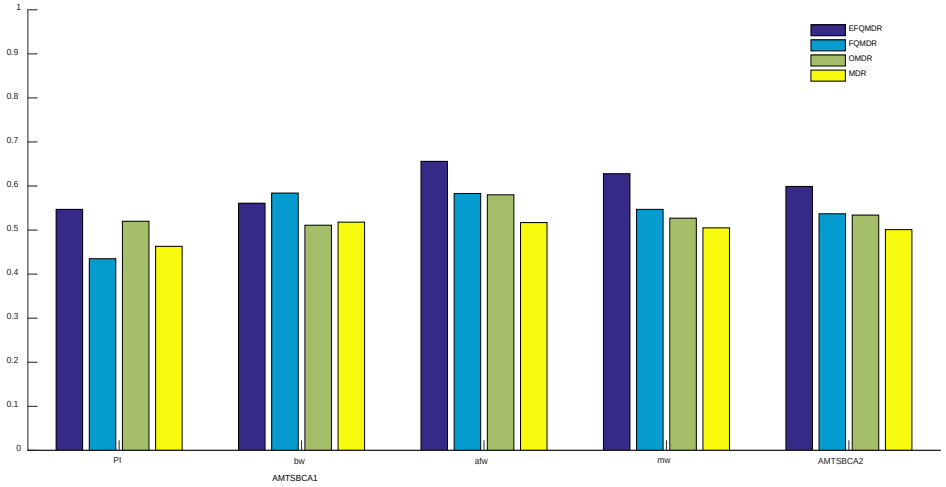

Figure 4.4. Comparison of AMTSBCA1(average maximum testing balanced classification accuracy on 2-way, 3-way and 4-way interactions) and AMTSBCA2 (average AMTSBCA1 on all four QTs) among EFQOMDR, FQMDR, OMDR and MDR when k=1.

Table 4.8 Comparison of MTSBCA and GCVC of PI classifiers among EFQMDR, FQMDR, OMDR and MDR when k=5. (a) For two loci. (b) For three loci. (c) For four loci.

(a)

Classifier	Two loci			
Method	EFQMDR	FQMDR	OMDR	MDR
MTSBCA1	0.588	0.639	0.6	0.476
MTSBCA2	0.563	0.597	0.583	0.456
MTSBCA3	0.488	0.542	0.5	0.45
MTSBCA4	0.476	0.458	0.467	0.417
MTSBCA5	0.472	0.456	0.45	0.413
GCVC1	5	6	5	1

GCVC2	5	8	6	3
GCVC3	5	4	2	6
GCVC4	3	2	2	1
GCVC5	1	4	4	3

(b)

Classifier	Three loci			
Method	EFQMDR	FQMDR	OMDR	MDR
MTSBCA1	0.686	0.583	0.857	0.533
MTSBCA2	0.657	0.486	0.625	0.514
MTSBCA3	0.542	0.45	0.562	0.5
MTSBCA4	0.488	0.4	0.478	0.45
MTSBCA5	0.389	0.388	0.458	0.431
GCVC1	2	2	7	1
GCVC2	6	2	9	2
GCVC3	2	4	4	1
GCVC4	2	1	1	2
GCVC5	2	2	2	1

(c)

Classifier	Four loci			
Method	EFQMDR	FQMDR	OMDR	MDR
MTSBCA1	0.783	0.543	0.651	0.583
MTSBCA2	0.783	0.514	0.651	0.514
MTSBCA3	0.617	0.5	0.613	0.5
MTSBCA4	0.55	0.475	0.590	0.5
MTSBCA5	0.533	0.467	0.583	0.5
GCVC1	8	2	1	5
GCVC2	7	2	1	1
GCVC3	4	1	2	2
GCVC4	5	2	5	1

GCVC5	3	1	5	1

Table 4.9 Comparison of MTSBCA and GCVC of bw classifiers among EFQMDR, FQMDR, OMDR and MDR when k=5. (a) For two loci. (b) For three loci. (c) For four loci.

(a)

Classifier	Two loci			
Method	EFQMDR	FQMDR	OMDR	MDR
MTSBCA1	0.6	0.719	0.667	0.667
MTSBCA2	0.6	0.688	0.517	0.533
MTSBCA3	0.588	0.625	0.483	0.458
MTSBCA4	0.552	0.533	0.444	0.433
MTSBCA5	0.55	0.5	0.433	0.433
GCVC1	9	7	9	8
GCVC2	6	6	6	6
GCVC3	1	3	4	3
GCVC4	4	7	3	5
GCVC5	4	8	5	4

(b)

Classifier	Three loci			
Method	EFQMDR	FQMDR	OMDR	MDR
MTSBCA1	0.640	0.656	0.65	0.659
MTSBCA2	0.610	0.568	0.559	0.583
MTSBCA3	0.599	0.567	0.533	0.55
MTSBCA4	0.567	0.560	0.533	0.523
MTSBCA5	0.558	0.533	0.459	0.522
GCVC1	5	4	5	7
GCVC2	1	6	3	5
GCVC3	7	4	6	3
GCVC4	4	1	5	1
GCVC5	4	4	2	3

(c)

Classifier	Four loci			
Method	EFQMDR	FQMDR	OMDR	MDR
MTSBCA1	0.640	0.659	0.695	0.608
MTSBCA2	0.632	0.628	0.653	0.608
MTSBCA3	0.611	0.628	0.595	0.547
MTSBCA4	0.556	0.620	0.582	0.541
MTSBCA5	0.525	0.604	0.558	0.532
GCVC1	1	7	5	4
GCVC2	1	5	2	2
GCVC3	6	3	1	2
GCVC4	2	1	1	1
GCVC5	1	1	4	2

Table 4.10 Comparison of MTSBCA and GCVC of afw classifiers among EFQMDR, FQMDR, OMDR and MDR when k=5. (a) For two loci. (b) For three loci. (c) For four loci.

(a)

Classifier	Two loci			
Method	EFQMDR	FQMDR	OMDR	MDR
MTSBCA1	0.739	0.588	0.614	0.644
MTSBCA2	0.739	0.537	0.574	0.494
MTSBCA3	0.717	0.537	0.556	0.459
MTSBCA4	0.717	0.515	0.537	0.455
MTSBCA5	0.717	0.5	0.524	0.441
GCVC1	3	7	2	9
GCVC2	1	6	3	5
GCVC3	4	4	6	3
GCVC4	1	4	4	5
GCVC5	1	4	5	3

(b)

Classifier	Three loci			
Method	EFQMDR	FQMDR	OMDR	MDR
MTSBCA1	0.826	0.639	0.634	561
MTSBCA2	0.783	0.574	0.620	541
MTSBCA3	0.636	0.562	0.611	532
MTSBCA4	0.597	0.554	0.520	511
MTSBCA5	0.576	0.541	0.506	0.5
GCVC1	7	10	9	5
GCVC2	8	5	7	4
GCVC3	1	4	1	3
GCVC4	1	5	3	4
GCVC5	4	5	2	2

(c)

Classifier	Three loci			
Method	EFQMDR	FQMDR	OMDR	MDR
MTSBCA1	0.826	0.639	0.634	561
MTSBCA2	0.783	0.574	0.620	541
MTSBCA3	0.636	0.562	0.611	532
MTSBCA4	0.597	0.554	0.520	511
MTSBCA5	0.576	0.541	0.506	0.5
GCVC1	7	10	9	5
GCVC2	8	5	7	4
GCVC3	1	4	1	3
GCVC4	1	5	3	4
GCVC5	4	5	2	2

Table 4.11 Comparison of MTSBCA and GCVC of mw classifiers among EFQMDR, FQMDR, OMDR and MDR when k=5. (a) For two loci. (b) For three loci. (c) For four loci.

(a)

Classifier	Two loci			
Method	EFQMDR	FQMDR	OMDR	MDR
MTSBCA1	0.691	0.681	0.572	0.650
MTSBCA2	0.567	0.630	0.544	0.544
MTSBCA3	0.556	0.630	0.544	0.544
MTSBCA4	0.547	0.576	0.535	0.505
MTSBCA5	0.545	0.505	0.505	0.499
GCVC1	9	9	3	7
GCVC2	1	6	3	5
GCVC3	7	4	2	2
GCVC4	1	1	6	9
GCVC5	5	7	6	3

(b)

Classifier	Three loci			
Method	EFQMDR	FQMDR	OMDR	MDR
MTSBCA1	0.699	0.593	0.586	0.586
MTSBCA2	0.606	0.526	0.520	0.535
MTSBCA3	0.6	0.526	0.491	0.520
MTSBCA4	0.569	0.514	0.483	0.491
MTSBCA5	0.558	0.513	0.451	0.467
GCVC1	3	2	4	3
GCVC2	6	4	3	2
GCVC3	5	4	2	6
GCVC4	6	2	1	2
GCVC5	1	1	6	2

(c)

Classifier	Four loci			
Method	EFQMDR	FQMDR	OMDR	MDR
MTSBCA1	0.594	0.646	0.661	0.618
MTSBCA2	0.581	0.574	0.565	0.563
MTSBCA3	0.576	0.564	0.556	0.475
MTSBCA4	0.566	0.552	0.542	0.45
MTSBCA5	0.533	0.544	0.511	0.433
GCVC1	3	5	2	4
GCVC2	1	1	1	2
GCVC3	4	1	1	1
GCVC4	4	1	1	1
GCVC5	1	1	1	1

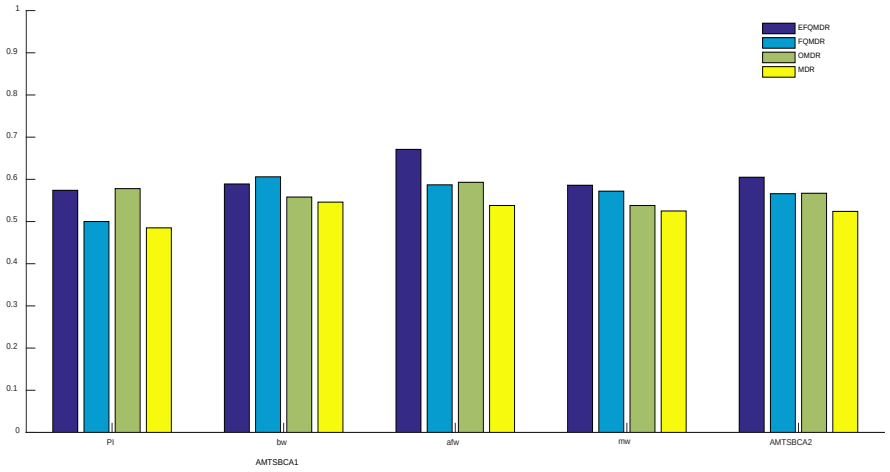

Figure 4.5. Comparison of AMTSBCA1(average maximum testing balanced classification accuracy on 2-way, 3-way and 4-way interactions) and AMTSBCA2 (average AMTSBCA1 on all four QTs) among EFQOMDR, FQMDR, OMDR and MDR when k=5.

In summary the performance of the proposed algorithm is better than that of FQMDR, OMDR and MDR.

4.5 Conclusion

In this chapter, we propose a new method to identify gene-gene interactions associated with complex quantitative traits based on extended fuzzy classification. To reduce loss of information contained in a quantitative trait, it is first divided into several (greater than two) ordinal levels. Then a new ordinal association measure, balanced accuracy based on extended fuzzy classification is employed to select multiple best sets of SNPs as having strongest association with the trait in our proposed EFQMDR algorithm. Experimental results on simulated datasets and real datasets show that our algorithm has better performance in identifying gene-gene interactions associated with a complex quantitative trait.

In step 3 and step 4 of EFQMDR Algorithm, an extended linear member function is used to compute the size of each category in a particular cell, while a traditional linear member function is used to compute the total size of each category in all cells. The reason is that when deciding the label or category of a particular cell, the difference among different categories when being tried to assign to that cell can be reflected by the size of different categories in that cell, rather than the total size of different categories in all cells. Such a difference can be better reflected by an extended linear member function. Experiments also show much better performance when using the extended linear member function and the traditional linear member function in different cases.

In EFQMDR Algorithm, fuzzification is not only applied to the computation of training and testing accuracies, but also applied to the classification of each cell or genotype combination. Experiments show better performance of such a double fuzzification than that of a single fuzzification in either the computation of training and testing accuracies or the classification of each cell or genotype combination.

Alternative methods could be to use balanced accuracy based on traditional member function of fuzzy sets, or balanced signed accuracy where 1 is used to denote that the predicted category is the same as the true category, 0 to denote that the predicted category is close to the true category, -1 to denote that the predicted category is far

from the true category. However our experiments show the performance of our algorithm is better than that of the above two methods.

To test the performance of the algorithm when other types of fuzzy membership functions are used, a sigmoid member function is used. Similar results are obtained, but not so good as the extended linear member function. The following are the traditional sigmoid member function and extended sigmoid member function respectively:

$$\mu_{L3}(x)=\begin{cases}1, & \text{if } x <= P_1\\ \dfrac{(x-P_2)^2}{(x-P_1)^2+(x-P_2)^2}, & \text{if } P_1 < x <= P_2\\ 0, & \text{otherwise}\end{cases} \tag{4.24}$$

$$\mu_{A3}(x)=\begin{cases}\dfrac{(x-P_1)^2}{(x-P_1)^2+(x-P_2)^2}, & \text{if } P_1 <= x <= P_2\\ \dfrac{(x-P_3)^2}{(x-P_2)^2+(x-P_3)^2}, & \text{if } P_2 < x <= P_3\\ 0, & \text{otherwise}\end{cases} \tag{4.25}$$

$$\mu_{H3}(x)=\begin{cases}0, & \text{if } x <= P_2\\ \dfrac{(x-P_2)^2}{(x-P_2)^2+(x-P_3)^2}, & \text{if } P_2 < x <= P_3\\ 1, & \text{otherwise}\end{cases} \tag{4.26}$$

$$\mu_{L4}(x)=\begin{cases}1, & \text{if } x <= P_1\\ \dfrac{2(x-P_3)^2}{(x-P_1)^2+(x-P_3)^2}-1, & \text{if } P_1 < x <= P_3\\ -1, & \text{otherwise}\end{cases} \tag{4.27}$$

$$\mu_{A4}(x)=\begin{cases}\dfrac{2(x-P_0)^2}{(x-P_0)^2+(x-P_2)^2}-1, & \text{if } P_0 < x <= P_2\\ \dfrac{2(x-P_4)^2}{(x-P_2)^2+(x-P_4)^2}-1, & \text{if } P_2 < x <= P_4\\ -1, & \text{otherwise}\end{cases} \tag{4.28}$$

$$\mu_{H4}(x)= \begin{cases} -1, & \text{if } x <= P_1 \\ \dfrac{2(x-P_1)^2}{(x-P_1)^2+(x-P_3)^2} - 1, & \text{if } P_1 < x <= P_3 \\ 1, & \text{otherwise} \end{cases} \quad (4.29)$$

where $P_0=P_1-(P_2-P_1)=2P_1-P_2$, $P_4=P_3+(P_3-P_2)=2P_3-P_2$.

We can also use parameters in the fuzzy membership functions. To adaptively estimate the parameters, we can compute hit ratios or testing balanced accuracy for new data and get overall average hit ratios or average testing balanced accuracies for all data for different parameters. Then we can select the parameter having the highest hit ratio or average testing balanced accuracy for all data.

Chapter 5
A Multi-stage Approach to Detect Gene-gene Interactions Associated with Multiple Correlated Phenotypes

5.1 Introduction

Genome-wide association studies (GWAS) which identify association between a genotype and a phenotype univariately with several hundred thousand to tens of millions SNPs may be underpowered to detect polygenetic effect of numerous genetic variants with small individual effects. On the other hand, much evidence has shown the correlation among quantitative phenotypes. For example, hypertension is evaluated using systolic and diastolic blood pressures; obesity is related to the increase of muscle weight and fat weight. Exploiting the correlation among these phenotypes may strengthen power to detect additional genetic variants with small effects across multiple phenotypes or pleiotropy effects. Identifying those interacting genetic factors shared by related multiple phenotypes will give us a deeper understanding on genetic mechanism on complex traits and complex diseases.

Therefore multi-locus analysis combined with multi-phenotype analysis has become a new tendency in the genome wide association study.

5.2 Related works

First, methods were proposed to consider multiple correlated phenotypes associated with genetic marker.

Multivariate analysis of variance (MANOVA) is the natural extension of the analysis of variance (ANOVA) for correlated multivariate phenotypic traits (Smith et al., 1962). Its assumption of the multivariate normal distribution provides many

good statistical properties for testing and estimation (Morrison, 1967).

Let n denote the total number of observation points in an experiment. If there are p outputs which are observed on each of the n observation points, then a p×n output matrix would be obtained.

$$\mathbf{Y}=\begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1n} \\ Y_{21} & Y_{22} & \cdots & Y_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ Y_{p1} & Y_{p2} & \cdots & Y_{pn} \end{bmatrix}=[\mathbf{Y_1, Y_2, \ldots, Y_n}] \tag{5.1}$$

where Y_i is the column vector of p outputs observed at the ith observation point. Then the usual fixed model of MANOVA can be written as:

$$\mathbf{Y'=A\xi+\varepsilon} \tag{5.2}$$

where **A** is an n×m input matrix, ξ is an parameter m×p matrix which decides the effects of the inputs, ε is an n×p error matrix following p-dimensional normal distribution, p is assumed to be $\leq$ (n-r).

From these assumptions, we have that $\mathbf{Y_1, Y_2, \ldots, Y_n}$ are independent samples with distribution N[E($\mathbf{Y_i}$), Σ]. Assumption (d) ensures that the sample error matrix is positive definite almost everywhere.

Since **A** has a rank r, we can partition **A** in the form $[A_I A_D]$, where A_I is a basis of A and form an n×r submatrix, A_D is a n×(n-r) submatrix. Therefore we can rewrite (1) as

$$\mathbf{Y'}=[\mathbf{A_I A_D}]\begin{bmatrix} \xi_I \\ \xi_D \end{bmatrix}+\boldsymbol{\varepsilon} \tag{5.3}$$

The object is to test the general linear hypothesis.

$$H_0: \mathbf{C\xi M}=O \tag{5.4}$$

where **C** is an s×m matrix of rank $s \leq r \leq m < n$, **M** is a p×u matrix of rank $u \leq p$ and O is an s×u null matrix. **C** is to used to state between treatments hypothesis, while **M** is used to state between responses hypothesis.

Let's consider a simple example. Suppose in an experiment with three treatments, two responses have been measured on each experimental unit, then the fixed effects of the three treatments on the two responses can be expressed in the following table:

Treatment	Response	
	$\mathbf{Y_1}$	$\mathbf{Y_2}$
T_1	ξ_{11}	ξ_{12}
T_2	ξ_{21}	ξ_{22}
T_3	ξ_{31}	ξ_{32}

Where $\boldsymbol{\xi_{ij}}$ represents the effect of treatment T_i on response $\mathbf{Y_i}$ (i=1,2,3 and j=1,2). The hypothesis of no difference between treatments can be stated as

$$H: \begin{bmatrix} \xi_{11} \\ \xi_{12} \end{bmatrix} = \begin{bmatrix} \xi_{21} \\ \xi_{22} \end{bmatrix} = \begin{bmatrix} \xi_{31} \\ \xi_{32} \end{bmatrix}$$

This is equivalent to $\xi_{11}=\xi_{21}=\xi_{31}$ and $\xi_{12}=\xi_{22}=\xi_{32}$. The alternative hypothesis H_1 is: not $H_{0,}$ i.e., at least one of the above equalities is violated. Here **M** is 2×2 identity matrix,

$$\mathbf{C}=\begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}, \qquad \xi=\begin{bmatrix} \xi_{11} & \xi_{12} \\ \xi_{21} & \xi_{22} \\ \xi_{31} & \xi_{32} \end{bmatrix}.$$

The hypothesis of no difference between responses can be stated as

$$H: \begin{bmatrix} \xi_{11} \\ \xi_{21} \\ \xi_{31} \end{bmatrix} = \begin{bmatrix} \xi_{12} \\ \xi_{22} \\ \xi_{32} \end{bmatrix}$$

This is equivalent to $\xi_{11}=\xi_{12}$, $\xi_{21}=\xi_{22}$ and $\xi_{31}=\xi_{32}$. Here **C** is 3×3 identity matrix, $\boldsymbol{\xi}$ is the same as above,

$$\mathbf{M} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

The alternative hypothesis may be stated as

$$H_1: \mathbf{C\xi M=n} \qquad (5.5)$$

where $\mathbf{n} \neq O$. Let **C** be partitioned in the form $[\mathbf{C_I C_D}]$ determined by the partitioning of $\boldsymbol{\xi}$ in the form

$$\begin{bmatrix} \xi_I \\ \xi_D \end{bmatrix}$$

Now we define two matrices. The first is the matrix due to the hypothesis,

$$\mathbf{S_H}(u\times u)=\mathbf{M'YA_I(A_I'A_I)^{-1}C'_I[C_I(A_I'A_I)^{-1}\ C'_I]^{-1}\ C_I(A_I'A_I)^{-1}A'_IY'M}. \quad (5.6)$$

The second is called the matrix due to error

$$\mathbf{S_E}(u\times u)= \mathbf{M'Y[I(n)-A_I(A_I'A_I)^{-1}A'_I]\ Y'M} \quad (5.7)$$

where **I(n)** denotes the identity matrix of order n.

$\mathbf{S_H}$ is symmetric at least positive semi-definite of rank k=min[rank(**M**),rank(**C**)], while $\mathbf{S_E}$ is symmetric positive definite since $u \le p \le (n-r)$.

To test (5.4) against (5.5), there are at least three alternative criteria, the largest-root criterion(C1), the product-of –the roots criterion(C2) and the sum-of–the roots criterion(C3):

(C1) $\text{chmax}(\mathbf{S_H}\ \mathbf{S_E}^{-1})/[1+\text{chmax}(\mathbf{S_H}\ \mathbf{S_E}^{-1})]$;

(C2) $\Lambda=|\ \mathbf{S_E}\ |/|\ \mathbf{S_H}+\mathbf{S_E}\ |=1/|\ \mathbf{S_H}\ \mathbf{S_E}^{-1}+\mathbf{I}\ |$;

(C3) $\text{tr}(\mathbf{S_H}\ \mathbf{S_E}^{-1})$;

where $\text{ch}_{\max}$ denotes the largest characteristic root, | | denotes determinant and "tr" denotes trace (sum of the diagonal elements).

In mixed effects models, the value of a phenotype is related to a mixture of the genetic marker effect and random effects caused by other correlated phenotypes (Laird and Ware, 1982; Fitzmaurice and Laird, 1993). Random effects are described using a multivariate normal distribution with elements of variance and covariance matrices as parameters. The parameters of these effects can be estimated using restricted maximum likelihood method.

Mixed effects models include linear mixed effects model (LME) and generalized linear mixed effects model (GLMM)

If y_{ij} represents the jth (j=1, . . . , J) component of the J-dimensional phenotype of the jth (j = 1, . . . , J) individual, g_i represents the genotype of a genetic marker of the ith individual and $X(g_i)$, its corresponding score, then the linear mixed effects model has the following form:

$$Y_{ij=}\beta_0+\beta_j\ X(g_i)+\eta_{ij}+e_{ij} \quad (5.8)$$

where β_0 represents effects caused by factors; β_k represents the effect size of $X(g_i)$ on the jth phenotype; $\eta_{ij}(j=1, \ldots, J)\sim N(0, \Sigma)$ are the random effects caused by multiple correlated components of a phenotype within ith person; e_{ij} is the random errors iid. ~

$N(0, \sigma_e^2)$.η_{ij} are independent between any two individuals. The null hypothesis of no association between the genetic marker and any phenotype component can be stated as $H_0 : \beta 1 =, \ldots, \beta_J = 0$.

This model can be extended to generalized mixed effects model (GLMM) for phenotypes consisting of categorical components:

$$E(y_{ij} \mid \eta_j) = \mu^{-1}(\beta_0 + \beta_j X(g_i) + \eta_{ij}) \quad (5.9)$$

where μ^{-1} is the inverse of a link function. μ takes the form of identity function when the components have Gaussian distribution and the model reduced to the linear mixed effects model; for binary components, it takes the form of logit function $\mu(x) = \ln(x/1 - x)$.

The likelihood ratio test or Wald chi-squared test can be used to test the null hypothesis under the LME or GLMM. The Wald chi-squared test statistic can be expressed as $\boldsymbol{\beta}^T \text{cov}(\boldsymbol{\beta})^{-1} \boldsymbol{\beta} \sim \chi_K^2$,where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ is estimated using (5.8) or (5.9). When the effect sizes are similar, the null hypothesis of $\beta_1 = \cdots = \beta_K = \beta$ can be tested with a a single degree-of-freedom (df) test $\hat{\beta}/\text{se}(\hat{\beta})$ and this test has better performance than the multi-df Wald chi-squared test in this case.

The principal component of quantitative trait locus heritability (PCQH) uses a linear combination of the phenotypes with coefficients which make the linear combination and the genetic marker have maximum correlation (Lange et al., 2003; Lange et al., 2004; Klei et al., 2008).

For a multivariate phenotype consisting of all continuous components that are approximately normal distributed, variable reduction approaches can be used. It uses a linear combination of the components:

$$Y' = a_1 Y_1 + a_2 Y_2 + \ldots + a_M Y_M \quad (5.10)$$

where $Y' = (Y_1, Y_2, \ldots, Y_M)$ is an m-dimensional phenotype. The principal component of quantitative trait locus heritability (PCQH) uses a linear combination of the components with coefficients making the linear combination and the genetic marker have maximum correlation, and therefore the phenotype variation of Y' reflected by the genetic marker (Lange et al., 2003; Lange et al., 2004; Klei et al., 2008).

Let y_{ij} denote the jth (j=1,…,M) component of Y' of the ith (i=1,…,N) subject, x_i denote the number of copies of the minor allele of a QTL for the ith subject. For each phenotype j, the relationship between y_{ij} and x_i can be approximated with a linear regression model:

$$y_{ij}=\mu_l+\beta_j x_i+\varepsilon_{ij} \tag{5.11}$$

where μ_l is the intercept of the model, β_j is the effect of the QTL on the jth trait and ε_{ij} is the residual error being normally distributed with mean 0.

The total phenotype variance can be partitioned as

$$V_P=V_Q+V_R, \tag{5.12}$$

where $V_Q=Var(\beta_1 x,\ldots,\beta_M x,)$ is the genetic variance due to the QTL and V_R is the residual variance. Accordingly the variance of Y' attributable to the QTL is

$$h_A^2=\frac{A^t V_Q A}{A^t V_P A}, \tag{5.13}$$

where A=(a_1,…, a_M).

In canonical correlation analysis, coefficients which maximize the squared correlation between Y' and the score of genetic marker X(g) are used. (Muller and Peterson,1984).

Here canonical correlation refers to $\hat{\rho}$ =corr(Y',X). $\hat{\rho}$ can be obtained by partitioning the covariance matrix of *Y* and *X* as follows:

$$\operatorname{cov}\begin{bmatrix} Y \\ X \end{bmatrix}=\begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix} \tag{5.14}$$

where Σ is the covariance-variance matrix. The sample covariance-variance matrix can be used to estimate each of the submatrix. The canonical correlation $\hat{\rho}$ can be expressed as $\hat{\rho}=\Sigma_{XY}A/(A^t\Sigma_{YY}A\,\Sigma_{XX})^{1/2}$ where A=(a_1,…, a_M) is the coefficients of Y' and has its maximum value as the squared root of the largest eigenvalue of $\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ when A is the corresponding eigenvector.

Similar to PCQH, canonical correlation analysis uses a linear combination of components to maximize its variation reflected by the genetic marker. Their

difference is that the former partition the sample into two subsets, one is used to estimate the coefficients, another is used to test the association, while the latter uses the whole sample to evaluate squared correlation.

Multivariate phenotypes can also be analyzing phenotype-genotype association for each phenotype and then combining their test statistics. This approach takes the advantage of simplicity of methods for analyzing univariate phenotypes, especially when multivariate phenotype consists of different types of components such as dichotomous, categorical and continuous. In addition, there are already many ready-made univariate phenotype analysis results available for many complex traits.

a) Methods for Homogeneous Genetic Effects across Phenotypes

Let $T=[T_1,\ldots, T_K]$ denote a vector of K test statistics for each individual phenotype analysis following a multivariate normal distribution with mean $\tau=(\tau_1,\ldots,\tau_k)^T$ and a nonsingular covariance matrix. The null hypothesis of no association for any phenotype is H_0: $\tau=(\tau_1,\ldots,\tau_k)^T=0$. O'Brien used the following linear combination of $T_1,\ldots, T_K$ which maximizes the power when$\tau_1=\ldots=\tau_k\neq 0$ to combine K individual test statistics (O'Brien, 1984):

$$S=e^T\Sigma^{-1}T \qquad (5.15)$$

where $\mathbf{e}=(1,\ldots,1)^T$ is the weight.

b) Methods for Heterogeneous Genetic Effects across Phenotypes

O'Brien's method may not be efficient when the means are not similar. In these cases, Yang et al. partitioned the sample into two subsets, one is used to estimate weights w to replace the uniform weight $\mathbf{e}^T$, another is used to estimate **T** in the above equation (Yang et al., 2010).

Another way is to use a quadratic form to combine individual association test statistics. For example, Xu et al. employed the following Wald chi-squared type test statistic (Xu et al., 2003):

$$S_w=T^T\Sigma^{-1}T \qquad (5.16)$$

It replaces e in (5.15) with T. The distribution of S_w is a linear combination of one degree-of-freedom chi-squared distribution with coefficients being the eigenvalues of

Σ. To prevent the decreased power of (5.16) when the number of phenotypes increases due to "curse of dimensionality", the variance-covariance matrix Σ is taken away from (2) to form the following test statistic (Yang and Wang, 2012):

$$S_{sq} = T^T T \tag{5.17}$$

It follows a a χ^2_d +b distribution with

$$a=\frac{\sum_{i=1}^{K} c_i^3}{\sum_{i=1}^{K} c_i^2}, \quad b=\sum_{i=1}^{K} c_i - \frac{(\sum_{i=1}^{K} c_i^2)^2}{\sum_{i=1}^{K} c_i^3}, \quad d=\frac{(\sum_{i=1}^{K} c_i^2)^3}{(\sum_{i=1}^{K} c_i^3)^2} \tag{5.18}$$

When there are highly correlated phenotypes, d may be less than K.

TATES (Trait-based Association Test that uses Extended Simes procedure) combines p-values of test of association with genetic variants for each phenotype of a multivariate trait to get an overall trait-based p-value to test the association of the trait with genetic variants by calculating through correlations between phenotypes and using the effective number of p-values (Van der Sluis S, et al., 2013).

Suppose there are m phenotypes contained in a trait. A statistically appropriate method (e.g., linear or logistic regression) is used to test the association between all m phenotypes and all n genotyped genetic variants (GVs) separately, rather than combining them into one general phenotype. Then p values of m phenotypes for a given GV are combined to obtain one overall trait-based p-value P_T as follows:

$$P_T=\min\left(\frac{m_e p_j}{m_{ej}}\right) \tag{5.19}$$

where $p_1 \ldots p_m$ are the ascending p-values of the m phenotypes for a given GV, m_e is the number of independent p-values which all m phenotypes for a given GV are equivalent to, m_{ej} is the number of independent p-values which the top j p-values with j running from 1 to m are equivalent to. Therefore P_T is the smallest weighted p-value used to test the null hypothesis that none of the phenotypes is associated with the GV, while the alternative hypothesis is that at least one of the phenotypes is associated with the GV.

m_{ej} can be estimated using eigenvalues of the m×m correlation matrix ρ between the m p-values. It is calculated as:

$$m_{ej} = j - \sum_{i=1}^{j} I(\lambda_i - 1) \quad (5.20)$$

where λ_i is the i^{th} eigenvalue, and I(λ_i -1) is an indicator func`tion which equals 0 if $\lambda_i \leq 1$ and 1 if $\lambda_i > 1$. This means m_{ej} equals j minus the number of eigenvalues which are greater than 1. m_{ej} equals m_e when all phenotypes are selected for top phenotypes.

The m×m correlation matrix ρ between the p-values can be accurately approximated through the m×m correlation matrix r between the phenotypes.

Later, methods combining multi-locus analysis and multi-phenotype analysis were proposed.

Low rank regularization adopts the tensor $l_{2,1}$ norm regularization to identify imaging markers having common effects on all regression tasks and time points and regularizes the unfolded coefficient tensor with the trace norm to identify interaction among SNPs by achieving low rank (Wang et al., 2012).

Unlike most studies where SNPs are selected and associated to disease status or imaging phenotypes, the authors use as input and SNP values as output to examine how phenotypic values influence SNP values, expecting to identify genetic associations with imaging phenotypes from a different angle.

Measuring the imaging markers on different time points may increase power to identify genetic markers associated with them. The longitudinal input imaging features at *T* consecutive time points can be described with a set of matrices $X=\{X_1,X_2,\ldots,X_T\} \in \mathbb{R}^{d\times n\times T}$, where X_t is the measurement matrix of imaging markers at time *t* ($1\leq t\leq T$) and therefore X is a tensor with *d* imaging features, *n* subject samples and *T* time points. The matrix $Y=[\mathbf{y}_1,\ldots,\mathbf{y}_n]^T \in \mathbb{R}^{n\times c}$ denotes the output genotypes of *c* SNPs for the *n* subject samples. Then the associations between the longitudinal imaging phenotypes *X* and the genotypes *Y* can be explored by learning a model from {*X*,*Y*}.

However this method has the limitation of ignoring valuable information contained in the longitudinal patterns of the phenotypic inputs. Therefore a unified longitudinal regression model called the task-correlated longitudinal sparse regression model that

unifies the measurement at different time points is proposed to learn a coefficient tensor $B=\{B^1,\cdots,B^T\} \in \mathrm{R}^{d\times c\times T}$ to explore temporal patterns of the coefficient matrices by using the low-rank structured sparse regularizations.

The regression coefficient matrix can be learned for each time point individually by solving the following optimization problem:

$$\min_B J_0=\min_B (L(B)+\gamma\|B\|_2^2)=\min_B (L(B)+\gamma\sum_{t=1}^{T}\sum_{k=1}^{d}\|\mathbf{b}_t^k\|_2^2) \tag{5.21}$$

where $\mathbf{b}_t^k$ denotes the kth row of coefficient matrix B_t and $L(B)$ is the longitudinal loss defined as follows:

$$L(B)=\sum_{t=1}^{T}\|\boldsymbol{X}_t^T B_t-Y\|_F^2 \tag{5.22}$$

Since J_0 can be decoupled for each individual time point, the temporal correlations between the imaging features and the SNPs are not reflected. To reflect such correlations, the structured sparse regularization is introduced into the longitudinal data regression and feature selection model:

$$\min_B J_1=\min_B (L(B)+\gamma_1\sum_{k=1}^{d}(\sum_{t=1}^{T}\|\mathbf{b}_t^k\|_2^2)^{1/2}) \tag{5.23}$$

With this expression, J_1 can not be decoupled over time dimension and thus the model can reflect temporal patterns of the phenotypic components. The second term in Equation (3) is actually a tensor extension of the widely used $l_{2,1}$-norm for matrices.

Since many SNPs are interrelated and their effects on phenotypic traits can overlap，the columns $(\mathbf{b}_t)_j$ of B_t should have some linear correlation, causing B_t $(1\leqslant t\leqslant T)$ to have low rank. The unfolding operation of an n-mode tensor $T\in \mathrm{R}^{I1\times I2\times\cdots\times In}$ along its kth mode can be denoted as $\mathrm{unfold}_k(T)= T_{(k)}\in \mathbb{R}^{Ik\times(I1\cdots I(k-1)I(k+1)\cdots In)}$. Then interrelation among SNPs can be reflected by minimizing the rank of $B_{(1)}=[B_1,B_2,...,B_T]\in \mathrm{R}^{d\times(c\times T)}$ and the optimization problem becomes:

$$\min_B J_1=\min_B (L(B)+\gamma_1\sum_{k=1}^{d}(\sum_{t=1}^{T}\|\mathbf{b}_t^k\|_2^2)^{1/2}+\gamma_2\|B\|_*) \tag{5.24}$$

where $\|\ \|_*$ denotes the trace-norm of a matrix, and the subscript of the matrix $B_{(1)}$ is

omitted for notation brevity. The trace-norm of a matrix$M \in \mathrm{R}^{n\times m}$ is defined as $\|\mathrm{M}\|_* = \sum_{i=1}^{\min(n,m)} \sigma_i = \mathrm{Tr}(MM^T)^{1/2}$ and has been shown to be the best convex approximation of the rank-norm.

Graph-guided fuse lasso incorporates the correlation structure among multiple phenotypes into a multivariate regression model using a threshold correlation graph to identify the genetic markers having common effects on a group of phenotypes having high correlation with high sensitivity and specificity (Kim S, *et al.*, 2009).

Let **X** be an $N\times J$ matrix of genotypes for N individuals and J SNPs. Each genotype has a value 0, 1 or 2 to represent its number of minor alleles. Let **Y** be an $N\times K$ matrix of K QT values for the above N individuals and $\mathrm{y_k}$ be the k-th column of **Y.** Then the relationship between multiple quantitative traits and multiple SNPs can be sought by fitting a linear regression model for each trait separately:

$$\mathbf{y_k} = \mathbf{X}\boldsymbol{\beta_k} + \boldsymbol{\varepsilon_k}, \qquad k=1,...,K, \tag{5.25}$$

where $\boldsymbol{\beta_k}$ is a vector of regression coefficients measuring significance of association, and $\boldsymbol{\varepsilon_k}$ represents N independent errors. Each column of **X** and **Y** has a mean value of zero, so there is no intercept in equation (1). The estimates of $\mathbf{B}=\{\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_K\}$ can be obtained by minimizing the following residual sum of squares:

$$\hat{B} = \mathrm{argmin} \sum_k (y_k - X\beta_k)^T \bullet (y_k - X\beta_k) \tag{5.26}$$

Since straightly applying Equation (5.25) to detect SNP markers with large J could make the estimated regression coefficients unstable and make many irrelevant markers have significant regression coefficients to cause difficulty in interpretation, sparse regression methods have been proposed that select a subset of markers having true association. Ridge regression adds a penalizing term in Equation (5.25) with the L_2 norm of regression coefficients to shrink them toward 0 but does not set them exactly to 0. Instead lasso regression adds a penalizing term with the L_1 norm of regression coefficients to set them exactly to 0 as follows:

$$\hat{B}^{lasso} = \mathrm{argmin} \sum_k (y_k - X\beta_k)^T \bullet (y_k - X\beta_k) + \lambda \sum_{k,j} | \beta_{kj} | \tag{5.27}$$

where λ is a parameter that regularizes the extent of sparsity of the estimation of regression coefficients. The larger λ is, the more penalization is imposed and the more regression coefficients which are set to 0.

Solving equation (5.26) is equivalent to solving K sets of regression coefficients independently for K traits and therefore they could not reflect correlation among traits which may be caused by common SNPs.

In order to identify these common SNPs having influence on multiple traits, pairwise Pearson correlation coefficients which measure the strength of correlation between two traits within multiple traits can be represented as a graph with weighted edge and then embedded in the lasso framework. If two traits have a correlation coefficient above the given threshold ρ, they are connected with an edge (m,l) whose weight is equal to the absolute value of correlation coefficient $|r_{m,l}|$. The authors assume that highly correlated traits may be influenced by common SNP markers and possibly have the same amount of influence from these markers. This can be reflected by an adding a penalty term which tends to make two regression coefficients β_{jm} and β_{jl} similar, where j represents any marker, m and l represent two highly correlated traits, and is weighted by the strength of their correlation to control the extent of their similarity. The strength of correlation can be generalized to its monotonically increasing function to get the following estimate of the regression coefficients:

$$\hat{B}^{GW} = \text{argmin} \sum_k (y_k - X\beta_k)^T \bullet (y_k - X\beta_k) + \lambda \sum_{k,j} |\beta_{kj}| + \gamma \sum_{(m,l)\in E} f(r_{ml}) \sum_j \left|\beta_{jm} - \text{sign}(r_{ml})\beta_{jl}\right| \quad . \tag{5.28}$$

where λ and γ control the amount of penalization. The last term controls the similarity between β_{jm} and $\text{sign}(r_{m,l})\beta_{jl}$. The larger γ and $f(r_{ml})$ is, the more similar they are. Two examples of $f(r_{ml})$ are $f_1(r)=|r|$ and $f_2(r)=r^2$. With the two penalty terms combined, equation (5.28) makes many regression coefficients become 0 and decreases the differences among multiple highly correlated phenotypes for each marker for the remaining non-zero regression coefficients so that each marker has similar influence on these correlated phenotypes.

For trait networks, identifying genetic markers that influence all traits in each sub-network where nodes are densely connected can strengthen the power of detecting genetic markers having small effects across multiple phenotypes. The fusion effect can propagate automatically to all nodes in such subgroups to make regression coefficients having similar values on correlated traits if a genetic marker has pleiotropic effect on those traits. On the contrary, if a group of nodes are sparsely connected, such fusion effect would not be obvious in the group.

However the above methods depend on a specific model and only use a linear combination of genetic markers in the regression model. Although they can detect multiple genetic loci associated with diseases or traits simultaneously, they can not detect interactive loci, because interaction should be represented by cross terms in the model and these cross term would make regression models much more complicated.

MDR was originally proposed as a method for detecting gene-gene interaction without significant main effects in case control studies and then extended to ordinal traits and quantitative traits (Yu, et al., 2015).

Multivariate quantitative MDR (Multi-QMDR) extended MDR to multivariate phenotypes by reducing multivariate phenotypes into a univariate score based on principle component analysis and then labeling the samples as high-risk and low-risk using this score according to MDR.

Let $Y=(Y_1,\ldots,Y_d)$ be a d-dimensional phenotype. The sample covariance matrix of d components of Y can be decomposed as

$$S=\sum_{j=1}^{d} \lambda_j \xi_j \xi_j^T \tag{5.29}$$

where λ_j is the j-th eigenvalue of S and ξ_j is its corresponding eigenvector.

Let $PC_{ij} = Y_i^T \xi_j$ be the j-th principal component (PC) score for the i-th subject. The following three summary scores can be employed to classify each cell of a genotype combination as a high-risk or low-risk group.

(1) Weighted Summation of PC (WPC):

$$S_{wi}=\sum_{j=1}^{d} PC_{ij}\sqrt{\lambda_j} \tag{5.30}$$

(2) First PC (FPC):

$$S_{Fi} = PC_{i1} \tag{5.31}$$

(3) Weighted Squared Summation of PC (WSPC):

$$S_{Si} = \sum_{j=1}^{d} PC_{ij}^2 \lambda_j \tag{5.32}$$

If the score of a genotype combination is greater than or equal to the global mean, it is classified into H group, otherwise, it is classified into L group. The HT2 statistic on the original traits or t statistic on the combined univariate trait for comparing H group and L group is then employed to choose the best m-order interaction in each K-fold cv and CV consistency (CVC) is employed to select the best overall m-order model.

However biological explanation of principle components for this method is difficult.

In this chapter, we propose a new method extended from MDR to identify interactive genetic loci associated with multiple correlated phenotypes by selecting the best classifier according to not only the training accuracy of the phenotype under consideration but also other phenotypes with weights determined mainly by their pair correlation with the phenotype under consideration. If a set of SNPs have interaction on different phenotypes, then these phenotypes would have correlation among each other. Conversely if some phenotypes have correlation among them, there may not be a common set of SNPs having interaction on these phenotypes. However current methods use all correlations among phenotypes to select common sets of SNPs having interaction on multiple phenotypes. The selected sets of SNPs may be very unreliable. To select more reliable sets of SNPs, we also identify interactive genetic loci associated with multiple correlated phenotypes through repeated selection. At first all correlated phenotypes are used to identify interactive genetic loci, but then less and less phenotypes are used to select more reliable ones.

In section 3 of this chapter, the procedure of our proposed Multivariate Quantitative trait based Ordinal MDR (MQOMDR) algorithm is described in detail. Then two real datasets are described and experimental results are given in section 4. Finally conclusion and discussion are made in section 5.

5.3. Methods

5.3.1. Deciding Weight of Correlated Phenotypes

For quantitative traits, most extended MDR also classified the outcome into two groups: high and low level groups, which results in the loss of the large variability of the quantitative outcome. Therefore in order to better use the information contained in the quantitative trait, we first classify the quantitative outcome into several (greater than two) ordinal levels. For MDR extended to ordinal traits, a natural way is to select the best classifier according to the training balanced accuracy of the phenotype under consideration. However for multiple phenotypes, the training balanced accuracy of other correlated phenotypes may also be useful in selecting the best classifier. Therefore we use a weighted sum of the phenotype under consideration and other correlated phenotypes (the absolute value of whose correlation coefficients with the phenotype under consideration are greater than or equal to 0.2) to select the best classifier. Since highly correlated phenotypes provide redundant information (Yu, et al.,2015), as a first step we take the weight of a phenotype j as:

$$\text{weight1}(j)=(0.5-\text{abs}(\text{abs}(\text{cov}(i,j))-0.5))/0.5 \quad (\text{abs}(\text{cov}(i,j)\geq 0.2) \qquad (5.33)$$

where cov(i,j) is the correlation coefficient of i: the phenotype under consideration, and j: another phenotype correlated with i, abs() is the absolute function. Weight1() function obtained its maximum value when abs(cov(i,j))=0.5, i.e., a correlated phenotype j has a maximum contribution to the phenotype i under consideration when their correlation coefficient is 0.5. The reason is that, when cov(i,j)=0, the two phenotypes have no correlation, when abs(cov(i,j))=1, the two phenotypes are actually the same phenotype, they have no contribution to each other in these two cases. Therefore it's natural to assume that when abs(cov(i,j))=0.5, the contribution is the biggest.

If another phenotype k correlated with i whose weight has been evaluated beforehand has a high correlation with j (cov(j,k)>0.5), then j will also become

redundant. So as a second step, we take the weight as:

$$\text{weight(j)}=\begin{cases}\text{weight1（j），} & \text{if cov(j,k)} \le 0.5 \\ (1-\max_{k\in A}(\text{abs(cov(j,k))}) \times \text{weight1(j)}/0.5, & \text{if cov(j,k)} > 0.5\end{cases} \quad (5.34)$$

where A is a set of phenotypes whose weights have been evaluated beforehand. 0.5 is used in (5.34) because it is in the middle of cov(j,k)=0, no redundancy, and cov(j,k)=1, biggest redundancy.

Since if the phenotype k having a smaller weight(k) evaluated before weight(j) is evaluated, then even if j has a high weight1(j), its weight(j) will also become small, resulting in the loss of impact of j on i without being compensated from k, therefore weight() function should be evaluated in a descending order.

5.3.2. Filtering of Correlated Phenotypes

Then we group phenotypes which have the same set of SNPs that has the largest CVC and calculate the average CVC of the same set of SNPs for each phenotype in each group using all phenotypes in the same group only. Remove a phenotype in each group which has the smallest CVC and calculate the average CVC again. This process is repeated until the average CVC is equal to or smaller than that in the last repetition or there are only two phenotypes left in the group.

5.3.3. The MQOMDR Algorithm

According to the above analysis, we have the following procedure for our proposed MQOMDR algorithm multiple quantitative phenotypes:

1. For each of q quantitative phenotypes, divide the range of the phenotype into *J* intervals and label them as categories 1,2,…,*J* respectively.
2. Partition the dataset into *L* subsets for *L*-fold cross-validation (CV). Use one of the *L* subsets as a testing set and the rest as a training set.
3. For each *m*-way interaction derived from *m* SNPs and each of q phenotypes, let n_{ij}

be the number of individuals belonging to category j with the ith multi-locus genotype in the training set, n_{+j} be the total number of individuals belonging to category j in the training set, where $i = \{1, 2,...,3^m\}$and $j = 1, 2,..., J$. Then all individuals with the ith multi-locus genotype will be assigned into the class $c(i)$ by the classifier corresponding to the m given SNPs as follows:

$$c(i)= \arg\max_{j\in\{1,\ldots,j\}}\left(\frac{n_{ij}}{n_{+j}}\right) \tag{5.35}$$

4. Compute the weighted sum of training balanced accuracies for each of q phenotype, e.g., for phenotype i, the weighted sum is

$$TA(i)+\sum_{j\neq i,\mathrm{cov}(i,j)\geq 0.2} weight(j)\times TA(j) \tag{5.36}$$

where $TA(\ )$ is the training balanced accuracy function and weight(j) is evaluated in a descending order.

5. Select the best classifier that has the largest value in (4) for each of q phenotypes.
6. Repeat steps 3-5 on all L CV dataset.
7. The strongest gene-gene interaction is selected according to the cross-validation consistency (CVC) for each of q phenotypes.
8. Group phenotypes which have the same set of SNPs that has the largest CVC for a fixed order of interaction.
9. For each phenotype in each group, using all phenotypes in the same group only to execute MQOMDR algorithm again and calculate CVC of the same set of SNPs. For each group, if the average CVC for all phenotypes becomes smaller, the group is abandoned.
10. For each remaining group, remove a phenotype which has the smallest CVC and repeat step 9 again. This process is repeated until the average CVC for phenotypes remaining in the group is equal to or smaller than that for the same phenotypes in the last repetition. If the average CVC in the last repetition is larger than or equal to that of MDR, then the corresponding set of SNPs are considered as having strongest interaction on all phenotypes left in the group. Or if there are only two phenotypes left in the group and the average CVC is larger than or equal to that of MDR, then

the corresponding set of SNPs are also considered as having strongest interaction for these two phenotypes. Otherwise, i.e. the average CVC is less than that of MDR, remove a phenotype which has the next smallest CVC in the beginning of step 10 and repeat this step again.

5.4 Experimental Results and Analysis

5.4.1. Experimental setup

Since mouse weight and body size provide analogy to human traits of adult weight and height, we use two real mouse datasets for our experiments.

In the first experiment, an intercross mouse population from intercross of DBA2 and NMRI8 is used to identify genetic determinants for body weight and its components, such as fat weight and muscle weight. NMRI8 is a long-term high body weight-selected mouse line and analyzed at the age of 6 weeks. It is extremely different in body composition from the control mouse line DBA2. For the DBA2 x NMRI8 intercross population, there are 275 mice (142 females, 133 males), 98 markers and 18 phenotypes. We use six phenotypes in the experiment which are body weight (bw), abdominal fat (afw), muscle weight (mw), the weight of liver (liver), the weight of kidney (kidney) and the weight of spleen (spleen). The data were downloaded from the QTL Archive curated by the Jackson Laboratories http://phenome.jax.org/db/q?rtn =projects/projdet&reqprojid=213.

Each of the above six continuous phenotypes is transferred to an ordinary phenotype with three categories. Let μ, σ be the mean value and variance of the quantitative phenotype, any phenotype value smaller than μ-σ/2 is classified as low category; any value between μ-σ/2 and μ+σ/2 is classified as middle category; any value larger than μ+σ/2 is classified as high category.

Our proposed MQOMDR method is used to select the best 2-way and 3-way common gene-gene interactions in the above dataset associated with multiple

phenotypes among the six ordinal phenotypes transferred from continuous phenotypes. As in [26], CVC and whether same sets of genetic markers are identified as the best models in different ordinal phenotypes are used to evaluate the performance of MQOMDR and compare with that of MDR, Quantitative MDR (QMDR) and Multi-QMDR.

In the second experiment, a mouse dataset of a 4-way cross between inbred strain BALB/cJ, C57BL/6J, C3H/HeJ and DBA/2J is used. The mouse samples are the result of the cross of two F_1 hybrid parents: the (BALB/cJ$\times$C57BL/6J) F1 maternal parent and the (C3H/HeJ$\times$DBA/2J) F1 paternal parent. The sample size is 505. 558 loci known to be polymorphic among the four founder strains were genotyped across the genome and 17 phenotype were measured. Genotyped locations are either single nucleotide polymorphisms (SNP) or simple sequence length polymorphisms. We use five body size phenotypes: (1) femur length (right femur, proximal–distal), (2) vertebra length (eighth caudal vertebra, cranio-caudal), (3) early adult weight at 3 months, (4) late adult weight at 13 months, and (5) the slope of the best-fit linear trajectory for each animal between 3 months and 13 months, and three trabecular bone morphology and microstructure phenotypes: (6)bone volume fraction (bone volume/total volume), (7) trabecular organization (plate number/millimeter) and (8) ultimate load to failure. The data were downloaded from the QTL Archive curated by the Jackson Laboratories http://qtlarchive.org/db/q?pg=projdetails& proj=burke_2012.

Each of the above eight continuous phenotypes is transferred to an ordinary trait with three categories as in the first experiment. Our proposed MQOMDR method is used to select the best 2-way and 3-way gene-gene interactions in the above real dataset associated with eight ordinal traits transferred from eight continuous phenotypes. CVC and whether same sets of genetic markers are identified as the best models in different ordinal traits are also used to evaluate the performance of MQOMDR and compare with that of MDR, QMDR and Multi-QMDR.

5.4.2. Experimental results

For the first experiment, the correlation coefficient values across different phenotypes are shown in a matrix of gray level values in Figure 5.1:

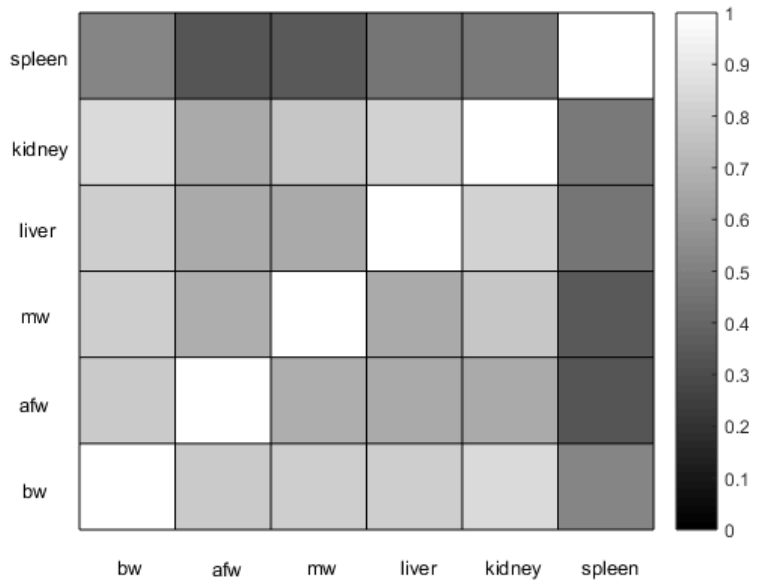

Figure 5.1. Absolute values of correlation coefficient between 6 phenotypes in experiment 1 presented as gray level values.

For the first stage of MQOMDR, all six phenotypes are used to select the best 2-way and 3-way gene-gene interactions.

As can be seen from Table 5.1, for 2-way interactions, bw, mw and spleen have the same set of SNPs (D1Mit68 and Dx9Mit192) as the best two-locus classifier, while afw, liver and kidney have another same set of SNPs (D7Mit26 and Dx9Mit192) as the best two-locus classifier. Therefore bw, mw and spleen form one group: group 1, afw, liver and kidney form another group: group 2. For each phenotype in each group, using all phenotypes in the same group only to execute MQOMDR algorithm again and calculate CVC of the same set of SNPs. CVCs are 4,3,6 for phenotypes in group 1 and 4,5,3 for phenotypes in group 2. Both average CVCs for two groups become larger. So both groups are remained for further treatment. For group 1, we remove mw which has the smallest CVC (2) from the group and use bw and spleen only to execute MQOMDR algorithm again for both bw and spleen and calculate CVCs of the same

set of SNPs. CVCs are both 6 for bw and spleen, resulting in a larger average of CVCs (6) than that in the last repetition (3.5) and also that for MDR (3.5). Since there are only two phenotypes left in group 1 now, SNPs D1Mit68 and Dx9Mit192 are considered as having strongest interaction on both bw and spleen. Similarly for group 2, liver which has the smallest CVC (2) is removed from the group firstly, however the resulting average CVC of afw and kidney is smaller than that of MDR, so afw is removed instead. The resulting CVCs are also both 6 for liver and kidney, resulting in a larger average CVC (6) than that in the last repetition (2.5) and also that for MDR (3.5). Since there are only two phenotypes left in group 2 now, SNPs D7Mit26 and Dx9Mit192 are considered as having strongest interaction on both liver and kidney.

Table 5.1 Best set of snps and corresponding cvc of the dba2×nmri8 dataset for mqomdr, mdr, qmdr and multi-qmdr

Method	Phenotype	Two-locus classifier			Three-locus classifier			
		SNPs		CVC	SNPs			CVC
MQOMDR	bw	D1Mit68	Dx9Mit192	4	D1Mit68	D7Mit21	DX9Mit95	3
	afw	D7Mit26	Dx9Mit192	3	D1Mit68	D7Mit21	DX9Mit95	3
	mw	D1Mit68	Dx9Mit192	2	D1Mit68	D7Mit21	DX9Mit95	3
	liver	D7Mit26	Dx9Mit192	2	D1Mit68	D7Mit21	DX9Mit95	6
	kidney	D7Mit26	Dx9Mit192	3	D1Mit68	D7Mit21	DX9Mit95	2
	spleen	D1Mit68	Dx9Mit192	3	D1Mit68	D7Mit21	DX9Mit95	4
MDR	bw	D14Mit87	DX9Mit95	2	D2Mit447	D13Mit130	Dx9Mit192	5
	afw	D2Mit266	Dx9Mit192	3	D1Mit49	14Mit257	Dx9Mit192	2
	mw	D14Mit87	Dx9Mit192	3	D14Mit87	D15Mit193a	Dx9Mit192	2
	liver	D2Mit447	DX9Mit95	3	D1Mit68	D7Mit21	DX9Mit95	2
	kidney	D9Mit229	Dx9Mit192	4	D9Mit229	D14Mit87	DX9Mit119	2
	spleen	D1Mit68	Dx9Mit192	5	D7Mit26	D9Mit64	DX9Mit95	3
QMDR	bw	D13Mit78	DX9Mit119	7	D7Mit246	D13Mit78	DX9Mit119	4
	afw	14Mit257	Dx9Mit192	6	D2Mit6	14Mit257	DX9Mit119	3
	mw	D10Mit16	Dx9Mit192	3	D8Mit4	14Mit257	Dx9Mit192	4
	liver	D2Mit447	DX9Mit119	5	D2Mit447	D9Mit136	DX9Mit119	5
	kidney	D13Mit78	DX9Mit119	2	D2Mit447	D9Mit229	DX9Mit119	5
	spleen	D1Mit46	DX9Mit95	3	D7Mit26	D9Mit64	DX9Mit95	3
Multi-QMDR		D13Mit78	DX9Mit119	7	D7Mit26	D9Mit229	Dx9Mit192	3

For 3-way interactions, all six phenotypes have the same set of SNPs (D1Mit68, D7Mit21 and DX9Mit95) as the best three-locus classifier and form a single group: group 3. Kidney which has the smallest CVC is removed from the group and the

remaining five phenotypes are used to execute MQOMDR algorithm again for each of these five phenotypes. Their CVCs of the set of SNPs D1Mit68, D7Mit21 and DX9Mit95 now are 4, 3, 3, 7, 3 respectively and the average is 4, greater than that of the last repetition (3.8) and for MDR (2.8). If we continue to remove any of the remaining five phenotype, the average CVC would become smaller. Therefore SNPs D1Mit68, D7Mit21 and DX9Mit95 are considered as having strongest interaction on each of bw, afw, mw, liver and spleen.

From the process of MQOMDR algorithm, the average CVCs of the sets of SNPs identified by MQOMDR for each of the final groups of phenotypes are certainly larger than or equal to that of the sets of SNPs identified by MDR for each of the same groups of phenotypes. Also the sets of SNPs identified by MDR within each group are different with each other. For QMDR, the average CVCs are also smaller than that for MQOMDR and the sets of SNPs identified by QMDR within each group are also different with each other. Multi-QMDR has similar average CVCs as MQOMDR in general, however biological explanation of the univariate score based on principle components is difficult (Figure 5.2).

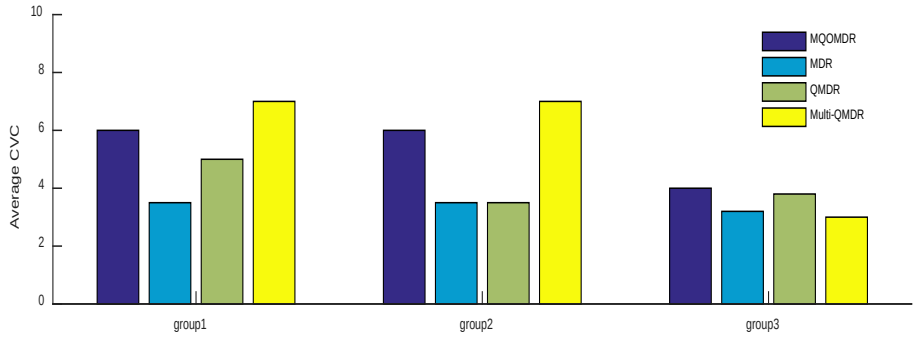

Figure 5.2 Comparison of average CVCs for three groups of phenotypes among MQOMDR, MDR, QMDR and Multi-QMDR for the DBA2╳NMRI8 dataset. Group 1 represents phenotypes: bw, mw and spleen for 2-way interactions. Group 2 represents phenotypes: afw, liver and kidney for 2-way interactions. Group 3 represents phenotypes: bw, afw, mw, liver and sleen for 3-way interactions.

For the second experiment, the correlation coefficient values across different phenotypes are shown in a matrix of gray level values in Figure 5.3.

For the first stage of MQOMDR, all eight phenotypes are used to select the best 2-way and 3-way gene-gene interactions.

As can be seen from Table 5.2, for 2-way interactions, femur length and vertebra length have the same set of SNPs (D1Mit105 and D2Mit58) as the best two-locus classifier, while plate number and 3 month weight have another same set of SNPs

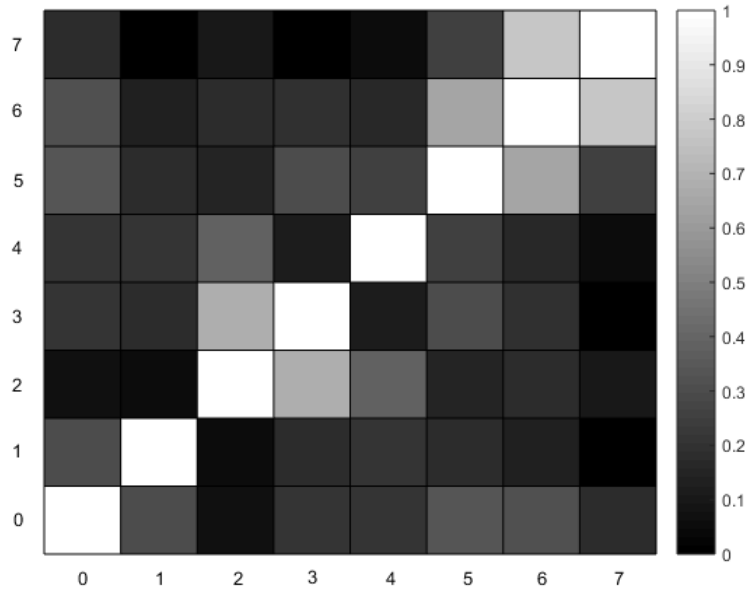

Figure 5.3. Absolute values correlation coefficient between phenotypes in experiment 2 presented as gray level values. (0 through 7 represent phenotype femur length, vertebra length, bone volume fraction, plate number, ultimate load to failure, 3 month weight, 13 month weight, weight slope)

(D14Mit170 and D15Mit100) as the best two-locus classifier. Therefore femur length and vertebra length form one group: group 1, plate number and 3 month weight form another group: group 2. For each phenotype in each group, using all phenotypes in the same group only to execute MQOMDR algorithm again and calculate CVC of the same set of SNPs. CVCs are 6,10 for phenotypes in group1 and 9,10 for phenotypes in group 2. Their average CVCs are larger or equal than that in the previous stage and much larger than that of MDR. So these two groups are both remained. Since there are now only two phenotypes in both groups, D1Mit105 and D2Mit58 are considered as having strongest interaction on both femur length and vertebra, while D14Mit170 and D15Mit100 are considered as having strongest interaction on both plate number and 3 month weight.

For 3-way interactions, no phenotypes have the same set of SNPs as the best three-locus classifier. Therefore there are no 3-way common gene-gene interactions

for these eight phenotypes.

As can be seen from Figure 5.4, MQOMDR has significantly larger average CVCs for each of the final groups of phenotypes than that of MDR, QMDR and Multi-QMDR.

Table 5.2 Best set of snps and corresponding cvc of the dba2╳du6i dataset for mqomdr, mdr, qmdr and multi-qmdr

Method	Phenotype	Two-locus classifier		
		SNPs		CVC
MQOMDR	Femur length	D1Mit105	D2Mit58	5
	Vertebra length	D1Mit105	D2Mit58	6
	Bone volume fraction	D5Mit251	D14Mit263	3
	Plate number	D14Mit170	D15Mit100	10
	Ultimate load to failure	D1Mit105	D15Mit100	5
	3 month weight	D14Mit170	D15Mit100	9
	13 month weight	D2Mit285	D15Mit100	7
	Weight slope	D3Mit127	D7Mit91	4
MDR	Femur length	D2Mit58	D5Mit95	6
	Vertebra length	D1Mit105	D2Mit58	4
	Bone volume fraction	D5Mit251	D14Mit263	3
	Plate number	D14Mit170	D15Mit63	6
	Ultimate load to failure	D9Mit12	D15Mit63	7
	3 month weight	D14Mit170	D15Mit100	7
	13 month weight	D13Mit26	D19Mit88	8
	Weight slope	D7Mit91	D13Mit57	3
QMDR	Femur length	rs4223558	rs3091203	4
	Vertebra length	D1Mit67	D1Mit105	2
	Bone volume fraction	rs13478469	D14Mit170	5
	Plate number	rs13480005	D14Mit263	4
	Ultimate load to failure	D3Mit64	D17Mit46	3
	3 month weight	D14Mit170	D15Mit100	6
	13 month weight	D14Mit170	D15Mit100	5
	Weight slope	D2Mit58	D13Mit64	2
Multi-QMDR		D13Mit64	D17Mit46	3

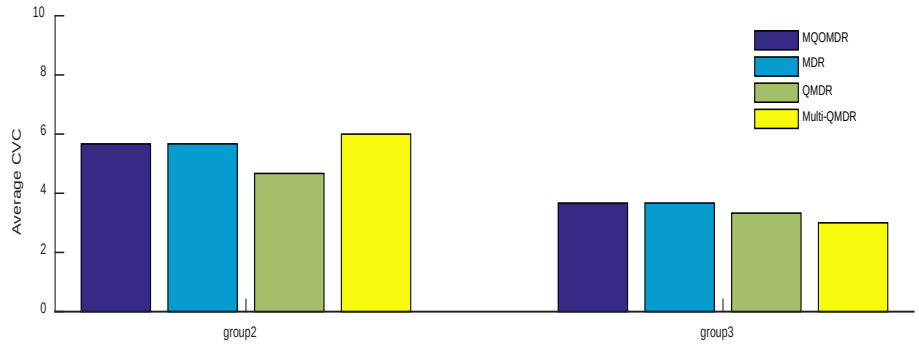

Figure 5.4 Comparison of average CVCs for three groups of phenotypes among MQOMDR, MDR, QMDR and Multi-QMDR for the DBA2╳DU6i dataset. Group 2 represents phenotypes: afw, mw and kidney for 2-way interactions. Group 3 represents phenotypes: afw, mw and kidney for 3-way interactions. Group 1 is abandoned.

5.5 Conclusion

In this chapter, a new approach to detect genetic factors associated with multiple correlated phenotypes is proposed. The best classifier is selected according to both the training accuracy of the phenotype under consideration and other phenotypes with weights determined mainly by their pair correlation with the phenotype under consideration. To select more reliable classifiers, a repeated selection process is adopted. All correlated phenotypes are used to identify interactive genetic loci at the beginning, then unreliable ones are gradually filtered out. Experimental results on two real datasets show better performance of our proposed algorithm than MDR, QMDR and Multi-QMDR.

For estimation of correlation coefficient ρ, we have the following statistics:

$$R=\frac{S_{12}}{S_1 \times S_2},$$

where $S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2$, $S_2^2 = \frac{1}{n}\sum_{i=1}^{n}(\eta_i - \bar{\eta})^2$, $S_{12}= \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi})(\eta_i - \bar{\eta})$, ξ and η are two phenotype random variables, n is the sample size.

The probability density function of R is complex. However when ρ=0, it can be reduced and

$$T=\sqrt{n-2}\,\frac{R}{\sqrt{1-R^2}}$$

follows a *t*-distribution with *n*-2 degrees of freedom. When significance level α=0.05, *n*=100, p value is 0.1946. In our experiments, *n*=275 and 505 respectively, so estimation bias for ρ=0 can be effectively controlled. From this, we can approximately conclude that other ρ values can also be effectively controlled.

In our proposed algorithm, only one best classifier is selected for each phenotype at first. Actually multiple best classifiers can be selected. This will not only increase the number of best classifiers, but also provide more opportunities to find out more reliable classifiers, since more groups which may overlap can be formed as long as the phenotypes in the group have one of multiple best classifiers in common. This will be our future work.

The MQOMDR extends MDR to multiple phenotypes. An alternative is to extend QMDR to multiple phenotypes in the same manner. Our experiments show that the performances of these two methods are similar. In addition, QMDR and its extension can only divide a quantitative trait into two categories, while our method can divided it into any number of categories as needed.

Chapter 6 Conclusions and Suggestions for Future Research

In this thesis, we have studied the problem of identifying gene-gene interactions associated with complex diseases and complex traits.

Gene-gene interaction is an important factor which needs to be considered when searching for genetic factors that influence complex diseases and complex traits. If a genetic factor influences a complex disease or complex trait primarily through interaction with other genetic factors or environmental factors, the effect might be missed if the gene is examined individually.

Although there are a variety of definitions of gene-gene interaction, they are potentially conflicting and not satisfying. Therefore we start our study by trying to provide a more reasonable definition of gene-gene interaction. We first derive an inequality describing the relationship between two genotype variables that represent the genotypes of two different genes, and a disease-status variable that represents the presence or absence of a complex disease and generalize it to n genotype variables. Based on this inequality, we provide a conditional independence and redundancy based definition of gene-gene interaction and the definition of an interaction group. We also derive a kai square statistic to measure gene-gene interaction.

Since with the increase of the number of interaction genes, the number of possible combinations of interaction genes increases exponentially and the number of sparse cells also increases, a new algorithm is proposed to efficiently detect high order gene-gene interactions after some properties and a theorem relating to these new definitions are given and proved which reveal the relation between high order interaction and low order interaction.

Experimental results on simulated and real datasets show the effectiveness of the new definition and measure of gene-gene interaction and the effectiveness and efficiency of the new algorithm to detect many high order gene-gene interactions.

Complex traits are very common in human bodies which exist in the majority of

human innate and acquired body and behavioral characteristics, many physiological characteristics and also are closely related to most diseases. Unlike complex diseases which have only two states, complex traits have continuous outcomes can provide more accurate analysis.

To better use the information contained in complex traits, we employ fuzzy logic in the selection of classifier. We propose extended member function in fuzzy classification which extend the range of traditional member function of fuzzy set from [0,1] to [-1,1] to better reflect the difference of different classifiers. The EFQMDR algorithm we proposed first transforms a quantitative trait into an ordinal trait by dividing it into several ordinal levels, then employs a new ordinal association measure, balanced accuracy based on extended member function to select multiple best sets of SNPs as having strongest association with the trait. Experimental results on simulated datasets and real datasets show that our algorithm has better performance in identifying gene-gene interactions associated with a complex quantitative trait.

Complex traits usually have several correlated phenotypes. These correlated phenotypes are useful to detect additional genetic variants with small effects across multiple phenotypes or pleiotropy effects.

To effectively identify gene-gene interactions associated with multiple correlated phenotypes, we propose MQOMDR algorithm which selects the best classifier according to not only the training accuracy of the phenotype under consideration but also other phenotypes with weights determined mainly by their pair correlation with the phenotype under consideration. Current methods use all correlations among phenotypes to select common sets of SNPs having interaction on multiple phenotypes. However these correlations may be caused by other factors. To make use of truly useful correlated phenotypes, we also employ a repeated selection process to filter out those phenotypes whose correlation with the phenotype under consideration is caused by other factors. Experimental results show that our algorithm has better performance in identifying gene-gene interactions associated with multiple correlated phenotypes.

In the future, in order to increase the hit ratio of CIR algorithm, more combinations of genes can be selected at the first step and true interaction genes can be identified by

executing permutation test at the second step. For EFQMDR algorithm, we will try to conduct mathematical analysis to explain the better performance of the extended fuzzy classification based on extended member functions and do more experiments to check whether our algorithm is still better when QTs are divided in larger number of categories. In step 3 and step 4 of MQOMDR algorithm, fuzzy logic can be adopted to better use the information contained in a quantitative trait.

References

Agresti, A., & Kateri, M. (2011). Categorical Data Analysis. Springer Berlin Heidelberg.

Amberger, J. *et al.* (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–7D96.

Angela, T., & Nieto, J. J., (2006). Fuzzy Logic in Medicine and Bioinformatics. *Journal of Biomedicine & Biotechnology*. 2, 91908.

Armitage, P., Berry, G, & Matthews, J. N. S. (2002). Statistical methods in medical research, 4th edn. Blackwell Science, Chichester.

Assareh, A. *et al*. (2012). Feature selections using AdaBoost: Application in gene-gene interaction detection. In: *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on*. 831-837.

Auffray, C. *et al.* (2009). Systems medicine: the future of medical genomics and healthcare. *Genome*, **1**, 2.

Aylor, D. L., & Zeng Z. B. (2008). From classical genetics to quantitative genetics to systems biology: modeling epistasis. *PLoS Genet.* **4**, e1000029.

Barro, S., & Marín, R. (2002). Fuzzy Logic in Medicine. Physica-Verlag HD.

Bateson, W. (1909). Mendel's principles of heredity. United Kingdom:Cambridge.

Breiman, L. (2001). Random forests. *Mach. Learn*. **45**, 5–32.

Brockmann, G. A., Tsaih, S. W., Neuschl, C., Churchill, G. A., & Li, R. (2009). Genetic factors contributing to obesity and body weight can act through mechanisms affecting muscle weight, fat weight, or both. *Physiological Genomics*, 36(2), 114-126.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P. *et al*. (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, 28(2), 171-182.

Busis, N. A. (2010). How can I choose the best electronic health record system for my practice? *Neurology*, **75**, S60–S64.

Calle, M. L., Urrea, V., Malats, N., & Van Steen, K. (2008). MB-MDR: Model based multifactor dimensionality reduction for detecting interactions in highdimensional genomic data. Annals of Human Genetics . 75.

Carlson, C. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics*.

Chanda, P. *et al*. (2009). Information-theoretic gene-gene and gene-environment interaction analysis of quantitative traits. *BMC Genomics*, 10(1).

Chanda, P. *et al*. (2008). AMBIENCE: a novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. *Genetics*, 180, 1191-1210.

Chanda, P. *et al*. (2007). Information-theoretic metrics for visualizing gene-environment interactions. *The American Journal of Human Genetics*, 81, 939-963.

Chen, L. *et al*. (2009). Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways. *Bioinformatics, 25*(2), 237-242

Chen, S., Sun, J., Dimitrov, L., Turner, A. R., Adams, T. S, Meyers, D. A. *et al*. (2008). A support vector machine approach for detecting genegene interaction. Genetic Epidemiology. 32(2):152-167.

Chen, Y., & Li, J. (2012). Generation of synthetic data and experimental designs in evaluating interactions for association studies. *Journal of Bioinformatics and Computational Biology*, 10, 1240005.

Chung, Y., Lee, S.Y., Elston, R., C., & Park, T. (2007). Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions. *Bioinformatics* **23**, 71–76.

Clemens, K. E., Churchill, G., Bhatt, N. *et al*. (2000). Genetic control of susceptibility to UV-induced immunosuppression by interacting quantitative trait loci. Genes & Immunity. 1(4):251-259.

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

Cordell, H. J. (2009). Detecting gene|[ndash]|gene interactions that underlie human diseases. *Nature Reviews Genetics.* 10(6), 392.

Cramér, H. (1945). *Mathematical methods of statistics*. Princeton university press.

Culverhouse, R. *et al.* (2004). Detecting epistatic interactions contributing to quantitative traits. *Genetic epidemiology*, 27, 141-152.

de Oliveira, J. V., & Pedrycz, W. (2007). Advances in Fuzzy Clustering and Its Applications. John Wiley & Sons.

Dembélé, D., & Kastner, P. (2003). Fuzzy C-means method for clustering microarray data. *Bioinformatics*. 19(8),973-980.

Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3,185-205.

Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L., Shi, T. et al. (2008). Exploration of gene–gene interaction effects using entropy-based methods. *European Journal of Human Genetics*. 16(2),229-35.

Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K.J., & Altman, R. B. (2011). Bioinformatics challenges for personalized medicine. *Bioinformatics*, *27*(13), 1741-1748.

Fitzmaurice, G. M., & Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, 80(1), 141–151,.

Franke, B. *et al*. (2009). Genome-wide association studies in ADHD. *Human genetics*, 126, 13-50.

Giacomini, K. M. *et al.* (2007). The pharmacogenetics research network: from SNP discovery to clinical drug response. *Clin. Pharmacol. Ther.*, 81, 328–345.

Greenland, S. (2009). Interactions in epidemiology: relevance, identification, and estimation. *Epidemiology* 20, 14–17.

Gui, J., Moore, J. H., Williams, S. M., Andrews, P., Hillege, H. L., van der Harst, P. *et al*. (2013). A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits. *PLoS One*. 8(6),e66545.

Gustafson, D., Kessel, W. (1978). Fuzzy clustering with a fuzzy covariance matrix. In: *Proc. of the IEEE Conference on Decision and Control Including the 17th Symposium on Adaptive Processes,* 761–766.

Hahn, L. W. *et al*. (2003). Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics*, 19, 376-382.

Hansen, N. T. *et al.* (2009). Generating genome-scale candidate gene lists for pharmacogenomics. *Clin. Pharmacol. Ther.*, 86, 183–189.

Jakulin, A., & Bratko, I. (2003). Analyzing attribute dependencies. Springer Berlin Heidelberg.

Jakulin, A. *et al*. (2003). Attribute interactions in medical data analysis. Springer Berlin Heidelberg.

Kang, G. *et al.* (2008). An entropy-based approach for testing genetic epistasis underlying complex diseases. *J. Theor. Biol*. 250, 362–374.

Keiser, M. *et al.* (2007). Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, 25, 197–206.

Kim, K., Kwon, M. S., Oh, S., & Park, T. (2013). Identification of multiple gene-gene interactions for ordinal phenotypes. *BMC medical genomics*. 6(Suppl 2), S9.

Kim, S. *et al.* (2009). A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics,* 25, i204-12

Klei, L., Luca, D., Devlin, B., & Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology*, 32(1), 9–19,.

Kohl, P. *et al.* (2010). Systems biology: an approach. *Clin. Pharmacol. Therap.*, 88, 25–33.

Kooperberg, C. & Ruczinski I. (2005). Identifying interacting SNPs using Monte Carlo logic regression. *Genetic epidemiology,* 28, 157-170.

Kooperberg, C. *et al*. (2001). Sequence analysis using logic regression. *Genetic epidemiology,* 21, S626-631.

Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4): 963–974.

Lange, C., van Steen, K., Andrew, T. *et al*. (2004). A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects," *Statistical Applications in Genetics and Molecular Biology*, 3(1), 1544–6115,.

Lange, C., Silverman, E. K., Xu, X, Weiss, S. T., & Laird, N. M. (2003). A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics*, 4(2), 195–206.

Li, J., & Chen Y. (2008). Generating samples for association studies based on HapMap data. *BMC bioinformatics*, 9, 44.

Lin, D., Li, J., Calhoun, V. D. *et al*. (2015). Detection of genetic factors associated with multiple correlated imaging phenotypes by a sparse regression model. *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on.*

Liu, J. *et al.* (2009). Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Human brain mapping,* 30, 241-255.

Li, Z. *et al*. (2007). Pattern-based mining strategy to detect multi-locus association and gene× environment interaction. *BMC proceedings*, *BioMed Central*, 1, S16.

Long, Q. *et al*. (2009). Detecting disease-associated genotype patterns. *BMC bioinformatics*, 10, S75.

Lou, X. Y. (2007). Chen GB, Yan L, Ma JZ, Zhu J, Elston RC et al. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *American Journal of Human Genetics*, 80(6), 1125-1137.

Lunetta, K. L., Hayward, L. B., Segal, J., & Van Eerdewegh, P. (2004). Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet.* 5, 32.

Malmberg, R. L. *et al*. (2005). Epistasis for fitness-related quantitative traits in Arabidopsis thaliana grown in the field and in the greenhouse. *Genetics*, 171, 2013-2027.

Manolio, T. A. *et al*. (2009). Finding the missing heritability of complex diseases. *Nature*, 461, 747-753.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London:Chapman & Hall/CRC.

Mceliece, R. J. (2002). *The theory of information and coding*. Cambridge University Press.

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

Moore, J. H. (2005). A global view of epistasis. *Nature genetics*, 37, 13-14.

Moore, J. H. (2004). Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert review of molecular diagnostics*, 4, 795-803

Moore, J. H. *et al*. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of theoretical biology,* 241, 252-261.

Motsinger, A. A. *et al*. (2006). GPNN: Power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. *BMC bioinformatics,* 7, 1.

Motsinger, A. A. *et al*. (2008). Comparison of approaches for machine‐learning optimization of neural networks for detecting gene‐gene interactions in genetic epidemiology. *Genetic epidemiology*, 32, 325-340.

Muller, K. E., & Peterson, B. L. (1984). Practical methods for computing power in testing the multivariate general linear hypothesis. *Computational Statistics and Data Analysis*, 2(2), 143–158.

Nelson, M. R. *et al*. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome research*, 11, 458-470.

Nunkesser, R. *et al*. (2007). Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics,* 23, 3280-3288.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40(4), 1079–1087.

Ogata, H. *et al*. (2000). A heuristic graph comparison algorithm and its application to

detect functionally related enzyme clusters. *Nucleic Acids Res.*, 28, 4021–4028.

Ohashi, W., & Tanaka, H. (2010). Benefits of pharmacogenomics in drug developmentearlier launch of drugs and less adverse events. *J. Med. Syst.*, 34, 701–707.

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

Phillips, P. C. (2008). Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Rev. Genet.*, **9**, 855–867.

Phuong, N. H., & Kreinovich, V. (2001). Fuzzy logic and its applications in medicine. *International Journal of Medical Informatics*, 62(2-3),165.

Procyk, T. J., & Mamdani, E. H. (1979). A linguistic self-organizing process controller. *Automatica* 15(1), 15–30.

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

Satagopan, J. M., & Elston, R. C. (2013). Evaluation of removable statistical interaction for binary traits. *Statistics in Medicine, 32*(7), 1164. Segre, D. *et al*. (2005). Modular epistasis in yeast metabolism. *Nature genetics*, 37, 77-83.

Sepulveda, N., Paulino, C. D., Carneiro, J., & Penha-Goncalves, C. (2007). Allelic penetrance approach as a tool to model two-locus interaction in complex binary traits. *Heredity* 99, 173–184.

Sepúlveda, N., Paulino, C. D., & Penha-Gonçalves, C. (2009). Bayesian analysis of allelic penetrance models for complex binary traits. *Computational Statistics & Data Analysis, 53*(4), 1271-1283.

Shang, J. *et al*. (2016). CINOEDV: a co-information based method for detecting and visualizing n-order epistatic interactions. *BMC bioinformatics*, 17, 1.

Shen, L., *et al.* (2014) Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. *Brain Imaging and Behavior*, 8, 183-207.

Sherry,S. T. *et al.* (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29, 308–311.

Shuldiner, A. R. *et al.* (2009). Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. *JAMA*, 302, 849–857.

Smith, H. (1962). Multivariate Analysis of Variance (MANOVA). *Biometrics* 18.1, 22-41

Tanaka, K., & Sugeno, M. (1992). Stability analysis and design of fuzzy control systems. *Fuzzy Sets System*. 45, 135–156.

Tanaka, K., Wang, H. O. (2004). Fuzzy Control Systems Design and Analysis: A Linear Matrix Inequality Approach. John Wiley & Sons.

The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature*, 447, 661–678.

Timm, H., Borgelt, C., Döring, C., & Kruse, R. (2004). An extension to possibilistic fuzzy cluster analysis. *Fuzzy Sets System*, 147(1):3–16.

van der Sluis, S. *et al.* (2013) TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet*, 9, e1003235.

Bateson, W. (1909). Mendel's principles of heredity. United Kingdom:Cambridge

Wang, H. *et al.* (2012). From phenotype to genotype: an association study of longitudinal phenotypic markers to Alzheimer's disease relevant SNPs. *Bioinformatics*, 28, i619-i625.

Weinshilboum, R. (2001). Thiopurine pharmacogenetics: clinical and molecular studies of thiopurine methyltransferase. *Drug Metab. Dispos.*, 29, 601–605.

Williams, T. N. *et al*. (2005). Negative epistasis between the malaria-protective effects of α+-thalassemia and the sickle cell trait. *Nature genetics*, 37, 1253-1257.

Yang, Q., Wu, H., Guo, H. C. Y., & Fox, C. S. (2010) Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic Epidemiology*, 34(5), 444–454.

Yang, Q., (1999) Khoury, M. J., Sun, F. & Flanders, W. D. Case-only design to measure gene–gene interaction. *Epidemiology* 10, 167–170.

Yang, Q., & Wang, Y. (2012). Methods for analyzing multivariate phenotypes in genetic association studies. *Journal of Probability & Statistics, 2012*(358), 652569.

Yee, J. *et al*. (2013). A modified entropy-based approach for identifying gene-gene interactions in case-control study. *PloS one*, 8, e69321.

Yee, J., Kwon, M. S., Jin, S., Park, T., & Park, M. (2015) Detecting Genetic Interactions for Quantitative Traits Using m-Spacing Entropy Measure. B*iomed Research International*. 2015(2), 523641.

Yu, W., Kwon, M. S., & Park, T. (2015). Multivariate quantitative multifactor dimensionality reduction for detecting gene-gene interactions.*Human Heredity*, *79*(3-4), 168-81.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8 (3),338–353.

Zhang,H. & Bonney,G. (2000). Use of classification trees for association studies. *Genetic epidemiology*, 19, 323-332.

Zhang, Z., Zhang, S., Wong, M., Wareham, N. J., & Sha, Q. (2008). An ensemble learning approach jointly modelling main and interaction effects in genetic association studies. *Genetic Epidemiology*, 32(4), 285-300.