이학박사학위논문

Gene set analysis for Genome-Wide Association Study and Next Generation Sequencing data

전장유전체연관성분석자료 및 차세대시퀀싱자료에 대한 유전자 집합 분석

2013년 2월

서울대학교 대학원

통계학과

이 재 훈

Gene set analysis for Genome-Wide Association Study and Next Generation Sequencing data

지도교수 박태성

이 논문을 이학박사 학위논문으로 제출함

2012 년 12 월

서울대학교 대학원

통계학과

이 재 훈

이재훈의 이학박사 학위논문을 인준함

2012 년 12 월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

학위논문 원문제공 서비스에 대한 동의서

본인의 학위논문에 대하여 서울대학교가 아래와 같이 학위논문 저작물을 제공하는 것에 동의합니다.

1. 동의사항

①본인의 논문을 보존이나 인터넷 등을 통한 온라인 서비스 목적으로 복제할 경우 저작물의 내용을 변경하지 않는 범위 내에서의 복제를 허용합니다.

②본인의 논문을 디지털화하여 인터넷 등 정보통신망을 통한 논문의 일부 또는 전부의 복제.배포 및 전송 시 무료로 제공하는 것에 동의합니다.

2. 개인(저작자)의 의무

본 논문의 저작권을 타인에게 양도하거나 또는 출판을 허락하는 등 동의 내용을 변경하고자 할 때는 소속대학(원)에 공개의 유보 또는 해지를 즉시 통보하겠습니다.

3. 서울대학교의 의무

①서울대학교는 본 논문을 외부에 제공할 경우 저작권 보호장치(DRM)를 사용하여야 합니다.

②서울대학교는 본 논문에 대한 공개의 유보나 해지 신청 시 즉시 처리해야 합니다.

논문제목 :Gene set analysis for Genome-Wide Association Study and Next Generation Sequencing data (전장유전체연관성분석 자료 및 차세대 시퀀싱자료에 대한 유전자 집합 분석)

학위구분 : **석사 □.박사 □**
학 과 : 통계학과
학 번 : 2007-30077
연 락 처 :02-880-9168
저 작 자 : 이재훈 (인)

제 출 일 : 2013 년 월 일

서울대학교총장 귀하

Gene set analysis for Genome-Wide Association Study and Next Generation Sequencing data

by

Jaehoon Lee

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

Department of Statistics

College of Natural Sciences

Seoul National University

Feb, 2013

Abstract

Gene set analysis for Genome-Wide Association Study and Next Generation Sequencing data

Jaehoon Lee
Department of Statistics
The Graduate School
Seoul National University

Genome-wide association study (GWAS) has successfully identified thousands of common genetic variants, mainly common single nucleotide polymorphisms (SNPs), associated with complex traits, including many common diseases. In general, many GWA methods only consider association of a single SNP and provide the list of the most significant SNPs or related genes due to computational burden.

However, complex diseases often result from compound action of multiple risk factors and therefore the single-SNP-based analysis may miss the genetic variants that affect risk effects jointly but have scarce individual effects. Also, it has been suggested that the associated variants can explain only a small fraction of the heritability of most common traits.

To resolve these issues, it was suggested to utilize prior biological knowledge or known pathway information, and thus to incorporate a set of related SNPs,

which leads to a smaller number of tests. This approach was motivated by the gene-set analysis (GSA), widely used in the analysis of microarray data. GSA focuses on gene-sets rather than individual genes, and combines weak signals from a number of individual genes in a set, when individual genes are weakly associated with the traits. Considering the multiple SNPs jointly within a gene-set in GWAS can increase power.

When multiple SNPs are jointly considered, the corresponding SNP-level association measures are likely to be correlated due to the linkage disequilibrium (LD) among SNPs. We proposed SNP-based parametric robust analysis of gene-set enrichment (SNP-PRAGE) method which handles correlation adequately among association measures of SNPs, and minimizes computing effort by the parametric assumption. SNP-PRAGE first obtains gene-level association measures from SNP-level association measures by incorporating the size of corresponding (or nearby) genes and the LD structure among SNPs. Afterward, SNP-PRAGE acquires the gene-set level summary of genes that undergo the same biological knowledge. This two-step summarization makes the within-set association measures to be independent from each other, and therefore the central limit theorem can be adequately applied for the parametric model.

In addition, rare variants study is another breakthrough to limitation of GWAS. Rare variants are defined as variants with minor allele frequency less than one percent. There are growing evidences that rare variants contribute to the etiology of complex disease. It has been argued that collections of rare variants could fill the missing heritability of common traits. Since frequencies of rare variants are very low, even with high penetrance, it will be difficult to detect association with any single rare variant. Hence, the most popular statistical test for

GWAS based on testing single SNPs is not expected to perform well.

Recently, there are growing methods has emerged to overcome this limitation. Three main strategies in this field are collapsing methods, weighting/prioritizing methods and distribution-based methods. However, assessing the association between rare variants and complex disease is still a challenging task and no single method gives consistently acceptable power across the range of these relationships, even in a large sample size.

For the consistently powerful association method under the various scenarios of act of rare variants, we propose some quadratic tests (QTests; Q_1, Q_2, and Q_3 for gene-level for quantitative traits. These methods are computationally efficient and have relatively high power in various patterns of disease rare variants. Also, in order to increase power to detect the genetic association from rare variants, we propose the QTest for the gene-set (Q_{GS}) by extending the unit of analysis from genes to gene-sets. When combining the gene-level statistic, we used a co-mutation based weight. The logic behind it is that highly interacted genes with other neighbor genes usually play an important role within the gene-set.

These association tests can cover the broad range of scenarios for joint action of rare variants including the existence of common disease variants. We demonstrate the performance of the proposed methods comparing with other gene-level and gene-set-level association methods in various simulation setting. These include collapsing methods (GRANVIL; Variable Threshold (VT) method; Weighted sum statistic (WSS)), a distribution-based approach (SKAT, SKAT-O) as gene-level association methods and GLOSSI and GlobalTest as traditional gene-set-level association methods.

We also applied our methods to real data. SNP-PRAGE is applied to two

GWA data sets: hypertension data of 8,842 samples from the Korean population and bipolar disorder data of 4,806 samples from the Wellcome Trust Case Control Consortium (WTCCC). The quadratic test (QTest) is applied to the sequence data and alanine aminotransferase (ALT) phenotype of 1058 samples from Korean population.

Keywords: Genome-wide association study (GWAS), Next Generation Sequencing (NGS), Gene-set analysis (GSA), rare variant association test, co-mutation

Student number: 2007-30077

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background of Genome-wide association study (GWAS)

1.1.1 History of genome-wide association study (GWAS)

The genome-wide association study (GWAS) is an approach that involves rapidly scanning single-nucleotide polymorphism (SNP) markers across genome (0.5 million or 1 million) of many people to find genetic variations associated with a particular disease [1]. A single-nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide (A, T, C, or G) in the gene differs between members of species. For example, two sequenced DNA fragments from different individuals, ATGAGCCTA to ATGAGCTA, contain a difference in a single nucleotide. In this case, we say that there are two alleles (C and T). Almost all common SNPs have only two alleles. We also say that this bi-allelic SNP has three genotypes: CC, CT, and TT.

Within a population, SNPs can be assigned a minor allele frequency (MAF) which is the lowest allele frequency at a locus that is observed in a particular population. This is simply the lesser of the two allele frequencies for single-nucleotide polymorphisms. There are variations between human populations, so a SNP allele that is common in one geographical or ethnic group may be much rarer in another.

Risch and Merikangas [2] noted that small genetic effects could be detected with greater power by association analyses and proposed that genome-wide linkage disequilibrium mapping (i.e., GWAS) could be applied if technologies were developed to study SNP frequencies in all genes, contrasting ill case subjects versus comparison subjects or case subjects and their parents (associated alleles are transmitted to ill offspring more often than expected by chance).

Lander [3] proposed the common-disease common-variant hypothesis (CDCV). Comparing any two people, most sequence differences are ancient, common SNPs (by convention, varying on at least 5% of chromosomes in a population), which Lander argued must confer at least some (not all) of the genetic risk for common diseases. He proposed cataloguing them and studying their association with disease in large samples. SNPs become common because they are neutral or favorable with respect to survival (e.g., evolutionary pressures can rapidly increase frequencies of adaptive SNPs in gene regulating regions). However, some have mildly harmful effects, perhaps depending on environmental conditions (e.g., preserving fat during an ice age but leading to obesity in the fast food era). The common-disease common-variant 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 enough people.

The HapMap Project [4] (http://www.hapmap.org) has validated more than 4 million SNPs, while creating competition among biotechnology companies to develop high-throughput genotyping technologies. The Affymetrix and Illumina

companies have developed chips (arrays of assays on glass slides) that assay large SNP sets with high accuracy at low cost and rapid speed [5].

This effort makes it possible to capture a good deal of the common genetic variation across the genome using a representative group of SNPs that can be affordably assayed by means of high-throughput technologies [6]. The current iteration of genotyping arrays assess ~1 million SNPs across the genome, providing coverage of ~80–90% of the HapMap SNPs with minor allele frequency (MAF) greater than 5%.

The genome-wide association study (GWAS) has been successful to investigate generic variants associated with some targeted phenotypes in complex diseases. Since 2005 when the first GWAS was reported, 1,400 studies have been added to the Catalog of Published Genome-Wide Association Studies (See Figure 1.1) (http://www.genome.gov/gwastudies/).

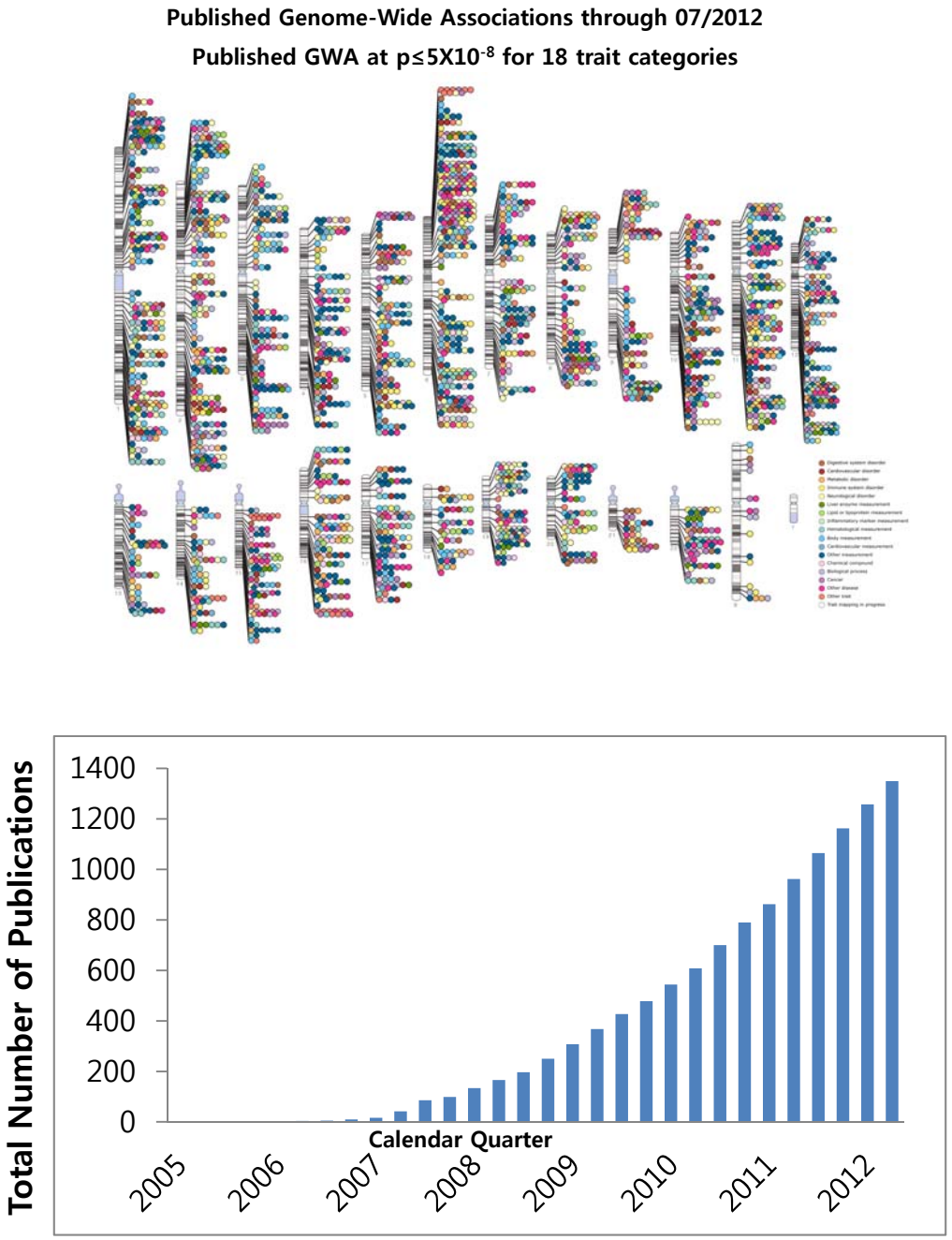

Figure 1.1 Published GWA Reports, 2005 – 6/2012

1.1.2 Single SNP-based analysis in GWAS

GWAS data analysis has largely focused on single marker discovery. At the stringent genome-wide significance level of p-val $< 5 \times 10^{-8}$, many markers that are truly but weakly associated with disease often fail to be detected [7]. Given the large number of markers typed and this stringent statistical criterion necessary to minimize false positive hits, so far only the most significant associations have been established [8]. It is likely that the genetic associations reported in GWAS represent only the tip of the iceberg of genes contributing to disease risk, and that the majority of genes still remain hidden within the statistical "noise" inherent in this approach [9].

Complex diseases often result from compound action of multiple risk factors and therefore the single-SNP-based analysis may miss the genetic variants that affect risk effects jointly but have scarce individual effects. Also, the locus heterogeneity, which implies that alleles at different loci target the same diseases in different individuals, would increase difficulty in replication of association of a single marker [10].

To resolve these issues, it was suggested to utilize prior biological knowledge or known pathway information, and thus to incorporate a set of related SNPs, which leads a smaller number of tests [11]. This approach was motivated by the gene-set analysis (GSA), widely used in the analysis of microarray data.

1.2 Background of next-generation sequencing (NGS)

1.2.1 Missing heritability and Next generation sequencing (NGS)

Only a small portion of heritability can be explained by loci identified from GWAS. Heritability is a concept that summarizes how much of the variation in a trait is due to variation in genetic factors. In fact, heritability is formally defined as the proportion of phenotypic variation that is due to variation in genetic values [12]. Traditionally, heritability was estimated from simple, often balanced, designs, such as the correlation of offspring and parental phenotypes, the correlation of full or half siblings, and the difference in the correlation of monozygotic (MZ) and dizygotic (DZ) twin pairs.

Many explanations for this missing heritability have been suggested, including much larger numbers of variants of smaller effect yet to be found; rarer variants (possibly with larger effects as you can see in Figure 1.2) that are poorly detected by available genotyping arrays that focus on variants present in 5% or more of the population; structural variants poorly captured by existing arrays; low power to detect gene–gene interactions [13].

Rare variants study is emerging as another breakthrough to limitation of GWAS. Rare variants are defined as variants with MAF less than one percent and low-frequency variants are defined as variants with MAF larger than 1% and less than 5%. Rare variants or low-frequency variants are not interrogated in GWAS and could explain a large fraction of the missing heritability of common diseases [14].

The primary technology for the detection of rare SNPs is sequencing, which may target regions of interest, or may examine the whole genome. Next-generation

sequencing technologies, which process millions of sequence reads in parallel, provide monumental increases in speed and volume of generated data free of the cloning biases and arduous sample preparation characteristic of capillary sequencing [13]. Detection of associations with low frequency and rare variants is facilitated by the comprehensive catalogue of variants being generated by the 1,000 Genomes Project (http://www.1000genomes.org/page.php).

There are growing evidences that these rare variants or low-frequency variants contribute to the etiology of complex disease. It is discovered rare variants associated with complex traits are sometimes causal through amino acid substitution and sequencing studies have shown that multiple rare variants can play an important role in complex disease [15].

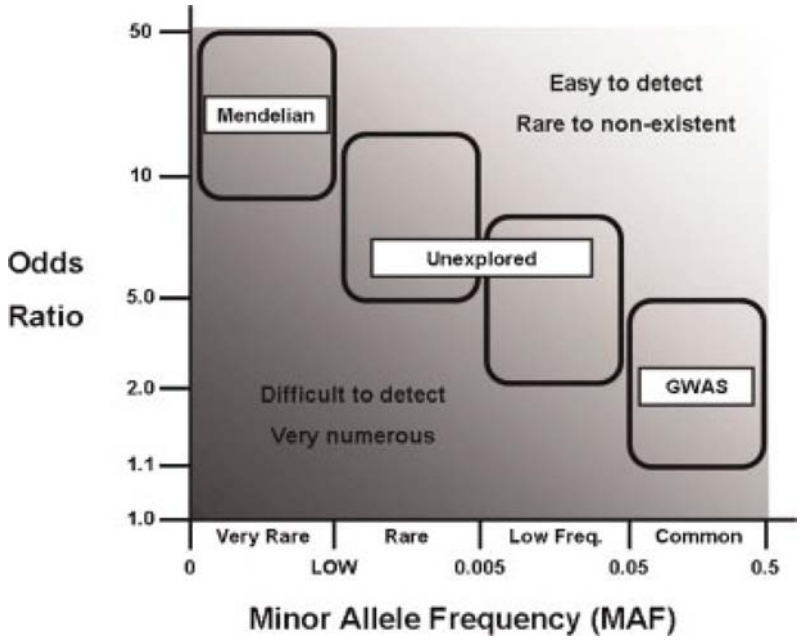

Figure 1.2 Rare and low-frequency variant not detected from GWAS

(Juran *et al.* [5])

1.3. Purpose of this study

The main purpose of this thesis is to develop the statistical methods to identify hidden association signals of genetic variants which were not detected by traditional GWAS. To overcome the limitations of traditional GWAS, two kinds of studies are designed. One is a study to use biological pathway information and conduct a gene-set analysis in GWAS. The other is a study to develop collapsing methods of multiple rare variants from next generation sequencing (NGS) data and conduct a gene-set analysis for rare variants.

In the first study, we proposed a parametric method for gene-set analysis in GWAS. This method is referred as SNP-PRAGE, a SNP-based parametric robust analysis of gene-set enrichment. SNP-PRAGE handles correlation adequately among association measures of SNPs, and minimizes computing effort by the parametric assumption. SNP-PRAGE first obtains gene-level association measures from SNP-level association measures by incorporating the size of corresponding (or nearby) genes and the LD structure among SNPs. Afterward, SNP-PRAGE acquires the gene-set level summary of genes that undergo the same biological knowledge. This two-step summarization makes the within-set association measures to be independent from each other, and therefore the central limit theorem can be adequately applied for the parametric model. We studied the performance of this method via simulation study and apply this method to real GWAS data.

In the second study, a new approach to conduct gene-level rare variant collapsing methods is suggested and parametric gene-set analysis for rare variants in NGS data is conducted. Proposed gene-level association tests first estimate

regression coefficients for quantitative traits in a multiple regression. Obtained regression coefficients are combined based on several different assumptions of disease rare variants. These methods are computationally efficient and have relatively high power in various patterns of multiple rare variants. We also conduct gene-set analysis for these rare variants. For the use of prior biological information about interactions among genes, gene-level statistics are combined using co-mutation-based weight.

1.4 Outline of the thesis

This thesis is organized as follows. Chapter 1 is an introduction of this study including review of GWAS and NGS. Chapter 2 presents the definition of gene-set and overview of gene-set analysis (GSA). Chapter 3 is the study of gene-set analysis for common variants in GWAS. Chapter 4 is the study of gene-level and gene-set-level association tests for rare variants. In both of Chapter 3 and Chapter 4, statistical methods, simulation study, and application to real data are introduced. Finally the summary and conclusion are presented in Chapter 5.

Chapter 2

Overview of gene set analysis (GSA) in microarray and GWAS

2.1 Overview of GSA in microarray

Gene-set analysis (GSA) is originally developed for microarray gene expression data analysis. In the analysis of microarray data, GSA has increased a power of detecting disease-related genes. By focusing on gene sets rather than individual genes, GSA increases a power by combining weak signals from a number of individual genes in a set, when individual genes have weakly associated with traits.

GSA can be classified into two distinct categories: non-parametric or parametric. The most popular non-parametric GSA method is gene set enrichment analysis (GSEA) [16]. GSEA uses the enrichment score (a weighted Kolmogorov-Smrinov statistic) which represents whether the members of gene set tend to occur toward top or bottom in ranked gene list based on a correlation. It permutes the phenotype label and repeats calculating the enrichment score for the test. This requires very expensive computational efforts.

On the other hands, the parametric GSA can reduce computing time by assuming a specific distribution. A hypergeometric distribution-based test [17, 18] is a typical choice in the parametric method and a binomial, a normal, and a chi-square distribution are also widely used [19, 20, 21]. We focus on the parametric analysis of gene set enrichment (PAGE) [20] that has a very simple parametric assumption. PAGE uses the mean of the association measures in a set as a summary measure and assumes that it follows a normal distribution by the central limit theorem when number of genes is large.

GSA can also be divided into competitive GSA and self-contained GSA according to the null hypothesis [22]. Competitive GSA identifies the genes in a gene set showing different pattern of associations with the phenotype compared with the genes in other gene sets. Many competitive methods are based on first identifying genes that are significantly associated with a trait, and then evaluating whether the significantly associated genes tend to cluster in predefined gene-sets. GSEA [16] is typical competitive GSA. The null hypothesis for competitive methods is

H_0: genes in the gene-set of interest are associated with the phenotype as much as genes in total gene-sets.

In contrast to competitive GSA, self-contained GSA is interested only in the genes in a gene set and identifies the gene set contains highly associated genes with a certain phenotype. The null hypothesis for self-contained GSA is

H_0: genes in the gene set of interest are not associated with the phenotype

One limitation of competitive GSA methods is that they cannot be applied to

studies of candidate gene sets for which only genes in the candidate gene set have been measured in expression value. Self-contained methods, on the other hand, can be used for genome-wide studies as well as candidate gene set studies. Because of these fundamental differences between competitive and self-contained methods, the appropriate approach should be selected based on a thoughtful consideration of the null and alternative hypotheses the researcher is interested in testing, and constraints imposed by the available data.

2.2 Overview of GSA in GWAS

During the past few years, many methods for gene-set analysis method have been developed as a complementary approach to GWAS [10-11, 23-26, 36-38]. For simplicity, we call all these methods as GSA-GWA. The workflow of GSA-GWA is presented in Figure 2.1. We address two issues regarding GSA-GWA. The first issue is that there has not been a widely agreed and accepted theory on how to combine the measures of multiple SNPs into one single gene-level measure, and moreover how to combine the gene-level measures into one single gene-set level measure. In original GSA in microarray, the gene-level measure is typically a fold-change or a correlation to represent the effect of a single gene. In GWAS data, however, it is often required to calculate association measures of genes by combining the SNP-level measures. The SNP-level measures include p-values, or chi-square test statistics from the univariate SNP-to-phenotype association tests. Once the SNP-level measure is decided, the gene-level summary statistics are then derived as the highest SNP-level statistics [11], the sum of SNP-level statistics [23],

or the combined p-value [10].

However, there are some substantial limitations in current GSA-GWA methods. First, in deriving the summary statistics the correlation among the SNP-level association measures has not been taken appropriately into account which is expected to play an important role. The SNP-level association measures are usually correlated because the linkage disequilibrium (LD) exists among SNPs. If this correlation is not correctly adjusted, the resulting gene-set-level measure would be inflated [11]. Unfortunately, many GSA-GWA methods have not considered the LD structures adequately.

Second, the computational burden is heavy. Once having the gene-level association measures, it is possible to apply different GSA methods to get various gene-set-level statistics and evaluate their performances. However, as explained later, the majority of GSA-GWA methods implement non-parametric permutation to calculate the observed significance, which takes a heavy computing time.

There have been several efforts to resolve these limitations. As the pioneering work of GSA-GWA, GSEA [16] was extended to GWA data by Wang *et al.* [11], which has been implemented in GenGen package (http://www. openbioinformatics. org/gengen/gengen_download.html). It repeats permutation of sample label and calculation of gene-set statistics 100~1,000 times [11, 23, 24]. This permutation-based testing can preserve a correlation among the SNP-level measures, but this is very computationally expensive in genome-wide scale.

In order to reduce computing time, some GSA-GWA studies use a parametric test. Peng *et al.* [10] used various kinds of the parametric test such as Fisher's combination test, Sidak's combination test, Simes' combination test, and a FDR-based test under the independence assumption of the SNP-level p-values. A GLOSSI method developed by Chai *et al.* [23] used Fisher′s combination test as a

parametric test under the assumption of correlated p-values. Nam *et al.* [25] proposed the Z-statistic method as an extension of the parametric analysis of gene-set enrichment (PAGE) [20].

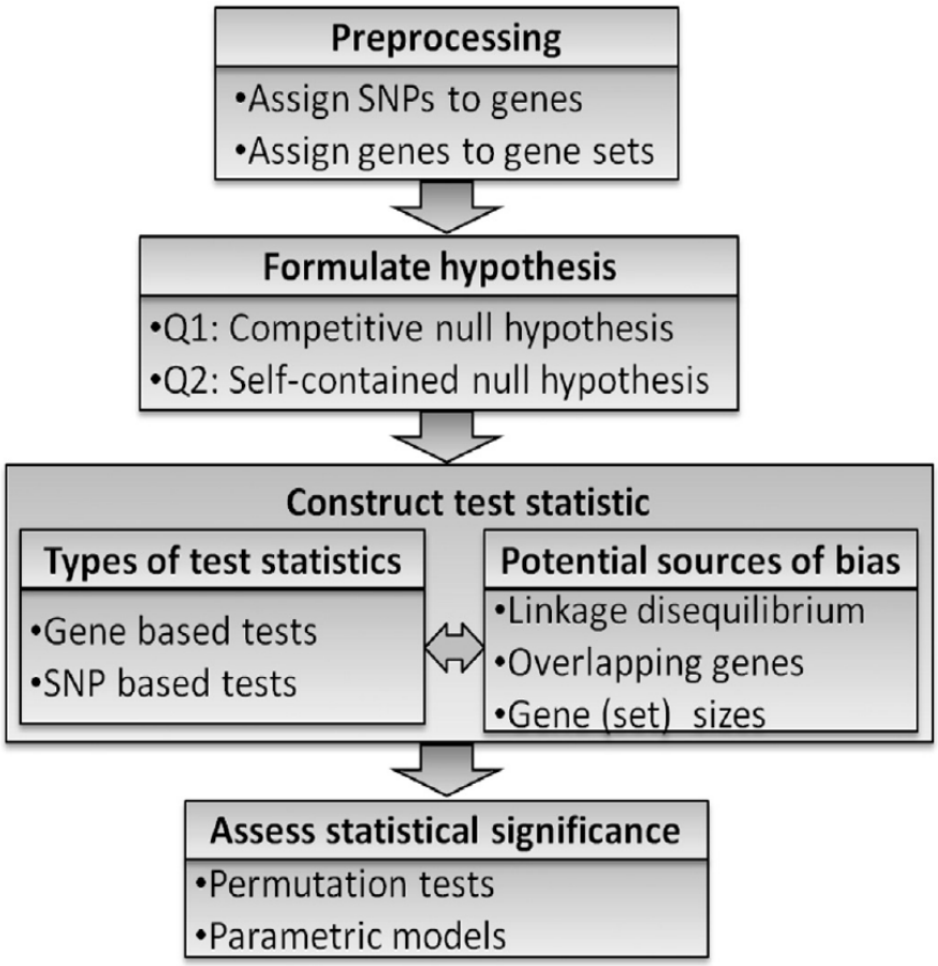

Figure 2.1 Work flow for gene set analysis of GWA dataset [26]

2.3 Gene-set definition

Gene sets are collections of genes with related function or characteristics. For example, gene-sets can be identified from manually drawn pathway maps representing molecular interaction and reaction networks. Gene-sets can be identified based on other criteria, such as a pre-specified region of the genome or similarity of function (eg, genes involved in DNA repair) [27].

A growing number of publically available resources provide descriptions of pathways, along with lists of genes that contribute to the processes making up the pathways. Pathguide (http://www.pathguide.org) [28] lists over 300 databases of information related to pathways, demonstrating the challenge of selecting a pathway resource. Several of these pathway resources, including the KEGG (http://www.genome.jp/kegg/) [29], the Gene Ontology project (http://www.geneontology.org/) [30], MetaCore (http://www.genego.com/metacore.php), BioCarta (http://www.biocarta.com/genes/index.asp), and MsigDB (http://www.broadinstitute.org/gsea/msigdb/index.jsp) are commonly used. Specialized pathway resources, such as the Pharmacogenetics and Pharmacogenomics Knowledge Base for pharmaco-genomics pathways (http://www.pharmgkb.org/), are also available. Additional information on pathway resources and gene-set definition can be found in Bader *et al.* [28], Bard and Rhee [31], and Viswanathan *et al.* [32].

When defining gene-sets for analysis, it is important to clearly state the scope of a gene-set, realizing that knowledge about the genome and definitions of gene-sets are evolving and that no single definition of a gene-set exists. Care should be taken in selecting a reliable ontology resource, as some resources are based on more rigorous curating of gene-sets (eg, KEGG), whereas others provide more

complete listings of biological pathways (eg, MetaCore).

Finally, it is important to recognize that current coverage of genes (and thus gene-sets) is not uniform, as the coverage of genes by SNPs on GWAS arrays is not uniform. This problem will diminish with the development of denser genome-wide SNP arrays, or with use of genotype imputation methods. However, at this point, interpretation of gene-set results should take into account coverage limitations for gene-sets of interest.

Once a set of genes is defined, questions remain regarding which SNPs should be included in the analysis of the gene-set. A commonly used approach is to include any SNP known to map to any gene or within a given distance of any gene, in the gene-set. Although it is not obvious how far up and downstream of each gene should be included in the mapping of SNPs to genes, ideally, the regulatory region(s) of each gene should be included and perhaps even regions in LD with any portion of the gene. Smith *et al.* [33] reported that the degree of disequilibrium for markers separated by ~30 kb in a Caucasian population was similar to the degree of disequilibrium between markers separated by ~10 kb in an African population, with the average level of LD decaying to less than r^2=0.10 after 50 kb. On the basis of these considerations, SNPs within 20-50 kb from the first and last exon should be included as part of a gene for GSA to cover the regulatory regions of the gene, as well as SNPs in LD with the gene.

Currently gene sets usually consist of SNPs in, or near, genes thought to contribute to a particular biological process. However, the definition of a gene set could be extended to use other knowledge related to gene function. For example, mRNA expression data has been used by Zhong *et al.* [34] to define gene sets that include eSNPs, that is, SNPs that have been shown to regulate the expression of a particular gene in either a cis- or trans-acting manner.

Chapter 3

A SNP-based parametric robust analysis of gene-set enrichment

3.1 Introduction

Recently, genome-wide association studies (GWAS), which typically test disease associations with half to a few million single nucleotide polymorphisms (SNPs) across the human genome in hundreds to thousands of samples, have successfully identified many genetic variants contributing to the susceptibilities of complex diseases. However, the variants identified so far, individually or in combination, account for only a small proportion of the inherited component of disease risk [13]. A possible explanation is that due to the large number of genetic polymorphisms examined in GWAS and the massive amount of tests conducted, real but weak associations are likely to be missed after multiple comparison adjustment (e.g., corrected by half a million tests in a typical GWAS).

To help prioritize association signals from GWAS and to better understand the biological themes underlying complex diseases, gene set analysis (GSA) has become increasingly popular. Instead of conducting analysis for single SNPs or single genes, GSA tests disease association with genetic variants in a group of functionally related genes, such as those belonging to the same biological pathway. GSA combines weak signals from a number of individual genes in a set, when individual genes are weakly associated with the traits. In this way, GSA increases a power of detecting disease-related genes and helps to interpret underlying genetic background

There are several prior works for applying GSA methods to GWA data (GSA-GWA). There are some substantial limitations in current GSA-GWA methods. First, in deriving the summary statistics the correlation among the SNP-level association measures has not been taken appropriately into account which is expected to play an important role. The SNP-level association measures are usually correlated because the linkage equilibrium (LD) exists among SNPs. If this correlation is not correctly adjusted, the resulting gene-set-level measure would be inflated. Unfortunately, many GSA-GWA methods have not considered the LD structures adequately.

Second, the computational burden is heavy. Once having the gene-level association measures, it is possible to apply different GSA methods to get various gene-set-level statistics and evaluate their performances. However, as explained later, the majority of GSA-GWA methods implement non-parametric permutation to calculate;pe the observed significance, which takes a heavy computing time.

The permutation-based testing can preserve a correlation among the SNP-level measures, but this is very computationally expensive in genome-wide scale. In order to reduce computing time, some GSA-GWA studies use a parametric test.

Peng *et al.* [10] used various kinds of the parametric test such as Fisher's combination test, Sidak's combination test, Simes' combination test, and a FDR-based test under the independence assumption of the SNP-level p-values. A GLOSSI method developed by Chai *et al.* [23] used Fisher's combination test under the assumption of correlated p-values.

Recently, Nam *et al.* [25] proposed the Z-statistic method that compares a specific gene-set to others. This method is the extension of the parametric analysis of gene set enrichment (PAGE) [20], which is the parametric and competitive GSA for microarray data. PAGE uses the mean of the association measures in a set as a summary measure and assumes that it follows a normal distribution by the central limit theorem when the number of genes is large.

However, these parametric methods including the Z-statistic method do not consider the LD structures adequately and assume no correlation between SNP-level p-values. In order to overcome these limitations of current GSA-GWA, we propose SNP-PRAGE [35], a SNP-based parametric robust analysis of gene-set enrichment, which is based on a simple normality assumption. SNP-PRAGE estimates the LD information among SNPs based on LD block-wise covariance structure to consider the correlation among SNP-level measures without taking the permutation step.

We compare our method to other GSA-GWA methods via the simulation study in terms of size, power and computing time. We also demonstrate SNP-PRAGE using two GWA data sets: hypertension data of 8,842 samples from the Korean population and bipolar disorder data of 4,806 samples from the Wellcome Trust Case Control Consortium (WTCCC).

3.2 Methods

3.2.1 Z-statistic method (GSA-SNP)

Nam *et al.* [25] implemented the Z statistic method in their software, GSA-SNP. The negative logarithm of the mth best p-value within each gene was used as the gene summary measure. Based on this gene summary measure, the Z-score was then calculated as gene-set-level summary. The Z-score was assumed to follow a normal distribution based on the central limit theorem (CLT).

$$t_{ij} = -\log(m\text{th best p - value in } j\text{th gene in } i\text{th set})$$

$$t_{ij} \sim i.i.d.\ (\mu_i, \sigma^2)$$

$$\bar{t}_{i\cdot} = \sum_j t_{ij} / N_i, \qquad \hat{\sigma} = \sqrt{\sum_i \sum_j (t_{ij} - \bar{t}_{\cdot\cdot})^2 / (N-1)}$$

$$Z_i = \frac{\bar{t}_{i\cdot} - \bar{t}_{\cdot\cdot}}{\hat{\sigma} / \sqrt{N_i}} \sim N(0,1) \quad \text{by CLT}$$

where N_i is the number of genes in ith set and N is total number of genes

In order to meet a normal distribution assumption, the gene-level order statistic is assumed to have an identical and independent distribution (i.i.d.). Let nij be the gene size which is the number of SNPs within the jth gene in the ith gene set. If we assume a p-value follows an independent uniform distribution, the mth order p-value $p_{(m)}$ follows a beta distribution with the mean $m/(n^{ij}+1)$ and the variance $m(n_{ij}-m+1)\ /\{(n_{ij}+1)^2(n_{ij}+2)\}$. This means that the gene with many SNPs have a lower $p_{(m)}$ than genes with a few SNPs. (See Figure 3.1(a).) So $p_{(m)}$ is not identically distributed over the gene size. To satisfy the identical distribution

assumption, the summary measures need some modifications.

The gene-level summary measure is also assumed to have a homogeneous variance. However, the variance of their summary measures also depends on the gene size. When the gene size is large, the variance of the summary measure of the gene tends to be small. This problem can be easily addressed by modifying Welch's *t* statistic [39] which is designed to handle for the heterogeneous variance of the two groups.

3.2.2 SNP-PRAGE

To address these issues of the Z-statistic method we mentioned above, we multiply $p_{(m)}$ by $(n_{ij}+1)$ to have an approximate identical distribution over the gene size. The moment generating function of $(n_{ij}+1)p_{(m)}$ does not depend on the gene size n_{ij} when p-values are independent from each other and n_{ij} is large enough. However, the SNP-level p-values are not independent from each other because of the LD structure. So $(n_{ij}+1)p_{(m)}$ has a non-identical distribution over the genes (See Figure 3.1(b).)

In SNP-PRAGE, we propose using the effective gene size n_{ij}^{*} instead of gene size n_{ij} to make sure that the gene-level summary measure has an approximate identical distribution over the gene size irrespectively of correlation among p-values. The effective gene size is computed by using the following equation [40].

$$n_{ij}^{*} = \frac{Var(\overline{p}_{IID})}{Var(\overline{p}_{CORR})} n_{ij}.$$

$Var(\overline{p}_{IID})$ is estimated under the independent covariance structure and $Var(\overline{p}_{CORR})$ under the haploblock-wise compound symmetric covariance structure.

Note that SNP-level measures within a LD block are correlated. The within-gene covariance matrix can be estimated by using maximum likelihood (ML) estimation. Among the several candidate covariance structures, the Akaike information criterion is used to choose the most appropriate covariance structure [41]. First, we construct the LD block among SNPs in GWA data so that any pair of SNPs from different LD blocks is independent from each other ($r^2 \leq 0.05$) [42]. Second, we obtain the ML estimator of the covariance matrix within the LD block for each gene set. The most appropriate covariance structure is then selected via AIC. In the Korean GWA data analysis, the LD-block-wise compound symmetric structure (LD-CS) was chosen as the appropriate covariance structure.

Within the gene, the highly ranked p-values tend to be correlated because of the LD structure. Through the simulation study, we found that the average of the top m p-values from a gene is a more robust gene-level summary measure than only the mth p-value (data are not shown). The following is the final gene-level summary measure proposed in SNP-PRAGE. In Figure 3.1(c), we can see this measure has the identical mean over the gene size.

$$t_{ij}^{*} = \frac{p_{(1)} + p_{(2)} + \cdots p_{(m)}}{m} \times (n_{ij}^{*} + 1)$$

However, our empirical study shows that gene-level measure t_{ij}^{*} does not have the common variance over the gene set especially with the small gene set size (See Figure 3.2). Thus, we assume that the gene-level measure has a heterogeneous variance over the gene sets:

$$t_{ij}^{*} \sim \text{i.i.d } (\mu_i, \sigma_i^2) \text{ for the } i\text{th set.}$$

The mean of the gene-level measures in a gene set follows a normal

distribution by the central limit theorem.

$$\bar{t}_{i\cdot\cdot} = \sum_j t^*_{ij} / N_i \sim N(\mu_i, (\sigma_i^2 / N_i))$$

We compute the sample variance distinctly over gene set and derive the following set-level test statistic:

$$T_i = \frac{\bar{t}_{i\cdot} - \bar{t}_{\cdot\cdot}}{\sqrt{s_i^2 / N_i + s^2 / N - 2 s_i^2 / N}} \sim t(df_i), \quad i = 1, \ldots, p$$

$$\text{where } s_i^2 = \sum_j (t_{ij} - \bar{t}_{i\cdot})^2 / (N_i - 1),$$

$$\bar{t}_{\cdot\cdot} = \sum_i \sum_j t^*_{ij} / N, \; s^2 = \sum_i (N_i - 1) s_i^2 / (N - p)$$

The degree of freedom (df_i) is computed by Welch-Satterthwaite equation [39].

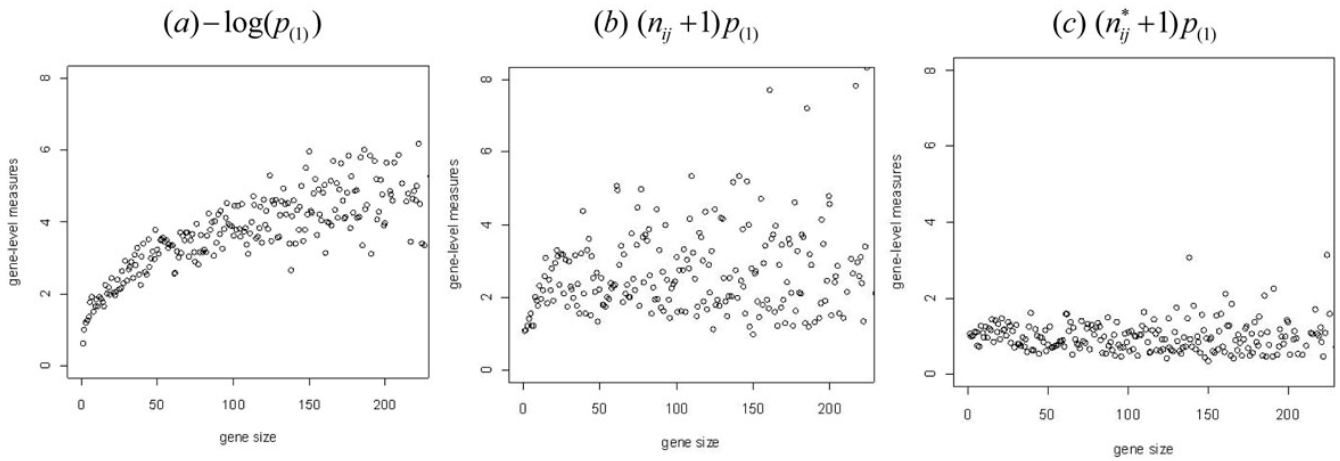

Figure 3.1 Distribution of gene-level measures over the gene size for hypertension data from Korean population The x-axis is gene size which is a number of SNPs within the gene and the y-axis is mean of gene-level summaries with the same gene size.

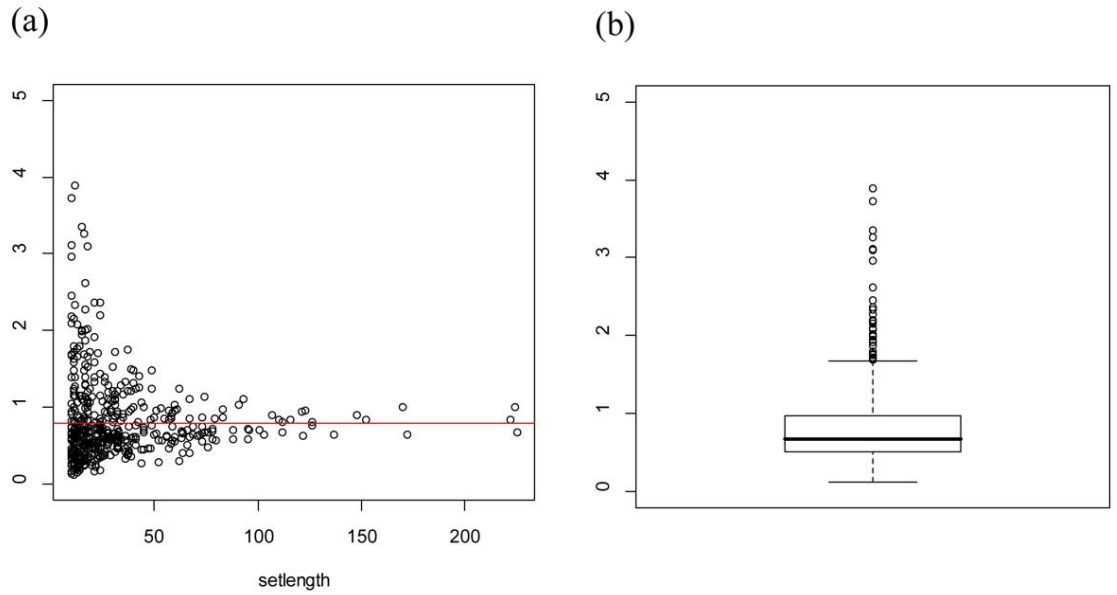

Figure 3.2. Variance of gene-level measure over the gene sets In the left plot (a), x-axis is gene set size (= number of genes in the gene set) and y-axis is sample variance of gene-level summaries in the gene set. The right plot (b) shows a boxplot of variance of gene-level measures. A red line represents total sample variance in the data.

3.3 Results

3.3.1 Hypertension data from the Korean GWA study

We used canonical pathways from MsigDB database (http://www.broadinstitute.org/gsea/msigdb/index.jsp). These canonical pathways are curated from other online database such as BioCarta, KEGG and GO and so on. MsigDB database contains 639 pathways and 4934 genes.

We applied SNP-PRAGE to GWA data set from the Korean GWA study which was initiated in 2007 to undertake a large-scale GWA analysis among 10,038 participants (aged between 40 and 69) of Ansung (n=5,018) and Ansan (n=5,020) population-based cohorts [43]. These cohorts, established as part of the Korean Genome Epidemiology Study (KoGES) in 2001 provide extensive phenotypic data for over 260 traits, but here we focus on analyses of hypertension. From the total of 10,038 participants, DNA was available for 10,004, all of whom were genotyped with the Affymetrix Genome-Wide Human SNP array 5.0 and the Bayesian Robust Linear Modeling using Mahalanobis Distance (BRLMM) algorithm. Markers with high missing gene call rate (>5%), low MAF (<0.01) and significant deviation from Hardy-Weinberg equilibrium ($P < 1 \times 10\text{-}6$) were excluded, leaving 352,228 SNPs. After removing samples with low call rates (< 96%, n = 401), sample contamination (n = 11), gender inconsistencies (n = 41), cryptic relatedness (n = 608) and serious concomitant illness (n = 101), GWA genotypes from 8,842 individuals were included. Hypertension phenotype was defined as a systolic blood pressure (SBP) ≥ 140 mm Hg or a diastolic blood pressure (DBP) ≥ 90 mm Hg. The logistic regression analysis with an additive model (1 *d.f.*) is conducted after adjustment for age, sex, and recruitment area (i.e. Ansung and Ansan). To correct for stratification, some methods that infer genetic ancestry, such as principal component analysis (PCA) and structured association can be used [44]. In our GWA data, there is no evidence of population stratification.

We obtained the SNP ID, rs ID, position information from dbSNP build 128 and gene ID, gene name, and position information from NCBI build 36. Each SNP is mapped to a gene closest to it. Only SNPs located within 500 Kb upstream or downstream of a gene are considered, because most enhancers and repressors are less than 500Kb away from genes, and

most LD blocks are within 500Kb [11]. As a result of mapping it covered 60% of all SNPs in our data. If the mapping range is larger, we could save more SNPs, but the risk of SNP's mapping to shared region of overlapping genes also increases.

Our proposed SNP-PRAGE was used to identify the significant gene sets associated with hypertension in Korean GWA data. We used the p-value from the logistic regression as SNP-level association measure for each SNP. We compared three kinds of gene-level measures, $-\log(p_{(1)})$, $(n_{ij}+1)p_{(1)}$, and $(n_{ij}^{*}+1)p_{(1)}$.

In Figure 3.1, each plot shows the mean of gene-level measures over the gene size. Figure 3.1(a) is from the gene-level measure used in the Z-statistic method. This measure tends to increase as the gene size increases. The non-causal gene set which has a larger number of genes tend to be detected as significant. Figure 3.1(b) shows the minimum p-values within the gene multiplied by (gene size + 1) over the gene size. Figure 3.1(c) shows the same plot but uses the effective gene size instead of the actual gene size. Figure 3.1(c) is most robust to the gene size showing the constant pattern.

Next, we checked the homogeneity assumption of variance of gene-level measure (t_{ij}^{*}) over the gene sets. Figure 3.2 represents whether t_{ij}^{*} has the homogeneous variance over the gene sets. We can see the sample variances are different over the gene sets especially for the gene sets with a small number of genes. Thus, it would be inappropriate to assume the homogeneous variance assumption for the gene-level measures. SNP-PRAGE allows the heterogeneous variance of gene-level measure.

Table 3.1 KARE result: Top 5 gene sets with smallest q-value associated with hypertension phenotype from Z-statistic method

Gene set	No. genes	No. SNPs	p-value	q-value
ST_JNK_MAPK_PATHWAY	36	2410	1.13E-04	6.38E-02
HSA00563_GLYCOSYLPHOSPHATIDYLINOSITOL_ANCHOR_BIOSYNTHESIS	18	700	2.67E-04	6.57E-02
FASPATHWAY	28	1489	8.82E-04	1.44E-01
HSA05060_PRION_DISEASE	117	762	2.42E-03	2.53E-01
HSA04520_ADHERENS_JUNCTION	64	4150	2.58E-03	2.53E-01

Table 3.2 KARE result: Top 5 gene sets with smallest q-value associated with hypertension phenotype from SNP-PRAGE

Gene set	No. genes	No. SNPs	p-value	q-value
ST_JNK_MAPK_PATHWAY	36	1701	2.40 E-05	9.48 E-03
ST_ERK1_ERK2_MAPK_PATHWAY	24	1765	1.61 E-04	3.16 E-02
HSA05214_GLIOMA	52	2301	3.92 E-04	5.16 E-02
HSA05050_DENTATORUBROPALLIDOLUYSIAN_ATROPHY	14	997	7.97 E-04	7.57 E-02
EXTRINSICPATHWAY	13	579	9.58 E-04	7.57 E-02

In order to handle multiple testing problems, the false discovery rate (FDR) was controlled [45]. The q-values were calculated to guard against the cost of multiple hypothesis testing [46]. The q-value provides an expected proportion of false positives among sets with unadjusted p-values at least as extreme as the current set of interest. Single SNP association test based on a logistic regression cannot detect SNP whose q-value is less than 0.05. Minimum SNP-level p-value is 2.043×10^{-6} and corresponding q-value is 0.4. Even though there is no significant SNP-level association in terms of q-values, multiple SNPs with moderate effects could affect the phenotype in the gene set-level.

Table 3.1 and Table 3.2 summarize the top 5 gene sets obtained by using the Z-statistic method and SNP-PRAGE, respectively. In Z-statistic method, minimum q-value is 0.06, which is not significant if we use 0.05 as q-value cut-off. SNP-PRAGE yielded 2 significant gene sets (q-values: 0.01, 0.03) based on q-value 0.05 as cut off, while Z-statistic method did not yield any significant gene sets.

The significant gene sets in SNP-PRAGE are ST_JNK_MAPK_Pathway and ST_ ERK1_ERK2_MAPK_Pathway. The MAPK signaling pathway is known to ultimately result in the dual phosphorylation and activation of terminal kinases, such as p38, c-Jun N-terminal kinases (JNKs), and extracellular signal-regulated kinases (ERK1/2 and ERK5), which are related to pressure-overload–induced cardiac hypertrophy [47]. Esposito *et al.* [47] mentioned the potential role of ERK activation in White Blood Cells (WBCs) as a novel molecular marker to identify uncontrolled human hypertension. In their study, JNK1 activation was also significantly induced in uncontrolled hypertension patients.

3.3.2 Bipolar disorder data from the WTCCC GWA study

We also applied SNP-PRAGE to bipolar disorder (BD) data from the Wellcome Trust Case Control Consortium (WTCCC) which was established in 2005 to conduct GWA analysis for group of 50 research groups across the UK [48]. In our analysis, 1868 BD cases and 2938 controls were included and markers with high missing gene call rate (>5%), low MAF (<0.05) and significant deviation from Hardy-Weinberg equilibrium ($P < 5.7 \times 10^{-7}$) were excluded, leaving 354,093 SNPs. The logistic regression analysis with an additive model (1 *d.f.*) was

conducted after adjustment for age, sex, region, and age × region.

SNP-PRAGE yielded 3 gene sets significantly associated with BD in terms of q-value at the 5% significance level (Table 3.3), while Z-statistic method did not detect any significant gene set (Table 3.4). The significant gene sets detected by SNP-PRAGE are AGPCR pathway, DREAM pathway, and CK1 pathway.

AGPCR pathway is G-protein coupled receptors (GPCRs) signaling pathway which transduces extracelluar signals across the plasma membrane. In a genome-wide linkage survey, the region of chromosome 22q12 containing the GRK3 gene was identified as a susceptibility locus for BD in humans and GRK3 is expected to play an important role in the regulation of any one of many GPCRs [49]. DREAM is a multifunctional Ca^{2+}-binding protein that can act as a transcriptional repressor for the prodynorphin gene. Subjects with BD were reported to show reduction of prodynorphin mRNA expression in discrete nuclei of the amygdaloid complex [50]. CK1 pathway is well known to be related to the circadian clock. Deregulation of this clock is involved in several human disorders. As a potent CK1 inhibitor, a imidazole derivative, PF-670462 could be used for therapy of cognitive deficits in mood changes in bipolar disorders [51].

Table 3.3 WTCCC Result: Top 5 gene sets with smallest q-value associated with bipolar disorder phenotype from Z-statistic method

Gene set	No. genes	No. SNPs	p-value	q-value
EICOSANOID_SYNTHESIS	15	669	6.85E-04	0.33
HSA04510_FOCAL_ ADHESION	171	10281	2.50E-03	1.00
HSA01030_GLYCAN_STR UCTURES_BIOSYNTHESI S _1	91	7475	4.01E-03	1.00
BADPATHWAY	17	1045	4.91E-03	1.00
HSA05223_NON_SMALL_ CELL_LUNG_CANCER	43	2933	5.49E-03	1.00

Table 3.4 WTCCC Result: Top 5 gene sets with smallest q-value associated with bipolar disorder phenotype from SNP-PRAGE

Gene set	No. genes	No. SNPs	p-value	q-value
AGPCRPATHWAY	12	616	5.2E-05	1.45E-03
DREAMPATHWAY	13	600	8.5E-05	1.45E-03
CK1PATHWAY	16	1079	3.1E-04	3.52E-03
BIOGENIC_AMINE_SYNT HESIS	16	914	1.0E-03	8.52E-03
BADPATHWAY	21	1045	5.6E-03	1.51E-01

3.4 Simulation study

In order to compare the performance of SNP-PRAGE with other GSA-GWA methods, we conducted the simulation study. Simulation data was generated based on a real GWA data. Using the subset of 5 gene sets from MsigDB canonical pathways, we constituted 5 gene sets so that each set has 20 genes. Over the gene sets, we varied the gene size which is the number of SNPs within a gene in order to study the effect of gene size on the gene set analysis. For example, one gene set consists of a small number of genes and other gene set consists of a large number of genes. The range of gene size is from 9 to 49 SNPs. Among 5 gene sets, we chose one causal gene set and selected 5 causal genes within the causal gene set. 500 individuals are randomly generated. For each causal gene, we selected one causal SNP whose minor allele frequency is about 0.2 for the selected individuals.

Given the genotype information of causal 5 SNPs and effect sizes, the case/control status was generated. Let SNP_{ij} denote *j*th causal SNP in *i*th individual and denotes effect size (=log odds ratio). Effect size of each causal SNP is given as 0, 0.3, or 0.6.

$$\text{logit}\{\Pr(Y_i = 1)\} = \sum_j \beta_j SNP_{ij}$$

Simulated gene sets and their gene sizes are given in Table 3.5. Either set 1, set 3, or set 5 is used as the causal gene set. For each causal get set, 1000 simulation datasets were generated for the effect size 0 to compute type I error and 100 simulation datasets for effect size 0.3 and 0.6 to compute powers.

In order to determine whether or not the central limit theorem works for relatively small gene set, we obtained a null distribution of set-level summary for reduced number of genes, say 5 and 10. We randomly chose 5 or 10 genes among 20 genes for each set. Figure 3.3 shows that the set level summary of small gene set follows a normal distribution when the number of genes is 10 and 20. However, there is a violation of normal approximation when the number of genes is 5. Thus, we expect that SNP-PRAGE would work well when the number of genes is at least 10. For practical applications, we recommend discarding the gene sets in the analysis if the number of genes is smaller than 10.

Table 3.5. Simulated gene set based on MsigDB pathways

Simulated gene set	No. genes	Gene Size	Reference gene set
Set1	20	9~12 SNPs	HSA04060_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION
Set2	20	12~20 SNPs	HSA04010_MAPK_SIGNALING_PATHWAY
Set3	20	20~30 SNPs	HSA04810_REGULATION_OF_ACTIN_CYTOSKELETON
Set4	20	26~40 SNPs	HSA04510_FOCAL_ADHESION
Set5	20	36~49 SNPs	HSA04080_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION

We compared the performance of SNP-PRAGE, Z-statistic method (Nam *et al.* [25], modified GSEA method (Wang *et al.* [11]) and GLOSSI (Chai *et al.* [23]). We used the GenGen package (http://www.openbioinformatics.org/gengen/ gengen_download.html) for GSEA and the R package for other methods. SNP-PRAGE, Z- statistic method and GLOSSI use parametric test and GSEA method use nonparametric test with 1000 permutations. GLOSSI permute the data 100 times to consider the correlation of p-values resulting from LD among SNPs.

Type 1 error is defined as the proportion of cases whose p-values is less than the significance level when the effect size of causal SNP is zero. Power is defined as the proportion of cases whose p-value is less than the significant level when effect size of causal SNP is 0.3 and 0.6. Tables 3.6 and 3.7 summarize the type 1 error and power of the methods compared.

Type 1 error and power of the Z-statistic depend largely on the gene size. When the causal gene set consisted of the genes with 9~12 SNPs, the Z-statistic method yielded low type 1 error and power. They tended to decrease, as m increased. We think it is because the genes with the smaller number of SNPs tend to have a larger minimum p-value and weaker LDs than the genes with a larger number of SNPs. When the causal gene set consists of the genes with 36~49 SNPs, on the other hand, the Z-statistic method yielded very high type 1 error and power. They tended to increase, as m increases. So the results from Z-statistic method can have high false positive errors, especially when the gene set has a larger number of genes.

Table 3.6. Type 1 error (when effect size is 0) in simulation studies

Causal gene Set	Geneset Size	Gene Size	Significance level	Z-statistic method					SNP-PRAGE					GLOSSI	GSEA
				m					*m*						
				1	2	3	4	5	1	2	3	4	5		
Set1	20 genes	9~12 SNPs	0.05	.005	.003	.004	.004	.003	.057	.053	.054	.054	.053	.052	.051
			0.01	.002	.002	.003	.002	.001	.013	.009	.010	.011	.010	.010	.011
Set2	20 genes	20~30 SNPs	0.05	.083	.087	.084	.080	.080	.051	.052	.052	.050	.052	.051	.049
			0.01	.033	.035	.034	.031	.031	.011	.011	.009	.008	.008	.009	.010
Set3	20 genes	36~49 SNPs	0.05	.430	.641	.760	.864	.891	.047	.049	.050	.050	.051	.049	.052
			0.01	.144	.294	.429	.634	.671	.008	.010	.010	.011	.011	.011	.012

Table 3.7 Power (when effect size is 0.3 or 0.6) in the simulation studies

Effect size (=β)	Causal gene set	Gene set size	Gene size	significance level	Z-statistic method					SNP-PRAGE					GLOSSI	GSEA
					M					*m*						
					1	2	3	4	5	1	2	3	4	5		
0.3	Set1	20 Genes	9~12 SNPs	0.05	.81	.81	.74	.67	.59	.92	.92	.94	.95	.95	.92	.95
				0.01	.78	.75	.66	.55	.38	.90	.91	.92	.92	.91	.89	.91
	Set3	20 Genes	20 ~30 SNPs	0.05	.85	.78	.79	.79	.76	.81	.81	.83	.82	.83	.82	.83
				0.01	.76	.75	.74	.74	.73	.71	.73	.73	.74	.74	.72	.71
	Set5	20 Genes	36~49 SNPs	0.05	.98	.99	.99	.99	.99	.74	.74	.75	.75	.76	.74	.73
				0.01	.95	.97	.97	.99	.98	.61	.62	.62	.62	.63	.60	.61
0.6	Set1	20 Genes	9~12 SNPs	0.05	.84	.83	.78	.69	.62	.97	.98	.98	.98	.97	.98	.98
				0.01	.80	.75	.69	.60	.48	.94	.95	.97	.97	.96	.96	.97
	Set3	20 Genes	20 ~30 SNPs	0.05	.86	.89	.86	.88	.88	.84	.85	.86	.86	.87	.84	.85
				0.01	.78	.82	.79	.80	.79	.75	.74	.75	.75	.76	.73	.74
	Set5	20 genes	36~49 SNPs	0.05	1.0	.99	1.0	1.0	1.0	.79	.80	.80	.82	.82	.79	.79
				0.01	.99	.97	.99	.99	.99	.69	.71	.72	.72	.73	.69	.68

Table 3.8 Computing time for simulation data analysis

Process	Z-statistic method	SNP-PRAGE	GLOSSI (100 permutations)	GSEA (1000 permutations)
Single SNP analysis	40sec	40sec	34 min	26 min 15sec
Gene set analysis	0.3 sec	52sec	0.5sec	2 min 10sec
Total analysis	40.3sec	1min 32sec	34 min 0.5sec	28 min 25sec

On the other hand, SNP-PRAGE gave the consistent results irrespective of gene size. As *m* goes from 1 to 5, SNP-PRAGE gets a little larger power. Based on these results, it is desirable to use the mean of top *m* p-values instead of the minimum p-value as the gene-level measure. If the top *m* p-values are from the SNPs in LD, the method using the top *m* p-values can yield larger power than that using only the minimum p-value. The computed power based on SNP-PRAGE with appropriate m was similar but slightly larger compared to one of GLOSSI and GSEA. In SNP-PRAGE, type 1 error is near 0.05 at the significance level 0.05. Table 3.8 summarizes the computing time of each method. Z-statistic method has the fastest computing time, because LD structure between SNPs is not taken into account. SNP-PRAGE has the fastest computing time among the methods which consider LD between SNPs, Specifically, our simulation results show that GSEA and GLOSSI methods take 18.5 and 22.1 times, respectively, of computational efforts than SNP-PRAGE.

The single SNP analysis for the Korean GWA data requires more than 1000 computing time compared to one set of simulation data. So, it would take a very long period of time if GSEA and GLOSSI are applied to our data, because both methods require permutation process. Thus, in practice it would not be easy to handle a large scale GWA data by GSEA and GLOSSI.

3.5 Conclusion

The power of SNP-PRAGE was computed for the several choice of *m*. When we choose appropriate m for the gene-level summary, the computed power based

on SNP-PRAGE was similar but slightly larger compared to one of GLOSSI and GSEA in the simulation study. Then how can we choose the appropriate m for the gene-level summary?

The best choice for the number of the top p-values used in gene-level summary depends on the LD structure among the SNPs within the causal genes. While we set a fixed m over the genes for the summary in SNP-PRAGE, setting different m over the genes according to each effective gene size can be considered in the future study.

Our SNP-PRAGE can be extended in several ways. In this study, we assume the gene sets are independent from each other. However, the gene sets often share some common genes because one gene can have multiple biological functions. So handling the overlapped common genes between gene set is another challenging issue.

SNP-PRAGE method is based on a normal distribution and similar to ANOVA (Analysis of Variance) model. In fact, SNP-PRAGE can be expressed as ANOVA model with some contrast and modified estimation of variance. As an extension, another well-defined parametric model can be applied. A nested ANOVA can be applied to the gene set analysis in terms of that gene effect is nested within gene-set effect. A mixed effect model can also be applied by treating the gene specific effects as random effects. Addressing these challenges we expect a more powerful GSA-SNP method in our near future.

3.6 Discussion

Single SNP analysis in GWAS offers only a limited understanding of complex diseases because the complex disease often arises from the joint action of multiple genetic variants. Single SNP analysis can find only a few most significant SNPs. GSA-GWA increases the power to detect the genetic variants which have a weak association but a meaningful biological association with a phenotype. GSA-GWA methods test the significance of gene set via permutation by generating permuted data more than hundred times, which requires expensive computational efforts. The use of a parametric test can reduce the computing time, because it needs to calculate the gene set statistic only once.

We compared the performance and computing time of three parametric test-based GSA-GWAs (Z-statistic method, GLOSSI, SNP-PRAGE) and one nonparametric test-based GSA-GWA (GSEA) in simulation study. The Z-statistic method does not consider the LD and has the shortest computing time but may have lots of false positive results because of overestimated gene set statistics when the gene set has many large genes. GLOSSI uses a parametric test but it needs to permute phenotype 100 times for an estimation of the correlation between association measures and GSEA requires much more permutations than GLOSSI. SNP-PRAGE reduces computing time much and has comparable performance to GLOSSI and GSEA without going through the permutation step.

We found that consideration of LD blocks between SNPs helps us to deal with the correlation between p-values more appropriately. The approach based on the mean of top m p-values provides more consistent and stable result than the approach based on the top mth p-value. Multiplying the effective gene size to the

minimum p-value for the gene-level summary of SNP-PRAGE can reduce the false positive errors when the gene size is large. We expect the SNP-PRAGE to play an important role in the parametric gene set analysis of large-scale GWA data.

.

Chapter 4

Gene-level and Gene-set-level association test of rare variants

4.1 Introduction

Genome-wide association studies (GWAS) have focused on the association between common complex traits and common genetic variants, and have reported an extensive list of the findings: single nucleotide polymorphisms (SNPs) or genes associated with traits. However, the associated variants can explain only a small fraction of the heritability of most common traits. This suggests that other genetic mechanisms, such as gene-gene interaction, gene-environmental interaction, joint play of multiple rare variants, and gene-set-level action could contribute to disease susceptibility.

Complex diseases often result from compound action of multiple risk factors and therefore the single-SNP-based analysis in GWAS may miss the genetic

variants that affect risk effects jointly but have scarce effects individually. In this reason, it was suggested to utilize prior biological knowledge or known pathway information, and thus to incorporate a set of related SNPs, which leads a smaller number of tests [11]. Gene set analysis (GSA) focuses on gene sets rather than individual genes or SNPs, and combines weak signals in a set. GSA can increase power of detecting disease association signals and help to interpret underlying biological background.

Rare variants or variants with a low minor allele frequency are not captured in traditional GWA strategies and are potential source for the missing heritability of common diseases [13]. Recent studies have shown that multiple rare variants could contribute to common diseases [52, 53]. Although each rare variant may has a small overall association on disease, the aggregate effects of these variants may yield a meaningful and strong association signal [54].

Since the frequencies of rare variants are very low, even with high penetrance, it will be difficult to detect association with any single rare variants. This has motivated the development of new statistical tests for detecting signals of rare variants over the past few years. These methods often employ the idea of collapsing multiple rare variants within a region. For example, the combined multivariate and collapsing (CMC) method [55] collapse variants within subgroups based on minor allele frequency (MAF) and test collapsed variants with Hotelling's T^2. These collapsing methods have good performance when rare variants within a region have mild to moderate association signals with the same direction even if there is no single variant of strong association signal. The assumption behind these methods is that the probability of being diseased is proportional to the number of rare minor alleles [15]. However, this might not explain the etiology of rare variants fully.

In some case, very small portion of rare variants might strongly affect the disease phenotype and in other cases, there could be protective alleles and deleterious alleles together within the same region. In order to cover these scenarios, there have been several methods which collapse rare variants with different weights across alleles [56] or focus on the changes in distribution of rare variants' effects [57, 58]. Weights assigned according to minor allele frequency or functional importance measures from SIFT [59], Polyphen2 [60], LRT [61] can help us to light a few candidates of high risk variants when a region include many noncausal rare variants. Distribution-based approaches such as C-alpha and SKAT have a great advantage when signals of different directions are mixed up in a region. Lee *et al.* [62] developed SKAT-O which combine a collapsing method and SKAT.

However, Ladouceur *et al.* [15] demonstrated that assessing the association between rare variants and complex diseases is still a challenging task, and no single method gives consistently acceptable power across the range of these relationships, even in a large sample size. Besides, gene set analysis for rare variants has not been well investigated yet in both a simulation study and real data application.

As another breakthrough to uncover the missing heritability, it was suggested to utilize prior biological knowledge or known pathway information, and thus incorporate a set of related SNPs, which leads to a smaller number of tests. Gene-set analysis (GSA) focuses on gene sets rather than individual genes or SNPs, and combines weak signals in a set. GSA can increase power of detecting disease association signals and help to interpret underlying biological background. So far, the gene set analysis for rare variants has seldom been investigated in both simulation studies and real data application.

For the consistently powerful association method under the various scenarios

for rare variants, we propose some quadratic tests (QTests; Q_1, Q_2, and Q_3) for gene-level focusing on quantitative traits. Also, in order to increase power to detect the genetic association from rare variants, we propose the QTest for the gene-set ($QTest_{GS}$) by extending the unit of analysis from genes to gene-sets. This association test can cover a broad range of scenarios for joint action of disease variants including the existence of common disease variants. We demonstrate the performance of the proposed methods comparing with other gene-level and gene-set-level association methods in various simulation setting. These include collapsing methods (GRANVIL [64], Variable Threshold method [65]), a weighting method (Weighted sum statistic [56]), a distribution-based approach (SKAT [59], SKAT-O [63]) as gene-level and GLOSSI [23] and Globaltest [21] as traditional gene-set-level analysis methods. We also applied our method to sequence data of 1058 samples from the Korean population.

GSA in rare variants needs two-step of summarizing rare SNPs for the gene and pathway. In first step, we first incorporated rare variants into a gene using proposed gene-level tests. In second step, from the result of gene-level based test, we conducted the gene-set analysis based on gene-level p-values. When combining the gene-level statistic, we used a co-mutation based weight. The logic behind it is that highly interacted genes with other neighbor genes usually play an important role within the gene-set. This gene-set analysis for rare variants is compared to 1) other traditional gene-set analysis methods for GWAS, and 2) the extension of rare variants association test from gene to gene-set via simulation study.

4.2 Methods

In this section, we first describe a preprocessing step for collapsing correlated rare variants and describe a series of methods to detect a joint association signal when there are rare variants and common variants together. We present three gene-level association methods for rare variants based on regression coefficients. Then the gene-set-level test for rare variants and common variants with gene-level weights is introduced.

4.2.1 Preprocessing Step for dependent rare variants within a LD block

For analysis, we first combine correlated rare variants within a region so that these combined variants are independent of each other. This combining scheme is commonly used in collapsing methods such as GRANVIL, Weighted Sum Statistic (WSS), and Variable Threshold (VT). We use this scheme only for correlated variants by linkage disequilibrium (LD), but not for all variants within a region. This is based on the assumption that correlated rare variants have association signals of the same direction. Rare variants usually meet this assumption especially when those variants have strong association signals.

Suppose there are some rare variants within a region. These variants are grouped into LD blocks based on LD measure like r^2 value. In order to generate independent variants across LD blocks, we use r^2 cut-off 0.05 [66]. Let $R_i = (r_{i1}, r_{i2}, \ldots, r_{in})^T$ be ith rare variant within a region. r_{ij} is defined as follows.

$$r_{ij} = \begin{cases} 1, & i\text{th marker has a mutation for individual } j \\ 0, & \text{otherwise} \end{cases}$$

Within a LD block, R_is' are summed into a collapsed variant S.

$$S = \sum_{i \in LD\,block} R_i$$

4.2.2 Gene-level collapsing methods based on the quadratic form statistic

As a gene-level association test for rare variants, we introduce several collapsing methods based on a unified regression framework. Regression coefficients from rare variants can be combined with three different methods according to relation among multiple rare variants.

As a gene-level association test for rare variants, we introduce the collapsing method of beta coefficients from a multiple regression framework. Beta coefficients for rare variants can be combined with three different methods according to the relation among multiple rare variants.

The followings are different assumptions among multiple rare variants.

i) All variants within a region are deleterious / All variants within a region are protective
ii) Some of variants are deleterious and other variants are protective
iii) Consider both i) and ii)

If there are *m* preprocessed rare variants within a region, we can obtain the estimated regression coefficients for rare variants from multiple linear regression with a particular phenotype *y* and covariates ***Z***.

$$y_i = \beta_0 + \sum_{k=1}^{m} \beta_k S_{ki} + \gamma Z_i + \varepsilon_i \quad \text{..... (1)}$$

We can first assume that rare variants within a region are all deleterious or all protective. Given this assumption, the collapsing method, which aggregates the effects with only one direction, gives powerful performance. We proposed $QTest_1$ by testing pooled effect size based on inverse variance weighting

method. Here, regression coefficients and their variances are estimated from equation (1). A chi-square statistic based on pooled β and its variance is computed as follows:

$$\text{For } k = 1,...,m,$$
$$\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_k\right)_{m\times 1}, \boldsymbol{\alpha} = \left(\alpha_k\right)_{m\times 1}, \text{ where } \alpha_k = \frac{1/\text{var}(\hat{\beta}_k)}{\sum_{k=1}^{m}\left(1/\text{var}(\hat{\beta}_k)\right)}$$
$$\hat{\beta}_{Pooled} = \boldsymbol{\alpha}\hat{\boldsymbol{\beta}} = \sum_{k=1}^{m}\alpha_k\hat{\beta}_k$$
$$\text{var}(\hat{\beta}_{Pooled}) = \boldsymbol{\alpha}^T\mathbf{V}\boldsymbol{\alpha} \quad \text{where } \mathbf{V} = \text{diag}_m\left(\text{var}(\hat{\boldsymbol{\beta}}_k)\right)$$
$$\mathbf{A}_1 = \left(\boldsymbol{\alpha}^T\mathbf{V}\boldsymbol{\alpha}\right)^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}^T$$
$$Q_1 = \hat{\boldsymbol{\beta}}^T\mathbf{A}_1\hat{\boldsymbol{\beta}} \sim \chi_1^2$$

We can also assume that some rare variants are deleterious and other rare variants are protective. In this assumption, QTest_1 has very poor performance because deleterious effects and protective effects yield no effect when combined. We proposed QTest_2 based on the following chi-square method in order to combine weak association signals with different directions.

$$\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_k\right)_{m\times 1}, \text{ for } k = 1,...,m,$$
$$\mathbf{V} = \text{diag}_m\left(\text{var}(\hat{\boldsymbol{\beta}}_k)\right)$$
$$Q_2 = \hat{\boldsymbol{\beta}}^T\mathbf{V}^{-1}\hat{\boldsymbol{\beta}} \sim \chi_1^2$$

Although QTest_2 can deal with the effect sizes of different direction, this method does not give better performance than QTest_1 when multiple rare variants have the aggregated effect in either direction. In order to cover the etiology of rare variants, we should consider all possible cases. Our newly proposed quadratic form statistic is computed as a weighted average of inverse

variance weighting and chi-square statistics. To combine the two statistics above, we used the Brown's approximation [57] for the sum of correlated chi-square statistics. If association signals of one direction dominate within a genomic unit, we give more weight (w_1) to the inverse variance weighting statistic, and, if the association signals are balanced with different directions, we give more weight (w_2) to the chi-square statistic.

$$\text{For } k = 1,\ldots,m,$$

$$\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_k\right)_{m\times 1},\ \boldsymbol{\alpha} = \left(\alpha_k\right)_{m\times 1}, \text{ where } \alpha_k = \frac{1/\operatorname{var}(\hat{\beta}_k)}{\sum_{k=1}^{m}\left(1/\operatorname{var}(\hat{\beta}_k)\right)}$$

$$\mathbf{V} = \operatorname{diag}_{(m+1)}\left(\operatorname{var}(\hat{\boldsymbol{\beta}}_k)\right)$$

$$\mathbf{A}_1 = \left(\boldsymbol{\alpha}^T\mathbf{V}\boldsymbol{\alpha}\right)^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}^T,\ \ \mathbf{A}_2 = \mathbf{V}^{-1}$$

$$Q = \hat{\boldsymbol{\beta}}^T\mathbf{A}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^T\left(w_1\mathbf{A}_1 + w_2\mathbf{A}_2\right)\hat{\boldsymbol{\beta}} \sim u\chi_v^2$$

$$\text{where } u = \frac{\operatorname{var}(Q)}{2E(Q)} = \frac{w_1^2 + mw_2^2 + 2w_1w_2}{w_1 + mw_2}$$

$$v = \frac{2E(Q)^2}{\operatorname{var}(Q)} = \frac{(w_1 + mw_2)^2}{w_1^2 + mw_2^2 + 2w_1w_2} = \frac{w_1 + mw_2}{u}$$

This quadratic test (QTest) can cover the assumptions i)~iii). When w_1=1, Q is reduced to Q_1 statistic for pooled effect size (β_{pooled}). When w_1=0, Q is reduced to the Q_2. We define Q3 statistic as Q statistic with w_1=0.5.

The above descriptions for the association test focus only on rare variants. In real sequencing data, there are common and rare variants together. The preprocessing step used for rare variants cannot be applied to common variants because the weakly correlated common variants often have different directions from each other. When considering common variants and rare variants, we need a more general

method which deals with a covariance structure resulted from LD among variants.

When the gene size is large (eg. more than 100), the chi-square statistic can lose power due to a large degree of freedom. In this case, we can use the gamma method instead of chi-square method in the quadratic form statistic.

For $k = 1,...,m$,

$$g(t) = \sqrt{2G_{a,1}^{-1}(2\Phi(|t|)-1))}$$

$$Q^* = w_1 Q_1^* + w_2 Q_2^* = w_1 g(\mathbf{A}^{1/2}\hat{\boldsymbol{\beta}})^T g(\mathbf{A}^{1/2}\hat{\boldsymbol{\beta}}) + w_2 g(\mathbf{V}^{-1/2}\hat{\boldsymbol{\beta}})^T g(\mathbf{V}^{-1/2}\hat{\boldsymbol{\beta}}) \sim u^* \chi^2_{v^*}$$

$$\text{where } u^* = \frac{\text{var}(Q^*)}{2E(Q^*)} = \frac{w_1^2 + 2maw_2^2 + w_1 w_2 \hat{\sigma}_{12}^2}{w_1 + 2maw_2}$$

$$v^* = \frac{2E(Q^*)^2}{\text{var}(Q^*)} = \frac{(w_1 + 2maw_2)^2}{w_1^2 + 2maw_2^2 + w_1 w_2 \hat{\sigma}_{12}^2} = \frac{w_1 + 2maw_2}{u^*}$$

4.2.3 Gene set analysis for rare and common variants

In order to maximize the genetic variation caused from moderate association, gene set analysis (GSA) which considers the effect of multiple variants jointly and use prior biological knowledge based on pathway information can be used.

We propose a combined analysis of rare variants and the gene-set based approach. The logic behind it is that if multiple mutations on the same functional class can influence the disease or trait, then a gene-set can be a key functional class as well as gene.

In our work, we first incorporated association signals of rare variants into a genomic unit. It is natural to choose a gene as a unit of structure, since most of gene-set studies use a gene that shares common functioning as a functional element. Then gene-level signals were used to compute one single gene-set-level signal.

Proposed gene set analysis for rare variants requires two steps.

1) Compute gene level chi-square statistic for rare variants
2) Combine gene-level chi-square statistics within a predefined gene set using gene-level weights

Given gene set S, gene-set level statistic can be computed using a following equation:

$$Q_{GS}= \quad u_{GS}\sum_{t\in GS}\tau_t Q_t(df_t)\cong\chi^2(v_{GS})$$

$$where\ u_{GS}=\frac{2\sum_{t\in GS}\tau_t df_t}{\hat{\sigma}^2},\ v_{GS}=\frac{2(\sum_{t\in GS}\tau_t df_t)^2}{\hat{\sigma}^2}$$

$Q_t(df_t)$ denotes the gene-level quadratic statistic for tth gene

τ_t is gene-level weight for tth gene

In the above statistic, Brown's approximation [56] is used for the sum of correlated chi-square statistic and the covariance between gene-level statistics $\hat{\sigma}^2$ is estimated from a small number of permutation (eg. 100 times). For the gene-level statistic $Q_t^2(df_t)$, we can use the proposed quadratic statistic or the gamma-transformed p-value. Given gene-level p-value (p_t), the Gamma-transformed p-value, $2G_{a,1}^{-1}(1-p_t)$ follows $\chi_i^2(2a)$. So we can substitute $2G_{a,1}^{-1}(1-p_j)$ for gene- level statistic $Q_t^2(df_t)$ when gene-set size is large. If we use the gamma-transformed p-value in the quadratic statistic, then any possible rare variants association test can be used for gene-level, and this gives great flexibility in our proposed method.

4.2.4 Gene-level weight based on co-mutation interaction

When combining gene-level statistics, many GSA do not consider the importance of hub genes which are highly interacted with neighboring variants. These hub genes usually play an essential role in a function and are excellent candidate genes for disease association studies [67]. It has been suggested that a close relationship exists between gene essentiality and network centrality in protein-protein interaction (PPI) networks. The simplest measure of centrality is degree centrality which is defined as the number of links incident upon a node. We can use some PPI databases such as PINA [68] and STRING [69] as the source of PPI-based centrality. Performance of gene set analysis with PPI-based weight depends on the quality of PPI information and there could be missing genes in the PPI database.

Another gene-level weight is calculated by using co-mutations among rare variants. Figure 4.1 shows the example of co-mutation-based interaction for a gene set from real data. If two rare variants share at least one mutation, then we can say two rare variants have co-mutation. If there are k co-mutated rare variants between two genes, then two genes are connected k times. The co-mutation-based weight is standardized for the gene size. This co-mutation-based weight can be calculated for all genes within a gene set and reflect a specific genotype structure from data.

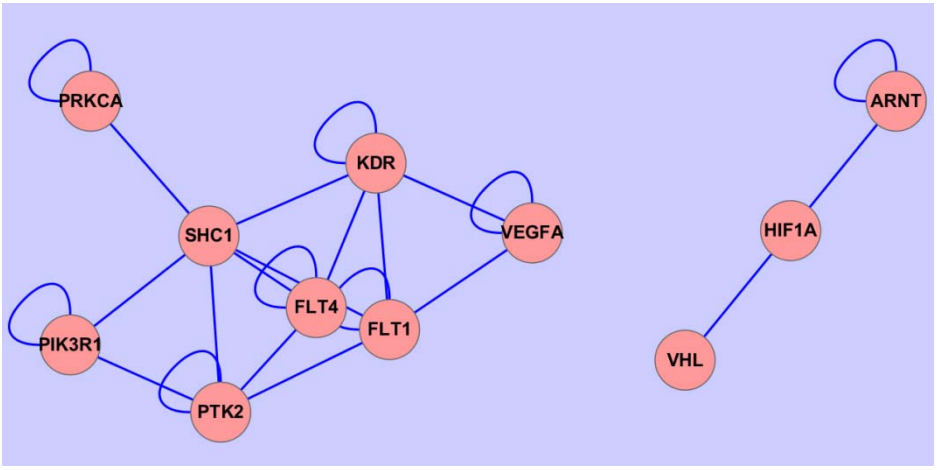

Figure 4.1 Co-mutation-based interaction for a gene set from real data

4.3 Simulation study

4.3.1 Simulation study for gene-level test

In order to compare the performance of proposed gene-level methods with other gene-level burden tests, we conducted the simulation study. Using SimRare program [70] which generates sequence-based data based on realistic population demographics and evolutionary scenarios. 1000 replicates of sequence and phenotype data for 3000 samples were generated for this simulation study. We varied the gene size which is the number of rare variants (10, 20, 30, 40, 50) within a gene and then used the average values of performance across the gene size. For each *i*th individual, we generated the phenotype value y_i given the effect sizes (βs) of causal rare variants along with standard normal-distributed error.

$$y_i = \beta_0 + \sum_{k=1}^{m} \beta_{S_k} s_{ki} + \varepsilon_i, \qquad \varepsilon_i \sim N(0,1)$$

If there are only protective variants or only deleterious variants in a gene, the signs of βs will all be the same and the collapsing methods will have good performance. Conversely, if there are protective variants and deleterious variants together in a region, the collapsing methods lose the power because of the mixed signs of βs.

Our proposed gene-level QTest has three versions of tests according to these scenarios of rare variants. The first and second version of QTest (Q_1, Q_2) is for the same signs of βs and mixed signs of βs, respectively. The third version of QTest (Q_3) compromises two methods (Q_1, Q_2) for the consistently powerful performance. So our proposed methods can deal with the case when there are protective variants and deleterious variants together in a region and also do not lose the power when there are only protective variants or only deleterious variants in a

region.

We compared the performance of methods under the various scenarios. Details for these scenarios are found in Table 4.1. The power in each scenario is defined as the number of p-values less than 0.05 among 1000 p-values from simulation of that scenario. In the first set of scenarios, the trait values are generated from 50 different scenarios by varying the gene size, the proportion of causal variants, and the effect size of causal variants. As expected, the average power of gene-level tests increases when a gene has a larger proportion of causal variants. (See Figure 4.2(a)). The power of SKAT, SKAT-O, Q_2 and Q_3 are sensitive to small change of proportion of causal variants. VT, WSS, GRANVIL, and Q_1 do not have comparable performance to other tests until a proportion of causal variants exceeds to some extent. The average power of gene-level tests tends to be increase as the gene size increases, even though even though they have the same proportion of causal variants within a gene (See Figure 4.2(b)). This shows that the power of a gene-level test depends on the number of causal variants as well as the proportion of causal variants.

In Figure 4.2(c), the power is averaged for all gene sizes and all proportions of causal variants. We can see that the power depends on effect size of causal variants. Overall, Q_2 and Q_3 perform best. When the effect size is larger than 0.75, VT outperforms SKAT and SKAT-O especially the proportion of causal variants is large (Figure 4.2(a)).

In the second set of scenarios, the effect sizes of 0.75 with different signs are assigned to causal rare variants. In Figure 4.3(a), as the proportion of protective variants becomes larger, the differences of power among gene-level tests increase. SKAT, SKAT-O, Q_2 and Q_3 have consistently powerful performance when causal variants have effects of different direction. VT, WSS, GRANVIL, and Q_1

do not seem to be good tests under this scenario.

In the third set of scenarios, we include common variants as well as rare variants for simulations. As can be seen in Figure 4.3(b), Q_2 and Q_3 are most robust to inclusion of common variants and the power increases slightly as the proportion of common variants increases. SKAT and SKAT-O are also robust to inclusion of common variants, but the power decreases slightly as the proportion of common variants increases. It may be because SKAT and SKAT-O gives more weight to rare variants with a function of MAF. If we use the uniform weight on SKAT and SKAT-O, then the power will be better for common variants. Among the collapsing methods, VT and WSS outperform GRANVIL and Q_1 because VT and WSS focus rare variants even if common variants exist.

Table 4.1 Summary of phenotype simulation scenarios for gene-level

	Proportion of causal variants	**Effect size (S.E.=1)**	**Gene size**	**Proportion of deleterious/ protective variants**	**Proportion of common/rare variants**	**Total number of scenarios**
Scenario set 1	10%, 20%, 30%	0 (for control), 0.5, 0.75, 1.0	10, 20, 30, 40, 50	0	0	45 + 5(control)
Scenario set 2	10%, 20%, 30%	0 (for control), 0.75 or -0.75	50	0.1, 0.2, 0.3, 0.4, 0.5	0	15 + 1(control)
Scenario set 3	10%, 20%, 30%	0 (for control), 0.75 (for rare) or 0.1 (for common)	50	0	0.1, 0.2, 0.3, 0.4, 0.5	15 + 1(control)

a) proportion of causal variants

b) different gene sizes

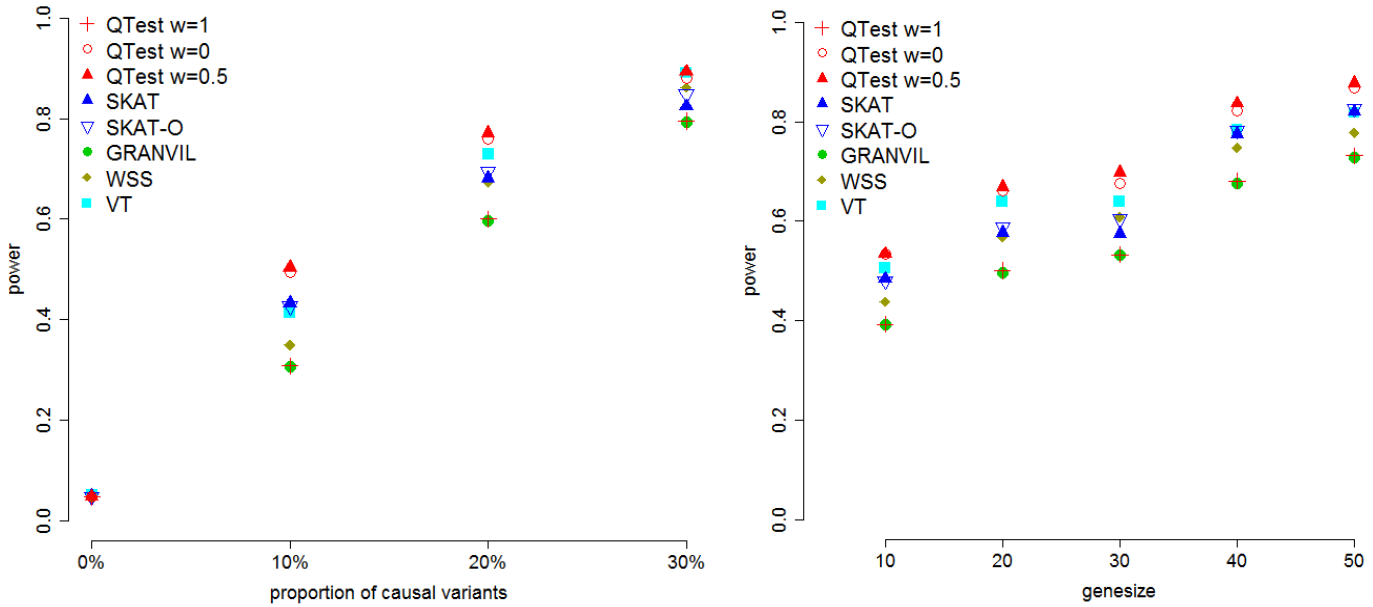

Figure 4.2 Comparison of power of gene-level tests under scenario set 1

c) different effect size

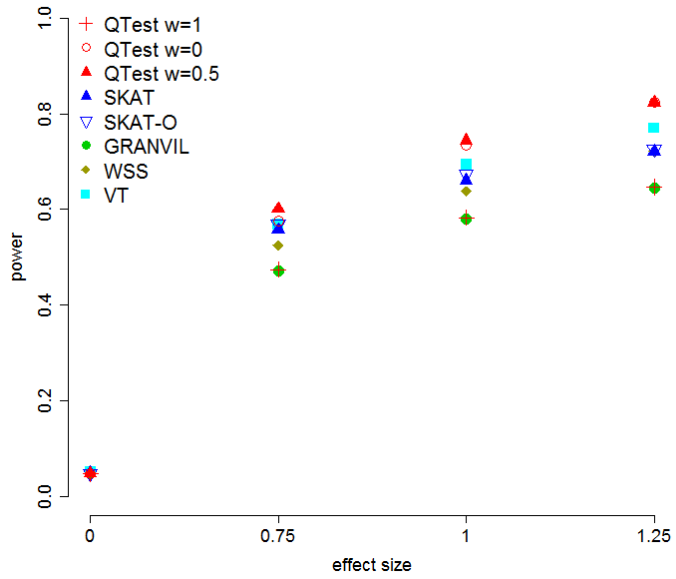

Figure 4.2 Comparison of power of gene-level tests under scenario set 1

(a) Proportion of protective variants

(b) Proportion of common variants

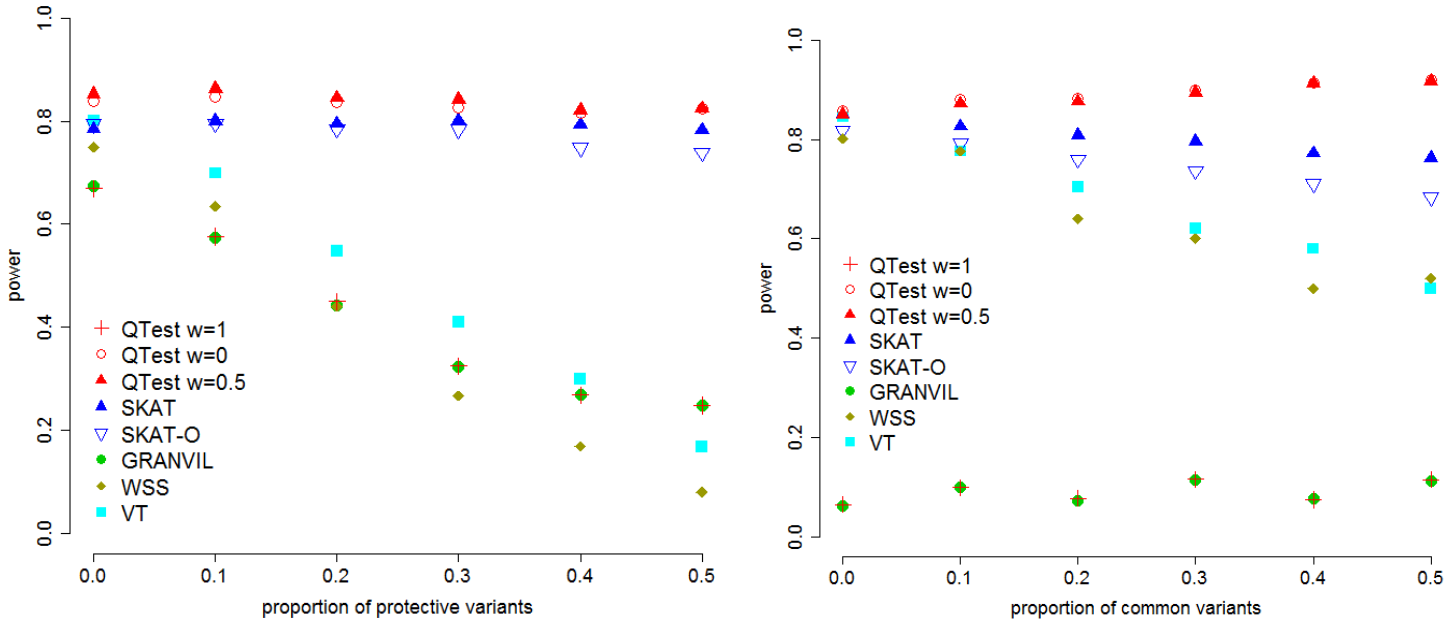

Figure 4.3. Comparison of power of gene-level tests under scenario set 2 and 3

4.3.2 Simulation study for gene set analysis

In order to investigate the performance of QTest for gene set (Q_{GS}), another simulation for gene-set analysis was conducted. Based on genotypes generated from SimRare, we constituted a gene set which has 12 genes. In scenario set 4, the phenotype values were generated under the 81 different scenarios by varying the number of causal genes within a set (2, 3, 4 genes), the proportion of causal variants within a gene (10, 20, 30%), effect size (0.75, 1.0, 1.25), and gene size (10, 30, 50 variants). Details for these gene-set level scenarios are found in Table 4.2.

We applied two traditional gene set analysis methods (Globaltest [21], GLOSSI [23]) and QTest for gene set (Q_{GS}) to simulated data. Q_{GS} considers gene-level association and can use any p-values for the gene-level association. In this study, we used Q_3 and SKAT as the gene-level p-values ($Q_{GS}(Q_3)$, Q_{GS}(SKAT)). We also conducted rare variants association tests (SKAT, SKAT-O, GRANVIL, Q_1, Q_2, Q_3) to simulated data by extending the collapsing unit from a gene to a gene-set. The gene-level weight is set to be 1 across all genes in this scenario.

In this scenario set, the overall trend of performances of the extended association test is similar to case of gene tests (See Figure 4.4). Among the extended burden tests, the power of Q_1, Q_2, Q_3 and GRANVIL tend to be smaller compared to case of gene-level. Simulated data for gene-set-level have larger size than for gene-level and have many non-causal variants within them. In order to remedy “large gene size” problem in Q_2 and Q_3, we used the transformed gamma statistic instead of the chi-square statistic. This brings an additional gain of power when gene size is large. SKAT and SKAT-O did not lose the power compared to the case of gene-level. We also found that the performance of Globaltest was very similar to performance of SKAT. We may conclude that SKAT, SKAT-O, Globaltest are very robust to inclusion of non-causal variants. As a result, SKAT and SKAT-O perform

best among extended association tests. The proposed gene-set-level test, Q_{GS}, has biggest overall power among all tests. In general, considering gene information makes the gene-set-level test become more powerful although there are exceptions. If causal variants are concentrated in a very few genes, summarizing for gene-level will give a loss of power.

Q_{GS} has the largest overall power among all tests. In general, considering gene information makes the gene-set analysis become more powerful than when not considering gene information, although there are some exceptions. If causal variants are concentrated in a very few genes, summarizing for gene-level will give a loss of power. Q_{GS} requires the permutation step for the estimation of covariance part in the gene-set-level statistic. We used only 200 permutations because we empirically found this number of permutation sufficient to attain a stable estimate of covariance. This leads to shorter computing time than other permutation-based methods.

When the effect size is 0.75 in Figure 4.4(d), we can see SKAT slightly outperforms QTests (Q_2, Q_3) and Q_{GS}(SKAT) outperforms $Q_{GS}(Q_3)$. However, their range of power (0.5~0.6) is not acceptable if we consider the multiple comparison of tests in application to real genetic data. When the effect size is larger than 0.75, $Q_{GS}(Q_3)$ outperforms Q_{GS}(SKAT).

We compared 100 permutations to 1000 permutations for estimating covariance among gene-level statistics (See Figure 4.5). We found that permuting a trait 100 times is sufficient to attain a stable estimate of covariance. For this reason, to save computing time, we used only 100 permutations for estimating covariance part in the gene-set-level statistic.

Table 4.2. Summary of phenotype simulation scenarios for gene-set-level

	Set size	**Gene size**	**Proportion of causal genes in a set**	**Proportion of causal variants in a gene**	**Effect size (S.E.=1)**	**Proportion of protective causal variants**	**Total number of scenarios**
Scenario set 4	12 genes	10 variants	2/12	10%	0	0	81 + 3 (control)
		30 variants	3/12	20%	0.75		
		50 variants	4/12	30%	1.0		
					1.25		

a) proportion of causal genes in a set

b) proportion of causal variants in a gene

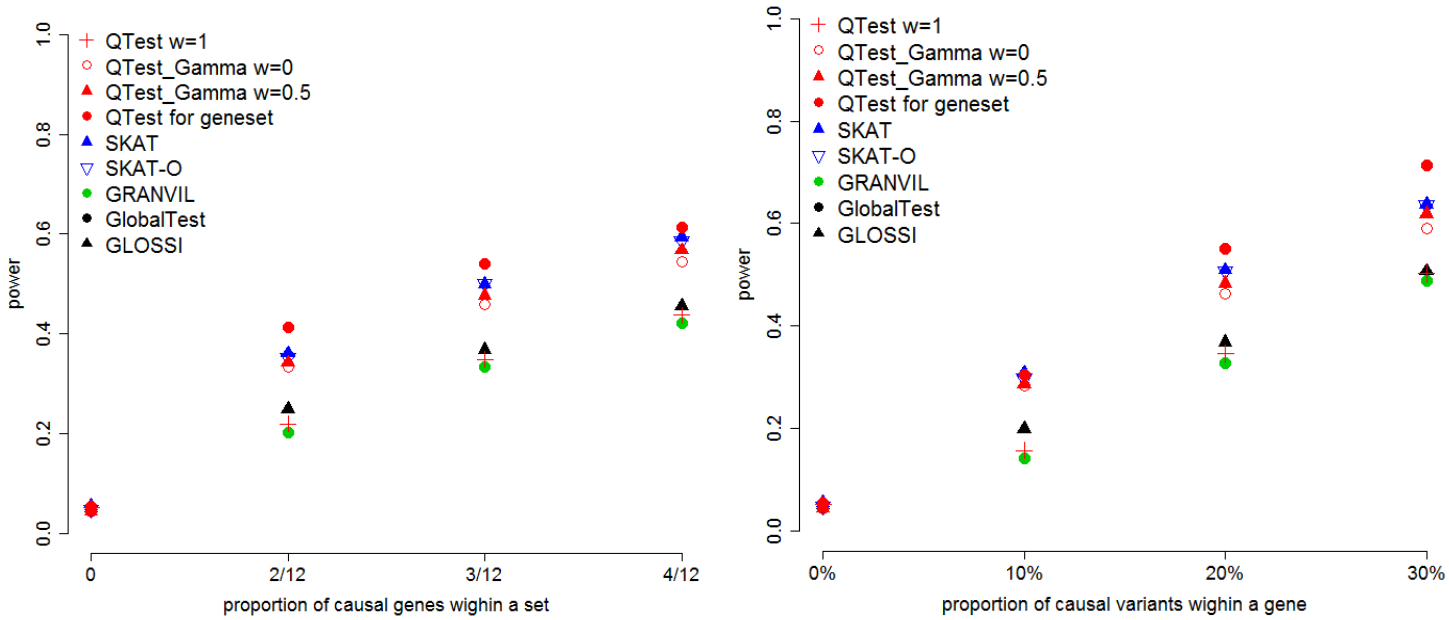

Figure 4.4 Comparison of power of gene-set-level tests under scenario set 4

c) different gene sizes

d) different effect sizes

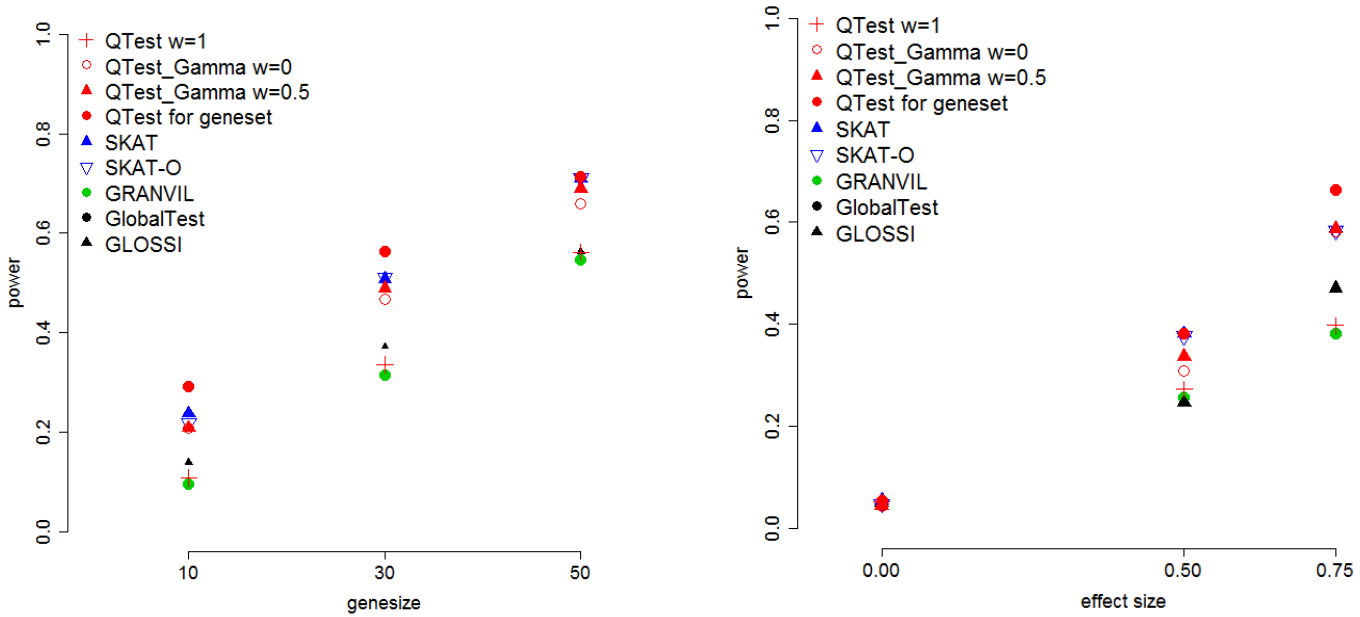

Figure 4.4 Comparison of power of gene-set-level tests under scenario set 4

a) –log(p-value)

b) p-value

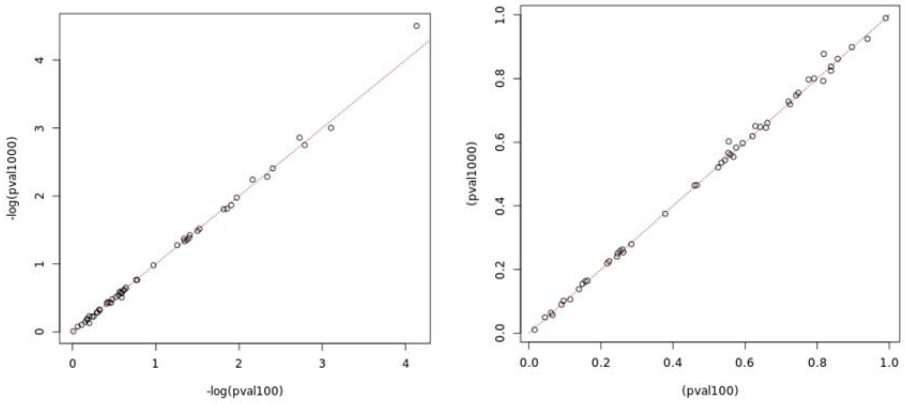

Figure 4.5 Gene-level p-value with 100 permutations and 1000 permutations

4.4 Application for Korean liver enzymes and Exome data

We applied the proposed tests to discover the association between the liver enzymes phenotype from the Korean Association REsource (KARE) [43] study and the whole exome data from 1,058 Korean individuals in KARE. As the liver enzymes, we used the alanine aminotransferase (ALT) trait. After excluding markers with high missing call rate (>5%), we used 398,098 common and rare markers and 18,851 corresponding genes. We used 1,452 canonical pathways from MsigDB database v3.0. These pathways are curated from other online database such as BioCarta, KEGG and GO, and so on. In the analysis, sex, age, area and interaction between age and sex are used as covariates. PLINK is used to conduct linear simple regression for each single variant given covariates. Table 4.3, 4.4 show the result of gene-level tests (SKAT, SKAT-O, GRANVIL, Q_1, Q_2, Q_3; See Figure 4.7-4.11) and Table 4.5, 4.6 show the result of gene-set test (Q_{GS}) with or without gene-level weights.

From the single variant test, the minimum q-value [46] was 0.311 and so there is no common or rare variant significantly associated with the ALT phenotype in terms of q-value.

From the gene test based on common and rare variants, SKAT, SKAT-O, Q_1 do not identify any significant genes. From Q_2 and Q_3, we can obtain 6 significant genes, KLF4, GPR22, PLAGL2, TRIM22, CTPS, CCDC129 in terms of q-value 0.05. It is known that there is a negative correlation between the KLF4 gene and the hepatocellular carcinoma (HCC) [71].

The association between the ALT and GPR22 genes was newly identified in previous Korean GWAS [72]. TRIM22 is also known to inhibit Hepatitis B Virus (HBV) gene expression and replication via transcriptional repression [73].

We also conducted the gene-level test using only rare variants with minor allele frequency (MAF) <1%. After multiple comparison adjustments, SKAT did not identify significant variant, and SKAT-O, GRANVIL, and $QTest_1$ detected one significant gene, ZFP161. QTest0 and QTest0.5 found 8 significant genes ZFP161, TONSL, FCRLA, MET, PLAGL2, GPR22, C19orf53, and TXNDC5. The protein encoded by TONSL gene is thought to be a negative regulator of NF-κB mediated transcription. NF-κB activity is well known to be enhanced in HBV-associated HCC [74].

When using co-mutation-based gene-level weights, highly associated gene sets with ALT were BIOCARTA SODD pathway, REACTOME transcriptional regulation of white adipocyte differentiation, SA caspase cascade, ST tumor necrosis factor (TNF) pathway, and so on. TNF-induced activation of NF-kappa-B is shown to be accelerated in SODD-deficient cells [75]. Adipocyte differentiation is also known to induce dynamic changes in NF-κB expression and activity. It has been found that the proinflammatory cytokine tumor necrosis factor alpha (TNF-α) induces ALT promoter transactivation, mRNA and protein synthesis, and enzymatic activity via NF-κB. In the SODD pathway, MAP3K8 and NFKBIA have the largest co-mutations with neighboring genes and also have the most significant p-values (8.18×10^{-3}, 1.96×10^{-5}, respectively).

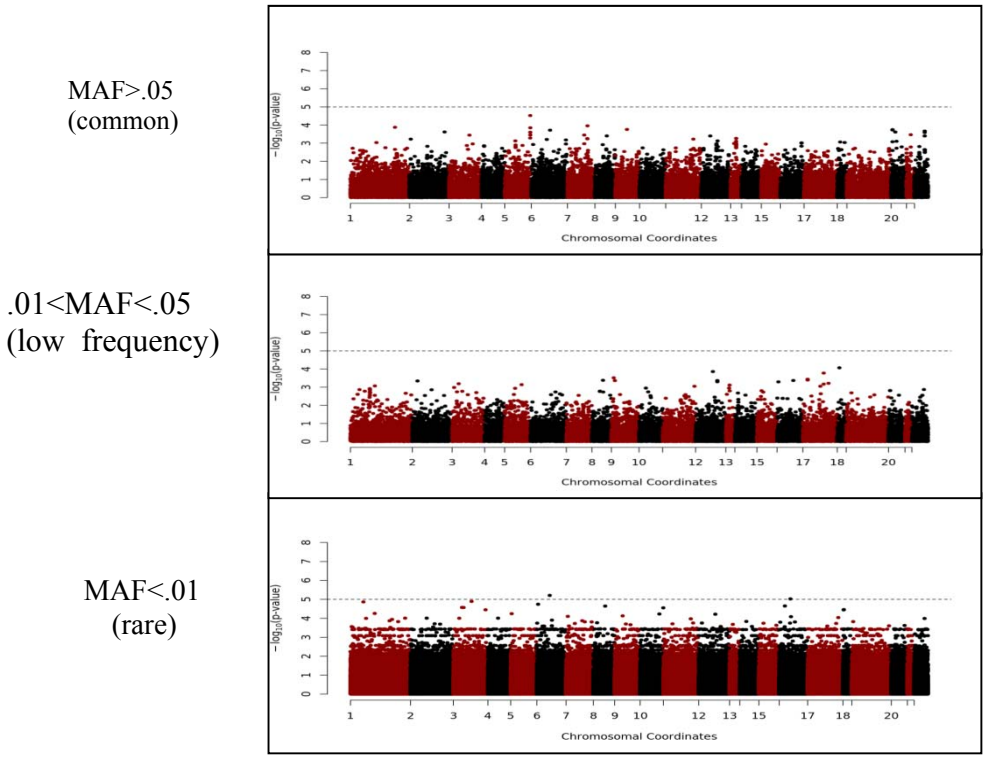

Figure 4.6 Manhattan plot for single variant-level association with linear regression

SKAT

SKAT-O

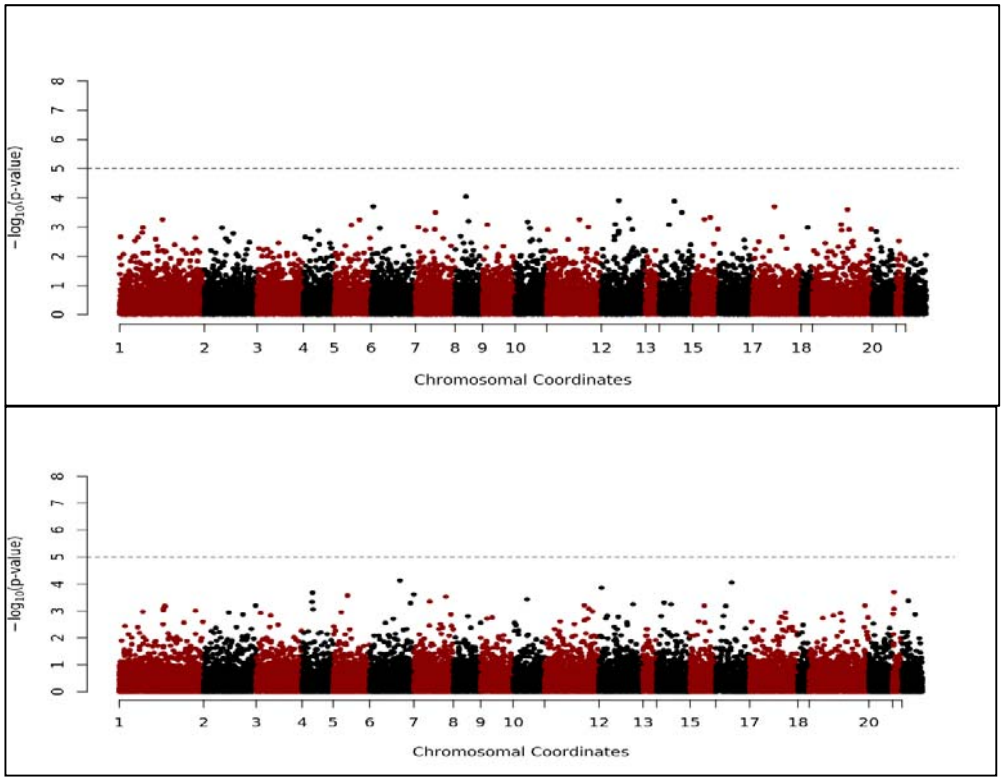

Figure 4.7 Manhattan plot for gene-level association with SKAT and SKAT-O for common and rare variants

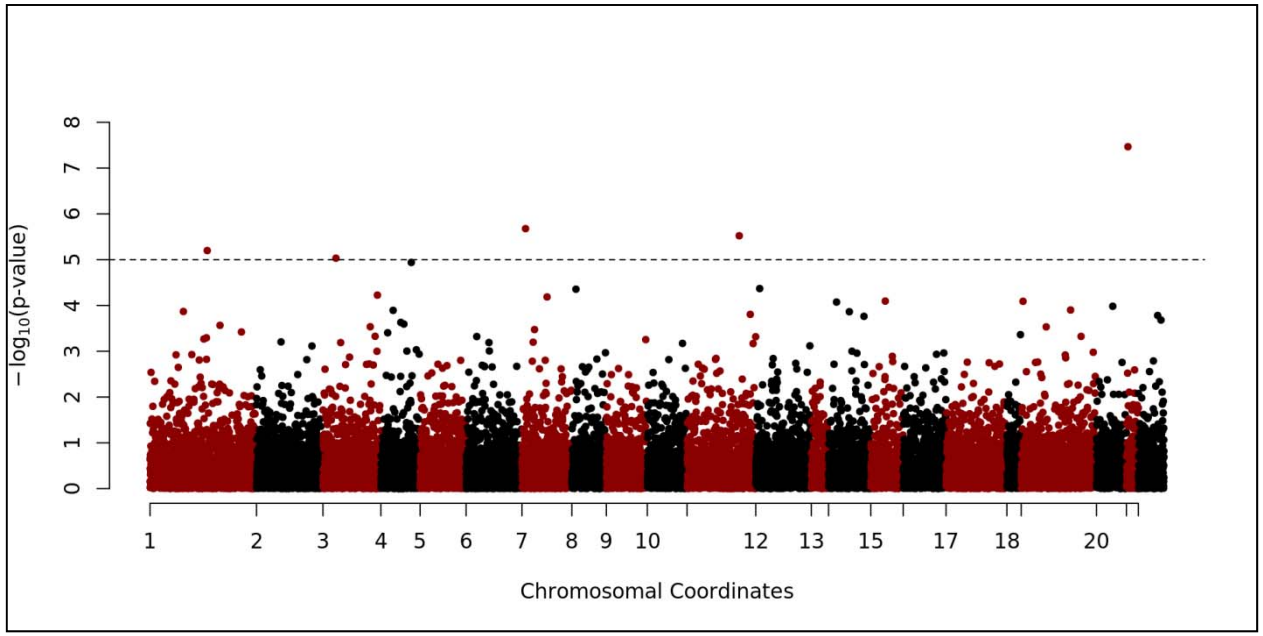

Figure 4.8 Manhattan plot for gene-level association with $QTest_3$ for common and rare variants

MAF<0.01

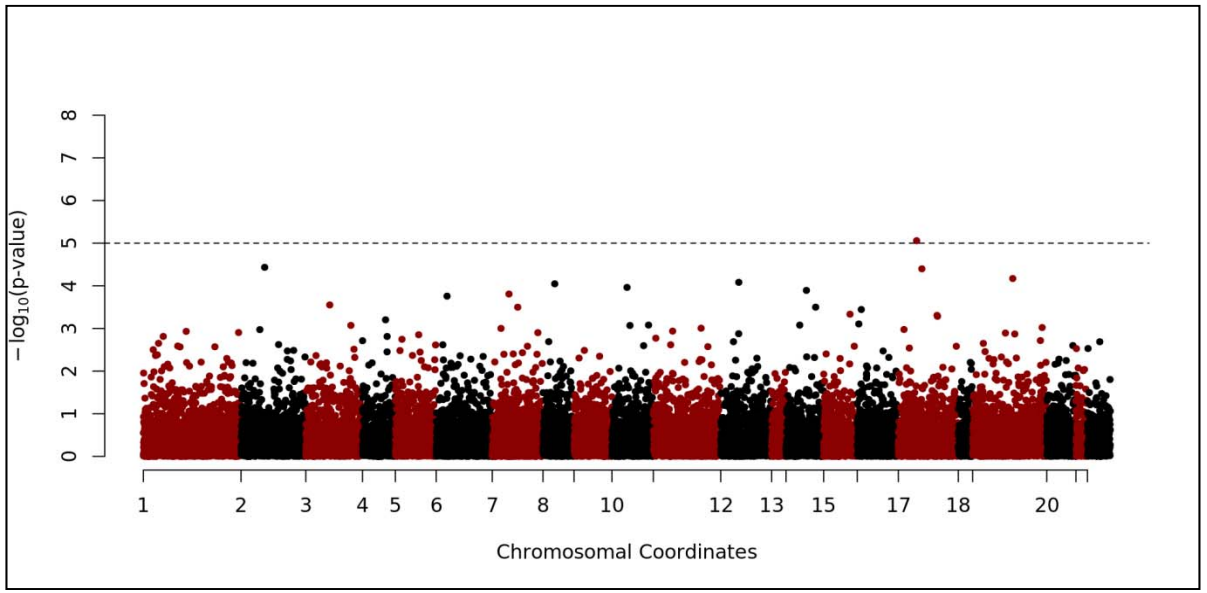

Figure 4.9 Manhattan plot for gene-level association with SKAT for rare variants

MAF<0.01

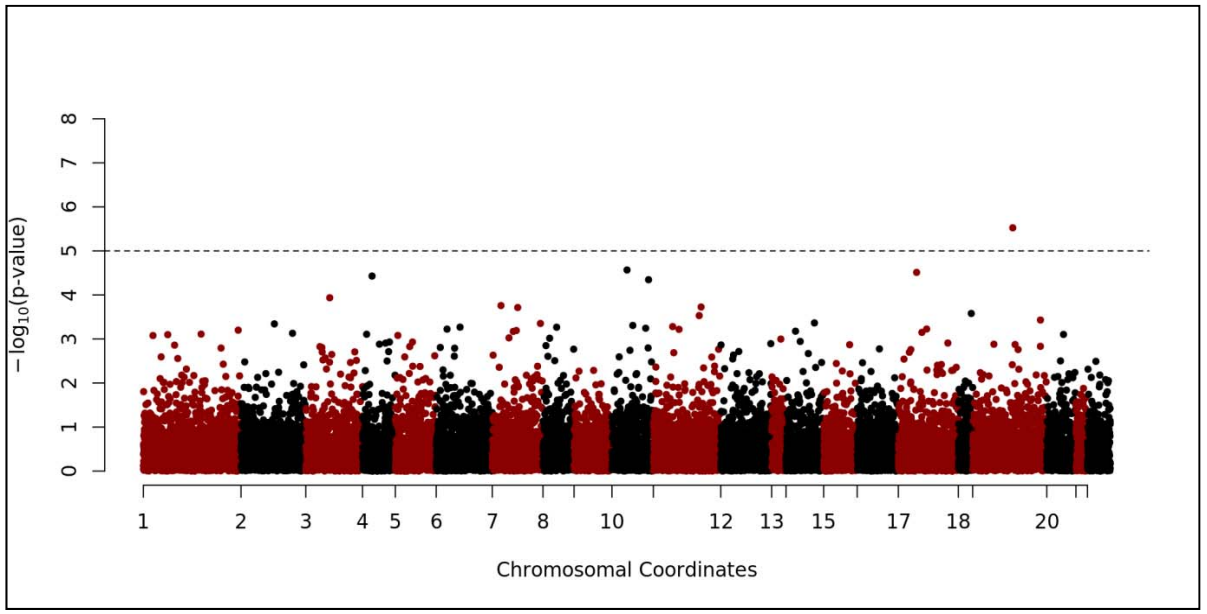

Figure 4.10 Manhattan plot for gene-level association with $QTest_3$ for rare variants

Table 4.3 Significantly associated genes with ALT phenotype for Korean individuals (using common and rare variants)

Gene name	Chr	Gene size	p-value (q-value)				
			SKAT	SKAT-O	$QTest_1$	$QTest_2$	$QTest_3$
KLF4	9	15	1.89E-04 (0.882)	1.37E-04 (0.644)	7.39E-05 (0.348)	6.24E-06 (**0.037**)	3.41E-08(**0.0006**)
GPR22	7	5	9.97E-04 (0.899)	1.11E-04 (0.644)	4.52E-05 (0.284)	9.77E-04 (0.321)	2.10E-06 (**0.018**)
PLAGL2	20	4	1.17E-03 (0.899)	5.23E-05 (0.493)	2.79E-05 (0.284)	5.68E-04 (0.280)	3.00E-06 (**0.018**)
TRIM22	11	14	7.05E-02 (0.976)	9.93E-02 (0.977)	4.44E-02 (0.927)	7.14E-06 (**0.037**)	6.33E-06 (**0.028**)
CTPS	1	29	4.92E-01(0.984)	6.70E-01(0.992)	3.28E-02(0.927)	7.77E-06(**0.037**)	9.28E-06(**0.033**)
CCDC129	7	32	4.80E-02(0.954)	6.77E-02(0.970)	4.55E-01(0.987)	4.51E-06(**0.037**)	1.15E-05(**0.034**)
AFTPH	2	32	5.41E-02(0.989)	1.98E-01(0.992)	1.10E-01(0.955)	3.34E-05(**0.099**)	4.27E-05(**0.098**)
BIRC3	11	22	4.74E-02(0.952)	7.56E-02(0.958)	7.21E-02(0.955)	4.37E-05(**0.099**)	4.41E-05(**0.098**)
RELT	11	33	1.23E-02(0.899)	2.49E-02(0.861)	4.82E-01(0.989)	3.79E-05(**0.099**)	1.26E-04 (0.136)
MUTED TXNDC5	6	25	2.72E-01(0.984)	4.08E-01(0.992)	1.98E-01(0.969)	3.84E-05(**0.099**)	1.35E-04(0.136)

Table 4.4 Significantly associated genes with ALT phenotype for Korean individuals (using rare variants with MAF<1%)

Gene name	Chr	Gene size	p-value (q-value)					
			SKAT	SKAT-O	GR	$QTest_1$	$QTest_0$	$QTest_{0.5}$
ZFP161	18	5	8.74E-06(0.156)	1.97E-06(**0.034**)	5.10E-06(0.089)	4.19E-06(**0.072**)	6.69E-05(0.187)	5.68E-07(**0.010**)
TONSL	8	51	8.33E-04(0.659)	1.88E-03(0.660)	3.49E-01(0.956)	3.17E-01(0.927)	8.46E-06(**0.015**)	1.60E-06(**0.018**)
FCRLA	1	12	6.75E-05(0.267)	3.20E-05(0.208)	8.36E-05(0.293)	5.43E-05(0.188)	4.21E-04(0.321)	6.71E-06(**0.040**)
MET	7	33	1.08E-04(0.277)	9.93E-02 (0.328)	6.89E-02(0.936)	3.97E-02(0.850)	1.52E-06(**0.091**)	1.19E-05(**0.045**)
PLAGL2	20	4	9.86E-04(0.677)	2.46E-04 (0.213)	3.02E-05(0.215)	2.36E-05(0.186)	5.47E-04(0.328)	1.26E-05(**0.048**)
GPR22	7	5	9.95E-04(0.677)	5.36E-05(0.213)	4.91E-05(0.215)	4.36E-05(0.188)	7.33E-04(0.328)	1.76E-05(**0.048**)
C19orf53	19	8	3.16E-04(0.435)	1.10E-04(0.208)	3.81E-05(0.215)	3.23E-05(0.186)	1.76E-03(0.404)	2.06E-05(**0.048**)
TXNDC5	6	15	6.59E-03(0.892)	7.56E-02 (0.869)	1.05E-01(0.939)	9.49E-02(0.875)	1.04E-05(**0.091**)	2.14E-05(**0.048**)

Table 4.5 Significantly associated gene-sets with ALT phenotype (no gene-level weights are used)

Gene set	# of genes	p-value(q-value)		
		$Q_{GS}(Q_1)$	$Q_{GS}(Q_2)$	$Q_{GS}(Q_3)$
REACTOME SYNTHESIS SECRETION AND DEACYLATION OF GHRELIN	10	2.40E-02(0.772)	3.04E-03(**0.060**)	1.57E-05(**0.002**)
SA CASPASE CASCADE	72	8.81E-02(0.772)	5.22E-05(**0.028**)	1.96E-05(**0.002**)
REACTOME TRANSCRIPTIONAL REGULATION OF WHITE ADIPOCYTE DIFFERENTIATION	19	1.53E-01(0.772)	8.07E-04(**0.044**)	2.23E-05(**0.002**)
REACTOME NRIF SIGNALS CELL DEATH FROM THE NUCLEUS	29	2.69E-02(0.772)	5.36E-04(**0.044**)	2.40E-05(**0.002**)
REACTOME TRANS GOLGI NETWORK VESICLE BUDDING	30	5.09E-02(0.772)	1.57E-04(**0.028**)	4.02E-05(**0.003**)
REACTOME GOLGI ASSOCIATED VESICLE BIOGENESIS	38	1.82E-01(0.772)	1.44E-04(**0.028**)	8.72E-05(**0.006**)
KEGG ACUTE MYELOID LEUKEMIA	46	3.33E-03(0.772)	5.70E-03(**0.078**)	1.02E-04(**0.006**)
REACTOME LYSOSOME VESICLE BIOGENESIS	60	4.14E-02(0.772)	6.54E-04(**0.044**)	1.76E-04(**0.008**)
REACTOME DOWNREGULATION OF ERBB2 ERBB3 SIGNALING	23	1.92E-01(0.772)	9.04E-04(**0.044**)	1.90E-04(**0.008**)

Table 4.6 Significantly associated gene-sets with ALT phenotype (gene-level weights are used)

Gene set	# of genes	p-value(q-value)		
		$Q_{GS}(Q_1)$	$Q_{GS}(Q_2)$	$Q_{GS}(Q_3)$
BIOCARTA SODD PATHWAY	10	1.49E-01(0.787)	2.35E-05(**0.004**)	1.67E-05(**0.002**)
REACTOME TRANSCRIPTIONAL REGULATION OF WHITE ADIPOCYTE DIFFERENTIATION	72	1.67E-01(0.787)	3.71E-04(**0.017**)	1.69E-05(**0.002**)
SA CASPASE CASCADE	19	1.40E-01(0.787)	1.74E-04(**0.009**)	1.89E-05(**0.002**)
ST TUMOR NECROSIS FACTOR PATHWAY	29	5.16E-01(0.787)	6.54E-05(**0.004**)	1.96E-05(**0.002**)
REACTOME NOD1 2 SIGNALING PATHWAY	30	4.87E-01(0.787)	6.56E-05(**0.004**)	2.06E-05(**0.002**)
PID FASPATHWAY	38	4.48E-01(0.787)	7.24E-04(**0.029**)	4.57E-05(**0.003**)
REACTOME NUCLEOTIDE BINDING DOMAIN LEUCINE RICH REPEAT CONTAINING RECEPTOR NLR SIGNALING PATHWAYS	46	4.82E-01(0.787)	6.85E-05(**0.004**)	4.97E-05(**0.003**)
REACTOME TRANS GOLGI NETWORK VESICLE BUDDING	60	1.06E-01(0.787)	3.07E-05(**0.004**)	5.24E-05(**0.003**)
REACTOME LYSOSOME VESICLE BIOGENESIS	23	9.16E-02(0.787)	5.94E-05(**0.004**)	5.56E-05(**0.003**)

4.5 Conclusion

As mentioned in the introduction, although many methods have been developed recently to assess the association between rare variants and complex diseases, no single method gives consistently acceptable power across the range of these relationships. In this chapter, for the consistent powerful association methods, we describe several methods to conduct gene-level test and gene-set-level test for sequencing data. The quadratic tests, QTests are proposed based on a multiple regression framework. In a various scenarios of simulation for gene-level, we can see QTest performed better than other previous methods in general.

We also proposed the gene-set-level rare variants association test, QTest for gene set in the same regression framework as the gene-level QTests. Then we demonstrate the performance of the proposed methods comparing with other gene-set-level association methods in various simulation scenarios. This demonstration shows that existing association test could work in gene-set-level in some extent if we do not go through gene-level summarization step. When a gene set has a large number of variants, these association tests often lose the power because of many noncausal variants in a gene set. Among existing methods, Globaltest and SKAT still perform well even if there are many noncausal variants in a gene-set. From the simulation studies, in almost cases, QTest for gene set is shown to outperform other methods in general.

We applied our method to Exome sequencing data of 1058 samples from Korean population. We also applied other methods such as SKAT, SKAT-O, GRANVIL in the gene-level. Proposed methods identified more genes than other methods in terms of multiple-comparison-adjusted p-value. The use of co-mutation-based weights gives prominence to the hub genes in a gene-gene interaction network. We could identify more biologically meaningful pathway by using gene-level weights

based on co-mutations.

4.6 Discussion

There are some issues for methods with consistent performance. The first issue is that the performance of association tests depends on the relation between the proportion and effect sizes of causal rare variants in a region. In some cases, a very small portion of rare variants could strongly affect the disease phenotype, and in other cases, multiple rare variants with mild to moderate effects could affect the disease phenotype. SKAT and GlobalTest can detect signals of the former case accurately, as well as the Bonferonni test. The latter is supported well in many collapsing methods such as GRANVIL, WSS, VT, and $QTest_1$. To compromise the two cases, combining methods such as SKAT-O and QTest are developed.

The second issue is the direction of causal variant effects. Many collapsing methods assume that rare variants have effects of the same direction in a region. However, this is not always the case. When there are deleterious alleles and protective alleles together, the collapsing methods may lose power if they do not consider the effect of opposite directions. On the contrary, the methods which consider the different directions could lose power when the causal variants have signals of the same direction. $QTest_1$ assumes that variants have the same direction and $QTest_0$ assumes that variants have different directions. The two methods need to be combined for a consistently powerful test.

The third issue is the existence of common variants within a region. Most association tests focus only on rare variants and give a larger weight to rarer variants with a function of minor allele frequency (MAF). For this reason, the common causal variants in sequencing data would not be focused on these association tests. To address this issue in collapsing methods, one ~~of~~ solution~~s~~ is to collapse multiple rare variants first, and then combine the collapsed variants with the common variants to test in gene-level or gene-set-level. SKAT and Globaltest

work well even without those steps because they do not collapse variants in their tests. We consider the case when there exist rare variants and common variants together within a region. If we can get the coefficients from each rare variant and each common variant under their independence assumption of those variants, this issue is no longer a problem. In ~~a~~ sequencing data, rare variants and common variants do not often show a high correlation among them. Our preprecessing step based on LD blocks makes $QTest_0$ robust to the existence of common variants in the same region.

We proposed the unified method, QTest, which deals with these three issues for consistently powerful tests. Our simulation study for gene-level tests shows QTest outperforms other burden tests in various scenarios in terms of average power. Of course, this does not mean the variants detected by QTest with specified parameters include all detected variants by other burden tests. However, if we use the proposed method with several parameter settings, we can cover the etiology of rare variants within a gene in great portions.

Recent rare variant association methods usually focus on gene-levels. However, joint play of rare variants affecting complex diseases does not arise only in a gene-level, but also in a gene-set or pathway level. We propose the gene-set-level rare variants association test, QTest, for gene sets (Q^{GS}). Then we also demonstrate the performance of the proposed methods compared with other gene-set-level association methods in various simulation scenarios. This demonstration shows that the existing burden test could work in a gene-set-level to some extent if we do not go through a gene-level summarization step. When a gene set has a large number of variants, these burden tests often lose the power because of many non-causal variants in a gene set. Among existing methods, Globaltest and SKAT still perform well even if there are many non-causal variants in a gene-set. From the simulation studies, in almost cases, Q_{GS} is shown to outperform other methods in general.

We applied our method to Exome sequencing data of 1058 samples from the Korean population. We also applied other methods such as SKAT, SKAT-O, and GRANVIL in the gene-level. Proposed methods identified more genes than other

methods in terms of q-value. The use of the co-mutation-based weights gives prominence to the hub genes in a gene-gene interaction network. By using gene-level weights based on co-mutations, we could identify more biologically meaningful pathways.

In this thesis, we deal only with the quantitative phenotype. If the phenotype is binary, then the association statistic (eg. p-value, chi-square statistic) for each rare variant is not stable. By collapsing rare variants in a sub-region first or by transforming the phenotype into a quantitative value, we can extend the proposed tests into a case of binary phenotypes.

Chapter 5

Summary & Conclusion

Genome-wide association studies (GWAS) — in which hundreds of thousands of single-nucleotide polymorphisms (SNPs) are tested for association with a disease in thousands of persons — have revolutionized the search for genetic influences on complex traits. In the past years, GWASs have reported an extensive list of the findings: SNPs or genes associated with traits. However, the associated variants can explain only a small fraction of the heritability of most common traits. This suggests that the other genetic mechanisms, such as gene-gene interaction, gene-environmental interaction, gene-set level effect, and multiple rare variants could contribute to disease susceptibility. In this thesis, we focus on gene-set-level effects of SNPs and aggregated effects of multiple rare variants as another source of heritability.

In Chapter 3, a parametric gene set analysis (GSA) for GWAS was studied. GSA in GWAS increases the power to detect the genetic variants which have a weak association but a meaningful biological association with a phenotype. Many GSA methods test the significance of gene set via permutation by generating

permuted data more than thousands times, which requires expensive computational efforts. The use of a parametric test can reduce the computing time, because it needs to calculate the gene set statistic only once. we proposed a parametric method for gene-set analysis in GWAS. This method is referred as SNP-PRAGE, a SNP-based parametric robust analysis of gene-set enrichment. SNP-PRAGE handles correlation adequately among association measures of SNPs, and minimizes computing effort by the parametric assumption. SNP-PRAGE first obtains gene-level association measures from SNP-level association measures by incorporating the size of corresponding (or nearby) genes and the LD structure among SNPs. Afterward, SNP-PRAGE acquires the gene-set level summary of genes that undergo the same biological knowledge. This two-step summarization makes the within-set association measures to be independent from each other, and therefore the central limit theorem can be adequately applied for the parametric model. We demonstrated the performance of this method via simulation study and applied the method to two GWAS data sets: hypertension data of 8,842 samples from the Korean population and bipolar disorder data of 4,806 samples from the Wellcome Trust Case Control Consortium. We found two enriched gene sets for hypertension and three enriched gene sets for bipolar disorder.

In Chapter 4, we focused on identifying the joint action of rare variants for complex diseases. Aggregated gene-level and gene-set-level effects of rare variants in next generation sequencing (NGS) could explain the missing heritability in GWAS. For the unified and robust association methods, we developed new regression coefficient-based collapsing methods in gene-level and gene-set-level when rare variants and common variants exist together within a region. We described proposed gene-level test which is referred as QTest in various setting and QTest for gene set. Then we demonstrated the performance of the proposed

methods comparing with other gene-level and gene-set-level association methods in various simulation setting. These methods were applied to the exome data and ALT phenotype from 1058 Korean samples. highly associated gene sets with ALT were BIOCARTA SODD pathway, REACTOME transcriptional regulation of white adipocyte differentiation, SA caspase cascade, ST tumor necrosis factor (TNF) pathway, and so on.

In summary, gene set analysis for GWAS and NGS data can increase the power to detect association signals in common complex diseases. Because the high dimensional genetic variants work together for disease and those variants are usually have the dependent structure, the gene set analysis requires a large amount of time for nonparametric tests. We developed the parametric gene set analysis for GWAS and NGS data and demonstrated the performance of the methods in simulation studies and real data.

Bibliography

1. Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., Zhu, J., Millstein, J., Sieberts, S., Lamb, J., GuhaThakurta, D., Derry, J., Storey, J.D., Avila-Campillo, I., Kruger, M.J., Johnson, J.M., Rohl, C.A., van Nas, A., Mehrabian, M., Drake, T.A., Lusis, A.J., Smith, R.C., Guengerich, F.P., Strom, S.C., Schuetz, E., Rushmore, T.H., Ulrich, R. (2008). Mapping the Genetic Architecture of Gene Expression in Human Liver. PLoS Biol. *6*, e107
2. Risch, N., Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516–1517.
3. Lander, E.S. (1996). The new genomics: global views of biology. Science *274*, 536–539.
4. International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
5. Psychiatric GWAS Consortium Coordinating Committee (2009). Genome-wide Association Studies: History, Rationale, and Prospects for Psychiatric Disorders. Am. J. Psychiatry *166*, 540-556.
6. Juran, B.D., Lazaridis, K.N. (2011). Genomics in the post-GWAS era. Semin. Liver Dis. *31*, 215-222.
7. Jia, P., Zheng, S., Long, J., Zheng, W., and Zhao, Z. (2010). dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics 27, 95-102.
8. Eleftherohorinou, H., Wright, V., Hoggart, C., Hartikainen, A.L., Jarvelin, M.R., Balding, D., Coin, L., Levin, M. (2009). Pathway Analysis of GWAS Provides New Insights into Genetic Susceptibility to 3 Inflammatory Diseases. PLoS one *4*, e8068. doi:10.1371/journal.pone. 0008068.
9. Kruglyak, L. (2008). The road to genome-wide association studies. Nat. Rev. Genet. *9*, 314–318.
10. Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J., Jin, L., Amos, C.I., Xiong, M. (2010). Gene and pathway-based second-wave analysis of genome-wide association studies. Eur. J. of Hum. Genet. *18*, 111-117.
11. Wang, K., Li, M., Bucan, M. (2007). Pathway-Based Approaches for Analysis of Genome-wide Association Studies. Am. J. of Hum. Genet. *81*, 1278-1283.
12. Wray, N., Visscher, P. (2008). Estimating trait heritability. Nat. Edu. *1*, 1-16.
13. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A., Kong, A., Kruglyak, L., Elaine Mardis, M.,

Rotimi, C., Slatkin, M., Valle, M., Whittemore, A., Boehnke, M., Clark, A., Eichler, E., Gibson, J., Haines, J., Mackay, T., McCarroll, S., Visscher, P. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

14. Orozco, G., Barrett, J.C., Zeggini, E. (2010). Synthetic associations in the context of genome-wide association scan signals. Hum. Mol. Genet. *19*, 137-144.
15. Ladouceur, M., Dastani, Z., Aulchenko, Y.S., Greenwood, C.M.T., Richards, J.B. (2012). The empirical power of rare variant association methods: Results from sanger sequencing in 1,998 Individuals. PLoS Genet. *8*, e1002496.
16. Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., Mesirov, J. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. of Sci. *102*, 15545–15550.
17. Tavazoie, S., Hughes, J., Campbell, M., Cho, R., Church, G. (1999). Systematic determination of genetic network architecture. Nat. Genet. *22*, 281-285.
18. Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C., Krawetz, S.A. (2003). Global functional profiling of gene expression. Genomics *81*, 98-104.
19. Jiang, Z., Gentleman, R. (2007). Extension to gene set enrichment. Bioinformatics *23*, 306-313.
20. Kim, S., Volsky, D. (2005). PAGE: parametric analysis of gene set enrichment. BMC bioinformatics *6*, doi: 10.1186/1471-2105-6-144.
21. Goeman, J.J., van de Geer SA, de K.F., van Houwelingen, H.C. (2004). A global test for groups of genes: Testing association with a clinical outcome. Bioinformatics. *20*, 93-99.
22. Tian, L., Greenberg, S., Kong, S.W., Altschuler, J., Kohane, I., Park, P. (2005). Discovering statistically significant pathways in expression profiling studies. Proc. Natl. Acad. of Sci. *102*, 13544-13549.
23. Chai, H.S., Sicotte, H., Bailey, K., Turner, S., Asmann, Y., Kocher, J. (2009). GLOSSI: a method to assess the association of genetic loci-set with complex diseases. BMC Bioinformatics *10*, doi:10.1186/1471-2105-10-102.
24. Holden, M., Deng, S., Wojnowski, L. and Kulle, B. (2008). GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics *24*, 2784-2785.
25. Nam, D., Kim, J., Kim, S., Kim, S.: GSA-SNP: a general approach for gene set analysis of polymorphisms, Nucleic Acids Res. *38(Suppl)*, 749-754.

26. Wang, L., Jia, P., Wolfinger, R.D., Chen, X., Zhao, Z. (2011). Gene set analysis of genome-wide studies: Methodological issues and perspectives. Genomics *98*, 1-8.
27. Fridley, B.L., Biernacka, J.M. (2011). Gene set analysis of SNP data: benefits, challenges, and future directions. Eur. J. of Hum. Genet. *19*, 837-843.
28. Bader, G.D., Cary, M.P., Sander, C. (2006). Pathguide: a pathway resource list. Nucleic Acids Res. *34,* D504–D506.
29. Kanehisa M, Goto S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.
30. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat. Genet. *25*, 25–29.
31. Bard, J.B., Rhee, S.Y. (2004). Ontologies in biology: design, applications and future challenges. Nat Rev. *5*, 213–222.
32. Viswanathan, G.A., Nudelman, G., Patil. S., Sealfon, S.C. (2007). BioPP: a tool for web-publication of biological networks. BMC Bioinformatics *8*, doi:10.1186/1471-2105-8-168.
33. Smith, A.V., Thomas, D.J., Munro, H.M., Abecasis, G.R. (2005). Sequence features in regions of weak and strong linkage disequilibrium. Genome Res. *15*, 1519–1534.
34. Zhong, H., Yang, X., Kaplan, L.M. (2010). Integrating pathway analysis and genetics of gene expression for genome-wide association studies. Am. J. of Hum. Genet. *86*, 581–591.
35. Lee, J., Ahn, S., Oh, S., Weir, B., Park,T. (2011). SNP-PRAGE: SNP-based parametric robust analysis of gene set enrichment. BMC Syst. Biol. *5(Suppl)*, doi:10.1186/1752-0509-5-S2-S1.
36. Chasman D: On the utility of gene set methods in genome-wide association studies of quantitative traits, Genetic Epidemiology 2008, 32:658-668.
37. Chen L, Zhang L, Zhao Y, Xu L, Shang Y, Wang Q, Li W, Wang H, Li X: Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways, Bioinformatics 2009, 25(2), 237-242
38. Yu K, Li Q, Bergen A, Pfeiffer R, Rosenberg P, Caporaso N, Kraft P and Chatterjee N: Pathway analysis by adaptive combination of P-values, Genetic Epidemiology 2009, 33(8): 700–709

39. Welch, B.L. (1947). The generalization of "Student's" problem when several different population variances are involved. Biometrika 34, 28-35.
40. Ko, K., Lee, J., Lund, R.B. (2008). Confidence intervals for long memory regression." Statistics and Probability Letters 78, 1894-1902.
41. Akaike, H. (1974). A new look at the statistical identification model, IEEE Transactions on Automatic Control *19*, 716-723.
42. Levinson, D., Holmans, P. (2004). The effect of linkage disequilibrium on linkage analysis of incomplete pedigrees, BMC genetics *6(Suppl 1)*, 10.1186/1471-2156-6-S1-S6.
43. Cho, Y.S., Go, M.J., Kim, Y.J., Heo J.Y., Oh, J.H., Ban, H-J, Yoon D., Lee, M.H., Kim, D-J., Park M., Cha, S-H., Kim, J.W., Han, B.G., Min, H., Ahn, Y., Park, M.S., Han, H.R., Jang, H.Y., Cho, E.Y., Lee, J.E., Cho, N.H., Shin, C., Park, T., Park, J.W., Lee, J.K., Cardon, L., Clarke, G., McCarthy, M.I., Lee, J.Y., Lee, J.K., Oh, B., Kim, H.L. (2009). A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative trait. Nat. Genet. *41*, 527-534.
44. Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. genet. *38*, 904-909.
45. Storey, J.D. (2002). Direct approach to false discovery rates, Journal of the Royal Statistical Sciety: Series B (Statistical Methodology) *64*, 479-498.
46. Storey, J.D. (2003). The positive false discovery rates: a Bayesian interpretation and the q-value. Ann. of Statist. *31*, 2013-2035.
47. Esposito, G., Perrino, C., Schiattarella, G.G., Belardo, L., di Pietro, E., Franzone, A., Capretti, G., Gargiulo, G. (2010). Induction of mitogen-activated protein kinases is proportional to the amount of pressure overload, Hypertension *55*, 137-143.
48. The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, Nature *447*, 661-678.
49. Barrett, T.B., Hauger, R.L., Kennedy, J.L., Sadovnick, A.D., Remick, R.A., Keck, P.E., McElroy, S.L., Alexander, M., Shaw, S.H., Kelsoe, J.R. (2003). Evidence that a single nucleotide polymorphism in the promoter of the G protein receptor kinase 3 gene is associated with bipolar disorder. Molecular Psychiatry. *8*, 546-557.
50. Hurd, Y.L. (2002). Subjects with major depression or bipolar disorder show reduction of prodynorphin mRNA expression in discrete nuclei of the amygdaloid complex. Molecular Psychiatry, *7*, 75-81.

51. Perez, D.I., Gil, C., Martinez, A. (2010). Protein Kinases CK1 and CK2 as New Targets for Neurodegenerative Diseases, Medical Research Reviews *31*, 924-954.
52. Robinson, R. (2010). Common disease, multiple rare (and distant) variants. PLoS Biol. *8*, e1000293.
53. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, K., Goldstein, D.B. (2010). Rare variants create synthetic genome-wide associations. PLoS Biol. 8, e1000294.
54. Bansal, V., Libiger, O., Torkamani, A., Schork, N.J. (2010). Statistical analysis strategies for association studies involving rare variants. Nat. Rev. Genet. *11*, 773-785.
55. Li, B., Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. Am. J. of Hum. Genet. *85*, 311-321.
56. Madsen, B.E., Browning, S.R. (2009). A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLoS Genet. *5*, e1000384.
57. Brown, M.B. (1975). A method for combining non-independent, one-sided tests of significance. Biometrics *31*, 987-992.
58. Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., Daly, M.J. (2011). Testing for an Unusual Distribution of Rare Variants. PLoS Genet. *7*, e1001322.
59. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke. M., Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. of Hum. Genet. *89*, 82-93.
60. Ng, P.C., Henikoff, S. (2001). Predicting deleterious amino acid substitutions. Genome Res. *11*, 863–874.
61. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
62. Chun, S., Fay, J.C. (2009). Identification of deleterious mutations within three human genomes. Genome Res. *19*, 1553-1561.
63. Lee, S., Lin, X., Wu, M.C. (2012). Optimal tests for rare variant effects in sequencing association studies. Biostatistics. *13*, 1-1
64. Morris, A.P., Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. *34*, 188-193.

65. Price, A.L., Kryukov, G.V., de Bakker, P.I, Purcell, S.M., Staples, J., Wei, L.J., Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am. J. of Hum. Genet. *86*, 832-838
66. Hoggart, C.J., Whittaker, J.C., De Iorio, M., Balding, D.J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet. 4, e1000130
67. Casci, T. (2006). Systems biology: Network fundamentals, via hub genes. Nat. Rev. Genet. *7*, 664-665.
68. Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Makela, T.P., Hautaniemi, S. (2009). Integrated network analysis platform for protein-protein interactions. Nat. Methods *6*, 75–77
69. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., von Mering, C. (2009). STRING: a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids. Res. *37*, 412–416
70. Li, B., Wang, G., Leal, S.M. (2012). SimRare: a program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits. Bioinformatics. *28*, 2703-2704
71. Matsuo, K., Hamajima, N., Shinoda, M., Hatooka, S., Inoue, M., Takezaki, T., Tajima, K. (2001). Krüppel-Like Factor 4, a Tumor Suppressor in Hepatocellular Carcinoma Cells Reverts Epithelial Mesenchymal Transition by Suppressing Slug Expression. Carcinogenesis *22*, 913-916.
72. Kim, Y.J., Go, M.J., Hu, C., Hong, C.B., Kim, Y.K., Lee, J.Y., Hwang, J-Y., Oh, J.H., Kim, D-J., Kim, N.H., Kim, S., Hong, E.J., Kim, J-H., Min, H., Kim, Y., Zhang, R., Jia, W., Okada, Y., Takahashi, A., Kubo, M., Tanaka, T., Kamatani, N., Matsuda, K., MAGIC consortium, Park, T., Oh, B., Kim, K., Kang, D., Shin, C., Cho, N.H., Kim, H-L., Han, B-G., Lee, J-Y., Cho, Y.S. (2011). Large-scale genome-wide association studies in east Asian identify new genetic loci influencing metabolic traits. Nat. Genet. *43*, 990-996.
73. Gao, B., Duan, Z., Xu, W., Xiong, S. (2009). Tripartite motif-containing 22 inhibits the activity of hepatitis B virus core promoter, which is dependent on nuclear-located RING domain. Hepatology *50*, 424-433.
74. Loew, M., Boeing, H., Stürmer, T., Brenner, H. (2003). Relation among alcohol dehydrogenase 2 polymorphism, alcohol consumption, and levels of gamma-glutamyltransferase. Alcohol *29*, 131-135.
75. Takada, H., Chen, N-J., Mirtsos, C., Suzuki, S., Suzuki, N., Wakeham, A., Mak, T.W., and Yeh, W.C. (2003) Role of SODD in Regulation of Tumor Necrosis Factor Responses. Mol. Cell. Biol. *23*, 4026-4033

초 록

전장유전체연관성분석(GWAS)은 주로 질환과 연관이 있는 개별적 단일염기변이(SNP)를 발굴하는 등 복합질환에 영향을 미치는 유전체 상의 공통변이 (common variant)를 찾는데 많은 기여를 하였다. 그러나 복합질환은 다중 인자의 복합적인 효과로 인해 발생하는 경우가 많아 기존의 방식으로는 그러한 유전요인을 효과적으로 검출할 수 없고 각종 질환의 유전율 (heritability) 의 일부분만을 설명할 수 있다는 한계점이 제시되었다.

이러한 문제점을 해결하기 위해 개별 유전인자를 검정하기보다는 생물학적인 Pathway정보를 활용하여 유전변이의 집합에 대한 검정을 하는 방법에 대한 연구가 활발히 진행되었다. 이러한 유전자집합분석(Gene set analysis)은 개별유전인자의 영향력이 작은 경우에 여러 인자의 효과를 병합함으로서 검정력을 높이는 데 기여를 하였다. 유전인자 상호간의 연관불균형(Linkage Disequilibrium)으로 인한 상관관계를 고려하기 위해 대부분의 유전자집합분석은 많은 계산량을 필요로 하는 비모수적 검정방법을 사용하였다. 본 연구에서는 유전인자 간의 상관관계를 고려하면서도 모수적 검정을 하는 방법을 제시하였다. 먼저 단일염기변이의 연관성측도 간의 상관관계를 고려하면서 유전자 단위로 요약하고 유전자 집합 내에서 평균을 이용한 통계량이 중심극한정리에 의해 정규분포를 따르는 것을 이용하였다.

또한 차세대 시퀀싱 기술의 발달로 희귀 변이(rare variant)에 대한 자료를 얻을 수 있게 되고 이러한 희귀변이가 복합질병에 미치는 영향력

에 대한 증거들이 제시되면서 기존에 일부분밖에 설명할 수 없었던 유전율을 설명하기 위한 희귀변이에 대한 연구가 활발히 진행되고 있다.

희귀변이의 특성상 소수의 사람들에게서 발견된다는 점에서 개별 인자 단위의 통계적 유의성을 얻기가 쉽지 않으므로 기존의 전장유전체 연관성분석방법을 그대로 적용하는데 한계가 있다. 이러한 희귀변이를 유전자 단위로 병합하여 검정하는 방법들이 많이 연구되고 있는데 각기 장단점을 가지고 있고 여러가지 환경 하에 일관된 검정력을 가지는 방법은 거의 없다. 이에 대해 여러 가지 희귀변이가 갖는 가정들을 복합적으로 고려하여 검정하는 방법을 제시하였고 기존의 희귀변이 연관성 분석 방법들과 성능을 비교하였다.

또한 유전자 단위로 병합된 희귀변이들을 유전자집합단위로 요약할 때 단백질간의 상호작용네트워크 (Protein-protein interaction) 또는 변이들간의 상호작용 (co-mutation)을 이용하여 가중치를 사용함으로써 질병에 영향을 미치는 요인을 검출하는 검정력을 높이고자 하였다.

이렇게 전장유전체연관성자료와 차세대 시퀀싱 자료 각각에 대한 유전자 집합분석 방법을 제시하였고 실제 데이터에 적용하였다. 본 연구에서 제시된 방법이 복합 질환에 영향을 미치는 유전인자의 집합을 효과적으로 발굴하고 질병이 발생하는 기작을 연구하는데 활용될 수 있을 것으로 기대된다.

주요어: 전장유전체연관성분석, 차세대시퀀싱자료분석, 유전자 집합분석, 희귀변이 연관성테스트 (rare variant association test), 공동변이 (co-mutation)

학 번: 2007-30077