이학박사학위논문

Gene-Gene Interaction Analysis of High-dimensional Genomic Data

고차원 유전체 자료에서의 유전자–유전자 상호작용 분석

2015년 2월

서울대학교 대학원

협동과정 생물정보학

권 민 석

Gene-Gene Interaction Analysis of High-dimensional Genomic Data

by

Min-Seok Kwon

**A thesis
submitted in fulfillment of the requirement
for the degree of Doctor of Philosophy
in
Bioinformatics**

Interdisciplinary Program in Bioinformatics

College of Natural Sciences

Seoul National University

Feb, 2015

Gene-Gene Interaction Analysis of High-dimensional Genomic Data

지도교수 박태성

이 논문을 이학박사 학위논문으로 제출함

2014 년 10 월

서울대학교 대학원

협동과정 생물정보학

권민석

권민석의 이학박사 학위논문을 인준함

2014 년 12 월

위원장 천종식 **(인)**

부위원장 박태성 **(인)**

위원 김선 **(인)**

위원 이승연 **(인)**

위원 원성호 **(인)**

Abstract

Gene-Gene Interaction Analysis of High-dimensional Genomic Data

Min-Seok Kwon
Interdisciplinary Program in Bioinformatics
The Graduate School
Seoul National University

With the development of high-throughput genotyping and sequencing technology, there are growing evidences of association with genetic variants and common complex traits. In spite of thousands of genetic variants discovered, such genetic markers have been shown to explain only a very small proportion of the underlying genetic variance of complex traits. Gene-gene interaction (GGI) analysis and rare variant analysis is expected to unveil a large portion of unexplained heritability of complex traits.

In GGI, there are several practical issues. First, in order to conduct GGI analysis with high-dimensional genomic data, GGI methods requires the efficient computation and high accuracy. Second, it is hard to detect GGI for rare variants due to its sparsity. Third, analysing GGI using genome-wide scale suffers from a computational burden as exploring a huge search space. It requires much greater number of tests to find optimal GGI. For *k* variants, we have *k*(*k*-1)/2 combinations for two-order interactions, and ${}_{n}C_{k}$ combinations for n-order interactions. The number of possible interaction models increase exponentially as the interaction order increases or the number of variant increases. Forth, though the biological interpretation of GGI is important, it is hard to interpret GGI due to its complex manner.

In order to overcome these four main issues in GGI analysis with high-dimensional genomic data, the four novel methods are proposed.

First, to provide efficient GGI method, we propose IGENT, Information theory-based GEnome-wide gene-gene iNTeraction method. IGENT is an efficient algorithm for identifying genome-wide GGI and gene-environment interaction (GEI). For detecting significant GGIs in genome-wide scale, it is important to reduce computational burden significantly. IGENT uses information gain (IG) and evaluates its significance without resampling. Through our simulation studies, the power of the IGENT is shown to be better than or equivalent to that of that of BOOST. The proposed method successfully detected GGI for bipolar disorder in the Wellcome Trust Case Control Consortium (WTCCC) and age-related macular degeneration (AMD).

Second, for GGI analysis of rare variants, we propose a new gene-gene interaction method in the framework of the multifactor dimensionality reduction (MDR) analysis. The proposed method consists of two steps. The first step is to collapse the rare variants in a specific region such as gene. The second step is to perform MDR analysis for the collapsed rare variants. The proposed method is applied in whole exome sequencing data of Korean population to identify causal gene-gene interaction for rare variants for type 2 diabetes (T2D).

Third, to increase computational performance for GGI in genome-wide scale, we developed CUDA (Compute Unified Device Architecture) based genome-wide association MDR (cuGWAM) software using efficient hardware accelerators. cuGWAM has better performance than CPU-based MDR methods and other GPU-based methods through our simulation studies.

Fourth, to efficiently provide the statistical interpretation and biological evidences of gene-gene interactions, we developed the VizEpis, a tool for visualizing of gene-gene interactions in genetic association analysis and mapping of epistatic interaction to the biological evidence from public interaction databases. Using interaction network and circular plot, the VizEpis provides to explore the interaction network integrated with biological evidences in epigenetic regulation, splicing, transcription, translation and post-translation level. To aid statistical interaction in genotype level, the VizEpis provides checkerboard, pairwise checkerboard, forest, funnel and ring chart.

Keywords: Gene-gene interaction (GGI), Genome-wide association study (GWAS), Massively parallel sequencing (MPS), rare variant, Graphic processing unit (GPU), Visualization

Student number: 2008-30830

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background of high-dimensional genomic data

1.1.1 History of genome-wide association studies (GWAS)

In genetic epidemiology, a genome-wise association study (GWAS) is an approach that detects causal variants rapidly examining genetic variants across genome of population [Visscher, et al. 2012]. The causal variant is sing-nucleotide polymorphism (SNP) that contributes to an increase or decrease in risk to disease arise in populations. A SNP is defined as a single nucleotide (A, T, C, or G) variation of DNA sequence occurring within more than 1% individuals in a population. For example, two sequenced DNA fragments from different individuals, GTCACGCTA to GTCATGCTA, contain a difference in a single nucleotide. In this case, the variation is called bi-allelic SNP which has two alleles (C and T) and three genotypes (CC, CT, and TT) (http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism).

Risch and Merikangas proposed that modest effective variant could be identified with greater power by association analyses if the modest effective variant and causal variant were in strong linkage disequilibrium (LD). Because variants in strong LD are likely to be inherited together, one can use a subset of ′tagging′ variants as proxies for the entire set. Using the advantage of LD between variants, ′indirect′ or ′map-based′ genome-wide association approach has the potential to identify real causal variant by investigating just limited number of tagging SNPs. [Risch and Merikangas 1996].

Lander proposed the common-disease common-variant hypothesis (CDCV) which predicted that common causal variants could be found in all populations which manifest a given disease [Lander 1996]. If SNPs are neutral or favorable with respect to survival, they become common over many generations. However, some common SNPs have a small additive or multiplicative effect on complex diseases. The CDCV GWAS strategy assumed that many different common SNPs have small effects on each disease and that some could be found by testing enough SNPs in large population [PGCC, et al. 2009].

For the indirect genome-wide association approach based on CDCV hypothesis, the Affymetrix and Illumina companies have competitively developed genotyping chips that assay large SNPs with high accuracy, low cost and rapid speed. Using high-throughput genotyping chips, International HapMap project has been conducted and validated more than 3.1 million variants in major 11 global ancestry groups, to develop a haplotype map of the human genome and discover the common patterns of human genetic variation [International HapMap, et al. 2007]. These efforts make it possible to capture the common genetic variations across the genome using a representative tagging SNPs.

The genome-wide association study (GWAS) has been successful in identifiying genetic variants associated with some targeted traits and complex diseases such as cardiovascular diseases, metabolic diseases, responses to drug and cancers. Since the first GWAS was reported in 2005 [Klein, et al. 2005], GWAS has rapidly grown in scale and complexity, and 2,051 studies and 14,836 causal variants (p-value $\leq 5.0 \times 10^{-8}$) have been added to the catalog of published Genome-Wide Association Studies (See Figure 1.1) (http://www.genome.gov/gwastudies/).

1.1.2 Missing heritability and proposed alternative methods

Since first successful approach of GWAS was published in 2005 investigating patients with age-related macular degenerations [Klein, et al. 2005], GWAS have been facilitated with the development of SNP arrays. Typically, 300,000 to 1 million common variants are captured in commercial SNP arrays. However it was shown that genetic etiology of complex diseases could be rarely explained by the genetic factors identified from GWAS. That is, the common causal variants can explain only a small fraction of heritability. For example, although about 80% of the variation in height among individuals is known to be due to genetic factor, associated causal 40 variants can explain only about 5 % of height variance [Visscher 2008]. Also, in the case of type 2 diabetes, identified 18 variants can explain about 15% of sibling recurrence risk [Manolio, et al. 2009].

To unveil the ′missing heritability′, Manolio et al. suggested that this missing heritability could be partly due to gene-gene interaction, rare variants and structural variants [Manolio, et al. 2009]. Specially, Zuk et al. showed that full heritability could not be explained without the gene-gene interaction effects based on simulation study [Zuk, et al. 2012].

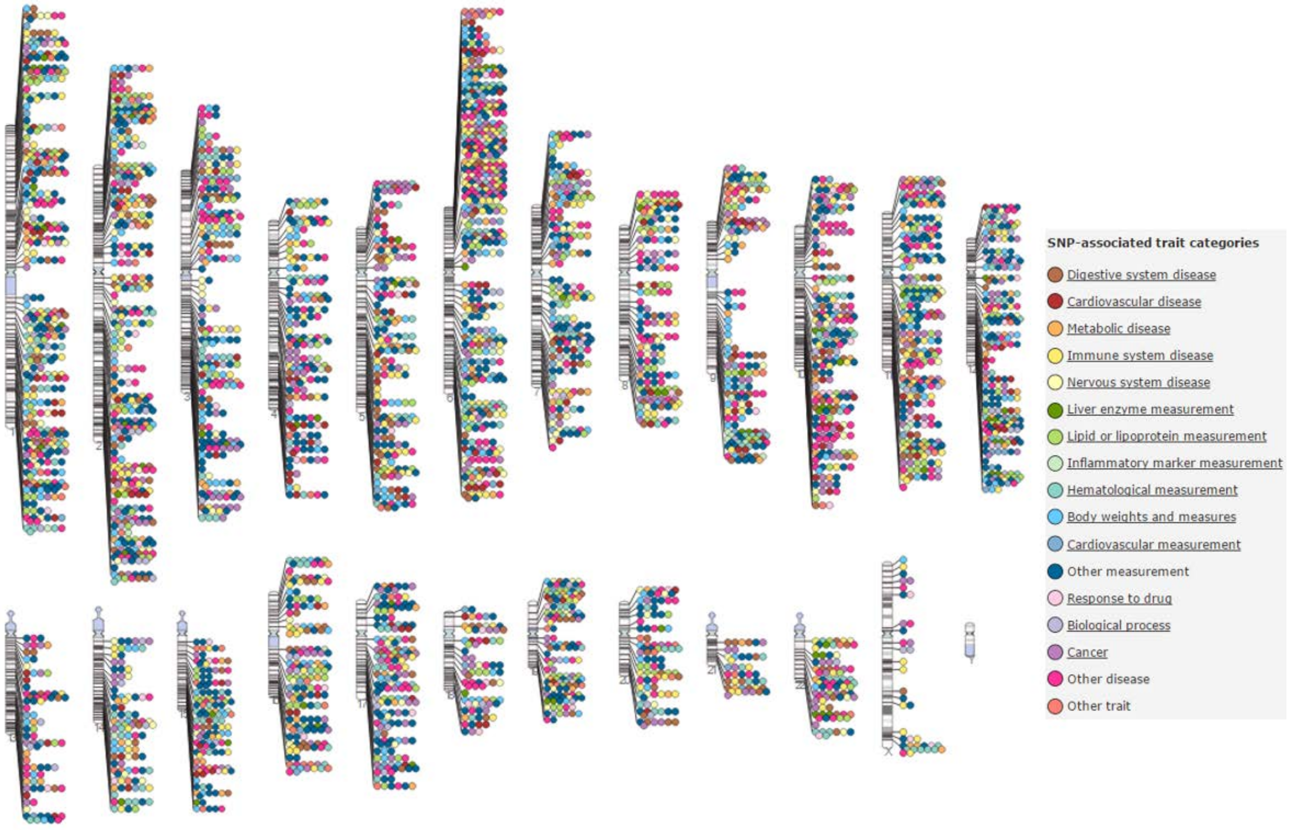


Figure 1.1 GWAS catalog. As of Nov/15/14, the catalog includes 2051 publications and 14836 SNPs. This diagram shows all SNP-trait associations with p-value ≤ 5.0×10^{-8}. (from http://www.genome.gov/gwastudies/)

1.1.3 Association studies of massively parallel sequencing (MPS)

Since massively parallel sequencing (MPS) technology was invented in 1996 [Ronaghi, et al. 1996], the MPS has been spotlighted as called ′next-generation sequencing (NGS)′. While SNP array can detect the genotype of common variants, MPS can genotype for both common and rare variants in entire genome.

Recent of the advances of MPS has facilitated the association study for rare causal variants in common complex diseases underlying the hypothesis of rare variant common disease (RVCD). To detect causal rare variants, several methods are proposed in which a collection of rare variants might show an association with a trait. There are two types of methods; one is burden test method and the other is non-burden test method. The burden test uses collapsed genotype of multiple rare variants for association test. These methods show good performance for many moderate signals with the same direction of effect. Several burden test-based methods such as CAST (Morgenthaler & Thilly 2007), CMC (Li & Leal 2008), WSS (Madsen & Browning 2009), KBAC (Liu & Leal 2010) and VT (Price et al. 2010) have been proposed. Non-burden test aggregates individual variant test statistics with weight when SNP effects are modeled linearly. This non-burden test methods such as C-alpha (Neale et al. 2011) , SKAT (Wu et al. 2011) and SKAT-O (Lee et al. 2012) show powerful performance when a genetic region has both protective and deleterious variants.

1.2 Purpose and novelty of this study

The main purpose of this thesis is to develop the methods to analyze gene-gene interaction of genetic variants. To overcome the limitations of traditional GWAS, four kinds of studies are designed.

In the first study, we proposed an entropy-based gene-gene interaction analysis method in genetic variant data. This method is referred as IGENT, an Interaction analysis method of Genetic variants using ENTropy. IGENT is an efficient algorithm for identifying genome-wide gene-gene interactions and gene-environment interaction. For detecting significant gene-gene interactions in genome-wide scale, it is important to reduce computational burden significantly. IGENT uses information gain and evaluates its significance without resampling. We studied the performance of our method through simulation study and apply to real genetic data.

In the second study, we propose a new gene-gene interaction method for the rare variants in the framework of the multifactor dimensionality reduction (MDR) analysis. The proposed method consists of two steps. The first step is to collapse the rare variants in a specific region such as gene. The second step is to perform MDR analysis for the collapsed rare variants. The proposed method is illustrated with 1080 whole exome sequencing data of Korean population to identify causal gene-gene interaction for rare variants for type 2 diabetes.

In the third study, for overcoming the computational problem in gene-gene interaction, we developed cuGWAM, CUDA (Compute Unified Device Architecture) based genome-wide association Multifactor dimensionality reduction (MDR) software using efficient hardware accelerators. cuGWAM has better performance than CPU-based MDR methods and other GPU-based methods though our simulation study.

In the fourth study, for the statistical interpretation and biological evidences of gene-gene interactions, we developed the VizEpis, a tool for visualizing of gene-gene interactions in genetic association analysis and mapping of epistatic interaction to the biological evidence from public interaction databases. Using interaction network and circular plot, the VizEpis explores the interaction network integrated with biological evidences in epigenetic regulation, splicing, transcription, translation and post-translation level. To aid statistical interaction in genotype level, the VizEpis provides checkerboard, pairwise checkerboard, forest, funnel and ring chart.

1.3 Outline of the thesis

This thesis is organized as follows. Chapter 1 is an introduction of this study with review of GWAS and MPS. Chapter 2 presents the definition of gene-gene interaction and the overview of gene-gene interaction (GGI) analysis. Chapter 3 is the study of entropy-based GGI. Chapter 4 is the study of GGI for rare variants. Chapter 5 is the study of CUDA-based computational enhancement for GGI using graphic processing unit (GPU). Chapter 6 is the study of visualization for GGI interpretation. Finally the summary and conclusion are presented in Chapter 7.

Chapter 2

Overview of gene-gene interaction

2.1 Definition of gene-gene interaction

The term ′gene-gene interaction′ has multiple meaning. Generally, the meaning of ′gene-gene interaction′ by biologists different is different from the one by statisticians. Biologists use the gene-gene interaction to refer to the deviation from the effect pattern expected by Mendelian inheritance model. Also some biologists refer the gene-gene interaction to the direct interaction physically between their protein products like the protein-protein interaction. This GGI can be resulted from protein-protein interaction, epigenetic regulation, chromosomal structural interaction, translational regulation, signal transduction, biochemical networks, and numerous other physiological and developmental pathways. This GGI referred to as either *biological*, *physical* or *functional* interaction. Statisticians often use the term to refer to the deviation from the additive effect of alleles from each individual locus. This GGI referred to as *statistical*, or *populational* GGI.

Koo et al. described three different biological interactions. As shown in Figure 2.1, there are three types of biological interactions. First one is ′synthetic-interaction′. Gene A and gene B have same function which produce the purple phenotype C and work independently. Actually, these two gene do not interact each other. Second one is ′epistatic-interaction′. In the example, the wild type produce both a purple phenotype C and green phenotype D. If gene B is knockouted, purple phenotype C cannot be seen, and if gene A is knockouted, both purple and green phenotype cannot be seen. Last one is ′suppressive-interaction′. In this example, gene A is act as antagonist to gene B [Koo, et al. 2013].

In this thesis, the gene-gene interaction means an ′epistatic-interaction′ between genes (non-allelic). It also is referred to as ′epistasis′. Generally, epistasis is when the effect of one gene depends on the presence of one or more genes. Genetic interaction is an interaction between multiple genes that impacts the expression of a phenotype. The concept of ′epistasis′ was firstly introduced into genetics by William Bateson and Punnett (1910) in order to describe a masking effect whereby a variant or allele at one locus prevents the variant at another locus from manifesting its effect. In Fisher′s 1918 definition, epistasis refers to a deviation from additivity in the effect of alleles at different loci with respect to their contribution to a quantitative phenotype.

The ultimate goals of GGIs are to recognize gene functions, identify pathways and discover potential drug targets.

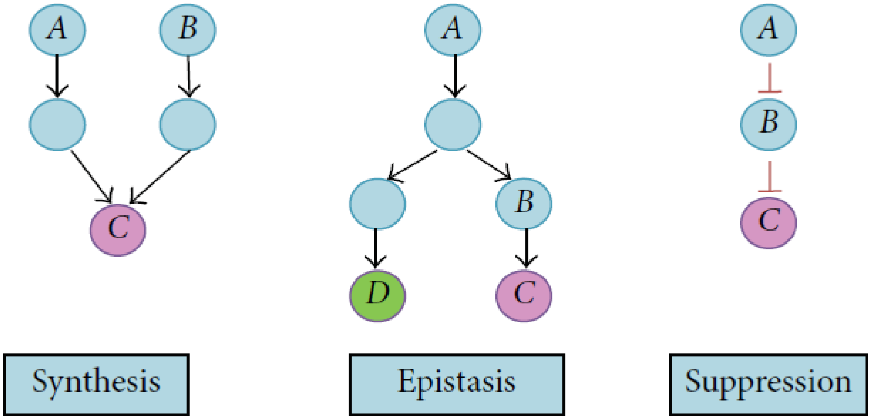


Figure 2.1 Types of genetic interactions. [Koo, et al. 2013]

2.2 Practical issues of gene-gene interaction

During the past few years, many methods for gene-gene analysis method have been developed as a complementary approach to GWAS (as described in chapter 2.3). However, there are some substantial limitations in current genome-wide scale gene-gene interaction methods.

First issue is that simple and powerful evaluation measure are required in GGI analysis, due to be rapidly computed across genome-wide scale high-dimensional genomic data.

Second issue is that it is hard to detect GGI for rare variants due to the sparcity of rare variants. Therefore, there are proposed few methods for GGI of rare variants. To overcome sparcity of rare variants in genetic interaction analysis, gene-wise interaction analysis method such as collapsing method is required.

Third issue is that the computational burden for gene-gene interaction analysis is heavy. For example, detection of 2^{nd} order interactions for 300,000 SNPs requires computing 4.5×10^{10} combinations for exhaustive searching. For 3^{rd} order interaction, 4.5×10^{15} combinations should be computed. Some interaction methods implement non-parametric permutation to calculate the observed significance, which takes a heavy computing time. Although MDR has a simple structure and fast computation, it is hard to find high-order interactions in large-scaled dataset because of its exhaustive searching scheme. For example, detection of 2^{nd} order interactions for 300,000 SNPs requires computing 4.5×10^{10} combinations by MDR. When we use 10-fold cross-validation or 1000-fold permutation test, it takes 10 times or 1000 times longer.

Fourth issue is that the statistical interaction is hard to be interpreted biologically. Phillips et al. defines three different forms of epistasis: compositional epistasis, statistical epistasis and functional epistasis [Phillips 2008]. Compositional epistasis describes the conventional usage of epistasis as the masking of one allelic effect by an allele at another locus. Functional epistasis is referred to the molecular interactions without a direct genetic link. Statistical epistasis is the usage of epistasis that is mentioned by Fisher, which is a deviation from additivity in the effect of alleles at different loci with respect to their contribution to a quantitative phenotype. In the cases of compositional epistasis and statistical epistasis without functional epistasis, it cannot take direct biological interpretation.

2.3 Overview of gene-gene interaction methods

2.3.1 Regression-based gene-gene interaction methods

In GWAS, it is widely used to fit a logistic regression model that includes both the main effects of variants and interaction effects between variants, and to test whether the interaction terms as equal to zero.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i}$$

where p_i is $\Pr(y_i = 1|X)$, α is the regression coefficient for intercept, coefficient β_1 and β_2 are main effect of two SNPs and β_3 represents the effect of an interaction term of two SNPs. The null hypothesis of test is $\mathrm{H}_0: \beta_3 = 0$.

The logistic regression model is a useful parametric method to model genetic or environmental factors with binary response such as disease status in case-control study. A similar approach can be applied for quantitative traits using linear regression. These analyses can be performed in almost any statistical analysis package such as SAS, SPSS and R.

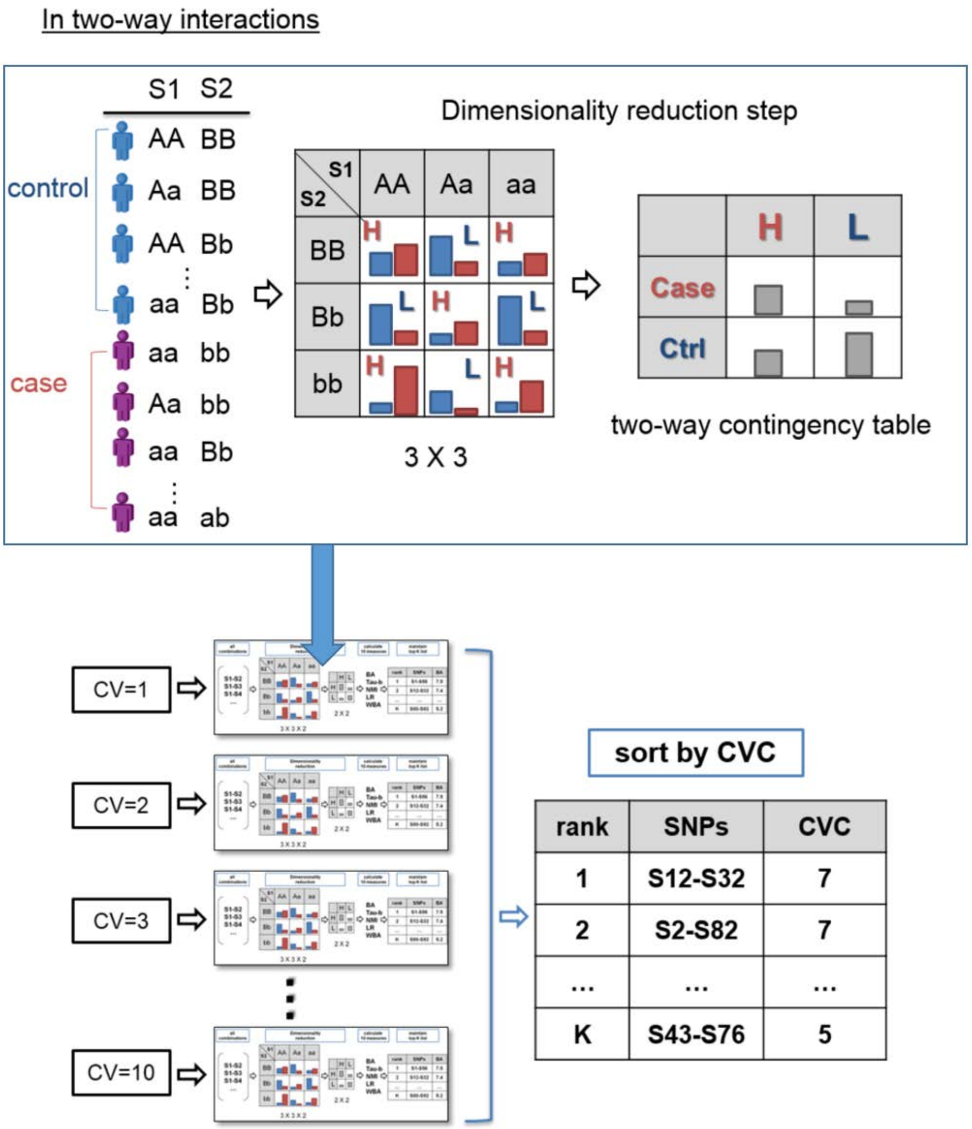
Zhao et al. proposed LD-based statistic to detect GGI between two unlinked loci [Zhao, et al. 2006]. The ′fast-epistasis′ option in PLINK performs a similar test. Zhao et al. found that their test statistic had more powerful than a conventional four degrees of freedom logistic regression of GGI.

Wan *et al.* proposed BOOST, which is a fast method for detecting gene-gene interaction using Boolean operation-based screening and testing [Wan, et al. 2010]. BOOST is computationally efficient and detects statistical significant interactions based on approximated likelihood ratio statistic. Their simulation study showed that BOOST has higher statistical power than PLINK.

2.3.2 Multifactor dimensionality reduction (MDR)

The multifactor dimensionality reduction (MDR) method proposed by Ritchie *et al.* [Ritchie, et al. 2001] is a non-parametric method that reduces the number of dimensions by converting a high-dimensional multi-locus model to a one-dimensional model to avoid the sparsity problem. MDR evaluates classifiers, which are SNP combinations associated with the disease of interest, to predict and classify disease status through cross-validation and permutation testing. The *k*-fold cross-validation splits the data into *k* subsets. The classifier is modelled on (*k*-1) subsets of the data and estimated by calculation of test accuracy on the remaining subset. This process is repeated for each subset. In addition to cross-validation, the permutation test can assess the statistical significance of MDR classifiers. However, it is unfeasible to use permutation tests for genome-wide scale interaction analysis because the permutation test is computationally intensive. To overcome this heavy computational burden, Pattin *et al.* proposed an efficient hypothesis test using extreme value distribution (EVD) [Pattin, et al. 2009]. Their simulation results showed that the proposed testing method requires at least 20 permutation data to keep up with similar power of 1000-fold permutation test.

A growing number of MDR extensions has been proposed since MDR firstly was introduced by [Ritchie, et al. 2001]. MDR can provide GGI results only for binary trait or phenotype such as case and control. Most MDR extensions focused on other types of trait.



rank	SNPs	CVC
1	S12-S32	7
2	S2-S82	7
…	…	…
K	S43-S76	5

Figure 2.2 MDR scheme

For GGI analysis of continuous trait, Lou et al. proposed generalized MDR (GMDR) [Lou, et al. 2007] which permits adjustment for discrete and quantitative covariates such as ethnicity, sex, weight, and/or age and is applicable to both dichotomous and continuous phenotypes. GMDR uses the classification scheme for high-risk and low-risk based on the residual score statistics of a generalized linear model. Gui et al. proposed Quantitative MDR (QMDR) that handles continuous trait by modifying MDR′s constructive induction algorithm to use a T-test [Gui, et al. 2013].

QMDR extends the MDR algorithm to working on continuous traits using comparing the mean value of each multi-locus genotype to the overall mean instead of comparing the case-control ratio of each multi-locus genotype to a fixed threshold.

For ordinal traits, Kim et al. suggested ordinal-MDR (OMDR) to facilitate gene-gene interaction analysis for ordinal traits [Kim, et al. 2013]. OMDR uses tau-b, a common ordinal association measure as an alternative to balanced accuracy to evaluate interactions. Kim et al. applied OMDR to ordinal obesity trait for body mass index (i.e., normal, pre-obese, mild obese and severe obese) of age-related eye disease study data.

For Family-based data, Martin et al. proposed the MDR-pedigree disequilibrium test (MDR-PDT), which is applicable to family-based designs [Martin, et al. 2006]. However, like the original MDR, the MDR-PDT method does not permit adjustment for covariates and is applicable only to dichotomous phenotypes. Lou et al. suggested a pedigree-based GMDR (PGMDR) method that extends GMDR to family type dataset [Lou, et al. 2008]. PGMDR can provide covariate adjustment and a unified framework for analyzing both continuous and dichotomous traits. Cattaert et al. proposed a FAMily MDR method (FAM-MDR) that are based on Wald-type statistics of association, whereas GMDR and PGMDR make use of score statistics [Cattaert, et al. 2010]. FAM-MDR showed to outperform PGMDR in their most of the considered simulation studies.

For survival traits, Gui et al. proposed Surv-MDR that is an extension of the MDR method to the survival phenotype using the log-rank test [Gui, et al. 2011]. Surv-MDR replaces balanced accuracy with log-rank test statistics as the score to determine the best models. Lee et al. proposed Cox-MDR which is an extension of the GMDR to the survival phenotype using a martingale residual as a score to classify multi-level genotypes as high- and low-risk groups [Lee, et al. 2012]. Cox-MDR provides covariate adjustment, but Surv-MDR cannot adjust the covariates.

For multiple phenotypes, Choi et al. proposed multivariate GMDR (multi-GMDR) that an extension of GMDR. Multi-GMDR determines high-risk from low-risk groups of GMDR framework by using the score vectors from generalized estimating equations with multivariate phenotypes to extend generalized linear models of GMDR [Choi and Park 2013].

2.3.3 Gene-gene interaction methods using machine learning methods

In machine learning methods, neural network (NN), support vector machine (SVM) and random forest (RF) have been applied to GGI analysis. These machine learning methods cannot investigate all combinations of multiple variants because of their intensive computation. Therefore, the most methods can be applied to feature selection method, or to selected candidates from feature selection algorithms such as recursive feature selection and recursive feature elimination.

NN is a one of the learning algorithms that have been widely applied in genetic data. Although its long history and well-structured algorithm, it is computationally infeasible to conduct exhaustive search of multilocus. Ritchie et al. had utilized genetic programming to optimize the architecture of neural network (GPNN) and back propagation neural network (BPNN) to model GGIs [Ritchie, et al. 2003b]. Motsinger-Reif et al. proposed grammatical evolution neural network (GENN) that identify gene-gene and gene-environment interactions.

SVM can be used to predict GGI by learning from the features which are known GGI. The training data of SVM has the two groups which are positive group with genetic interaction and negative group without genetic interaction. In order to detect gene-gene interactions, Chen et al. had applied SVM to various kinds of combinatorial optimization methods such as recursive feature addition (SVM-RFA), recursive feature elimination (SVM-RFE), local search (SVM-Local), and genetic algorithm (SVM-GA) [Chen, et al. 2008]. Fang and Chiu had proposed extended SVM and SVM based pedigree-based generalized multifactor dimensionality (PGMDR) for family structured genomic data [Fang and Chiu 2012]. de Oliveira proposed SVM extension to simultaneously select the most relevant SNPs markers by a continuous variable using Support Vector Regression with Pearson Universal kernel as fitness function of a binary genetic algorithm [de Oliveira, et al. 2014]

Jiang et al. applied RF method to transform GGI from all possible combinations of genetics variants into a manageable set of candidates by reducing the search space for GGI [Jiang, et al. 2009]. Schwarz et al. introduced a random jungle using permutation importance measures to detect important SNP [Schwarz, et al. 2010]. Winham et al. proposed that the RF method is used to detect high-dimensional GGI effects and their potential effectiveness for detecting GGI [Winham, et al. 2012].

Davis et al. applied RF and random jungle to filter the candidates with main effect for enrichment of interaction effect [Davis, et al. 2013].

2.3.4 Entropy-based gene-gene interaction methods

Recently, several approaches based on information theory for modelling GGI have been proposed [Chanda, et al. 2007; Dawy, et al. 2006; Ruiz-Marin, et al. 2010]. Shannon started the information theory in 1948 by introducing the entropy that is a measure for complexity in mathematical theory of communications [Shannon 1948].

Dawy *et al.* [Dawy, et al. 2006] proposed a relevance-chain method to identify the strongly associated lower-order interactions and build high-order interaction with the use of conditional mutual information. This method can provide fast detection of high-order interaction but shows poor performance for GGI with no strong marginal effects. Chanda *et. al.* [Chanda, et al. 2007] proposed the k-way interaction information (KWII) metric and the total correlation information (TCI) for GGI identification. These entropy-based measures represent the amount of information of redundancy and dependency between SNPs and an environmental variable. This method performs a permutation test for statistical significance of detected interaction models. Ruiz-Marín *et al.* [Ruiz-Marin, et al. 2010] proposed an entropy-based test for identification of single-locus association analysis. Although it showed a more powerful performance than the conventional Fisher tests, this method needs to be extended to handle GGI analysis. Yee *et al.* [Yee, et al. 2013] proposed a modified entropy based method to evaluate the interactions between single SNP combinations.

Their method was shown to be superior to the MDR method in most simulation cases. However, applying this entropy based method directly to the genome-wide scale data would be infeasible because of computationally intensive permutations.

Chapter 3

Entropy-based gene-gene interaction

3.1 Introduction

In this chapter, we develop a fast and efficient method, named IGENT, Information theory-based GEnome-wide gene-gene iNTeraction method, using entropy to identify the gene-gene interaction in genome-wide scale. IGENT supports two types of strategies to identify gene-gene interactions related with diseases in genome-wide scale. One is an exhaustive search approach for lower-order interactions such as 2nd order interaction, and the other is a stepwise selection approach for higher-order interaction. With tens of thousands of SNPs from thousands of samples, it is difficult to calculate higher-order interaction exhaustively because the computational burden is too heavy. IGENT provides a stepwise approach for higher-order interactions. The evaluation is based on the approximated gamma distribution of information gain without using permutation procedure, which allows us to overcome the computation burden for the GGI analysis in genome-wide scale [Goebel, et al. 2005]. In our two-way interaction simulations, we compared the performances of IGENT, BOOST, MDR, and SVM. IGENT showed better performance than BOOST, MDR and SVM in most simulation settings. Also, IGENT is as fast as BOOST, and presented stable performance is various epistasis models even with low MAF. We successfully applied IGENT to age-ralted macular degeneration (AMD) data and WTCCC bipolar disorder data.

3.2 Methods

3.2.1 Entropy-based gene-gene interaction analysis

For detecting GGI associated with phenotypes, our measure is based on basic concept of information theory. The entropy, which measures the quantity of an uncertainty, is defined as

$$H(Y) = -\sum_{j} p\big(Y = y_j\big) \log_2 p\big(Y = y_j\big),$$

where the entropy *H(X)* of a discrete random variable *Y* is a function of the probability distribution *p(Y=y$_j$)* which measures the average amount of information contained in *Y*, or equivalently, the amount of uncertainty removed upon revealing the outcome of *Y*.

Conditional entropy of *Y* given another discrete random variable *X* is

$$H(X|Y) = -\sum_{i} p(X = x_i) H(Y|X = x_i)$$

The information gain (IG) is defined as follows,

$$IG(Y|X) = H(Y) - H(Y|X)$$

IG which is also called mutual information (MI) can be explained as the reduction in entropy (or uncertainty) of one random variable given another. It is known that the IG follows gamma distribution with parameter $a = (|X| - 1)(|Y| - 1)$ and $b = 1/(N \ln 2)$ approximately for the independent X and Y random variables [Goebel, et al. 2005].

$$\widehat{IG}(Y|X) \sim \Gamma\left(\frac{1}{2}(|Y| - 1)(|X| - 1), \frac{1}{N \ln 2}\right) \qquad (1)$$

where N is the sample size and $|X|$ and $|Y|$ denote the number of levels of the random variables X and Y. For example, two-way interaction in case and control dataset, $|X|$ is 9 and $|Y|$ is 2.

We use the information gain to detect GGI associated with phenotype. Given a case-control study with n individuals, let Y be the disease status and X be the SNP combinations, then

$$H(Y) = H(disease\ status)$$
$$H(Y|X) = H(disease\ status|SNP\ combination)$$

IG is given as

$$IG = H(Y) - H(Y|X)$$

The value of IG represents the true association strength. Since, under the null hypothesis of no association, IG follows a gamma distribution approximately by (1), we can assess the statistical significance of the association of SNP combinations and disease.

3.2.2 Exhaustive searching approach and stepwise selection approach

We propose IGENT, an entropy-based gene-gene interaction method for genome-wide interaction analysis. IGENT supports exhaustive search (IGENT_exhaust) for lower-order interaction and stepwise search (IGENT_stepwise) for higher-order interaction. In Figure 3.1, our exhaustive search approach and stepwise selection approach are described graphically.

IGENT_exhaust performs an exhaustive search for all possible combinations of variants for the given low order. IGENT_stepwise selects higher-order interactions in a stepwise manner. The detailed steps are summarized as the follows.

1. Initial step: for all SNPs, calculate 1st order IG^k when k is order (in 1st order, k = 1.).
2. Select SNP or SNP combinations with $p^k < t$, when p^k is p-value of hypothesis testing using the gamma distribution and t is significant threshold.
3. Calculate IG^{k+1} for k+1 order interactions for the combinations with selected SNP or combinations adding additional other single SNP.
4. If there are significant interactions in k+1 order, let $k = k + 1$ and repeat step 2~4.Otherwise, stop forward selection and repeat 2~4 step with the next ranked combinations.

This IGENT_stepwise selection approach reduces search space dramatically. With large genome-wide scale data, this approach makes it feasible to discover higher-order interactions. Although this stepwise algorithm is not guaranteed to find the global optimum interaction model, it provides at least a local optimum interaction model with some marginal effects. Therefore, this stepwise approach may have a limitation in detecting the gene-gene interactions without any marginal effects.

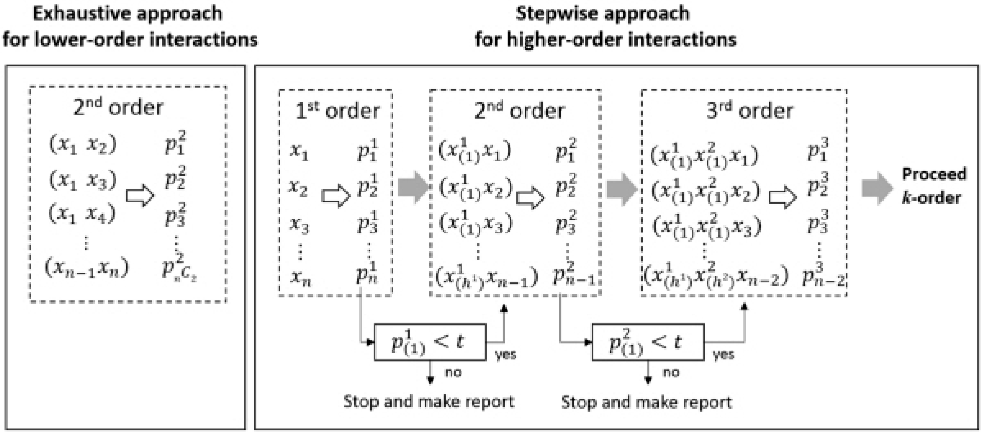


Figure 3.1 Exhaustive approach and stepwise approach in IGENT. *t* is threshold, p_j^k is p-value for *j*th combination in *k*-order interaction. $p_{(i)}^k$ is *i*th ordered p-value among p-values of all combinations in *k*-order interaction. h^k is the number passing threshold in *k*-order interaction.

3.2.3 Simulation setting

The main purpose of our method is to identify epistatic interactions from genome-wide data. In order to detect gene-gene interaction for genome-wide data, computational efficiency is a key issue. In simulation 1, we compared the computational efficiency of IGENT and other methods such as BOOST, MDR, RF and SVM. Among these methods, only IGENT and BOOST was shown to be feasible to analyze gene-gene interaction in genome-wide scale, as shown in simulation 1 of Results section. Thus, we mainly compared IGENT and BOOST in genome-wide scale with regard to the power of identifying causal gene-gene interaction through simulations 2, 3, and 4. In simulation 5, we compared IGENT_exhaust and IGENT_stepwise.

For these simulation studies, we use following three epistatic models:

1) Epistatic model set 1 : Eight interaction models

Models 1-1, 1-2, and 1-3 have different strength of genetic effects while fixing the interaction structure, the minor allele frequencies (MAF) and prevalence which have been used by Namkung *et al.* [Namkung, et al. 2009b]. Models 1-4, 1-5, and 1-6 have different interaction structures and penetrance functions which were used by Ritchie *et al.* [Ritchie, et al. 2003a]. Models 1-7 and 1-8 were used by Bush *et al.* [Bush, et al. 2008]. Eight interaction models are summarized in Table 3.1.

Table 3.1 Eight interaction model

	Model 1-1			**Model 1-2**			**Model 1-3**			**Model 1-4**		
Prevalence	0.05			0.05			0.05			0.046		
MAF	0.1			0.1			0.1			0.1		
	AA	**Aa**	**aa**	**AA**	**Aa**	**aa**	**AA**	**Aa**	**aa**	**AA**	**Aa**	**aa**
BB	1.21	0.20	0.20	1.23	0.33	0.33	1.22	0.40	0.40	0.55	1.75	1.33
Bb	0.20	5	5	0.33	3	3	0.40	2.50	2.50	1.54	0.18	0.74
bb	0.20	5	5	0.33	3	3	0.40	2.50	2.50	1.75	0.18	0

	Model 1-5			**Model 1-6**			**Model 1-7**			**Model 1-8**		
Prevalence	0.026			0.017			0.052			0.048		
MAF	0.1			0.1			0.2			0.4		
	AA	**Aa**	**aa**	**AA**	**Aa**	**aa**	**AA**	**Aa**	**aa**	**AA**	**Aa**	**aa**
BB	1.16	0.38	0.76	1.15	0.40	0.17	0.84	1.35	0.80	0.52	1.07	1.89
Bb	0.38	3.70	1.97	0.28	4.23	4.89	1.30	0.39	1.45	1.30	0.92	0.59
bb	0.76	1.97	2.92	1.15	0.06	5.56	1.45	0.13	1.04	1.21	1.08	0.33

2) Epistatic model set 2 : four interaction models with main effects

Model 2-1 is a multiplicative model. Model 2-2 is an epistasis model that has been used to describe handedness and the colour of swine. Model 2-3 is a classical epistasis model. Model 2-4 is the XOR model. The details of these four models have been described by Wan *et al.* [Wan, et al. 2010].

3) Epistatic model set 3 : Seventy interaction models without main effects

Seventy Disease models without main effects have been proposed by Velez *et al.* [Velez, et al. 2007]. These 70 epistatic models are distributed across six heritability values (0.01, 0.025, 0.05, 0.1, 0.2, and 0.4) and two different MAFs (0.2 and 0.4).

Using these epistatic model sets, we conduct the following five simulation studies.

Simulation 1: Comparing computational efficiency for genome-wide gene-gene interaction analysis

To compare computational efficiency with IGENT, BOOST, MDR, SVM and RF, we construct simulation data using the epistatic model set 1. Each epistatic models contains 2000 individuals balanced between cases and controls. Various numbers of SNPs (50, 100, 500, 1K, 2K, 5K, 10K, 100K, 350K, and 500K) are considered. All analysis are carried out on single core of a 3.16 GHz CPU with 4G memory on LINUX.

Simulation 2: Estimating type I error in null simulation

To take an assessment in terms of type I error, we construct 1000 replicates of null simulation data with 1000 SNPs and 1000 individuals based on the epistatic model set 1. In this null simulation data, all SNPs have no association with disease status. Using null simulation, we compare false positive rates of IGENT and BOOST.

Simulation 3: Comparing the power of gene-gene interaction with main effects

To compare the power of IGENT and BOOST in gene-gene interaction with main effects, we use the epistatic model set 2. The MAFs of disease-associated SNPs is set to be 0.1, 0.2, and 0.4. Each data set has 1000 SNPs with two different sample sizes of 800 and 1600 respectively. We generate 100 replicate data sets under each setting. Using this simulated data, we compare the power of IGENT and BOOST for gene-gene interaction with main effects.

Simulation 4: Comparing the power of gene-gene interaction without main effects

For evaluation of finding causal gene-gene interaction with no marginal effects, we use the epistatic model set 3. Using these 70 epistasis models in the set, we generate 100 replicate sets with 1000 SNPs (one pair is causal interaction, others are non-causal SNPs), and four sample sizes (200, 400, 800, and 1600 individuals).

Simulation 5: Comparing the efficiency of stepwise search approach

For comparison of the efficiency of IGENT_stepwise, we use the epistatic model set 1. We generate 100 replicate set with 50 SNPs from 400 individuals. Through this simulation, we compare the power and computational efficiency between IGENT_stepwise and IGENT_exhaust.

3.2.4 Genome-wide data for Biopolar disorder (BD)

Using bipolar data from the Wellcome Trust Case Control Consortium (WTCCC) [Wellcome Trust Case Control 2007], we demonstrated genome-wide gene-gene interaction analysis for 2nd-order and higher-order interaction. SNPs with call rates <95% were excluded from the analysis. SNPs showing Hardy-Weinberg equilibrium (HWE) p-value<5.7□10^{-7} were filtered out. Of the remaining SNPs, only SNPs showing MAF of at least 5% were carried forward for further analysis. All quality control steps were conducted using PLINK version 1.07 [Purcell, et al. 2007] and R scripts. We performed imputation using fastPHASE version 1.2 [Scheet and Stephens 2006] to increase the density of interrogated SNPs. After quality control and imputation process, WTCCC-BD dataset contained 354,022 SNPs and 4,806 samples.

IGENT was applied to exhaustive two-way interaction analysis of 6.27×1010 pairs of SNPs for WTCCC-BD data and stepwise selection approach for higher-order interactions.

3.2.5 Genome-wide data for Age-related macular degeneration (AMD)

For real data application, we used the AMD data set which contains 116,209 SNPs genotyped with 96 cases and 50 controls from the Age-Related Eye Disease Study (AREDS) [Klein, et al. 2005]. We conducted the same quality control process as in the BD data analysis except for MAF < 0.01. All quality control steps were conducted using PLINK version 1.07 [Purcell, et al. 2007] and R scripts. After quality control process, we used remained 102,504 SNPs from 146 individuals. Pair-wise interaction analysis of all 5,253,483,756 pairs was conducted with IGENT_exhaust and BOOST. Also, IGENT_stepwise was performed for higher-order interactions.

3.3 Results

3.3.1 Simulation results

In this section, we perform simulation studies to evaluate the properties of IGENT and to compare it with other previous proposed methods. In order to detect gene-gene interaction with genome-wide data, computational efficiency is a key issue. In simulation 1, we compared the computational efficiency of IGENT and other methods such as BOOST, MDR, RF, and SVM. Among these methods, only IGENT and BOOST were shown to be feasible to analyze gene-gene interaction in genome-wide scale in simulation 1. We mainly compared IGENT and BOOST in regard to the power of identifying causal gene-gene interaction in simulations 2, 3, and 4. In simulation 5, we compared IGENT_stepwise and IGENT_exhaust.

Simulation 1: comparing computational efficiency for genome-wide gene-gene interaction analysis

In order to compare the computational efficiency of IGENT and other methods including BOOST, MDR, RF, and SVM, we conducted 2nd order interaction analysis with various the number of SNPs (50 to 500K). We used LIBSVM library [Chang and Lin 2011] and “randomforest” R package [Breiman 2001] for SVM and RF methods, respectively. All methods used an exhaustive search strategy for fair comparison.

Table 3.2 Computation time of IGENT, BOOST, MDR, RF, and SVM

Computation time is measured in simulation 1 dataset which have 2000 individuals. All methods used an exhaustive search strategy for 2nd order interaction analysis.

SNP size	IGENT_exhaust	BOOST	MDR	RF	SVM
50	<1s	<1s	1s	11s	13s
100	<1s	<1s	4s	46s	53s
500	<1s	<1s	1m 8s	20m	23m
1K	3s	1s	4m 25s	1h 15m	1h 29m
2K	8s	6s	19m 52s	5h	5h 50m
5K	38s	30s	2h 4m	1d 6h	1d 12h
10K	2m 34s	2m 7s	*8h 16m	*5d 5h	*6d 3h
100K	4h 23m	3h 32m	*35d	*520d	*614d
350K	2d 4h	1d 19h	*422d	*6366d	*7524d
500K	4d 10h	3d 15h	*861d	*12992d	*15353d

* This computing time is estimated from the computing time in simulation data with 5000 SNPs.

All analysis are carried out on single core of a 3.16 GHz CPU with 4G memory on LINUX.

Table 3.3 Comparison of the type I error in null simulation

Thresholds	**False Positive Rate**	
after Bonferroni correction	**IGENT**	**BOOST**
0.01	0.012	0.011
0.05	0.057	0.054
0.10	0.112	0.108
0.15	0.166	0.153
0.20	0.219	0.215
0.25	0.270	0.264
0.30	0.321	0.318

Table 3.2 presents computation times to finish 2nd order interaction analysis by each method. In simulation data with 350K SNPs, IGENT_exhaust and BOOST can finish the interaction analysis within about 2.17 days and 1.8 days, respectively. However, due to their heavy computation times, MDR, RF, and SVM are not feasible to conduct the gene-gene interaction analysis with genome-wide dataset. For genome-wide interaction analysis, we focus on comparing the power of IGENT and BOOST in simulations 2, 3, and 4.

Simulation 2: estimating type I error in null simulation

The type 1 error rates of IGENT_exhaust and BOOST are shown in Table 3.3. Although the type I error rates of IGENT_exhaust and BOOST seem to be slightly higher than the nominal value, it can be shown that the type I errors of IGENT and BOOST agree with the nominal value lying within the confidence interval.

Simulation 3: comparing the power of gene-gene interaction with main effects

In simulation 3, we compared the IGENT_exhaust, IGENT_stepwise, and BOOST for detecting causal gene-gene interactions with main effects. In simulation data, IGENT used both exhaustive mode and stepwise mode, and BOOST used an exhaustive mode for searching the 2nd order interactions. The power is calculated as the proportion of 100 data sets in which the interactions of the disease-associated SNPs are detected. In all simulation data, we counted the interaction with its p-value (after multiple comparison procedure by Bonferroni correction) < 0.05. In stepwise mode, only variants with marginal p-value < 0.05 were proceeded to the next step for calculating the 2nd order interactions. In simulation 3, the detection probability of IGENT_exhaust showed the best performance in most models except for Models 2-4 (Figure 3.2). The performance of BOOST became worse in the simulation models with low minor allele frequency (MAF 0.1 and 0.2). In simulation 3, the average power of IGENT_stepwise was about 60% relative to IGENT_exhaust, but its computing time was less than 1%(only 0.43%) of IGENT_exhaust.

Simulation 4: comparing the power of gene-gene interaction without main effects

In simulation 4 which has causal gene-gene interaction without main effects, IGENT_exhaust performed better than or equivalent to BOOST in most simulation models. In simulation model with lower MAF and small sample size, BOOST showed poor performance. However, they provided equivalent results for models with a MAF of 0.4 or large sample sizes (Figure 3.3).

Simulation 5: comparing the efficiency of stepwise analysis and exhaust analysis of IGENT

We evaluated the performance of IGENT_stepwise in simulation 5 based on epistatic model set 1. All models were designed with the 2nd order interaction effects and no marginal effects. Although these simulation models do not include the higher-order interaction effects over the 2nd order, it is possible for spurious higher-order interaction to show the large effects on phenotype. To allow for finding spurious higher-order interactions, we exhaustively identified interactions from 1st to 4th orders. By comparing the identified interactions from IGENT_exhaust to those from IGENT_stepwise, we were able to evaluate the performance of IGENT_exhaust.

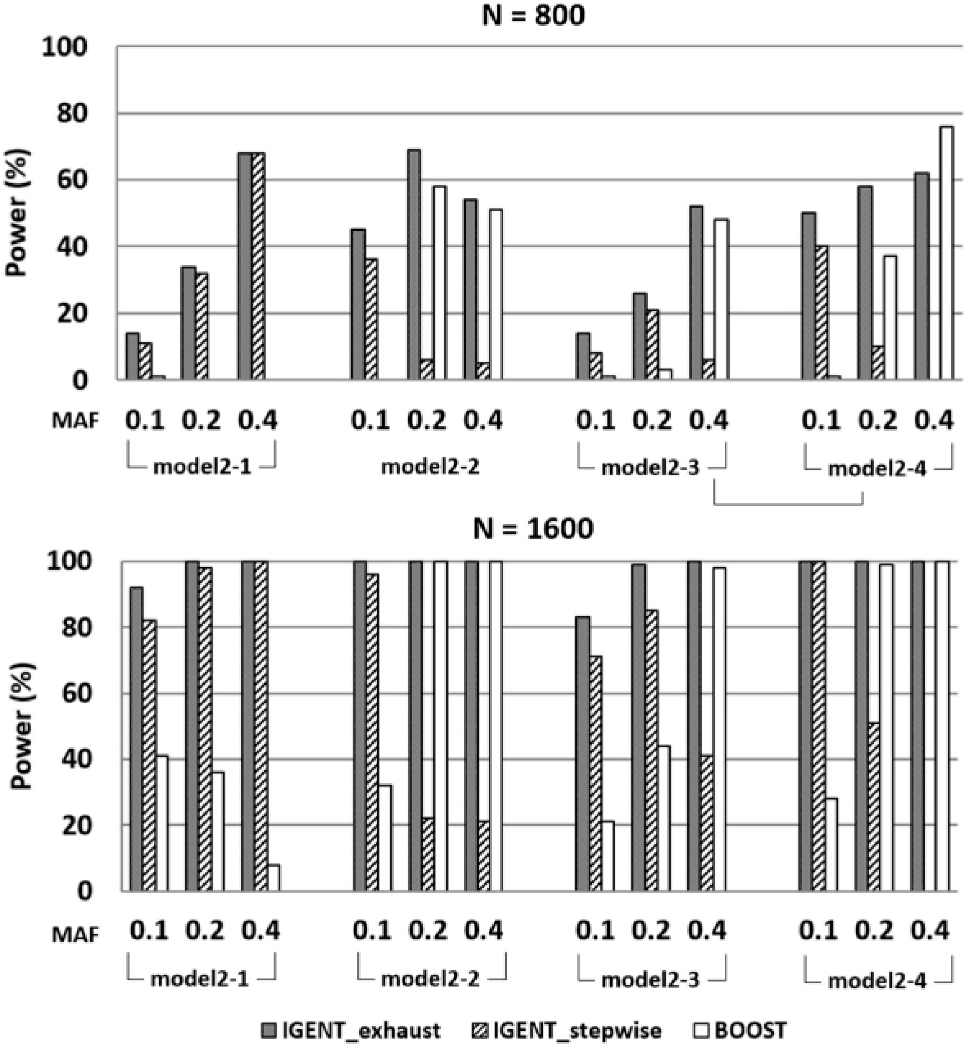


Figure 3.2 The power comparison between IGENT and BOOST on four disease models with main effects. Results are shown in separate panels for each sample size (800 and 1600). MAF are presented on the X-axis. Model 2-1 is a multiplicative model. Model 2-2 is an epistasis model that has been used to describe handedness and the colour of swine. Model 2-3 is a classical epistasis model. Model 2-4 is the XOR model.

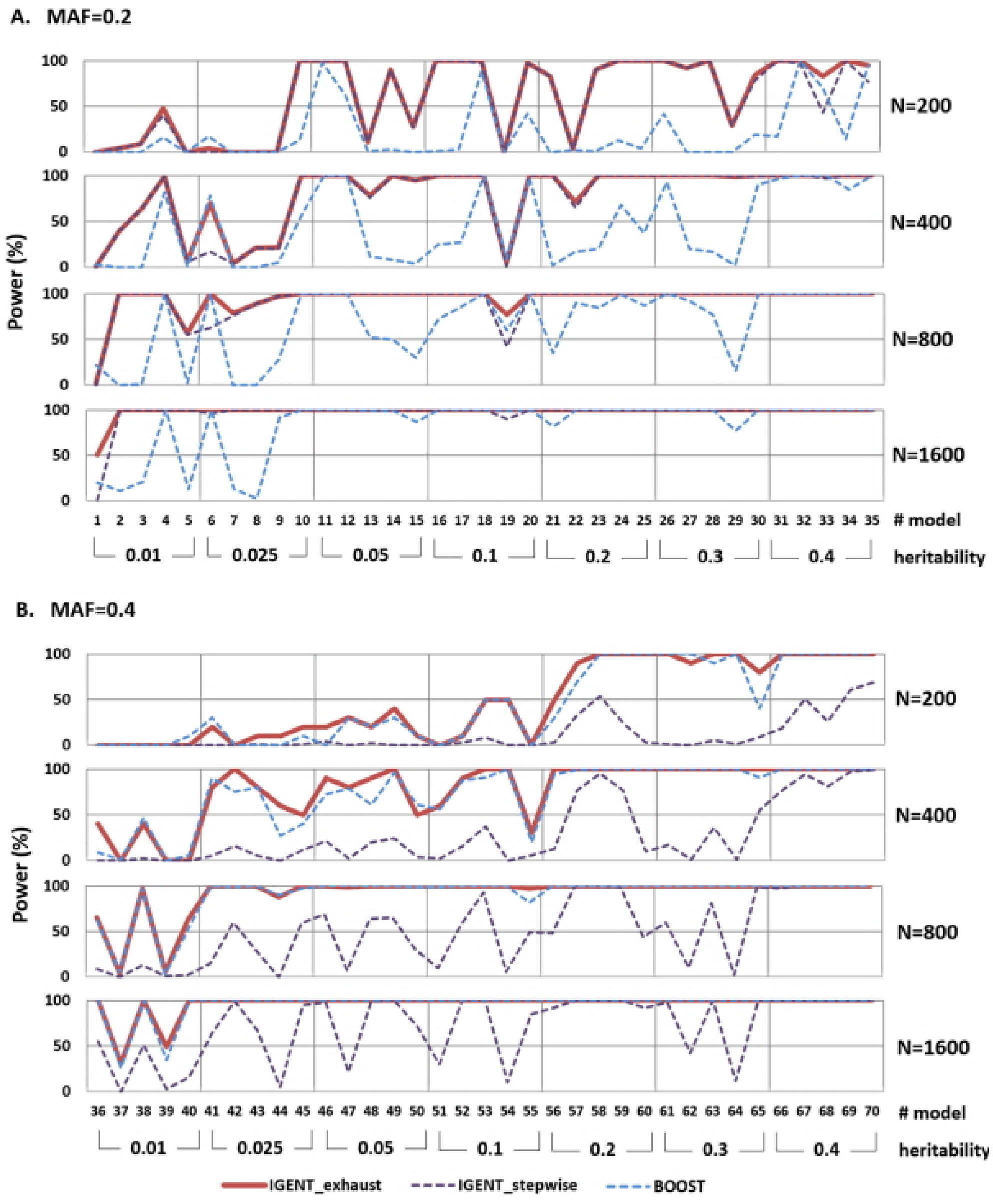


Figure 3.3 Performance comparison with IGENT, BOOST in 70 simulation models

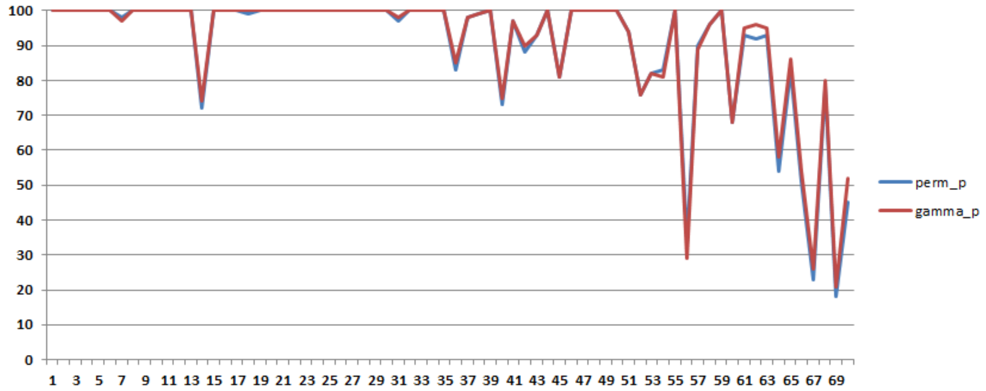


Figure 3.4 Performance comparison with permutation method (perm_p) and gamma distribution approximation based method (gamma_p) in IGENT in 70 simulation models. In permutation method (perm_p), we conducted 1000 permutations with 100 repeated dataset. Y-axis is detection probability of causal interaction pair with p<0.05.

Table 3.4 Efficiency of stepwise analysis

Model	**Power[a] in Stepwise approach**	**Power in exhaustive approach**	**ratio of power[b]**	**Computation in stepwise approach[c]**	**ratio of computation[d]**
1	0.69	1.00	**0.69**	148.4	**0.12**
2	0.71	0.92	**0.77**	149.7	**0.12**
3	0.67	0.80	**0.84**	154.7	**0.13**
4	0.87	0.94	**0.93**	147.6	**0.12**
5	0.62	0.88	**0.70**	147.0	**0.12**
6	0.63	0.96	**0.66**	145.3	**0.12**
7	0.19	0.25	**0.76**	167.3	**0.14**
8	0.15	0.17	**0.88**	445.6	**0.36**

[a] Detection probability,

[b] the ratio of power between stepwise approach and exhaustive approach

[c] Average number of combinations to be computed in stepwise approach

[d] Computation ratio is the ratio of computation amount of stepwise approach and computation amount of exhaustive approach. The computation of exhaustive approach is calculated using ${}_{50}C_2$ = 1225.

Table 3.4 shows IGENT_stepwise has the 66~93% of power of the IGENT_exhaust by using only 12~36% computation of the IGENT_exhaust. For the genome-wide interaction analysis, IGENT_stepwise can perform high-order interaction analysis very efficiently.

3.3.2 Analysis of WTCCC bipolar disorder (BD) data

We conducted genome-wide two-way interaction analysis and higher-order interactions with WTCCC-BD dataset [Wellcome Trust Case Control 2007]. The IGENT_exhaust completed all two-way interaction pairs (6.25×10^9) in about 74 hours on a 3.16 GHz CPU with 4G memory on LINUX. IGENT_stepwise took about 1.5 hour in higher order interactions on the same system. Through exhaustive two-way interactions, IGENT_exhaust reported 39 significant interactions. Among these 39 interactions, 26 pairs were also reported by IGENT_stepwise. Among these hub genes, LOC390730, DPP10, and CDC25B have been reported with strong marginal effects in a previous study [Wellcome Trust Case Control 2007] (Table 3.5). B2GALT5, PI15, TLE4, AKAP10, and CHST2 did not show significant associations in single locus analysis but showed strong interactions. These genes have been reported as causal genes associated with bipolar disorder in other studies [Djurovic, et al. 2010; Hamshere, et al. 2009; Iwamoto, et al. 2011; Laje, et al. 2009; Martinowich, et al. 2009; van Winkel, et al. 2011].

In Figure 3.5, using two-way interaction analysis by IGENT, we constructed the interaction network of WTCCC-BD. In two-way interaction network, a node represents a gene with SNP(s), edge is interaction reported by IGENT analysis. Node size shows the degree of the node and edge width shows the number of SNP-SNP interactions. All significant interactions were annotated by HuGE navigator database [Yu, et al. 2008] and GWAS catalog [Hindorff, et al. 2009]. This network graph represents two-way interactions of genome-wide association with bipolar disorder and facilitates biological interpretations.

3.3.3 Analysis of age-related macular degeneration (AMD) data

We conducted 2nd order interaction analysis and high-order interaction analysis using IGENT and BOOST for AMD data. Table 3.6 shows the top 5 interactions or SNPs identified by IGENT. In the case of AMD data, there are SNPs (rs380390 (CFH) and rs1329428 (CFH)) with strong marginal effect. These SNPs were also reported previously that they have strong association with AMD disorder [Klein, et al. 2005]. IGENT also detected two interactions (CFH (rs380390) - SGCD (rs931798) and CFH (rs1329428) - MED27 (rs9328536)). These two interactions also have a SNP with a strong marginal effect.

Table 3.5 Hub genes (degree of nodes ≥ 10) in two-way interactions of WTCCC-BD

Hub gene	degree	location	SNP(s)	Reference[a]
B3GALT5	115	21q22.2b	rs980184	[Hamshere, et al. 2009]
LOC442261	98	6q23.2d	rs4896044	
PI15	32	8q21.11b	rs2954873	[Martinowich, et al. 2009]
LOC390730	26	16q12.2a	rs7188309 rs11640993 rs8056052 rs2192859 rs1344484 rs10521275 rs11647459 rs2387823	[Wellcome Trust Case Control 2007]
PHF20	24	20q11.23a	rs6060710	
TLE4	13	9q21.31b	rs914715 rs11138278	[Laje, et al. 2009]
DPP10	12	2q14.1b	rs11123306 rs708647 rs1375144 rs6741692	[Djurovic, et al. 2010; Wellcome Trust Case Control 2007]
AKAP10	10	17p11.2d	rs203466 rs203457 rs119672 rs2108978	[Iwamoto, et al. 2011]
CHST2	10	3q23d	rs4683457	[van Winkel, et al. 2011]

a. Reference is literature related with bipolar disorder.

Table 3.6 Interaction analysis result using AMD data set

rank	SNP	*P*
1	CFH(rs380390) SGCD(rs931798)	8.454×10^{-12}
2	CFH(rs1329428) MED27(rs9328536)	1.943×10^{-10}
3	CFH(rs380390)	2.087×10^{-7}
4	INPP4B(rs3775640)	3.128×10^{-7}
5	CFH(rs1329428)	1.166×10^{-6}

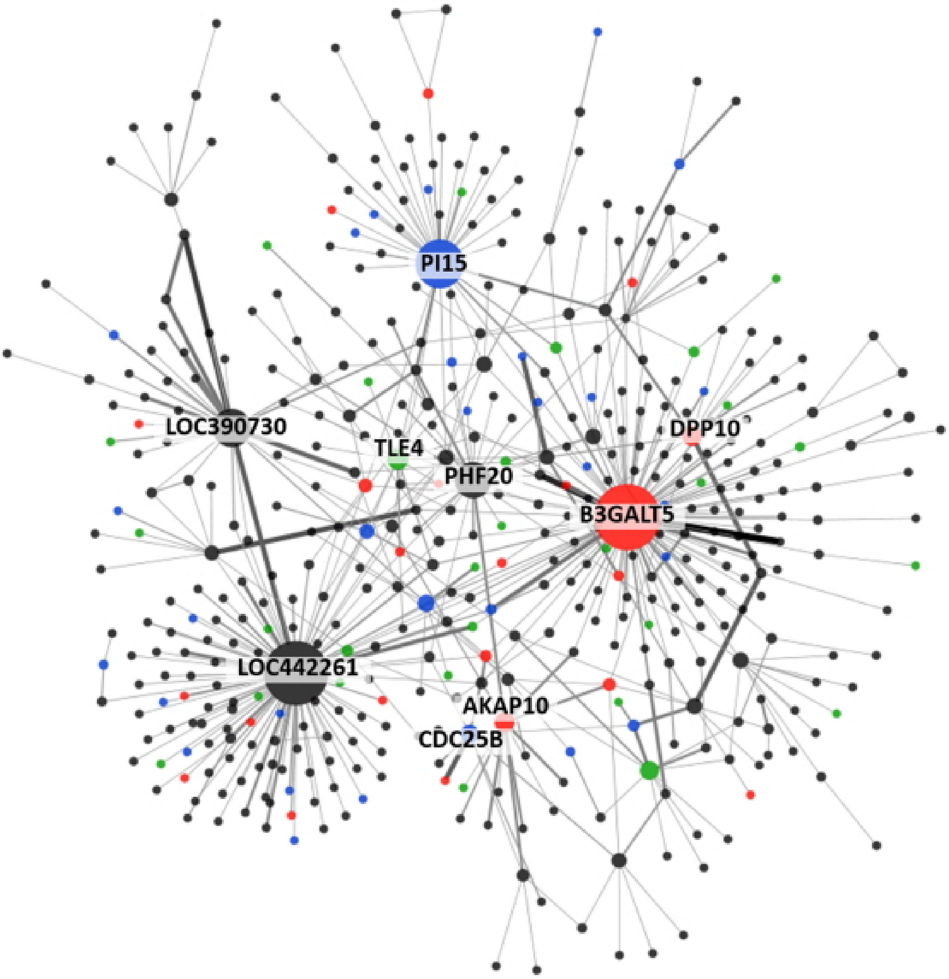


Figure 3.5 Gene-gene interaction network for WTCCC-BD dataset. Red nodes represent genes reported in previous GWAS literature with bipolar disorder dataset. Blue nodes are the genes related with bipolar disorder in previous literature. Green nodes are the genes related with other psychiatric disorders (schizophrenia and depression disorder). Width of edge is the significance level of interaction.

3.4 Discussion

In this chapter, we proposed a fast analysis for searching for high-order interactions associated with complex diseases. IGENT uses information gain which represents association strength with GGI and phenotype without a specific genetic model. The IG measure can be used to compare the association strength across different order of interactions. IGENT adopts an exhaustive search scheme that investigates all possible interactions in lower-order interactions and a stepwise search scheme for higher-order interactions. The permutation and exhaustive search schemes of the previous GGI methods are computationally too intensive to be employed in large genome-wide scale data set for high-order interactions.

Note that IGENT is as fast as BOOST and shows better performance than BOOST. BOOST has been known to have a limitation that the degree of freedom of the statistical test should be reduced when the contingency table is too sparse due to low MAF [Wan, et al. 2010]. IGENT, however, presents stable performance in various epistasis models even with low MAF.

To evaluate significance of IGENT's result, we used hypothesis testing framework by approximating the gamma distribution. It is known that IG follows the gamma distribution under the null hypothesis. Using approximation to the gamma distribution instead of permutation, we can easily calculate statistical significant interactions and save the computation time remarkably. To find more accurate significant causal interactions, this approximation of the gamma distribution can be used for screening step. After screening step, the permutation for selected pairs gives more accurate significant value.

A stepwise approach is more efficient than exhaustive approach in terms of computation. However, this stepwise approach has a trade-off between computational efficiency and detection of optimal gene-gene interactions. Our stepwise approach, IGENT_stepwise, reduced a search space extremely for detecting

GGI with marginal effects. Although GGI without marginal effects can be generated mathematically [Botstein and Risch 2003; Culverhouse, et al. 2002; Kotti, et al. 2007], it is still unclear in practice how the GGI model without marginal effect is biologically associated with a complex disease [Cordell 2009].

In MDR and IGENT methods, a pair with strong marginal effects and week interaction effect can be detected as a significant interaction pair. In this case, we excluded the pair by calculating the marginal effect and interaction effect.

In an exhaustive search scheme, our simulation result showed that IGENT_exhaust consistently had better performance than BOOST, as shown in Figures 2 and 3. Although both BOOST and IGENT showed efficient and fast computational performances, IGENT showed power higher than or equivalent to that of BOOST.

3.5 Conclusion

In conclusion, we proposed a fast and efficient enhanced entropy-based GGI analysis method. Due to its fast and efficient computation scheme, it can easily identify the gene-gene interaction in genome-wide scale. Through real GWAS data analysis, IGENT successfully identified low order and high order interactions. IGENT has been implemented with C++.

Chapter 4

Gene-gene interaction for rare variants

4.1 Introduction

In recent years, studies support the ‘common-disease rare-variants’ (CDRV) hypothesis which claims that complex disorders are caused by multiple rare variants. Type 2 diabetes mellitus (T2D) is a complex disease which is caused by both genetic composition and environmental factors. The exact biochemical mechanism is yet to be unveiled, however, impairments in insulin action and secretion certainly take parts. Unlike Type 1 diabetes, T2D is characterized primarily by ‘insulin resistance’; and a vast majority of this resistance is shown as defects at the postreceptor level (Kroc et al.). Heterogeneity in T2D’s pathological and physiological symptoms leads to a variety of complications such as coronary heart disease, retinopathy, nephropathy, etc. [Flannick, et al. 2014]

Due to rare variants having low frequencies and existing in large number, traditional single-marker association tests generally lacked power in these variants. In recent studies, several methods have been developed based on collapsing rare variants in specific regions of interest, i.e. a gene or genes from a specific pathway. This is followed by a region-based test rather than association tests on individual loci. Methods such as Combined Multivariate and Collapsing (CMC) method, Weighted Sum (WS) method, Variable Threshold (VT), etc. and many other variations of these methods have been published.

In this chapter, we proposed the GGI method for rare variants. The proposed method consists of collapsing step and MDR step. In collapsing step, the rare variants were collapsed and recoded in gene-based genotype. Using the gene-based collapsed genotypes, we performed the MDR step. In simulations, MDR with information gain (IG) showed better performance than SPA and MDR with other measures. And, we conducted GGI for rare variants in type 2 diabetes data.

4.2 Methods

4.2.1 Collapsing-based gene-gene interaction

As a gene-level association test for rare variants, we proposed collapsing-based MDR method. In collapsing step, we used three collapsing method, MAF-based collapsing (MDRcol), functional-region-based collapsing (MDRcol_func) and weight-based collapsing (MDRcol_weight).

We used balanced accuracy (BA) and information gain (IG) as the evaluation measure in MDR framework.

$$BA = \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)/2$$

$$IG(Y|X) = H(Y) - H(Y|X)$$

Also, for weight of variants′ interaction, we used four interaction effect models which are extended by Marchini′s three gene-gene interaction effect models (Figure 4.1) (Marchini et al. 2005).

4.2.2 Simulation setting

The statistical efficiency of the proposed method was evaluated through a set of gene-gene interaction simulation settings. We incorporated Marchini’s four interaction models for rare variants: multiplicative, additive, maximum and minimum threshold effects models. The genotypes were generated under HWE with

20 rare SNPs in each genes. As for the phenotypes, 2 cases were considered; GGI with and without marginal effects. Also, to handle the directions of SNP effects, deleterious vs. protective, we analyzed under unidirectional and bidirectional conditions. Lastly, different weighting schemes were also applied to the phenotypes to reflect the characteristics of real data. Here, MAF based and conservation score based weighting schemes have been considered, and various combination of the above parameters have been used for simulation models. The following are the details of each simulation settings:

In order to compare with our method and SPA, we conducted the simulations with the genotype of two genes in 2000 individuals. We generated the genotype of 20 rare SNPs (MAF < 0.01) in each gene under Hardy–Weinberg equilibrium.

The phenotype was generated as followed equation.

$$logit\, P(y=1) = \alpha_0 + \left(\sum_{i=1}^{k\cdot m1} (-1)^i \beta_i X_i + \sum_{j=1}^{k\cdot m2} (-1)^j \beta_j X_j \right) + \sum_{i=1}^{k\cdot m1} \sum_{j=1}^{k\cdot m2} (-1)^i \beta_{ij} X_i X_j$$

which $m1$ and $m2$ are the number of rare SNPs in gene1 and gene2, respectively, k is the proportion of effective rare SNPs which is 0.5 or 0.7, β_i , β_j and β_{ij} are weight for SNP$_i$ of gene1 and SNP$_j$ of gene2.

In this simulation, we used MAF weight and conservation weight as weight of variants.

We generated MAF weight as following equation.

$$\beta_i = \ln \frac{5}{4} \cdot |\log_{10} MAF_i|$$

We generated the conservation weight of a variant using PhastCons score distribution (Figure 4.2) in T2D Korean data as following equation.

$$\beta_i = c|\log_{10}(1 - PhastConsScore_i)|$$

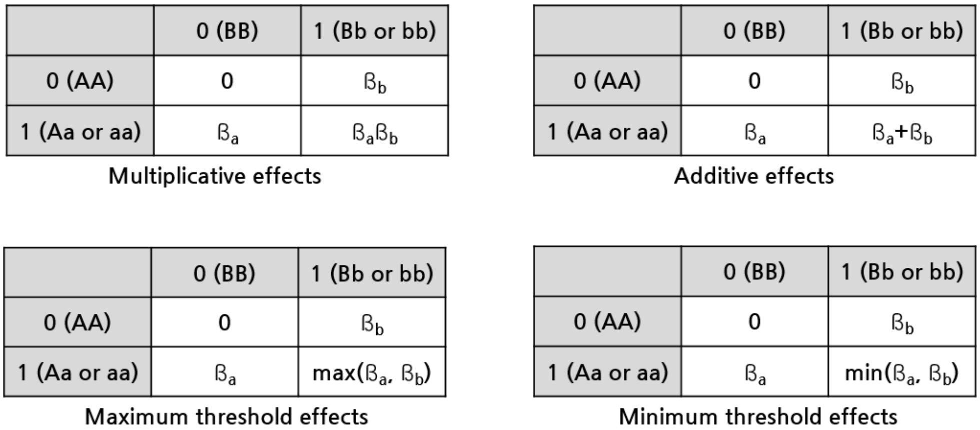
For MAF and conservation weight, we used linear combination with ratio (γ is [0,1]) as followed.

$$\beta_i = \gamma \cdot \left(\ln\frac{5}{4} \cdot |\log_{10} MAF_i|\right) + (1-\gamma)(c|\log_{10}(1 - PhastConsScore_i)|)$$

In this simulations, we generated five different simulation scenarios.

- scenarios 1 - no effect weight, only interaction effect, unidirectional
- scenarios 2 - no effect weight, interaction + marginal effect, unidirectional
- scenarios 3 - no effect weight, only interaction effect, bi-directional
- scenarios 4 - MAF weight, only interaction effect, uni-directional
- scenarios 5 - CONS weight (0.5), only interaction effect, uni-directional

In order to measure type I error, we generated 100,000 repeated dataset which has no effective causal SNP interactions.



	0 (BB)	1 (Bb or bb)
0 (AA)	0	β_b
1 (Aa or aa)	β_a	$\beta_a\beta_b$

Multiplicative effects

	0 (BB)	1 (Bb or bb)
0 (AA)	0	β_b
1 (Aa or aa)	β_a	$\beta_a+\beta_b$

Additive effects

	0 (BB)	1 (Bb or bb)
0 (AA)	0	β_b
1 (Aa or aa)	β_a	$\max(\beta_a, \beta_b)$

Maximum threshold effects

	0 (BB)	1 (Bb or bb)
0 (AA)	0	β_b
1 (Aa or aa)	β_a	$\min(\beta_a, \beta_b)$

Minimum threshold effects

Figure 4.1 Four effect models of variant interaction

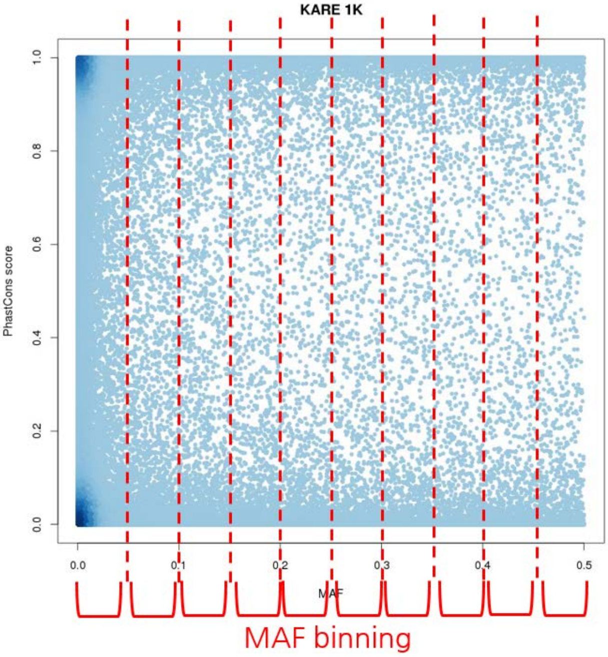


Figure 4.2 PhastCons score distribution as MAF and MAF bin

4.3 Results

4.3.1 Simulation study

In most simulation settings, the power of MDRcol outperformed than SPA method. (Figure 4.3 ~ Figure 4.12)
In table 4. , our methods with IG and BA can control type I error in null dataset which has 100,000 repeated dataset with no effective causal SNP interaction.

4.3.2 Real data analysis of the Type 2 diabetes data

We studied 13,124 individuals from multiple ancestries as part of five whole-exome sequencing studies: the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) study wave 3. Here, we utilized the Korean subjects in the project, which consists of 1,072 individuals and 488,457 autosomal variants. After quality control with HWE, missingness filtering, and MAF<0.01 (rare variants); 414,193 variants remained for the analysis. The rare variants were collapsed under three different schemes: MAF-based, Functional region-based, and Weight-based collapsing. In MAF-based collapsing, the variants inside the genes are collapsed based on their MAFs; 0.01 < MAF < 0.05 and MAF<0.01. For the Functional region-based collapsing, the variants in a gene are collapsed to their annotated functional regions, such as coding region, splice junctions, etc. This reflects the importance of gene structure and region specific variants. And, the weight-based method collapses the variants to their according genes and multiply functional weight information, such as risk scores or conservation scores to each genotypes. The MAF has been used as weights on the hypothesis that "disease-promoting variants should be rare" (Gibson, 2011). The two functional risk weights are from annotated scores of PolyPhen2 and SIFT, however, these scores have poor coverage (only 60 and 81% of human proteome, respectively).

Finally, the conservation scores calculated from PhastCons and phyloP are utilized; this is based on the claims of Ng and Henikoff that “disease-causing mutations are more likely to occur at positions that are conserved throughout evolution”.

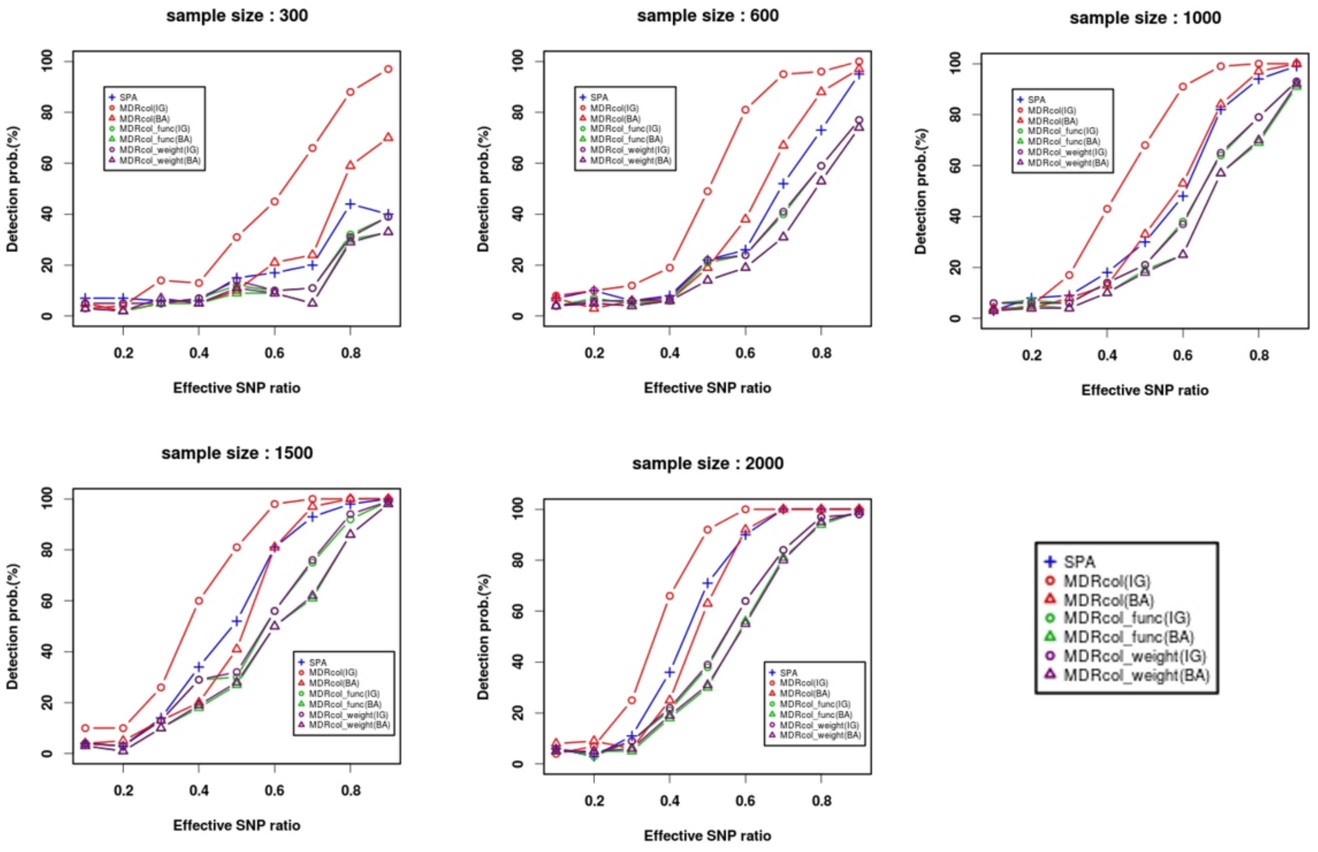


Figure 4.3 Detection probability in simulation 1 with no effect weight and only unidirectional interaction effect

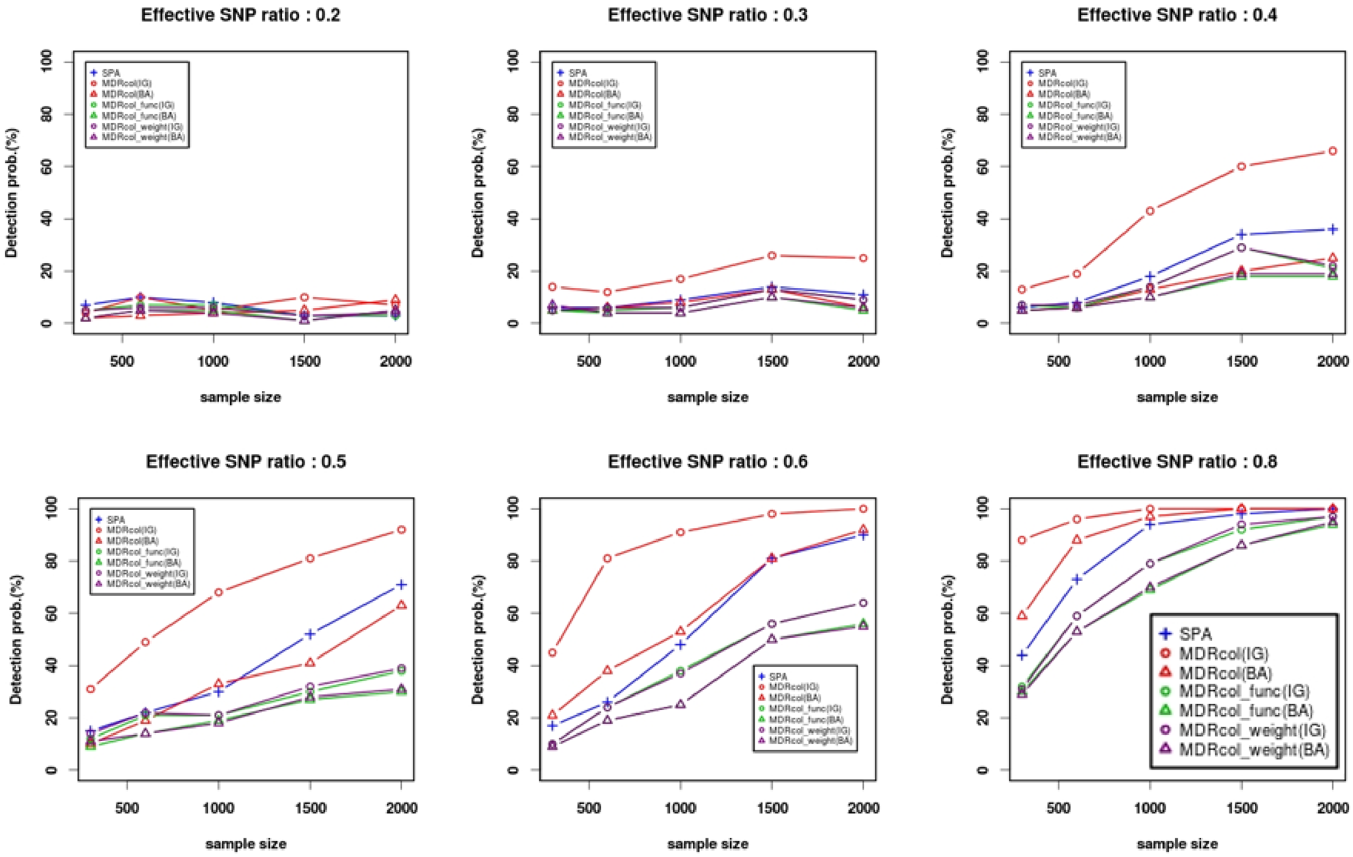


Figure 4.4 Detection probability in simulation 1 with no effect weight and only unidirectional interaction effect

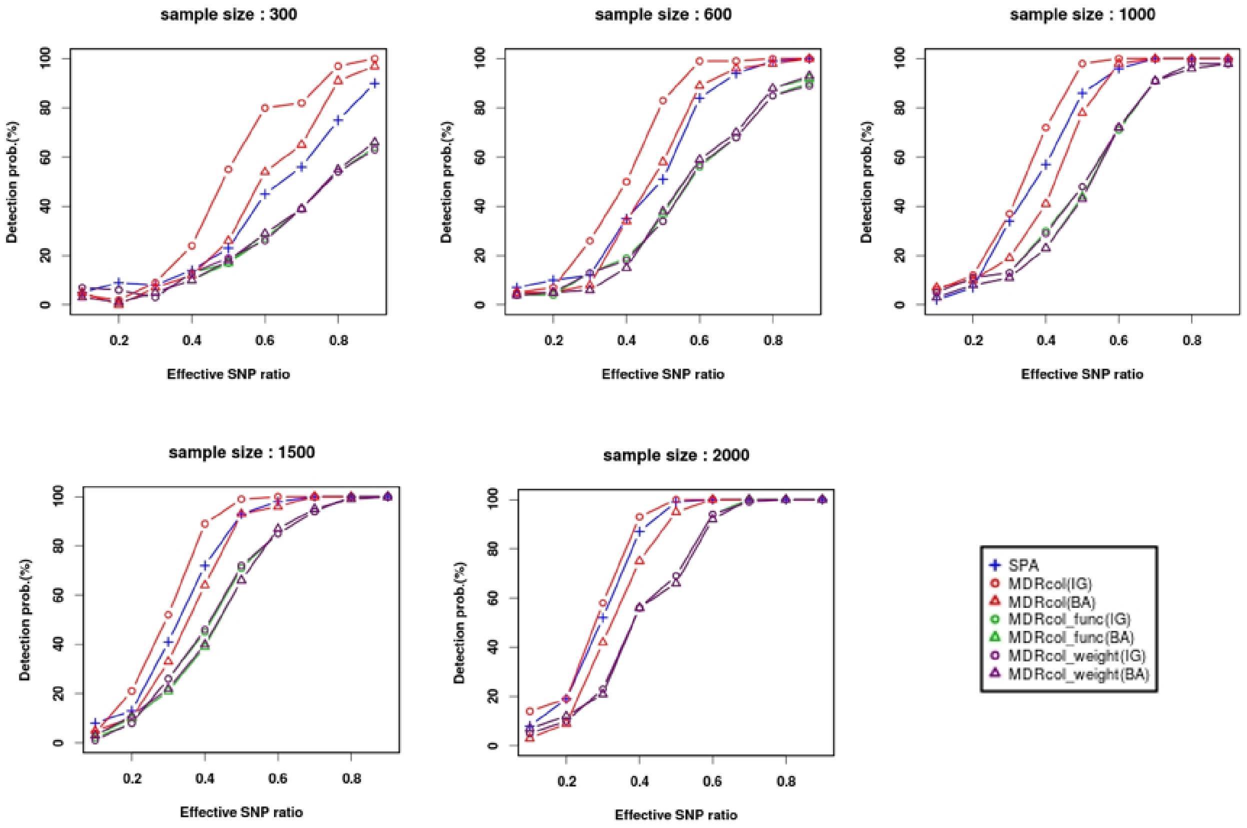


Figure 4.5 Detection probability in simulation 2 with no effect weight and unidirectional interaction effect and marginal effect

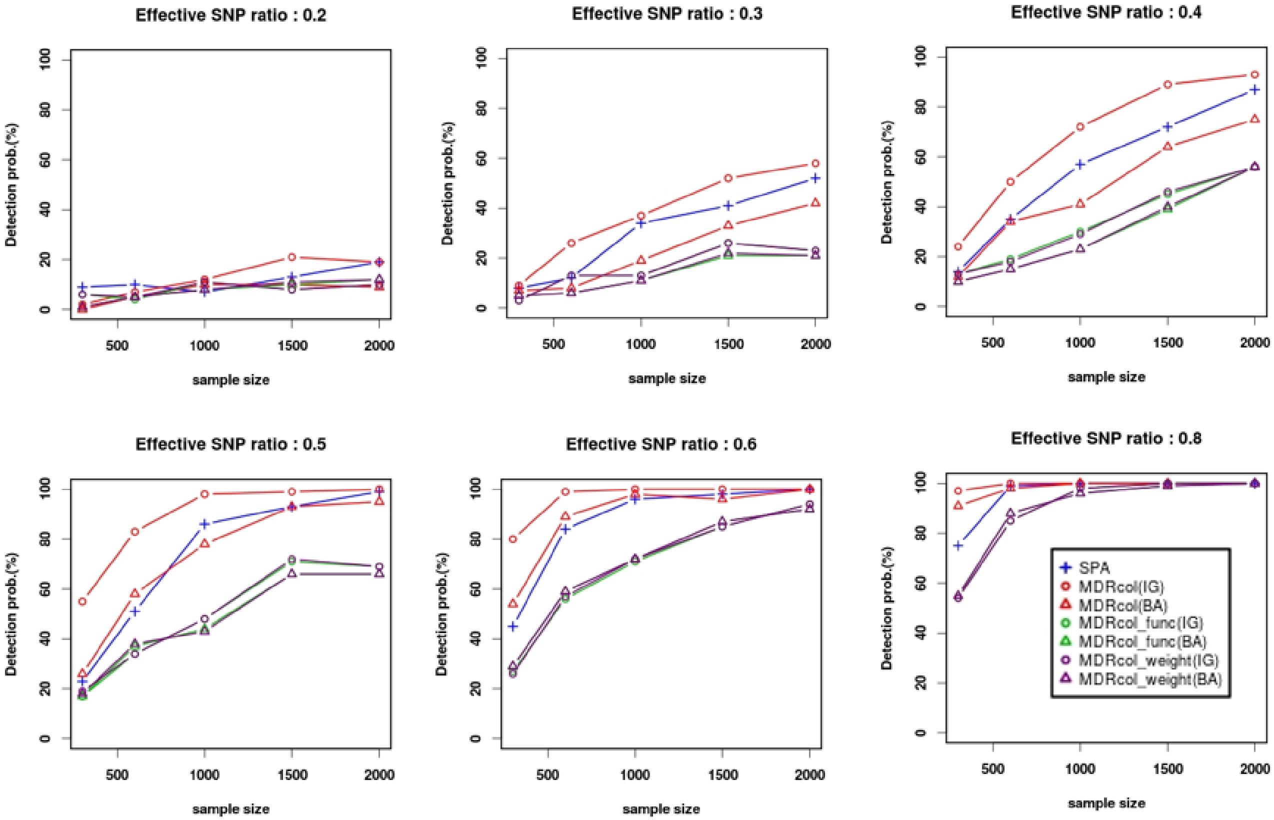


Figure 4.6 Detection probability in simulation 2 with no effect weight and unidirectional interaction effect and marginal effect

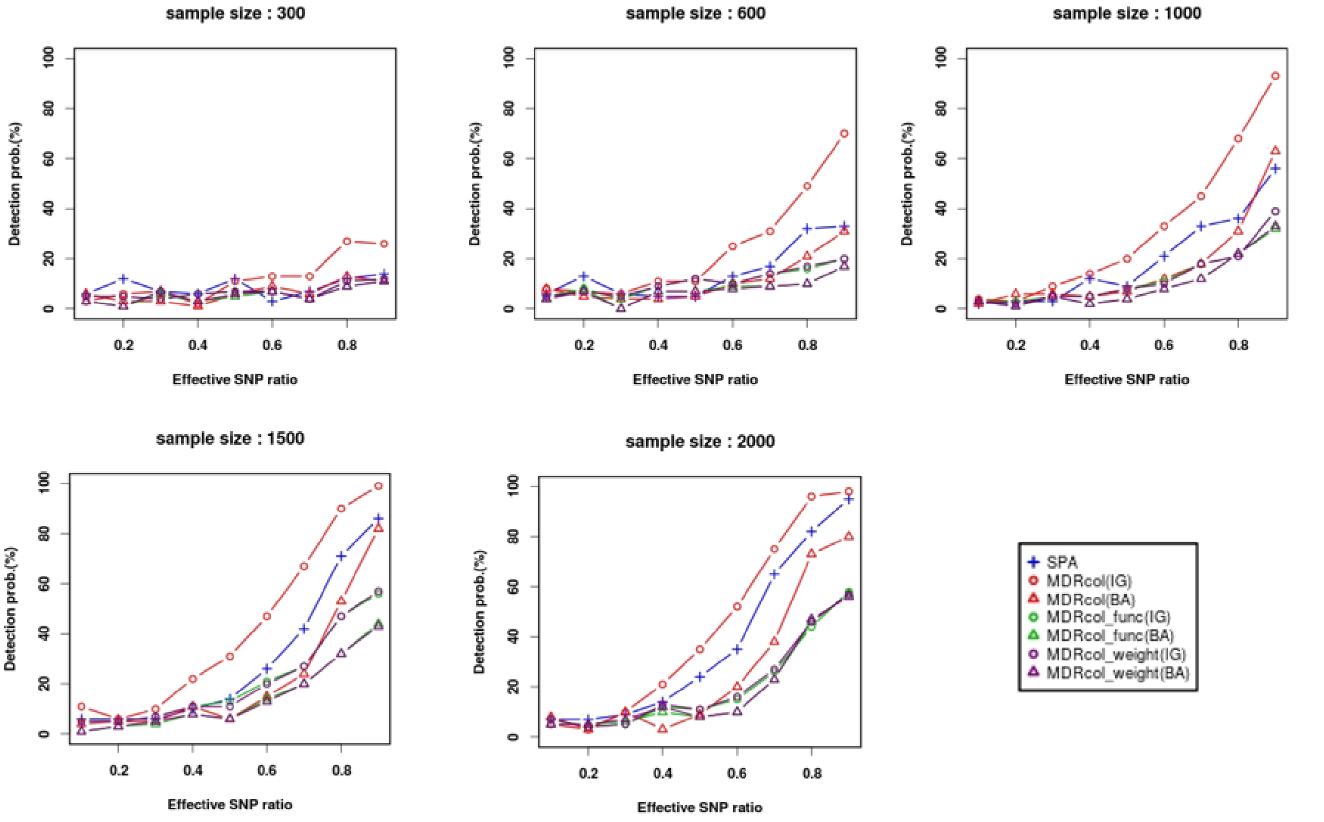


Figure 4.7 Detection probability in simulation 3 with no effect weight and bidirectional interaction effect and marginal effect

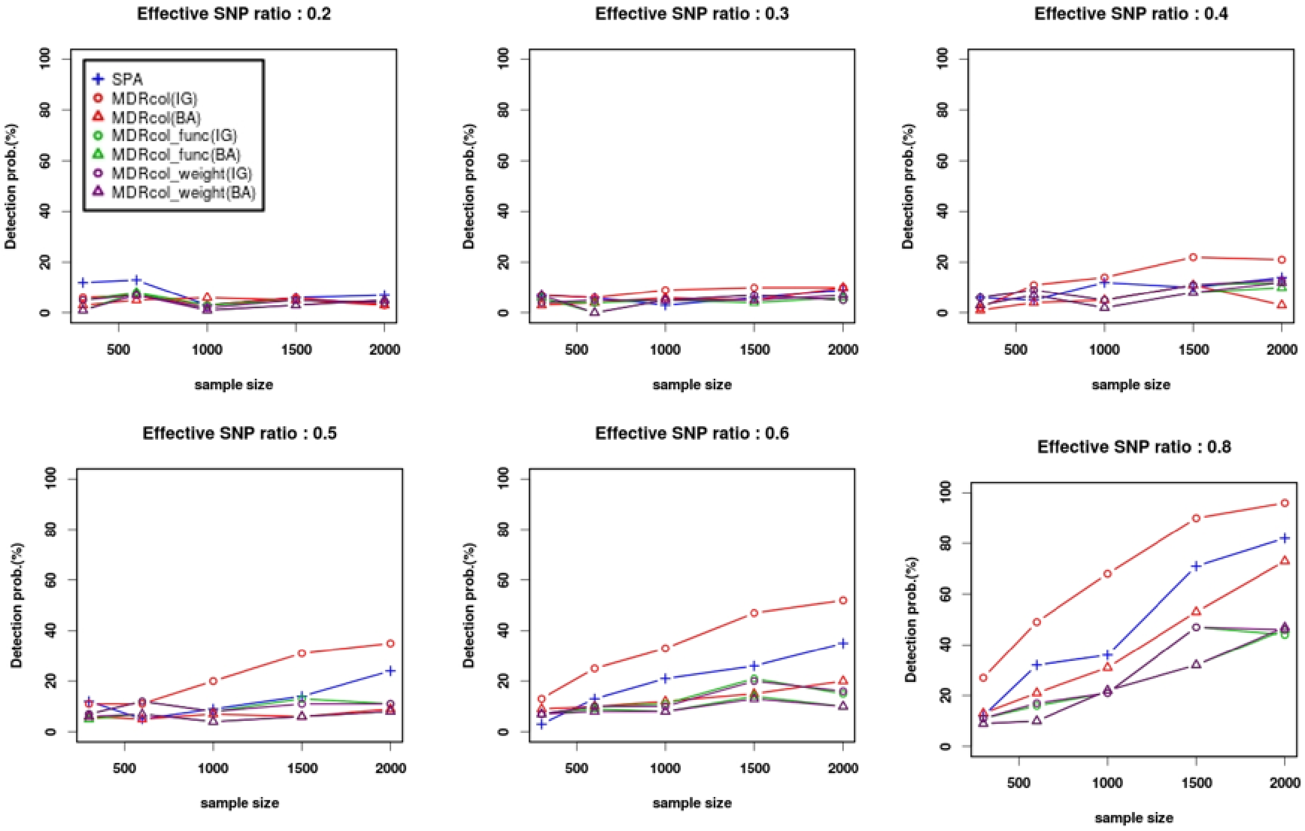


Figure 4.8 Detection probability in simulation 3 with no effect weight and bidirectional interaction effect and marginal effect

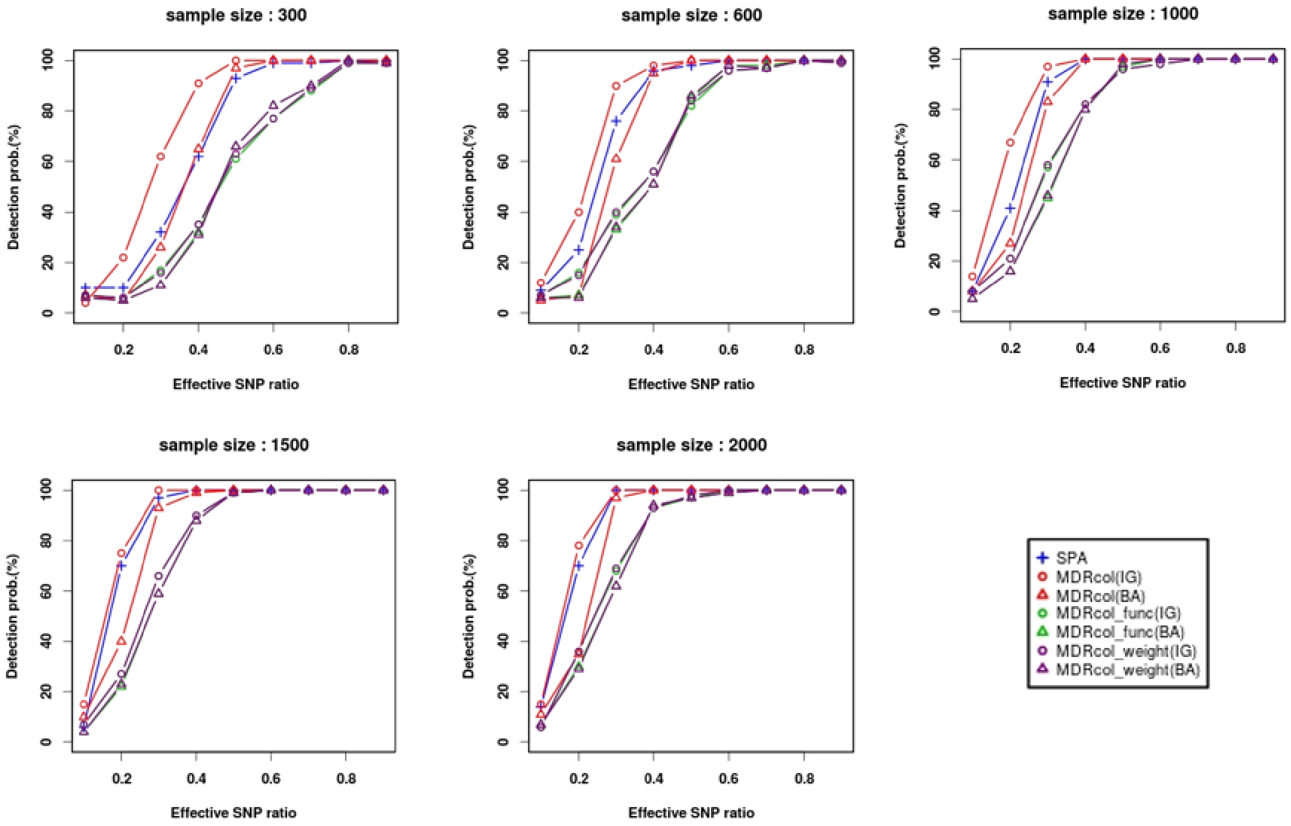
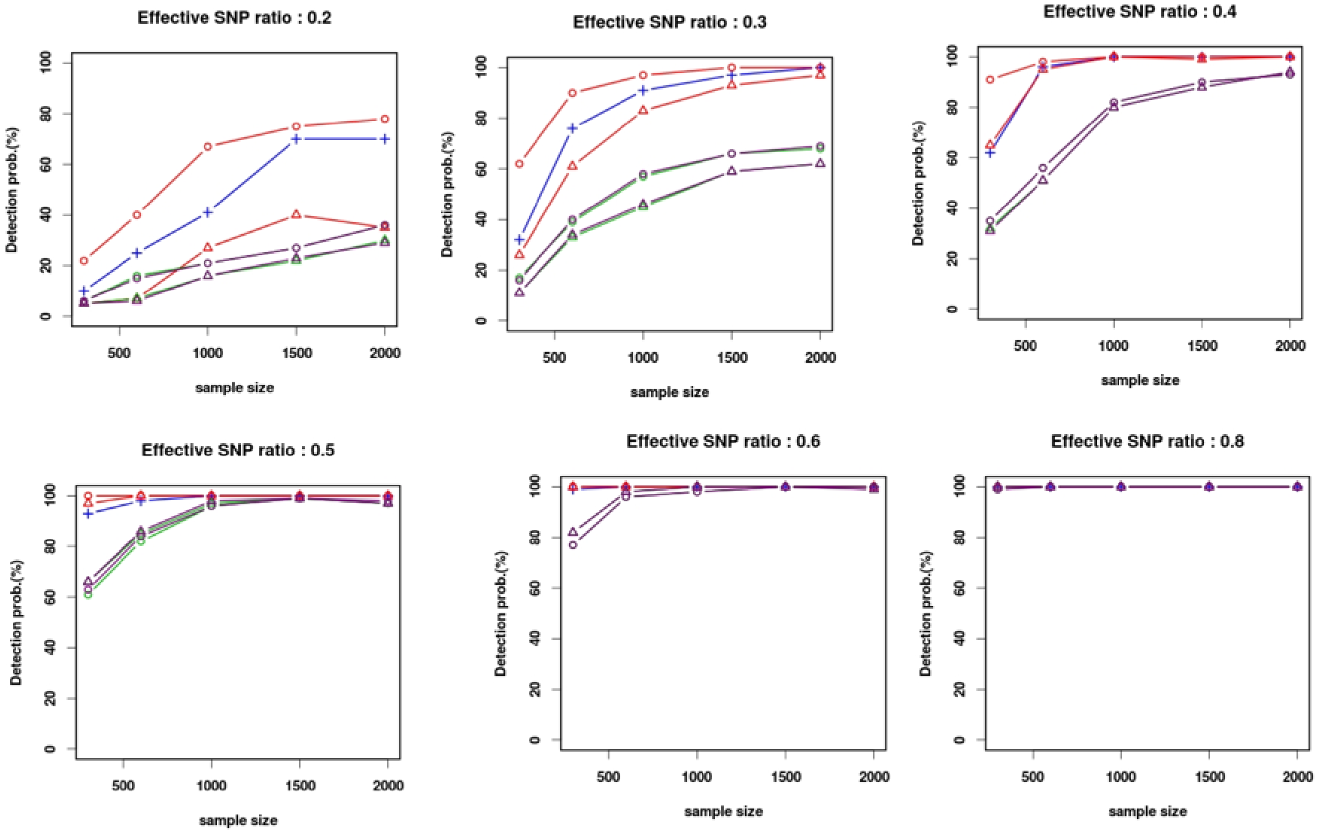


Figure 4.9 Detection probability in simulation 4 with MAF weight and only unidirectional interaction effect



Effective SNP ratio : 0.2

Effective SNP ratio : 0.3

Effective SNP ratio : 0.4

Effective SNP ratio : 0.5

Effective SNP ratio : 0.6

Effective SNP ratio : 0.8

Figure 4.10 Detection probability in simulation 4 with MAF weight and only unidirectional interaction effect

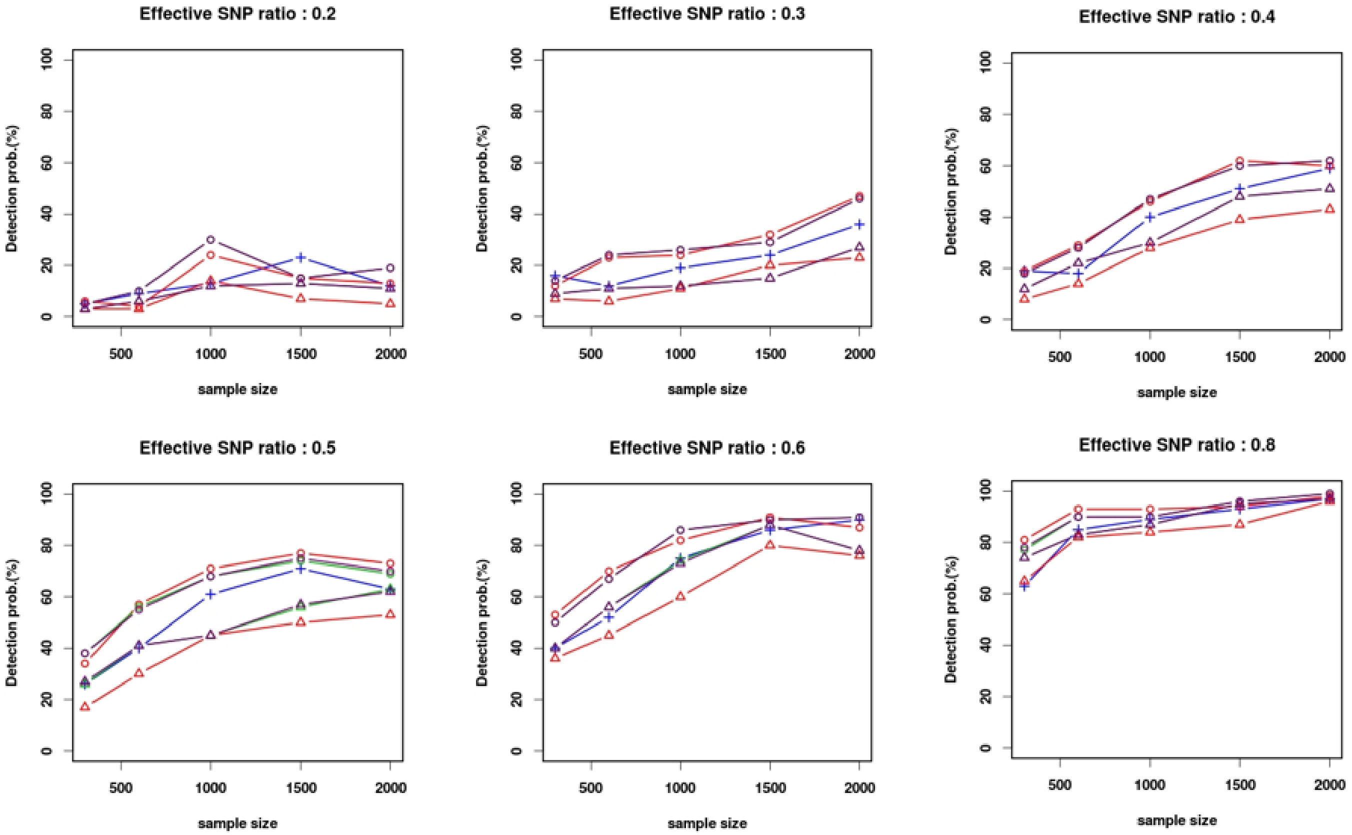


Figure 4.11 Detection probability in simulation 5 with conservation weight and only unidirectional interaction effect

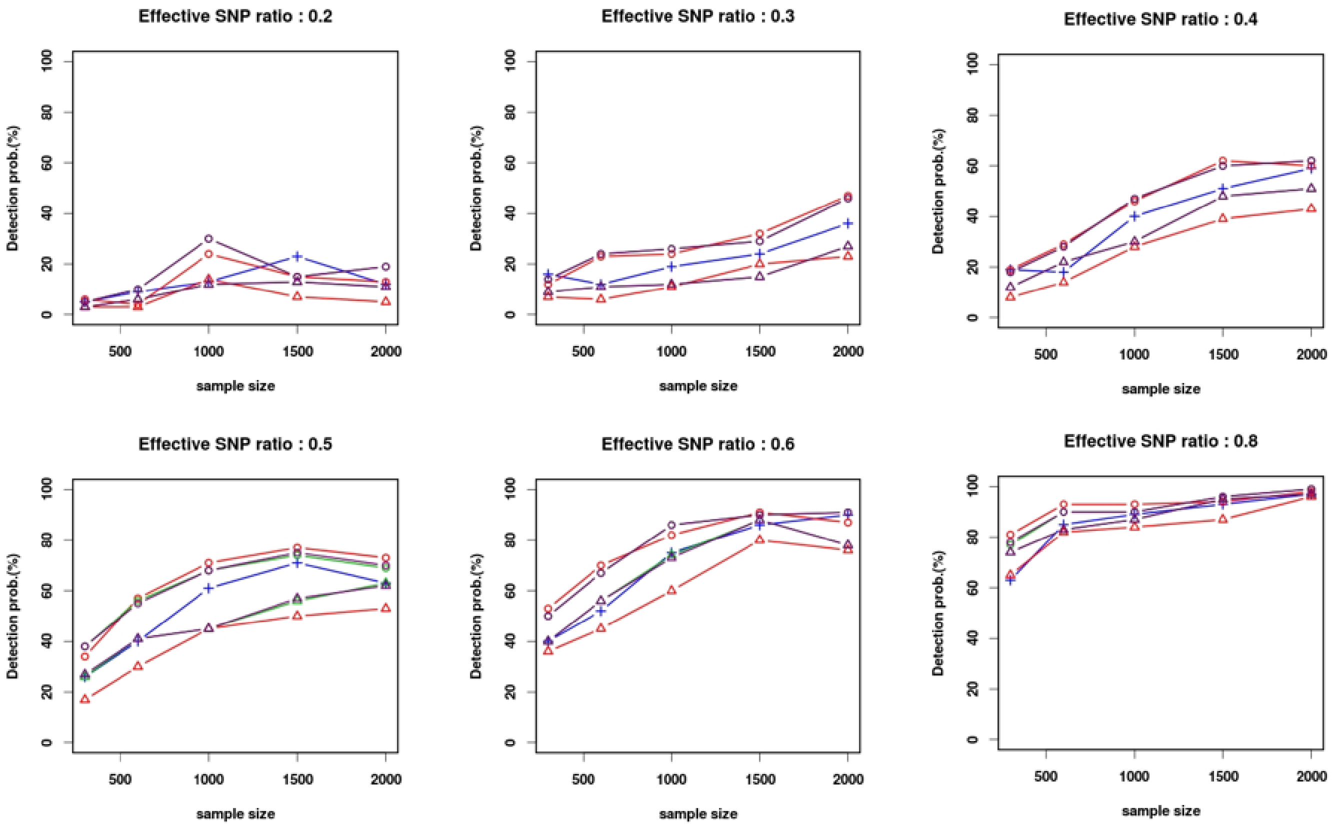


Figure 4.12 Detection probability in simulation 5 with conservation weight and only unidirectional interaction effect

Table 4.1 Type I error

	MDRcol (IG)	MDRcol (BA)	MDRcol Func (IG)	MDRcol Func (BA)	MDRcol Weight (IG)	MDRcol Weight (BA)
Type I error	0.047	0.055	0.045	0.044	0.045	0.044

4.5 Discussion and Conclusion

In this paper, we proposed a collapsing-based gene-gene interaction method which was extended from MDR method. Using five different simulation scenarios, we showed that our method has better performance than SPA method. Using our methods, we applied to T2D dataset with 1072 Koreans.

We used annotation information of all variants predicted by SNPeff, SIFT, Polyphen2 and Condell for functional region-based collapsing.

Functional weight-based collapsing method is expected to remove artifact rare variants. Because the sparcity of rare variant, artifact false variants are called sometimes in the sequencing alignment procedure or variant calling process. To filter out the called false variants, we used functional weight-based collapsing method because the variants tend to have low functional scores. All variants cannot have an effect on their related protein function. Only variants on the specific functional region can affect their related protein function. There are specific genomic regions that affect their protein function. A variant on the specific functional region has an effect on the function of protein which is encoded in the region or regulated by the region. Therefore, the deleterious effects of variants are closely related with the genomic region.

In our simulations, MDRcol (IG) method has better performance than SPA. The IG measure is known to have better performance than BA.

Chapter 5

Computation enhancement for gene-gene interaction

5.1 Introduction

The original MDR software was implemented by JAVA programming language and had graphical user interface [Hahn, et al. 2003]. Because this MDR JAVA version is designed to run on only single workstation, it is suitable for candidate gene studies with a small number (<500) of genetic variables, such as single nucleotide polymorphisms (SNPs). In order to handle large dataset, Bush et al. proposed parallel MDR (pMDR) which employs parallel computing environment and can analyze pair-wise interactions for GWAS [Bush, et al. 2006]. In this approach, the large scaled cluster system is required for interaction analysis of genome-wide data.

Recently, an efficient alternative computing method is proposed by general-purpose scientific computation on graphics processing units (GPGPUs). GPUs are specialized microprocessors developed for accelerating graphic rendering. Their highly parallel structure makes them more effective for a range of complex algorithms than general-purpose CPUs. In fact, the computational power of the most recent GPUs is comparable to that of a cluster with hundreds of CPU cores [Dematte and Prandi 2010]. Using this modern GPUs, MDRGPU implemented MDR method based on pyCUDA library [Sinnott-Armstrong, et al. 2009]. However, MDRGPU has some limitations in direct application to GWAS in that it reports only one single best MDR classifier, while most complex traits are multifactorial.

To address this limitation, we have developed cuGWAM with many special features for GWAS; effective memory handling, top-K report, 3-methods for missing genotype and various performance measures. Furthermore, cuGWAM is much faster than previous MDR software. In our estimation, cuGWAM performs 2~5 times faster than MDRGPU and up to 500 times faster than CPU-based MDR software for one million variants.

5.2 Methods

5.2.1 MDR implementation

The MDR method is accompanied with a classification procedure to classify samples into high and low risk groups based on their genotype combinations of genetic variables. A cross-validation (CV) approach is implemented in MDR to detect the best classifiers via classification performance and predictability. For example, the whole data is partitioned into 10 equal-sized sub-datasets in 10-fold CV. At each CV step, one subset is used as a testing set while the remaining nine subsets form a training set. Then, MDR classifiers are fitted based on the training set for each order of interactions. Their classification and prediction performances can be evaluated with training and test sets, respectively, via an evaluation measure, such as classification accuracy (CA), balanced accuracy (BA), and cross-validation consistency (CVC; Ritchie et al. 2001). Then, the best MDR classifier is selected based on the performance evaluation. A voting algorithm, such as CVC, is used to suggest the single best MDR classifier that is most strongly supported in CV.

Along with CA and BA, we implemented eight additional evaluation measures, including tau-b, likelihood ratio, and normalized mutual information, three of which are known to improve the MDR performance [Namkung, et al. 2009b]. Also, three popular approaches were implemented for handling missing genotypes inside cuGWAM by specifying one of the missing-handling options, such as ‘complete’ (deleting all samples with missing data and using a complete dataset), ‘available’ (using all available samples for each combination under consideration), and ‘missing category’ (treating missing genotypes as a new genotype category) [Namkung, et al.

2009a].

In order to report multiple candidates, each candidate is evaluated respectively with training and testing datasets via user-selected evaluation measure. Then, a pre-defined number (K) of the best candidates with largest evaluation measures with training set are selected.

5.2.2 Implementation using high-performance computation based on GPU

There are two types of GPU based developing platforms, CUDA by NVidia (http://developer.nvidia.com) and CTM by AMD (http://ati.amd.com). In order to design MDR implementation, we selected NVidia graphic card with CUDA programming environment which has been successfully applied in scientific field [Dematte and Prandi 2010; Liu, et al. 2010; Stivala, et al. 2010].

The MDR implementation scheme in cuGWAM, shown in Figure 5.1, has four steps as follows.

Step1. Calculate all combinations. In this step, all possible combinations are constructed by function ʹcombiʹ running on host side.

Step2. Reduce dimensionality. The function which is running on GPU device is called kernel. The kernel ʹcountTableʹ constructs an m-dimensional contingency table where each of all possible multi-locus genotypes of m given SNPs is represented. The binary MDR classifier with two levels of high/low risk groups reduces the m-dimensional space to one dimensional space. After this dimensionality reduction, the contingency table is stored to host memory from shared memory. In this time, all parallel job streams on GPU, which are called threads, are synchronized to prevent misallocation of data.

Step3. Calculate evaluation measures. Each MDR classifier is evaluated, respectively, with training and testing datasets via an evaluation measure (e.g. training BA and test BA). The evaluation measures of each model are estimated in kernel ʻcalcʹ. After this step, all threads were synchronized.

Step4. Store top-K classifiers. All MDR classifiers are ordered by their evaluated measures and a user-specified number (K) of the best MDR classifiers is stored.

Because host device memory has higher latency and lower bandwidth than GPU chip memory, host device memory accesses should be minimized to increase performance of most CUDA applications. But, since GPU chip memory is of limited size, it is important to allocate proper data in proper memory space. There are six memory spaces (e.g. global, constant, shared, local, texture and register memory) to access from chip. cuGWAM was designed to optimize global, constant, and shared memory as follows.

Global memory. The global memory is large (1 GB), but has high latency, low bandwidth and is not cached. In cuGWAM, whole genotype data is stored in this global memory because of its large size.

Constant memory. The constant memory can be only read by kernels. In our implementation, individualʹs phenotype information and address of contingency table from dimensional reduction are stored in this constant memory.

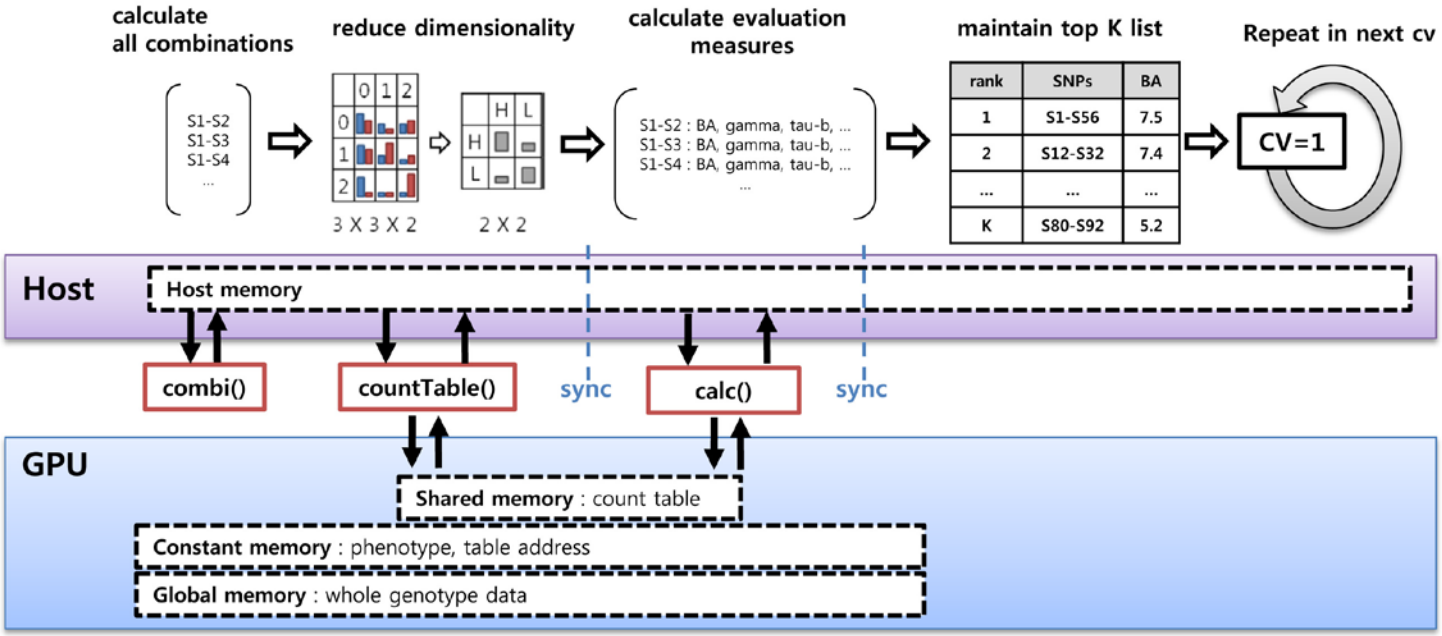

calculate all combinations
S1-S2
S1-S3
S1-S4
...
reduce dimensionality
0
1
2
3 X 3 X 2
H
L
2 X 2
calculate evaluation measures
S1-S2 : BA, gamma, tau-b, ...
S1-S3 : BA, gamma, tau-b, ...
S1-S4 : BA, gamma, tau-b, ...
...
maintain top K list
rank
SNPs
BA
1
S1-S56
7.5
2
S12-S32
7.4
...
K
S80-S92
5.2
Repeat in next cv
CV=1
Host
Host memory
combi()
countTable()
sync
calc()
sync
GPU
Shared memory : count table
Constant memory : phenotype, table address
Global memory : whole genotype data

Figure 5.1. Implementation scheme in cuGWAM

Shared memory. The shared memory space is much faster than the local and global memory spaces, but it has small size (16K bytes). After dimensional reduction step by ʹcountTableʹ kernel, the contingency table data is produced and stored in this shared memory.

Our cuGWAM was developed with focus on memory optimization and effective memory allocation for large scaled genome-wide data.

5.2.3 Environment of performance comparison

We installed MDR (JAVA version) [Ritchie, et al. 2001] and pMDR [Bush, et al. 2006] on 2-GHz Dual Core AMD Opteron(tm) Processor with 8 GB RAM in Linux cluster system with openMPI. MDR (JAVA) was run on single node. For pMDR, we used 100 cores in the cluster system. We tested with different sample sizes (500, 1K, 2K and 5K) and numbers of SNPs (500, 1K, 2K and 5K). In this test, we tested all MDR applications for two-way interaction.

In order to compare the cuGWAM with MDRGPU, we have tested on a workstation, having Intel Core i7 2.66GHz Processor, 12GB RAM and three NVIDIA GeForce GTX285 graphic card in Linux system with various sample sizes (500, 1K, 2K, 5K and 10K) and numbers of SNPs (100, 500, 1K, 2K, 5K and 20K). Two-way interaction was considered for comparing all programs and three-way interactions were considered for comparing GPU based programs.

To be optimized in various hardware systems, cuGWAM provides user-defined number of threads and blocks. In three GTX 285 cards, we set 145 threads and 200 blocks.

5.3 Results

5.3.1 Computational improvement

Exhaustive searching tests were performed to compare our implementation of MDR in CUDA with MDR (JAVA), pMDR and MDRGPU. As published by Sinnott-Armstrong et al. [Sinnott-Armstrong, et al. 2009], this performance result shows that the both GPU solutions, cuGWAM and MDRGPU, are faster than standard CPUs (Table 1 and 2). The execution time has linear relationship with the combinations of markers as shown Table 5.1. Especially, cuGWAM has faster ~650 times than MDR (JAVA version) in test data with 2000 markers and 5000 samples. Also two GPU applications (cuGWAM and MDRGPU) perform better than cluster application (pMDR) on 100 cores cluster. Because this test shows that execution time has linear relationship with sample size and combinations of markers, we estimate that cuGWAM has equivalent performance with cluster system with about 200 cores.

For three-way interaction based on single GPU devices, cuGWAM showed better performance than MDRGPU in both different marker size (100, 500 and 1K) and different sample size (500, 1K, 2K, 5K and 10K), as shown in Figure 5.2 and 5.3. The performance was improved by ~2.9 fold when all three-way interactions of 100 markers were executed with a sample size of 10K. Also, since the linear increment in execution time was observed as sample size or marker size increases, the more samples and markers were examined, the larger benefits of performance was gained by cuGWAM.

Table 5.1 Execution time (sec) in marker size 500, 1K, 2K and 5K for sample size of 2K

system	methods	marker size			
		500	1K	2K	5K
CPU-based	MDR (JAVA)	195	1054	4997	17540
Cluster [a]	pMDR	6	10	48	289
1GPU	MDRGPU	9	16	35	115
	cuGWAM	2	3	10	55

a. 100 cores were used in this performance testing

Table 5.2 Execution time (sec) in sample size 500, 1K, 2K and 5K for marker size of 2K

system	methods	sample size			
		500	1K	2K	5K
CPU-based	MDR (JAVA)	910	1725	4997	11090
Cluster [a]	pMDR	10	23	48	126
1GPU	MDRGPU	10	19	35	86
	cuGWAM	6	7	10	17

a. 100 cores were used in this performance testing

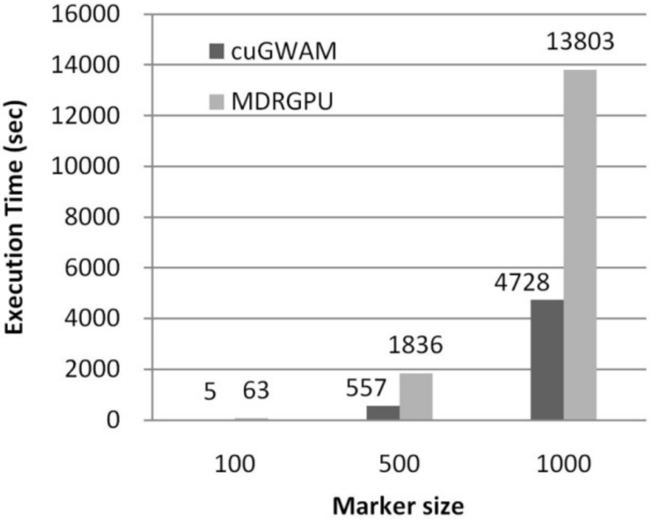


Figure 5.2 Performance comparison with sample size 10K on single GPU for three-way interaction

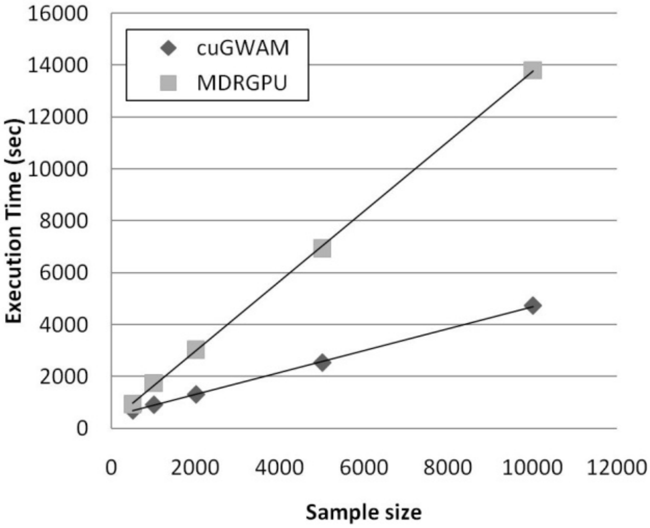


Figure 5.3 Performance comparison with marker size 1K on single GPU for three-way interaction

We have measured the performance by running the applications on single, two and three GPU configurations. In both cuGWAM and MDRGPU, we observed a nearly linear performance improvement as adding more graphic boards to the system. But cuGWAM is slightly better performance than MDRGPU (Figure 5.4).

For high order interactions, more memory space is needed after calculation of count table. In two-way interaction, one count table uses 3×3×2 bytes = 18 bytes but, in three-way interaction, single count table uses 3×3×3×2 bytes = 54 bytes. Since graphic card has limited shared memory space (16K bytes), the number of threads loaded on shared memory at once is 888 (16K bytes / 18 bytes) in two-way interaction, but 296 (16K bytes/ 54 bytes) in three-way interaction. As this calculation, the performance will be decreased 1/3 times from two-way to three-way interaction. We observed the reduced performance from two-way interaction to three-way interaction in both cuGWAM and MDRGPU (Figure 5.5). The performance losses of cuGWAM and MDRGPU are about 45.2% and 63.4% respectively. Note that cuGWAM has lower decrease rate of performance than MDRGPU from two-way to three-way interaction.

To evaluate in real data set with 327,632 SNPs and 6,417 samples, cuGWAM completed 5.4 × 1010 classifiers (e.g., pairwise interactions between 327,632 SNPs) in ~4.1 days on 3 GPUs. We expect that cuGWAM can evaluate 5 × 1011 classifiers (e.g., pairwise interactions between one million SNPs) with 10k samples and 10 cross validation in ~17.7 days on 3 GPUs while MDRGPU completed the same evaluation in ~69.3 days.

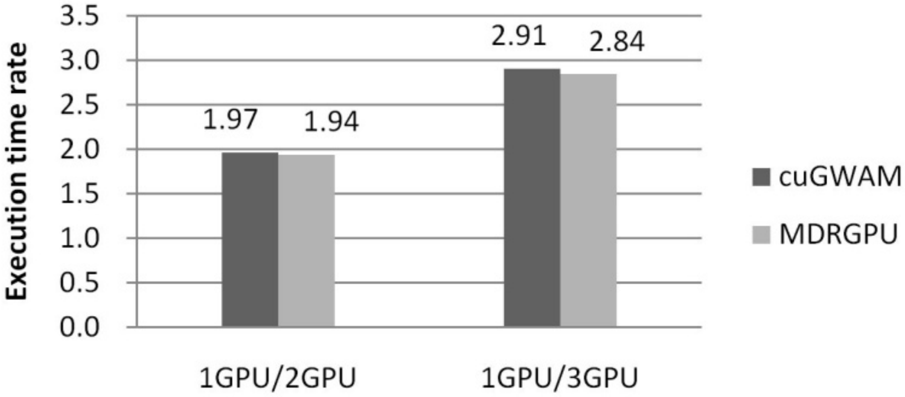


Figure 5.4 Performance increment as adding graphic boards (sample size 10K, marker size 1K)

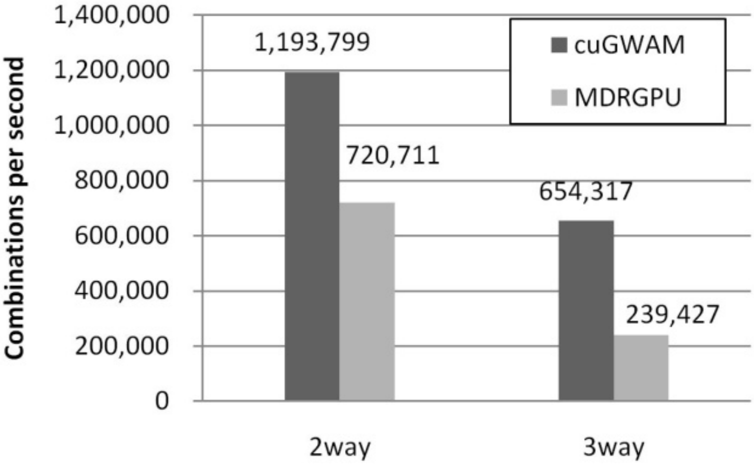


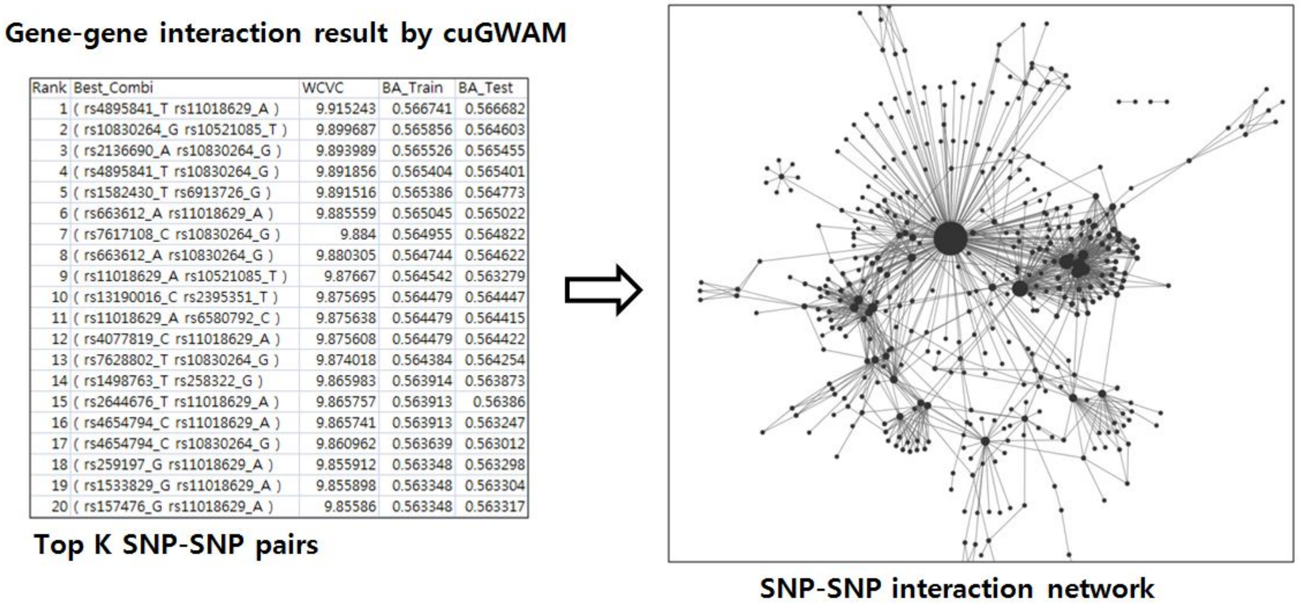
Figure 5.5 Performance loss in high-order interaction (sample size 5K, marker size 20K on single GPU)

Table 5.3 Computational and construction cost for 1M SNPs and 2K individuals

system	Construction cost ($)	MDR applications	Expected execution time (day)	Computational cost ($)
1 node	1000	MDR (JAVA)	8122	7417
Cluster	50000	pMDR	134	6111
1GPU	2300	MDRGPU	53	112
		cuGWAM	25	53
2GPU	2900	MDRGPU	27	71
		cuGWAM	13	34
3GPU	3500	MDRGPU	18	57
		cuGWAM	8	27

The expected execution time for 1M SNPs and 2K individuals derive from the execution time for 5K markers and 2K samples in table 1. We assume that life span of every system is three years. The computation cost is calculated as (computation cost) = (execution time) × (construction cost) / (3 years × 365 days). We do not consider any management cost.

In Table 5.3, we compared the computational cost for 1million SNPs and 2000 individuals. If we assumed that life span of every system is three years, we calculated computational cost as depreciation cost. The three GPU (GTX-285) system costs approximately $3500, but cluster system with 100 cores costs approximately $50000. If we calculate two-way interaction for 1 million SNPs and 2000 individuals, computational cost is $7417 for MDR(JAVA), $6111 for pMDR, $57 for MDRGPU (3GPU) and $27 for cuGWAM (3GPU). In this test, cuGWAM is more cost-effective 274 times than MDR(JAVA), 226 times than pMDR (100 cores cluster) and 2 times than GPUMDR.


Gene-gene interaction result by cuGWAM
Top K SNP-SNP pairs
SNP-SNP interaction network

Rank	Best_Combi	WCVC	BA_Train	BA_Test
1	(rs4895841_T rs11018629_A)	9.915243	0.566741	0.566682
2	(rs10830264_G rs10521085_T)	9.899687	0.565856	0.564603
3	(rs2136690_A rs10830264_G)	9.893989	0.565526	0.565455
4	(rs4895841_T rs10830264_G)	9.891856	0.565404	0.565401
5	(rs1582430_T rs6913726_G)	9.891516	0.565386	0.564773
6	(rs663612_A rs11018629_A)	9.885559	0.565045	0.565022
7	(rs7617108_C rs10830264_G)	9.884	0.564955	0.564822
8	(rs663612_A rs10830264_G)	9.880305	0.564744	0.564622
9	(rs11018629_A rs10521085_T)	9.87667	0.564542	0.563279
10	(rs13190016_C rs2395351_T)	9.875695	0.564479	0.564447
11	(rs11018629_A rs6580792_C)	9.875638	0.564479	0.564415
12	(rs4077819_C rs11018629_A)	9.875608	0.564479	0.564422
13	(rs7628802_T rs10830264_G)	9.874018	0.564384	0.564254
14	(rs1498763_T rs258322_G)	9.865983	0.563914	0.563873
15	(rs2644676_T rs11018629_A)	9.865757	0.563913	0.56386
16	(rs4654794_C rs11018629_A)	9.865741	0.563913	0.563247
17	(rs4654794_C rs10830264_G)	9.860962	0.563639	0.563012
18	(rs259197_G rs11018629_A)	9.855912	0.563348	0.563298
19	(rs1533829_G rs11018629_A)	9.855898	0.563348	0.563304
20	(rs157476_G rs11018629_A)	9.85586	0.563348	0.563317

Figure 5.6. SNP-SNP interaction network using two-way interaction results by cuGWAM.

5.4 Discussion

In this chapter, we presented the computing capability of CUDA-enabled GPUs for accelerating MDR algorithm. Our cuGWAM has various features that distinguish from existing MDR application for GWAS as follows. First, it implements an effective memory handling algorithm and efficient procedures for MDR to make joint analysis of multiple genes feasible for GWAS. Second, it can report multiple candidates for causal gene-gene interactions. Third, various performance measures, including tau-b, likelihood ratio, and normalized mutual information, were implemented to evaluate MDR classifiers [Namkung, et al. 2009b]. Finally, it implements three methods for handling missing genotypes: complete, available and missing category [Namkung, et al. 2009a].

This features of our method can lead to vast speed-up and, enable gene-gene interaction analysis on millions variant data that presently cannot be performed due to computing time limitations in traditional computing systems. We obtained ~500 fold performance improvement over original MDR and equivalent performance of about 200 cores cluster system in one GPU system. Also, by porting the MDR onto high-performance graphic cards using the CUDA environment, we obtained up to ~5 fold acceleration compared with MDRGPU. Especially, cuGWAM has better performance than MDRGPU in every types of testing (sample size, marker size and high-order interaction). Even though a Python-based GPU compute code was expected to show the same full performance of GPU hardware as a C-controlled GPU compute code (http://mathema.tician.de/software/pycuda/), our test results showed that there was different performance between Python-base MDRGPU and C-base cuGWAM. We expect that this difference was derived from optimal utilization of hardware memory. Because MDR is exhaustive algorithm with exploring all combinations of variants, memory optimization is critical in performance of software. Especially in GPU system, the optimization of shared memory determines the performance of the application. Because cuGWAM uses binarized count values in shared memory to save memory space, it can run more threads with restricted shared memory space and input/ouput (IO) transaction load between threads and memory is minimized.

When higher order interaction is searched exhaustively, MDR algorithm maintains higher dimensional count table, which occupies more memory space and restricts the number of threads running on GPU device in parallel. The decrease of the number of threads means the decrease of performance. The performance loss in both cuGWAM and MDRGPU was observed in our results. But, cuGWAM showed less performance loss than MDRGPU from two-way to three-way interactions. This better performance of cuGWAM is due to loop unrolling technique which reduces loop procedure to increase a program’s speed (Nvidia, CUDA programming guide, http://developer.nvidia.com).

Our optimized GPU-based MDR application produced reasonably stable performance even in large sample size (e.g. 10K) and large variants size (e.g. 320K). Unlike other MDR applications, cuGWAM converts text-typed data file to binarized and compressed data format which facilitates load and process large data set stably.

One of the distinguishing features of cuGWAM is to report multiple candidates for causal gene-gene interactions. It is inapplicable to report one single best candidate when causal gene-gene interactions are searched for complex phenotypes in a genome-wide scale. User cannot know only the best interaction but also possible strong candidate interactions via cuGWAM. Also, using multiple candidate pairs, SNP-SNP interaction network can be generated with node (SNP) and edge (interaction). Figure 5.6 shows SNP-SNP interaction network with top 1000 interaction pairs reported from two-way interaction analysis by cuGWAM. In this network, since node size means number of interactions, we can easily identify hub SNPs which interact with many SNPs and deserve to be inspected more. This interaction network is potentially useful for the biological interpretations.

5.5 Conclusion

cuGWAM is high-performance software for gene-gene interaction analysis of large genome-wide data. It is C++ parallel implementations of MDR method using CUDA runtime application programming interface. With reduced data transaction by binarization, efficient usage of the global memory bandwidth and optimized loop transformation technique by loop unrolling, cuGWAM showed best performance both in our simulation data and 320K real dataset. Since our results on GPU show that it is possible to detect gene-gene interaction in genome-wide scale with one million variants, our optimized GPU implementation is especially encouraging. Furthermore, we have investigated optimized GPU-based MDR implementation which reports a list of candidate causal gene-gene interactions and various performance measures to evaluate MDR classifiers, including tau-b, likelihood ratio, normalized mutual information as well as balanced accuracy. Executable cuGWAM are freely available at http://bibs.snu.ac.kr/cugwam from system with CUDA-enabled GPU devices.

Chapter 6

Visualization for gene-gene interaction interpretation

6.1 Introduction

Many methods have been proposed to analyze gene-gene interactions, including logistic regression, logic regression, recursive partitioning, multifactor dimensionality reduction, ReleifF, and Bayesian model selection [Cordell 2009]. However, interpretation of identified gene-gene interactions is not straightforward.

Visualization or graphical representation can be a powerful tool for biological characterization of interactions because it provides an effective way to recognize multi-locus genotype combinations that enhance/repress a trait and to display the polygenic structure of interactions. In addition, visualization can aid prior exploration of interaction patterns among genes of interest. Unfortunately, there is no software specifically designed for visualizing gene-gene interactions. In order to aid the interpretation of such gene-gene interactions, we developed the VizEpis, for use in genetic association analysis.

When gene-gene interactions are analyzed for a binary trait, such as disease susceptibility, biological interpretation of identified interactions is often based on odds ratios (ORs) between case and control groups for each multi-locus genotype combination. Therefore, visualization of genotype-wide ORs can aid in exploring and/or characterizing (especially high-order) interactions among multiple genes. The VizEpis can effectively visualize gene-gene interactions based on genotype-wide ORs as well as raw data.

In high-dimensional multivariate data, visualization is essential to pattern recognition and characterization. With certain modification and/or extension, some of those tools can be employed to graphical representation of high-order genetic interactions. Examples include techniques for displaying multi-way contingency tables (e.g., mosaic maps), for plotting effects and their significance in meta-analysis (e.g., forest and funnel plots), and for an effective representation of raw high-dimensional multivariate data (e.g., parallel coordinate plots). In the VizEpis, various graphical tools used for high-dimensional multivariate data were tailored for gene-gene interaction analyses. Also, 3D lattice plot was developed especially to investigate 3-way interactions. Based on case studies, we illustrated the usage of the VizEpis and demonstrated its benefits.

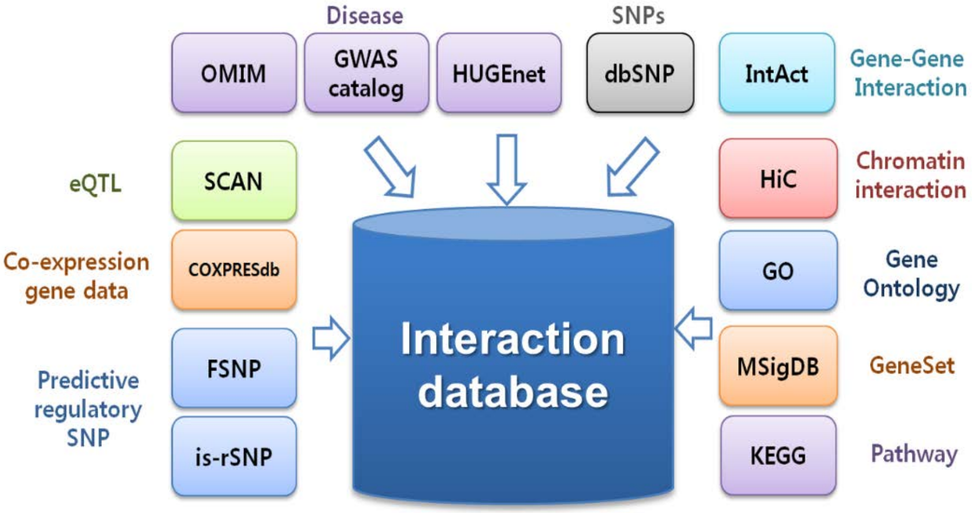


Figure 6.1 Interaction data sources of VizEpis for biological interpretation of statistical interaction

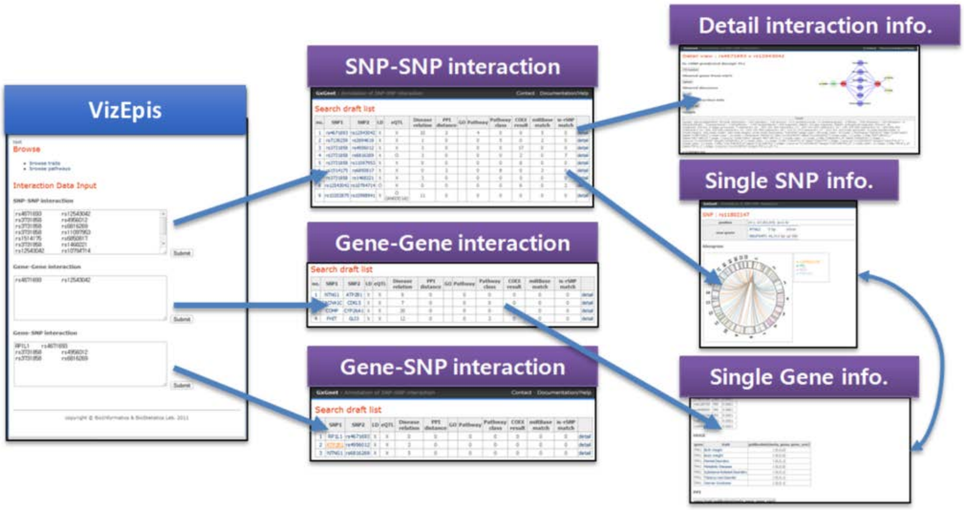


Figure 6.2 VizEpis working flow

6.2 Methods

In the VizEpis, we implemented six plots to visualize patterns in gene-gene interactions: checkerboard (CB), pairwise CB, forest, funnel, ring chart. These plots display the summarized information, including ORs for each multi-locus genotype of multiple variants, such as single nucleotide polymorphisms (SNPs). Note that red and blue are used consistently to indicate multi-locus genotypes with high and low risks or to specify cases and controls. The gradation in colors illustrates the magnitude of ORs.

6.2.1 Interaction mapping procedure

VizEpis is web-based system to provide integrated biological relation data which are provided by 13 resources (Figure 6.1). These resources are for disease relations (OMIM, GWAS catalog and HUGEnet), eQTL relations (SCAN), Co-expression relations (COXPRESdb), regulatory relations (is-rSNP), protein-protein interaction (IntAct), chromatin interaction (HiC), gene ontology (GO), MSigDB (GeneSet) and pathway (KEGG). VizEpis searches the biological relation in collected relation data for user′s inputted SNP list (Figure 6.2).

6.2.2 Checker board plot

A heat map was designed to display the quantitative values of one variable as colors on a 2D map of other variables, and its variants have been popular for visualization in many areas. For example, in a mosaic plot, a tiled heat map has been employed for graphical representation of contingency tables and hence for visual investigation of multi-way interactions in categorical data analysis [Friendly 1994]. However, the mosaic plot may be not practical for visual investigation of more than three variables because it displays multivariate data via multiple heat maps that are hierarchically organized.

We developed a new variant of the mosaic plot, called a CB plot, in which high-order gene-gene interactions can be investigated on a single 2D heat map. In CB plots, all possible multi-locus genotypes are represented in two dimensions, and their corresponding ORs are encoded as colors. For instance, when an m-way interaction among m SNPs is under investigation, users can specify row and column dimensions with k and (m-k) SNPs, respectively to view the m-way interaction on a 3k×3(m-k) CB plot, where k is user-defined number of rows on CB plot. Optionally, hierarchical clustering can be done over each dimension to group genotypes having similar OR patterns. This feature provides users with further help in capturing underlying patterns within gene-gene interactions. We also implemented a pairwise CB plot that consists of 3×3 CB plots for all two-way interactions of m SNPs. The pairwise CB plot provides a quick scan of all possible two-way interactions in one shot (Figure 6.3).

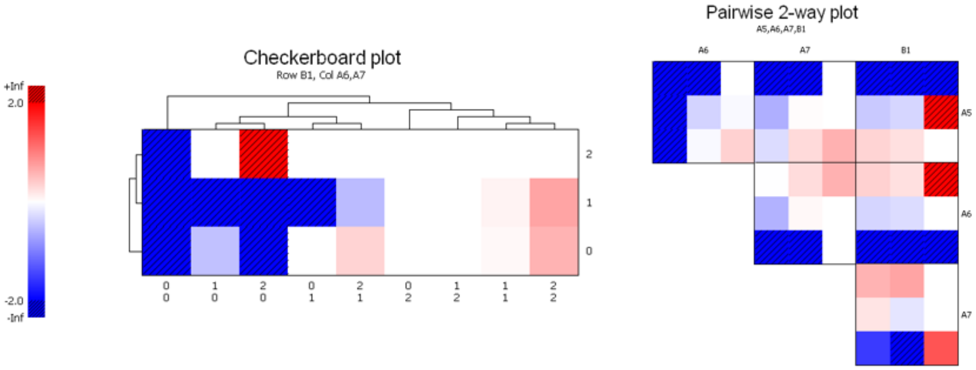


Figure 6.3 Checkerboard plot in VizEpis

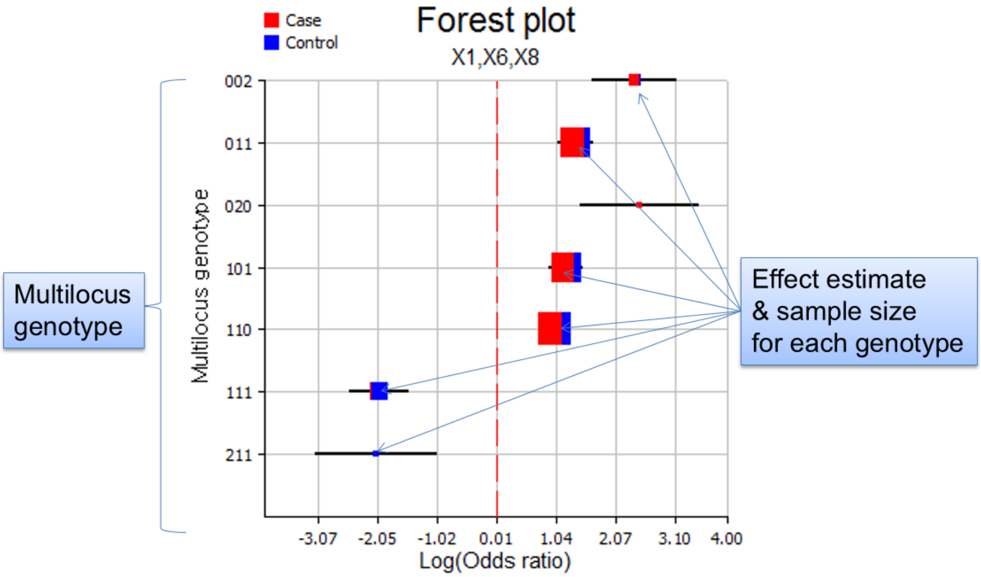


Figure 6.4 Forest plot in VizEpis

6.2.2 Forest and funnel plot

The forest and funnel plots were originally developed to display a meta-analysis of multiple studies in medical research [Lewis and Clarke 2001; Sterne and Egger 2001]. For example, in a meta-analysis of epidemiological studies, the result is summarized with an OR and its precision (e.g., inverse standard error) for each study. We adopted these plots to represent ORs with their significance for multi-locus genotypes. In the forest plot, the 95% confidence interval of the OR for each multi-locus genotype is represented by a horizontal line (Figure 6.4). The graph is plotted on a natural logarithmic scale so that the confidence intervals are symmetrical about the observed ORs, which are indicated with squares. The funnel plot is a scatter plot of the OR against its precision (Figure 6.5 and Figure 6.6). In both plots, the area of each square is proportional to the number of samples having each multi-locus genotype. Whereas the forest plot is useful in investigating ORs for each multi-locus genotype, the funnel plot is apt for detecting multi-locus genotypes with a specific feature (e.g., a large OR with high precision). In these two plots, each multi-locus genotype is displayed as a box, whose size corresponds to the sample size. The red box is for number of cases and, the blue box is for number of controls.

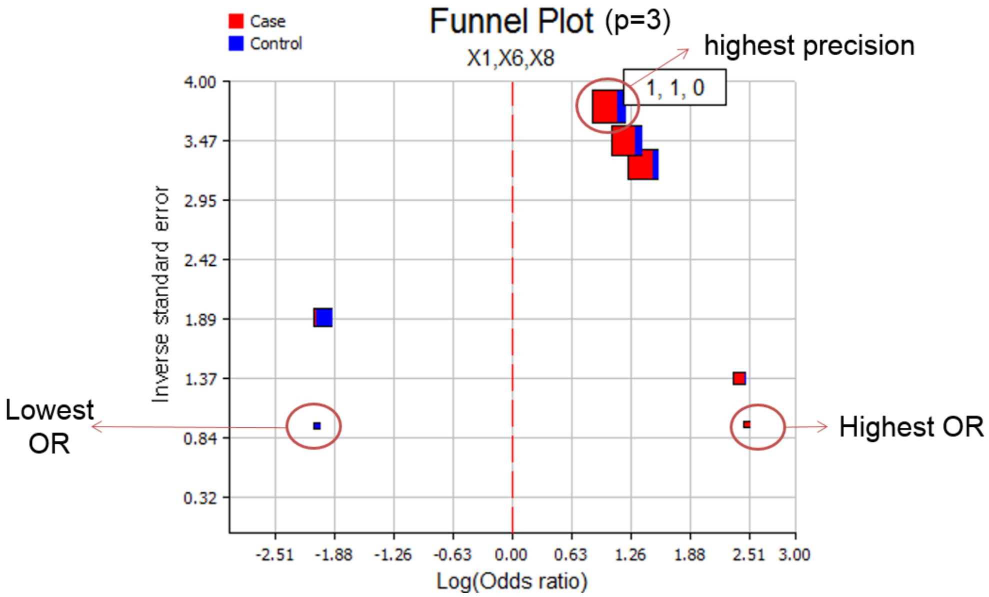


Figure 6.5 Funnel plot in VizEpis

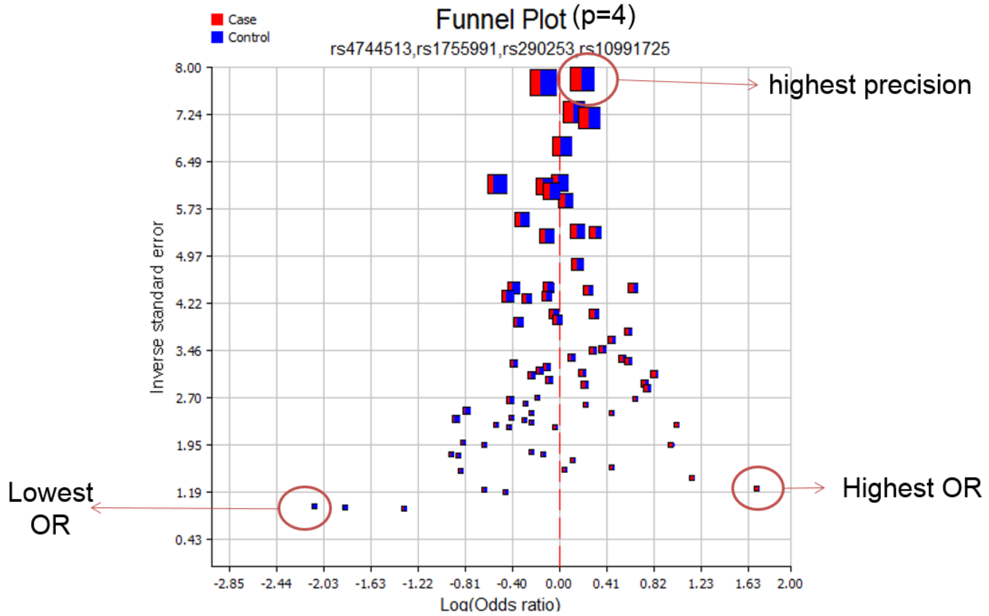


Figure 6.6 Funnel plot in VizEpis

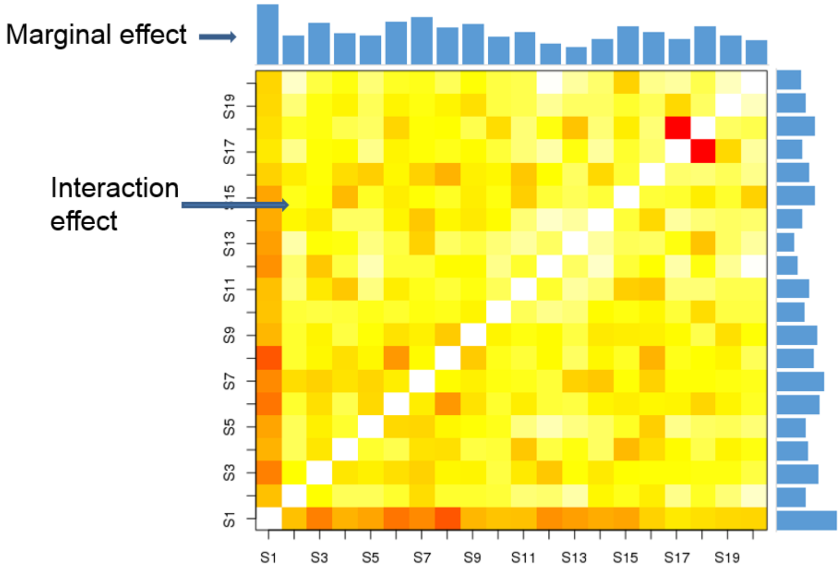


Figure 6.7 Interaction heapmap in VizEpis

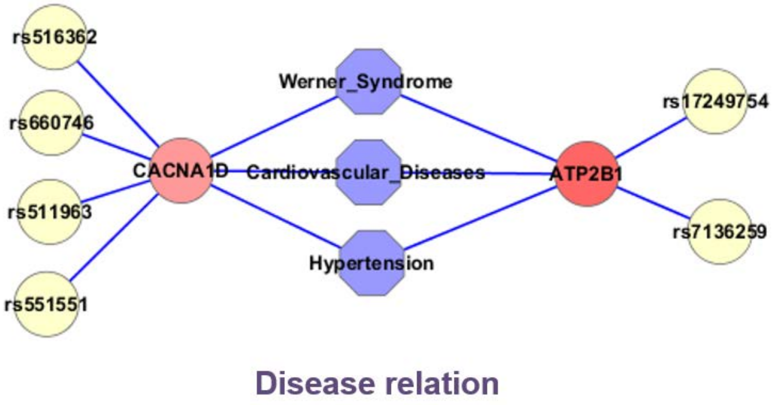


Figure 6.8 Diesease relation plot in VizEpis

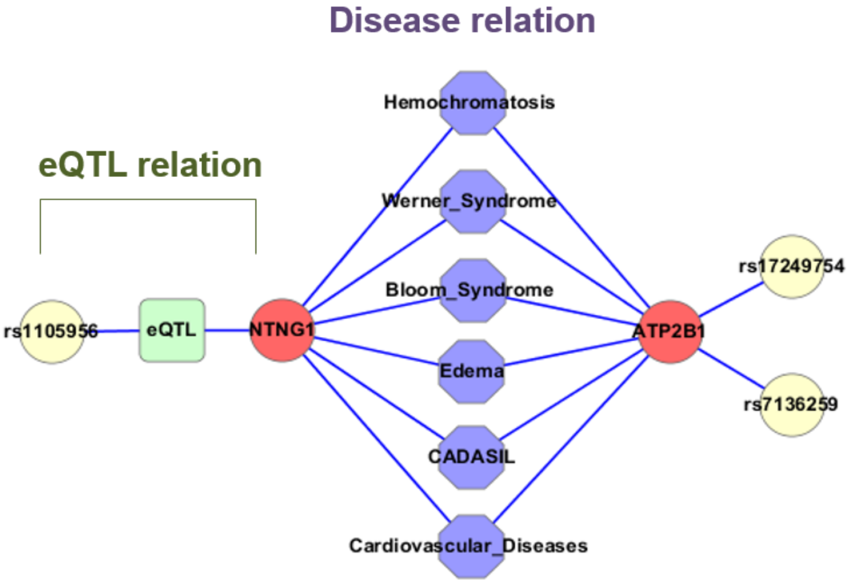


Figure 6.9 Diesease relation plot in VizEpis

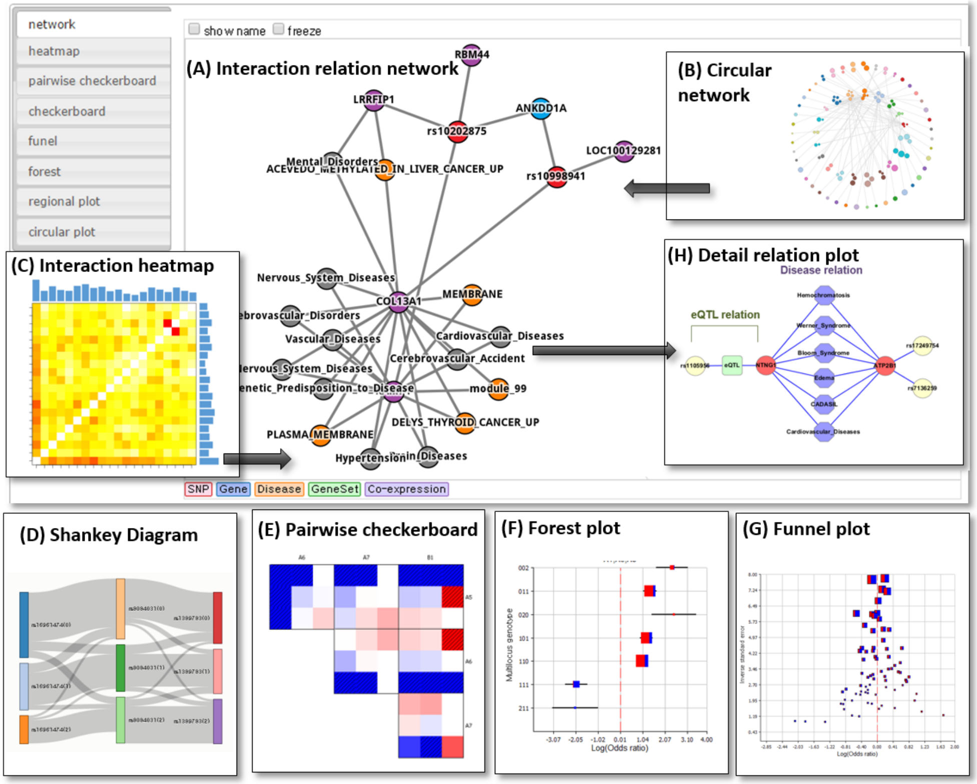


Figure 6.10 workflow of VizEpis

Table 6.1 Comparison of genotype-wise plot

	CB	**Pairwise CB**	**Forest**	**Funnel**
represent OR	color	color	X-axis	X-axis
represent sample size	X	X	box size	box size
represent high-order interaction	O	2nd order	O	O
detect genotype with highest/lowest OR	by color	by color	easy to detect	easy to detect
represent CI	X	X	O	X

6.3 Case Study

6.3.1 Interpretation of gene-gene interaction in WTCC bipolar disorder data

In previous studies [Kwon, et al. 2011; Oh, et al. 2012], we conducted a genome-wide interaction analysis with WTCCC-BD dataset by using GWAS-GMDR. The LOC390730 (rs2192859) and MYH13 (rs2320796) were reported to show the best two-way interaction. This two-way interaction was visualized via checkerboard and forest plots (Figure 3). The checkerboard plot suggests a multiplicative epistatic pattern of the two SNPs. Recently, MYH13 has been reported to be associated with formal thought disorder (FTD) or disorganized speech which is one of the central signs of schizophrenia [12]. No functional annotation has not yet been assigned to LOC390730 . It is worth investigating this LOC390730 further as a potential candidate gene for bipolar disorder.

In Figure 6.6, PC and funnel plots display four-locus genotypes of SNPs (rs4744513, rs1755991, rs290253, rs10991725) in SYK which has been reported to be related with dopamine receptors-related diseases, such as Schizophrenia, Parkinson's disease or bipolar disorder [Oh, et al. 2012; Seol, et al. 2004].The PC plot shows that (1,1,1,1) is the most common genotype in case samples, and (1,1,1,0) is the one in control sample. In the funnel plot, we can easily discover that (1,1,1,0) and (1,1,1,1) have high precision and, while(1,2,0,2) and (1,0,2,2) have large and small OR, respectively.

6.3.2 Interpretation of gene-gene interaction in Age-related macular degeneration (AMD) data

In AMD dataset [Klein, et al. 2005], CFH has been reported to have strong association with AMD. Figure 6.5 presents the two-way interaction patterns between CFH and other genes, such as (CFH, SGCD) and (CFH, MED27). We observed that the marginal effect of CFH is predominant in both two-interactions with SGCD and MED27. While SGCD has been reported to be associated with AMD [Tang, et al. 2009], the relationship between MED27 and AMD has been uncovered.

6.4 Conclusion

In the analysis of gene-gene interaction, the characterization is as important as the identification because it can provide insight into a mechanism for how a gene-gene interaction influences a trait of interest. However, such characterization can be a challenge, especially for high-order gene-gene interactions. The VizEpis provides visualization tools that facilitate graphical representation of gene-gene interactions in various ways, and hence enables ones to effectively characterize and interpret (especially high-order) gene-gene interactions in details.

Chapter 7

Summary and Conclusion

With high-throughput detection technique of genetic variants, main focuses of genetic research for unveiling complex etiology of disease go over to gene-gene interaction and rare variant association. Although many gene-gene interaction methods have been proposed, there are still some unsolved critical issues including heavy computation, biological misunderstanding and absence of method for rare variant's interaction for statistical genetic interactions. In this thesis, we focus on gene-gene interaction of common variants and rare variants.

In chapter 3, we proposed IGENT, a fast analysis for searching for high-order interactions associated with complex diseases. IGENT can detect gene-gene interactions using information gain which represents association strength with phenotype and gene-gene interaction without the assumption of a specific genetic model. IGENT adopts an exhaustive search scheme and stepwise search scheme. In the exhaustive search, IGENT investigates all possible interactions in lower-order interactions. And, IGENT can identify high-order interaction using the stepwise search scheme. The permutation and exhaustive search schemes of the previous GGI methods are computationally too intensive to be employed in large genome-wide scale data set for high-order interactions. In our simulation, IGENT is as fast as BOOST and shows better performance than BOOST. Also, IGENT can evaluate significance of the interaction result using hypothesis testing framework by approximating the gamma distribution that information gain value follows under the null hypothesis. Using approximation to the gamma distribution instead of permutation, IGENT can easily calculate statistical significant interactions and save the computation time remarkably. Through real WTCCC and AMD data analysis, IGENT successfully identified low order and high order interactions.

In chapter 4, we propose a new gene-gene interaction method for the rare variants in the framework of the multifactor dimensionality reduction (MDR) analysis. The proposed method consists of two steps. The first step is to collapse the rare variants in a specific region such as gene. The second step is to perform MDR analysis for the collapsed rare variants. The proposed method is illustrated with 1080 whole exome sequencing data of Korean population to identify causal gene-gene interaction for rare variants for type 2 diabetes.

In chapter 5, we presented the computing capability of CUDA-enabled GPUs for accelerating MDR algorithm. Our cuGWAM has various features that distinguish from existing MDR application for GWAS as follows. First, it implements an effective memory handling algorithm and efficient procedures for MDR to make joint analysis of multiple genes feasible for GWAS. Second, it can report multiple candidates for causal gene-gene interactions. Third, various performance measures, including tau-b, likelihood ratio, and normalized mutual information, were implemented to evaluate MDR classifiers (Namkung et al., 2009a). Finally, it implements three methods for handling missing genotypes: complete, available and missing category (Namkung et al., 2009b). These features of our method can lead to vast speed-up and, enable GGI analysis on millions variant data that presently cannot be performed due to computing time limitations in traditional computing systems. We obtained ~500 fold performance improvement over original MDR and equivalent performance of about 200 cores cluster system in one GPU system. Also, by porting the MDR onto high-performance graphic cards using the CUDA environment, we obtained up to ~5 fold acceleration compared with MDRGPU. Especially, cuGWAM has better performance than MDRGPU in every types of testing (sample size, marker size and high-order interaction). Even though a Python-based GPU compute code was expected to show the same full performance of GPU hardware as a C-controlled GPU compute code (Klockner et al., 2009), our test results showed that there was different performance between Python-base MDRGPU and C-base cuGWAM. We expect that this difference was derived from optimal utilization of hardware memory. Because MDR is exhaustive algorithm with exploring all combinations of variants, memory optimization is critical in performance of software. Especially in GPU system, the optimization of shared memory determines the performance of the application. Because cuGWAM uses binarized count values in shared memory to save memory space, it can run more threads with restricted shared memory space and input/ouput (IO) transaction load between threads and memory is minimized.

In chapter 6, we developed the VizEpis, a tool for visualizing of GGIs in genetic association analysis and mapping of epistatic interaction to the biological evidence from public interaction databases. Using interaction network and circular plot, the VizEpis provides to explore the interaction network integrated with biological evidences in epigenetic regulation, splicing, transcription, translation and post-translation level. To aid statistical interaction in genotype level, the VizEpis provides checkerboard, pairwise checkerboard, forest, funnel and ring chart.

In summary, we developed GGI analysis methods of high-dimensional genomic data. First, for more efficient analysis, we developed the entropy-based gene-gene interaction method. Second, we suggested gene-gene analysis method of rare variants. Third, we implemented CUDA-based gene-gene interaction software for high performance computation. Finally, we proposed the visualization methods for interpretation of gene-gene interaction.

There are some more future study issues in GGI analysis. Although IGENT and cuGWAM are fast and high-performed methods, they still cannot cover high-order interaction in whole genome-wide scale due to tremendous computational burden. To conduct high-order GGI analysis, the practical alternatives are statistical feature selections or biological feature selections. The examples of the statistical feature selections are marginal effect-based forward selection, regularization-based selection, genetic algorithm-based selection and ant-colony algorithm-based selection. The instances of the biological feature selections are functional region based selection, gene based selection and pathway based selection. Although the efforts of the mapping and interpretation of GGI have been tried, the interpretation of GGI is not simple. Specially, lots of variants are located in inter-genic region. The variants in inter-genic region cannot be annotated to their functions. It means that unknown or uninterpretable genomic regions have larger portion than the unknown or interpretable regions. As increasing of biological evidence of GGI, the biological databases update their data or change their data format. For persistently providing of the interpretation of GGI in VizEpis, we need to update the biological knowledge, constantly.

Bibliography

Botstein D, Risch N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33 Suppl:228-37.

Breiman L. 2001. Random forests. Machine Learning 45(1):5-32.

Bush WS, Dudek SM, Ritchie MD. 2006. Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions. Bioinformatics 22(17):2173-4.

Bush WS, Edwards TL, Dudek SM, McKinney BA, Ritchie MD. 2008. Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. BMC Bioinformatics 9:238.

Cattaert T, Urrea V, Naj AC, De Lobel L, De Wit V, Fu M, Mahachie John JM, Shen H, Calle ML, Ritchie MD and others. 2010. FAM-MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals. PLoS One 5(4):e10304.

Chanda P, Zhang A, Brazeau D, Sucheston L, Freudenheim JL, Ambrosone C, Ramanathan M. 2007. Information-theoretic metrics for visualizing gene-environment interactions. Am J Hum Genet 81(5):939-63.

Chang CC, Lin CJ. 2011. LIBSVM: A Library for Support Vector Machines. Acm Transactions on Intelligent Systems and Technology 2(3).

Chen SH, Sun J, Dimitrov L, Turner AR, Adams TS, Meyers DA, Chang BL, Zheng SL, Gronberg H, Xu J and others. 2008. A support vector machine approach for detecting gene-gene interaction. Genet Epidemiol 32(2):152-67.

Choi J, Park T. 2013. Multivariate generalized multifactor dimensionality reduction to detect gene-gene interactions. BMC Syst Biol 7 Suppl 6:S15.

Cordell HJ. 2009. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10(6):392-404.

Culverhouse R, Suarez BK, Lin J, Reich T. 2002. A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet 70(2):461-71.

Davis NA, Lareau CA, White BC, Pandey A, Wiley G, Montgomery CG, Gaffney PM, McKinney BA. 2013. Encore: Genetic Association Interaction Network centrality pipeline and application to SLE exome data. Genet Epidemiol 37(6):614-21.

Dawy Z, Goebel B, Hagenauer J, Andreoli C, Meitinger T, Mueller JC. 2006. Gene mapping and marker clustering using Shannon's mutual information. IEEE/ACM Trans Comput Biol Bioinform 3(1):47-56.

de Oliveira F, Borges CC, Almeida F, F ES, da Silva Verneque R, da Silva MV, Arbex W. 2014. SNPs selection using support vector regression and genetic algorithms in GWAS. BMC Genomics 15(Suppl 7):S4.

Dematte L, Prandi D. 2010. GPU computing for systems biology. Brief Bioinform 11(3):323-33.

Djurovic S, Gustafsson O, Mattingsdal M, Athanasiu L, Bjella T, Tesli M, Agartz I, Lorentzen S, Melle I, Morken G and others. 2010. A genome-wide association study of bipolar disorder in Norwegian individuals, followed by replication in Icelandic sample. J Affect Disord 126(1-2):312-6.

Fang YH, Chiu YF. 2012. SVM-based generalized multifactor dimensionality reduction approaches for detecting gene-gene interactions in family studies. Genet Epidemiol 36(2):88-98.

Flannick J, Thorleifsson G, Beer NL, Jacobs SB, Grarup N, Burtt NP, Mahajan A, Fuchsberger C, Atzmon G, Benediktsson R and others. 2014. Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. Nat Genet 46(4):357-63.

Friendly M. 1994. Mosaic Displays for Multiway Contingency-Tables. Journal of the American Statistical Association 89(425):190-200.

Goebel B, Dawy Z, Hagenauer J, Mueller JC. An approximation to the distribution of finite sample size mutual information estimates; 2005 16-20 May 2005. p 1102-1106 Vol. 2.

Gui J, Moore JH, Kelsey KT, Marsit CJ, Karagas MR, Andrew AS. 2011. A novel survival multifactor dimensionality reduction method for detecting gene-gene interactions with application to bladder cancer prognosis. Hum Genet 129(1):101-10.

Gui J, Moore JH, Williams SM, Andrews P, Hillege HL, van der Harst P, Navis G, Van Gilst WH, Asselbergs FW, Gilbert-Diamond D. 2013. A Simple and Computationally Efficient Approach to Multifactor Dimensionality Reduction Analysis of Gene-Gene Interactions for Quantitative Traits. PLoS One 8(6):e66545.

Hahn LW, Ritchie MD, Moore JH. 2003. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19(3):376-82.

Hamshere ML, Green EK, Jones IR, Jones L, Moskvina V, Kirov G, Grozeva D, Nikolov I, Vukcevic D, Caesar S and others. 2009. Genetic utility of broadly defined bipolar schizoaffective disorder as a diagnostic concept. Br J Psychiatry 195(1):23-9.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106(23):9362-7.

International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P and others. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851-61.

Iwamoto K, Ueda J, Bundo M, Kojima T, Kato T. 2011. Survey of the effect of genetic variations on gene expression in human prefrontal cortex and its application to genetics of psychiatric disorders. Neurosci Res 70(2):238-42.

Jiang R, Tang W, Wu X, Fu W. 2009. A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinformatics 10 Suppl 1:S65.

Kim K, Kwon MS, Oh S, Park T. 2013. Identification of multiple gene-gene interactions for ordinal phenotypes. BMC Med Genomics 6 Suppl 2:S9.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST and others. 2005. Complement factor H polymorphism in age-related macular degeneration. Science 308(5720):385-9.

Koo CL, Liew MJ, Mohamad MS, Salleh AH. 2013. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. Biomed Res Int 2013:432375.

Kotti S, Bickeboller H, Clerget-Darpoux F. 2007. Strategy for detecting susceptibility genes with weak or no marginal effect. Hum Hered 63(2):85-92.

Kwon M-S, Kim K, Lee S, Chung W, Yi S-G, Namkung J, Park T. GWAS-GMDR: A program package for genome-wide scan of gene-gene interactions with covariate adjustment based on multifactor dimensionality reduction; 2011 12-15 Nov. 2011. p 703-707.

Laje G, Allen AS, Akula N, Manji H, John Rush A, McMahon FJ. 2009. Genome-wide association study of suicidal ideation emerging during citalopram treatment of depressed outpatients. Pharmacogenet Genomics 19(9):666-74.

Lander ES. 1996. The new genomics: global views of biology. Science 274(5287):536-9.

Lee S, Kwon MS, Oh JM, Park T. 2012. Gene-gene interaction analysis for the survival phenotype based on the Cox model. Bioinformatics 28(18):i582-i588.

Lewis S, Clarke M. 2001. Forest plots: trying to see the wood and the trees. BMJ 322(7300):1479-80.

Liu Y, Schmidt B, Maskell DL. 2010. CUDASW++2.0: enhanced Smith-Waterman protein database search on CUDA-enabled GPUs based on SIMT and virtualized SIMD abstractions. BMC Res Notes 3:93.

Lou XY, Chen GB, Yan L, Ma JZ, Mangold JE, Zhu J, Elston RC, Li MD. 2008. A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. Am J Hum Genet 83(4):457-67.

Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC, Li MD. 2007. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am J Hum Genet 80(6):1125-37.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy

MI, Ramos EM, Cardon LR, Chakravarti A and others. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747-53.

Martin ER, Ritchie MD, Hahn L, Kang S, Moore JH. 2006. A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. Genet Epidemiol 30(2):111-23.

Martinowich K, Schloesser RJ, Manji HK. 2009. Bipolar disorder: from genes to behavior pathways. J Clin Invest 119(4):726-36.

Namkung J, Elston RC, Yang JM, Park T. 2009a. Identification of gene-gene interactions in the presence of missing data using the multifactor dimensionality reduction method. Genet Epidemiol 33(7):646-56.

Namkung J, Kim K, Yi S, Chung W, Kwon MS, Park T. 2009b. New evaluation measures for multifactor dimensionality reduction classifiers in gene-gene interaction analysis. Bioinformatics 25(3):338-45.

Oh S, Lee J, Kwon MS, Weir B, Ha K, Park T. 2012. A novel method to identify high order gene-gene interactions in genome-wide association studies: gene-based MDR. BMC Bioinformatics 13 Suppl 9:S5.

Pattin KA, White BC, Barney N, Gui J, Nelson HH, Kelsey KT, Andrew AS, Karagas MR, Moore JH. 2009. A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction. Genet Epidemiol 33(1):87-94.

PGCC PGCCC, Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J, Lehner T, Levinson DF, Moran A and others. 2009. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. Am J Psychiatry 166(5):540-56.

Phillips PC. 2008. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9(11):855-67.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and others. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559-75.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273(5281):1516-7.

Ritchie MD, Hahn LW, Moore JH. 2003a. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24(2):150-7.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69(1):138-47.

Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH. 2003b. Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. BMC

Bioinformatics 4:28.

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 242(1):84-9.

Ruiz-Marin M, Matilla-Garcia M, Cordoba JA, Susillo-Gonzalez JL, Romo-Astorga A, Gonzalez-Perez A, Ruiz A, Gayan J. 2010. An entropy test for single-locus genetic association analysis. BMC Genet 11:19.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78(4):629-44.

Schwarz DF, Konig IR, Ziegler A. 2010. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics 26(14):1752-8.

Seol IW, Kuo NY, Kim KM. 2004. Effects of dopaminergic drugs on the mast cell degranulation and nitric oxide generation in RAW 264.7 cells. Arch Pharm Res 27(1):94-8.

Shannon CE. 1948. A Mathematical Theory of Communication. Bell System Technical Journal 27(4):623-656.

Sinnott-Armstrong NA, Greene CS, Cancare F, Moore JH. 2009. Accelerating epistasis analysis in human genetics with consumer graphics hardware. BMC Res Notes 2:149.

Sterne JA, Egger M. 2001. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol 54(10):1046-55.

Stivala AD, Stuckey PJ, Wirth AI. 2010. Fast and accurate protein substructure searching with simulated annealing and GPUs. BMC Bioinformatics 11:446.

Tang W, Wu X, Jiang R, Li Y. 2009. Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy. PLoS Genet 5(5):e1000464.

van Winkel R, Genetic R, Outcome of Psychosis I. 2011. Family-based analysis of genetic variation underlying psychosis-inducing effects of cannabis: sibling analysis and proband follow-up. Arch Gen Psychiatry 68(2):148-57.

Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, Moore JH. 2007. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet Epidemiol 31(4):306-15.

Visscher PM. 2008. Sizing up human height variation. Nat Genet 40(5):489-90.

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. Am J Hum Genet 90(1):7-24.

Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W. 2010. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet 87(3):325-40.

Wellcome Trust Case Control C. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature

447(7145):661-78.

Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, Huebner M, Biernacka JM. 2012. SNP interaction detection with Random Forests in high-dimensional genetic data. BMC Bioinformatics 13:164.

Yee J, Kwon MS, Park T, Park M. 2013. A modified entropy-based approach for identifying gene-gene interactions in case-control study. PLoS One 8(7):e69321.

Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. 2008. A navigator for human genome epidemiology. Nat Genet 40(2):124-5.

Zhao J, Jin L, Xiong M. 2006. Test for interaction between two unlinked loci. Am J Hum Genet 79(5):831-45.

Zuk O, Hechter E, Sunyaev SR, Lander ES. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 109(4):1193-8.

초록

지노타이핑(genotyping)과 시퀀싱(sequencing)기술이 발전하면서 복합질환 및 복합형질과 연관성이 있는 유전 변이(genetic variants)를 점차 밝혀 가고 있다. 이러한 연구를 통해 이미 형질과의 연관성이 있는 수 천개 이상의 유전변이를 발굴하였음에도 불구하고, 발굴된 유전변이만으로는 유전율(heritabiltiy)의 전체를 설명하지 못하는 것으로 밝혀졌다. 따라서 아직 설명이 안된 복합형질의 유전율의 많은 부분을 설명하기 위해 유전자-유전자 상호작용(gene-gene interaction) 분석과 희귀유전변이 분석의 중요성이 대두되었다.

유전자-유전자 상호작용에는 여러 해결해야 할 이슈가 있다. 첫째로 유전자 상호작용 분석은 많은 수의 테스트(test)를 수행해야 한다. 만약 k개의 변이자료가 있다면, k(k-1)/2개의 상호작용을 계산해야 한다. 따라서 상호작용의 차수가 증가하게 된다면 계산해야 할 상호작용 수는 기하급수적으로 늘어난다. 그리므로 상호작용 분석에서 효율적인 알고리즘이 필요하다. 둘째로 희귀유전변이는 그 특징상 발생 이벤트가 극히 드물기 때문에 희귀유전변이간의 상호작용을 찾기가 쉽지 않다. 셋째로 상호작용 분석에서 빠른 계산력을 가진 시스템이 필요하다. 넷째로 유전자 상호작용의 생물학적인 해석이 중요함에도 불구하고, 상호작용의 복잡성 때문에 해석이 쉽지 않다.

본 논문에서는 정보학 이론에 기반한 전장유전체를 포괄할 수 있는 상호작용 분석 방법인 Information theory-based GEnome-wide

gene-gene iNTeraction(IGENT)를 개발하였다. IGENT는 유전자-유전자 상호작용과 유전자-환경 상호작용을 분석할 수 있는 효과적인 알고리즘이다. 전장유전체 수준에서 유전자 상호작용을 분석하기 위해서는 계산시간을 단축시키는 것이 중요하다. IGENT는 순열(permutation) 방법과 같은 리샘플링(resampling) 방법을 사용하지 않고, 보다 계산이 간단한 IG(information gain)을 사용한다. 시뮬레이션 연구를 통해서 IGENT의 파워가 BOOST와 비슷하거나 더 좋은 것을 밝혔다. 그리고 제안된 방법을 사용하여 WTCCC의 조울증 자료와 노인성 황반변성(AMD)자료에서 성공적으로 유전자 상호작용 분석을 수행하였다.

유전자-유전자 상호작용 분석은 공통 변이(common variant)에 대해 방법론과 분석이 많이 이루어 졌으나, 상대적으로 희귀변이(rare variant)에 대해서는 상호작용 분석 방법이 발전이 거의 없다시피 하다. 본 논문에서 Multifactor dimensionality reduction(MDR) 프레임을 사용한 희귀유전변이를 위한 새로운 유전자 상호작용 방법을 제안한다. 첫번째 단계에서는 유전자 영역 안의 희귀변이들을 유전자 단위로 변환(collapsing)시키고, 두번째 단계에서는 변환된 유전자 단위변이를 이용하여 MDR 분석을 수행한다. 제안된 방법은 1072명 한국인의 제2형 당뇨 자료를 이용하여 희귀유전변이의 상호작용을 분석하였다.

전장유전체 수준에서 상호작용 분석을 하기 위해, 본 논문에서는 CUDA를 적용하여 MDR 기반의 상호작용을 분석할 수 있는 cuGWAM을 개발하였다. cuGWAM은 모든 시뮬레이션에서 CPU 기반으로 하는 MDR 소프트웨어나 다른 GPU기반의 상호작용 분석 소프트웨어보다 우수한 성능을 보였다.

유전자 상호작용을 보여주는 많은 방법이 나왔음에도 불구하고, 상호작용의 해석은 그리 쉽지 않다. 이러한 유전자 상호작용은 유전자의 결과물과의 생화학적 상호작용을 의미하지는 않다. 이러한 유전자 상호

작용은 단백질-단백질 상호작용, 후성유전체적인 조절, 크로모좀상의 구조적 상호작용, 번역과정의 조절(translational regulation), 신호 전단, 생화학적 네트워크 그리고 발달상의 신호전달과정(developmental pathways)의 원인이 될 수 있다. 따라서 효과적으로 유전자 상호작용의 통계적인 해석과 생물학적인 증거를 제공하기 위해서, 본 논문에서는 VizEpis을 개발하였다. VizEpis은 유전자 상호작용을 공개적인 상호작용 데이터베이스에 맵핑(mapping)함으로써 시각화하는 프로그램이다. VizEpis은 상호작용 네트워크를 이용하여 후성유전적 조절, 스플라이싱(splicing), 전사, 변역, 변역후 과정에서 생물학적 상호작용을 표시해준다. 유전자형(genotype) 수준에서 통계적 해석을 돕기 위해, VizEpis은 checkerboard plot, forest, funnel and ring chart를 제공해준다.

이렇게 본 연구에서는 효과적인 유전자 상호작용의 방법들을 제시하였고 실제 데이터에 적용하였다. 본 연구에서 제시된 방법은 복합질환에 영향을 미치는 유전인자의 집합을 효과적으로 발굴하고 질병이 발생하는 기작을 연구하는데 활용될 수 있을 것으로 기대된다.

주요어: 유전자-유전자 상호작용, 전장유전체연관성분석, 대량병렬시퀀싱, 희귀변이, 그래픽 연산 유닛(GPU), 시각화

학 번: 2008-30830