이학박사학위논문

# Statistical analysis for large-scale sequencing dataset using pathway information

## 패스웨이 정보를 이용한 대용량 유전체 자료의 통계적 분석

**2018 년 2 월**

서울대학교 대학원

협동과정 생물정보학과

이성영

# Statistical analysis for large-scale sequencing dataset using pathway information

by

**Sungyoung Lee**

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

**Interdisciplinary Program in Bioinformatics**
**College of Natural Sciences**
**Seoul National University**
**Feb, 2018**

# Abstract

## Statistical analysis for large-scale sequencing dataset using pathway information

Sungyoung Lee
Interdisciplinary Program in Bioinformatics
The Graduate School
Seoul National University

In the past two decades, rapid advances in DNA sequencing technology have enabled extensive investigations into human genetic architecture, especially for the identification of genetic variants associated with complex traits. In particular, genome-wide association studies (GWAS) have played a key role in identifying genetic associations between Single Nucleotide Variants (SNVs) and many complex biological pathologies. However, the genetic variants identified by many successful GWAS have explained only a modest part of heritability for most of phenotypes, and many hypotheses have been proposed to address so-called “missing heritability” issue, such as rare variant association, gene-gene interaction or multi-omics integration.

Methods for rare variants analysis arose from extending individual variant-level approaches to those at the gene-level, and extending those at the gene level to multiple phenotypes. In this trend, as the number of publicly available biological resources is increasing, recent methods for analyzing rare variants utilize pathway knowledge as *a priori* information. In this respect, many statistical methods for pathway-based analyses using rare variants have been proposed to analyze pathways individually. However, neglecting correlations between multiple pathways can result in misleading solutions, and pathway-based analyses of large-scale genetic datasets require massive computational burden. Moreover, while a number of methods for pathway-based rare-variant analysis of multiple phenotypes have been proposed, no method considers a unified model that incorporate multiple pathways.

In this thesis, we propose novel statistical methods to analyze large-scale genetic dataset using pathway information, Pathway-based approach using HierArchical components of collapsed RAre variants Of High-throughput sequencing data (***PHARAOH***) and ***PHARAOH-multi***. ***PHARAOH*** extends generalized structural component analysis, and implements the method based on the framework of generalized linear models, to accommodate phenotype data arising from a variety of exponential family distributions. ***PHARAOH*** constructs a single hierarchical model that consists of collapsed gene-level summaries and pathways, and analyzes entire pathways simultaneously by imposing ridge-type penalties on both gene and pathway coefficient estimates; hence our method considers the correlation of pathways and handles an entire dataset in a single model. In addition, ***PHARAOH-multi*** further extends the original model into multivariate analysis, while keeping the advantages of our previous approach. We extend ***PHARAOH*** to enable analysis of multiple traits using hierarchical components of genetic variants. In addition, ***PHARAOH-multi*** can identify associations between multiple phenotypes and multiple pathways, with a single model, in the presence of subsequent genes within pathways, as a hierarchy.

Through simulation studies, ***PHARAOH*** was shown to have higher statistical power than the existing pathway-based methods. In addition, a detailed simulation study for ***PHARAOH-multi*** demonstrated advantages of multivariate analysis, compared to univariate analysis, and comparison studies showed the proposed approach to outperform existing multivariate pathway-based methods. Finally, we conducted an analysis of whole-exome sequencing data from a Korean population study to compare the performance between the proposed methods with the previous pathway-based methods, using validated pathway databases. As a result, ***PHARAOH*** successfully discovered 13 pathways for the liver enzymes, and ***PHARAOH-multi*** identified 8 pathways for multiple metabolic traits. Through a replication study using an independent, large-scale exome chip dataset, we replicated many pathways that were discovered by the proposed methods and showed their biological relationship to the target traits.

**Keywords**: Genome-wide association studies (GWAS), Next-Generation Sequencing (NGS), rare-variant association test, multivariate analysis, generalized linear model

**Student number**: 2010-23162

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1. The background on genetic association studies

### 1.1.1. Genome-wide association studies and the missing heritability

Over the past two decades, drastic advances in biological sequencing technology have enabled extensive and comprehensive investigations into genetic architecture of many organisms including human, especially for the identification of genetic variants associated with complex traits. In this trend, Genome-Wide Association Study (GWAS) has played a key role in identifying genetic associations between Single Nucleotide Variants (SNVs) and many complex biological pathologies. In general, the term SNV is defined as a variation in a specific position of DNA that commonly occurs in a subset of a population. Due to its bi-allelic property, a single SNV can be coded as the number of minor alleles: 0, 1 and 2.

The ultimate goal of the GWAS is to identify so-called "causal variants" that are statistically associated with the interested traits in the population (Visscher, et al., 2012). In the context of GWAS, the causal variant is defined as a genetic variant that is statistically associated with increased or decreased individual risk of an interested complex trait.

The first concept of GWAS was proposed to overcome the limitations of linkage analysis (Risch and Merikangas, 1996), and they predicted that the future of the researches on complex traits will require massive number of testing by GWAS, by showing the association studies are more powerful than the linkage study design in terms of detection of common variants. The GWAS has successfully identified more than 50,000 common genetic variants associated with many complex human traits (Altshuler, et al., 2008; Manolio, et al., 2008; McCarthy, et al., 2008). Since the first major GWAS was reported in 2007 (Sladek, et al., 2007), such studies have rapidly grown in scale and complexity, and 3,197 curated publications of 53,020 SNPs have been added to the catalog of published Genome-wide Association Studies, as of November 2017 (**Figure (1.1)**).

However, even while the number of detectable genetic variants increases, the proportion of the variance of complex traits explained by common variants has been generally very small. At the beginning, two groups of researchers were claimed this problem at first (Maher, 2008; Manolio, et al., 2009). The studies showed that more than 40 genetic variants that are newly identified via the GWAS for human heights accounted for approximately 5% of the heritability of height (Maher, 2008), and found that the same problem occurs in the GWAS on many other complex traits such as type 2 diabetes (T2D) or early-onset myocardial infarction (Manolio, et al., 2009).

In detail, this so-called "missing heritability" became a serious problem for two reasons. First, at the early stage of GWAS, it was expected to unveil the underlying genetic architecture of several complex diseases, from the earlier estimates on their heritability. However, some unsuccessful results of the unprecedented scale of GWAS led to admit the limitation of traditional GWAS. Second, although steep decline of sequencing costs has been enabled extensive validations of the original findings through larger GWAS, only a number of the genetic variants has been replicated, while the others were found to be false positives (Duncan and Keller, 2011).

In that regard, three major hypotheses have been proposed to explain the missing heritability issue: polygenic effect, gene-gene interaction, and rare variant. For the polygenic effect, it assumes that an actual genetic effect of common SNVs behaves a polygenic way, rather than an additive way. This resolution was demonstrated by a series of studies, which showed that trait similarity could be predicted by the genetic similarity of a population on common SNVs treated additively, and for many traits the SNP heritability was indeed a substantial fraction of the overall heritability (Yang, et al., 2011). In the other hand, for the gene-gene interaction, one study showed that nonignorable amount of missing heritability could arise from overestimation of the denominator and led to create “phantom heritability” (Zuk, et al., 2012). The phantom heritability was found to be accounted for the presence of gene-gene interaction, since the statistical model that was used to estimate the heritability assumes the target trait has no effect of genetic interaction (or epistasis) among genetic loci. Finally, some technical limitation of the early stage GWAS and the common disease-common variant (CDCV) hypothesis obstructed an access to the rare variants. Despite an existence of missing heritability, the rare variant study was shown to be not feasible by GWAS (Li and Leal, 2008; Wu, et al., 2011). However, the advance of sequencing technology now enabled deeper level of extensive investigations into human genetic architecture using rare variants via whole-genome or whole-exome sequencing. Recent efforts on the analysis of rare variants have been successfully identified statistically significant genes associated with several complex traits, including high-density lipoprotein levels, obesity, schizophrenia, and multiple cancer types (Ahituv, et al., 2007; Brunham, et al., 2006; Cohen, et al., 2004; Slatter, et al., 2008; Walsh, et al., 2008).

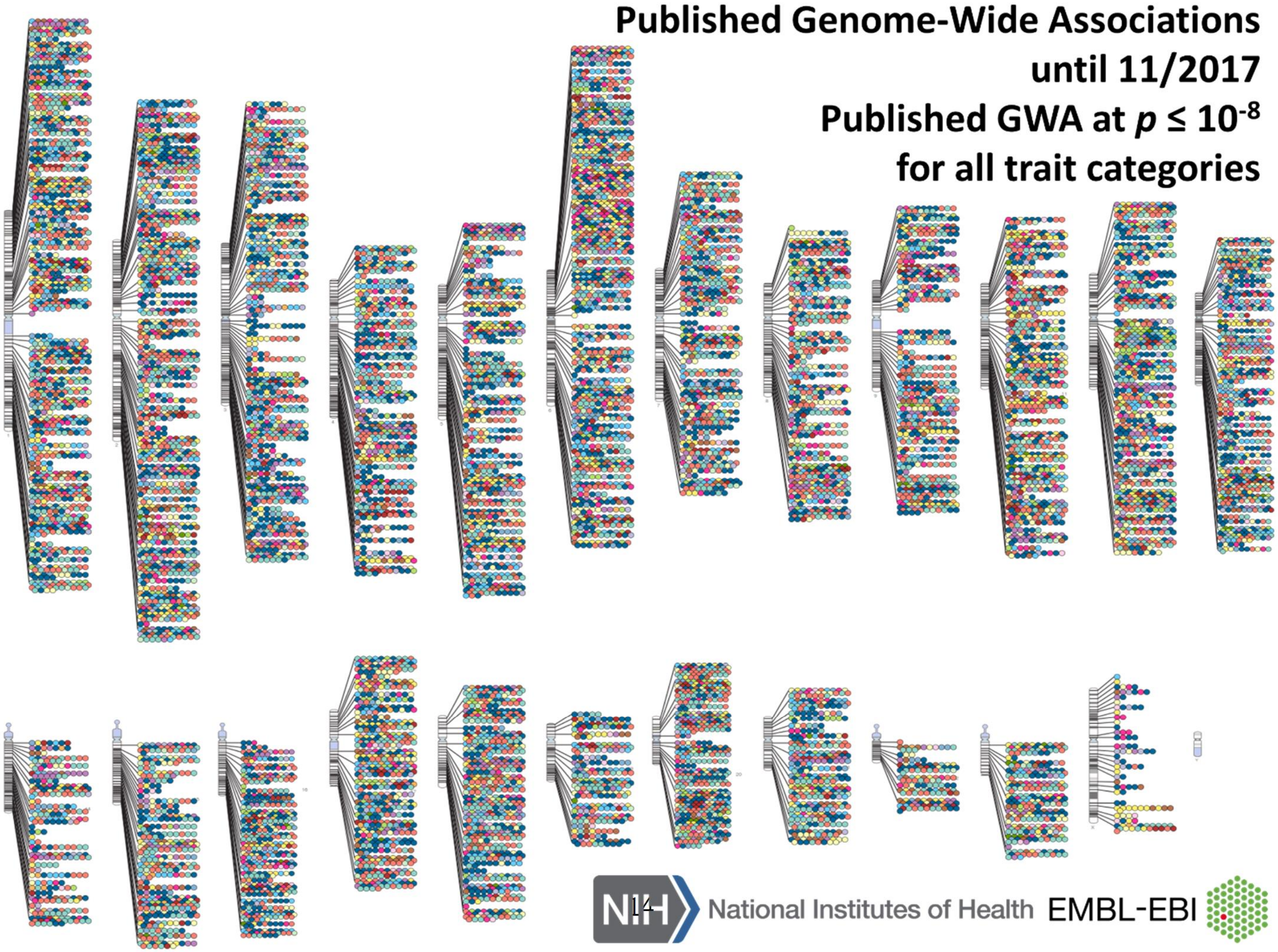


**Figure 1.1. GWAS catalog as of 11/2017.** This diagram shows all SNP-trait association with *p*-value $\leq 5.0\times10^{-8}$.

### 1.1.2. Rare variant analyses

To deal with the sparseness of rare variants, early approaches simply aggregated multiple rare variants of a gene by the existence of minor alleles or by summation of minor alleles (Li and Leal, 2008; Price, et al., 2010). By contrast, more recent methods seek to consider biological information, such as linkage disequilibrium, and the biological effects of genetic variants, to enhance biological interpretation (Hu, et al., 2013; Wu, et al., 2011). These approaches have been useful for identifying statistically significant genes associated with several complex traits, including high-density lipoprotein levels, obesity, schizophrenia, and multiple cancer types (Ahituv, et al., 2007; Brunham, et al., 2006; Cohen, et al., 2004; Slatter, et al., 2008; Walsh, et al., 2008).

Most approaches for identifying rare variants focus mainly on individual gene analysis. However, it has now been recognized that a majority of biological behaviors manifest from a complex interaction of biological pathways (Costanzo, et al., 2010; Hirschhorn, 2009). For instance, the pathway-based approaches have been developed for analysis of cancer-based dataset, by investigating patterns of somatic mutations in the cancer samples (Vandin, et al., 2011). In this respect, using pathway or gene-set information to analyze next generation sequencing data has several advantages in addressing the multiple testing problem and improving biological interpretation. First, it is possible to dramatically reduce the number of tests, because tens of millions of SNVs or tens of thousands of genes are grouped into hundreds of pathways. By grouping such large numbers of SNVs into pathways, pathway-based analysis is much less restricted by multiple testing problems, even compared to gene-based analyses. Second, interpreting statistically significant pathways can be easier than interpreting individual SNVs or genes. By analyzing pathway information that associates with biological processes, components or structures, the underlying bases for biological traits can be characterized more intuitively than by examining individual genes (Khatri, et al., 2012; O'Dushlaine, et al., 2009). Moreover, many successful discoveries of pathways that underlie complex traits have proven the utility of pathway-based analysis (Askland, et al., 2009; International Multiple Sclerosis Genetics, 2013; Lesnick, et al., 2007).

However, these methods are mainly designed for the analysis of common variants and are not suitable for analysis of rare variants including the most recent pathway-level analyses using genetic information such as linkage disequilibrium or gene-environmental interaction (Lamparter, et al., 2016; Qian, et al., 2016).

Recent pathway-based methods for the analysis of rare variants have extended gene-based analysis methods for rare variants by aggregating *P* values from each gene-based test, or extending existing powerful gene-based tests to pathways (Wu and Zhi, 2013; Yan, et al., 2014; Zhao, et al., 2014). For example, the Weighted Kolmogorov-Smirnov (WKS) method, the Direct Region-Based (DRB) method (Wu and Zhi, 2013), and Smoothed Functional Principal Component Analysis (SFPCA) (Zhao, et al., 2014) are approaches that extend pathway-based analyses of GWAS data to pathway-based analyses of high-throughput sequencing data. The WKS method, a modification of Gene Set Enrichment Analysis (GSEA) (Wang, et al., 2007), uses the results of single-variant analysis. Moreover, DRB methods have extended existing gene-based methods, including the Burden type (Li and Leal, 2008), C-alpha type (Neale, et al., 2011; Wu, et al., 2011) and Optimal type (Lee, et al., 2012), to pathway analysis for rare variants.

However, there are several limitations to using current pathway approaches to identify rare variants. First, a substantial number of genes are shared by pathways, potentially leading to high correlations between pathways. Thus, neglecting these correlations can result in misleading solutions. For example, high correlations between pathways can yield highly correlated results or confound the interpretation of significant pathways (Alexa, et al., 2006; Jiang and Gentleman, 2007; Skarman, et al., 2012). Second, the multiple testing problem is another challenge for current pathway-based analyses. Although the number of pathway-based tests is far less than that of variant-level or gene-based tests, the required *P* value threshold by Bonferroni correction is quite small, leading to low statistical power. In addition, methods using permutation tests suffer from a heavy computational burden to obtain more precise *P* values, when the *P* value threshold is very small.

Another effort to enhance the power of rare variants is to develop multivariate analysis methods. In general, many complex diseases arise from multiply correlated traits. For example, according to American Diabetes Association guidelines, diabetic status is diagnosed based on four traits: fasting glucose, two hours after plasma glucose, random plasma glucose, and HbA1c (American Diabetes, 2014). In that regard, simultaneous analysis of those correlated traits offer two substantial advantages over univariate analysis. First, multivariate analysis can elevate statistical power to identify additional causal biomarkers, which are not discovered by single phenotype analysis. Second, by analyzing multiple traits at once, the required number of statistical tests can be reduced, compared to those of univariate analysis. Those advantages have been well documented in past studies of large-scale sequencing datasets (O'Reilly, et al., 2012; Yang and Wang, 2012).

There have now been many applications of multivariate analysis to large-scale datasets. In particular, for variant- and gene-level analysis, many multivariate methods, for common and rare variants, have been proposed (O'Reilly, et al., 2012; Zhou and Stephens, 2014). Despite those efforts, only a number of pathway-based multivariate analyses have been deemed feasible. Recently, three multivariate approaches, for region-level analyses, were proposed: MARV, aSPU, and MURAT. MARV (Kaakinen, et al., 2017) uses a statistical approach, reverse regression, to investigate associations between genetic regions and multiple phenotypes, by treating phenotypes as independent variables, hence enabling rapid multivariate analysis of large-scale datasets. On the other hand, aSPU (Kwak and Pan, 2016), extends an original concept, data-adaptive sum of powered score test, to multivariate analysis, using summary statistics from single SNVs. For multivariate extension of powerful gene-based tests, MURAT (Multivariate Rare-variant Association Test) extended the original SKAT (sequence kernel association test) method to multiple phenotypes (Sun, et al., 2016). However, it might not be adequate to apply SKAT-based methods to pathway-based analysis, as we have previously demonstrated (Lee, et al., 2016). Moreover, none of the above methods are available for multivariate pathway-based association tests for rare variants with multiple pathways. Since the established pathway databases have substantial overlap among their pathways, they may ignore significant correlations between pathways, leading to misleading biological interpretations (Alexa, et al., 2006; Skarman, et al., 2012).

## 1.2. The purpose of this study

In this study, the primary purpose is to develop a statistical methods for investigation of rare variants using biological pathway information, to identify causal pathologies of the complex traits. To this end, the thesis is focused on two studies that are closely relevant each other. The first one is a study to develop a novel statistical method that is based on Generalized Linear Model (GLM) and utilizes an entire pathway information. The second one is an extended study of the first one and addresses the multivariate analysis.

In the first study, we introduce a novel statistical approach for the analysis of rare variants using pathways, named **P**athway-based approach using **H**ier**A**rchical components of collapsed **RA**re variants **O**f **H**igh-throughput sequencing data (***PHARAOH***). Our method has several unique distinctive features. First, ***PHARAOH*** can examine associations between a phenotype and entire pathways with a single model, using collapsed rare variants derived from gene information. Using this model, ***PHARAOH*** can evaluate effects of pathways to the phenotype, in addition to effects of genes to the phenotype via the pathway. Thus, ***PHARAOH*** provides an expansive view of biological processes underlying the trait of interest by examining entire pathways. Second, ***PHARAOH*** can account for potential correlations between pathways by imposing a ridge penalty on the effects of pathways on a phenotype. ***PHARAOH*** also adds another ridge penalty on the weights of genes to their corresponding pathways, allowing consideration of potential correlations between genes.

In the second study, we introduce a new method, "***PHARAOH-multi***" (**P**athway-based approach using **H**ier**A**rchical component of collapsed **RA**re variants **O**f **H**igh-throughput sequencing data), for analyzing **multi**ple phenotypes. Here, while keeping the advantages of the first approach, we extend it to enable analysis of multiple traits using hierarchical components of genetic variants. In addition, the proposed model can identify associations between multiple phenotypes and multiple pathways, with a single model, in the presence of subsequent genes within pathways, as a hierarchy.

## 1.3. Outline of the thesis

This thesis is organized as follows. Chapter 1 is an introduction to this study with a brief history from the GWAS to the beginning of rare variant analysis that was inspired by the missing heritability problem. Chapter 2 introduces the existing pathway-based methods for large-scale sequencing dataset and the generalized structured component analysis that motivated our analysis. Chapter 3 describes the pathway-based analysis based on the hierarchical components of rare variants. Chapter 4 is about the multivariate pathway-based analysis using the hierarchical components of rare variants. In Chapters 3 and 4, a detailed description of the proposed method, simulation studies, and the real dataset analyses are introduced. Finally, the summary and conclusions of this thesis are presented in Chapter 5.

# Chapter 2

## An overview of existing methods

### 2.1. Review of pathway-based methods

It has now been recognized that a majority of biological behaviors manifest from a complex interaction of biological pathways (Costanzo, et al., 2010; Hirschhorn, 2009). In this respect, using pathway or gene-set information to analyze next generation sequencing data has several advantages in addressing the multiple testing problem and improving biological interpretation. First, it is possible to dramatically reduce the number of tests, because tens of millions of SNVs or tens of thousands of genes are grouped into hundreds of pathways. By grouping such large numbers of SNVs into pathways, pathway-based analysis is much less restricted by multiple testing problems, even compared to gene-based analyses. Second, interpreting statistically significant pathways can be easier than interpreting individual SNVs or genes. By analyzing pathway information that associates with biological processes, components or structures, the underlying bases for biological traits can be characterized more intuitively than by examining individual genes (Khatri, et al., 2012; O'Dushlaine, et al., 2009). Moreover, many successful discoveries of pathways that underlie complex traits have proven the utility of pathway-based analysis (Askland, et al., 2009; International Multiple Sclerosis Genetics, 2013; Lesnick, et al., 2007).

Many pathway-based analysis methods have been proposed by using existing pathway databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) by the molecular pathway, Gene Ontology (GO) by the function of genes, Biocarta or Molecular Signatures Database (MSigDB). Since the term “pathway-based analysis” covers very broad context according to the literatures (Green and Karp, 2006), Wang et al. classified existing pathway analyses into two categories according to their null hypothesis: self-contained and competitive (Wang, et al., 2010). This classification also can be applied to classify pathway-level analysis methods using genetic variants. According to the two categories of gene-set (pathway) analysis, there have been two well-addressed reviews of the pathway analysis (de Leeuw, et al., 2016; Khatri, et al., 2012). Based on the previous efforts, the current pathway-based analysis can be categorized into three categories, as follows.

*Competitive test* – This type of test assumes that the genes in a gene-set are associated with the phenotype with same degree to the other gene-sets. Hence, the gene-sets need to “compete” each other to be statistically significant.

*Self-contained test* – This type of tests do not consider the other gene-sets. Instead, it only assumes the genes in a gene-set are not associated with the target phenotype.

*Overall test* – While the former two approaches test the association of an individual gene-set, this type of tests assumes that any gene-set within the dataset is associated with the phenotype of interest. Note that the proposed methods in this thesis are included in this type.

Owing to the previous efforts, many researchers have proposed the pathway-based methods for analysis of large-scale sequencing dataset by extending those three types of gene-set analysis methods. However, GWAS and large-scale sequencing has a major difference. In general, the sequencing dataset consists of a complete set of the qualified genetic variants that are found in the sequenced samples. In contrast, the genetic variants of GWAS are often predefined by the chip design which is intended to maximize the genomic mapping, by tagging the representative variants. In this respect, there are some issues that need to be properly addressed prior to the extension. First, despite the advantages of the tagging approach, the GWAS might be less accurate for the rare variants. Second, the sequencing dataset has a chance to incorporate many *de novo* rare variants that might have true causal effects, while the GWAS dataset does not. The ability of *de novo* rare variant discovery is a major advantage of the sequencing dataset and must be considered in the extension.

In this chapter, we would like to introduce some existing approaches for each category to substantiate the issues we tried to overcome.

### 2.2.1. Competitive and self-contained tests: WKS and DRB

Recent pathway-based methods for the analysis of rare variants have extended gene-based analysis methods for rare variants by aggregating *P* values from each gene-based test, or extending existing powerful gene-based tests to pathways (Wu and Zhi, 2013; Yan, et al., 2014; Zhao, et al., 2014). Among those methods, the Weighted Kolmogorov-Smirnov (WKS) method is a modification of Gene Set Enrichment Analysis (GSEA) (Wang, et al., 2007) and uses the results of single-variant analysis. From the nature of GSEA, WKS is a competitive method for pathway-based analysis. In the other hand, the Direct Region-Based (DRB) method (Wu and Zhi, 2013) extended existing gene-based methods, including the Burden type (Li and Leal, 2008), C-alpha type (Neale, et al., 2011; Wu, et al., 2011) and Optimal type (Lee, et al., 2012), to pathway analysis for rare variants.

Wu and Zhi (2013) proposed a framework to extend existing gene-based methods to the pathway-based methods. The research stated that there are two variations to extend the region-based rare variant association methods to pathway-level analysis. Although these rare variant association methods are originally designed for gene-based and region-based tests, these methods can be extended to the pathway-level by simply relaxing the definition of regions to all variants nested within whole genes in one pathway, regardless of contiguity. The second one is to run these methods at gene-based level and then rank gene-based $p$-values to form the WKS-based enrichment score for pathway-level analysis. Moreover, those two applications of region-based rare variant methods differ with their null hypotheses: the former tests the self-contained hypothesis and the latter tests the competitive hypothesis.

For the calculation of WKS-based extension, only a small modification is required to the original WKS. In the original WKS algorithm (Wang, et al., 2007), for each SNP $V_i$ ($i = 1, \cdots, L$, where $L$ is the total number of SNPs in a GWA study), a test statistic from any variant-level test is calculated at first. Then, the algorithm associates SNP $V_i$ with gene $G_j$ ($j = 1, \cdots, N$, where $N$ is the total number of genes represented by all SNPs) according to the gene-variant mapping. For each gene, we assigned the minimum $p$-values among all SNPs mapped to the gene as the statistic value of the gene. For all $N$ genes that are represented by SNVs in the GWA study, we sorted their statistic values from the largest to the smallest, denoted by $r_{(1)}, \cdots, r_{(N)}$. For any given gene set $S$, composed of $N_H$ genes, then a WKS statistic is calculated. Here, the WKS statistic reflects the overrepresentation of genes within the set $S$ at the top of the entire ranked list of genes in the genome:

$$N_R = \Sigma_{G_{j^*} \in S} \left| r_{(j^*)} \right|^p , \tag{2.1}$$

$$ES(\mathrm{S}) = \max \left( \sum_{G_j^* \in S, j^* \leq j} \frac{\left| r_{(j^*)} \right|^p}{N_R} - \sum_{G_j^* \notin S, j^* \leq j} \frac{1}{N - N_H} \right), \tag{2.2}$$

where $p$ is a parameter that gives higher weight to genes with extreme statistic values. When $p = 0$, this test statistic reduces to a regular Kolmogorov-Smirnov statistic, which identifies groups of genes whose $r_j$ distribution differs from that of a random gene set. The enrichment score, $ES(S)$, measures the maximum deviation of concentration of the statistic values in gene set $S$ from a set of randomly picked genes in the genome. Therefore, if the association signal in $S$ is concentrated at the top of the list, then $ES(S)$ will be high. Finally, Statistical significance and adjustment for multiple hypothesis testing are done by the permutation-based procedure (Wang, et al., 2007).

As shown in **Equation (2.1)** and **(2.2)**, the WKS statistic can be adapted to the pathway-based analysis by altering the descending-sorted gene-level statistics $r_{(1)}, \cdots, r_{(N)}$ to those for analysis of the sequencing dataset. In this respect, Wu and Zhi (2013) proposed several WKS-based tests, in addition to the direct-type (i.e., altering the region to multiple regions of the pathway) extensions of existing gene-level analysis.

### 2.2.2. Self-contained test: aSPU

Similar to other pathway-based tests for large-scale sequencing dataset, adaptive sum of powered score (aSPU) test was originally proposed for analysis of rare variants (RVs). The key idea of the aSPU is constructing a class of tests overweighting a sequence of increasingly smaller sets of the top-ranked SNVs, then selecting the test with the most significant result. By using this strategy, aSPU showed that it outperforms other tests when the set size is small (Pan, et al., 2014). In the extension of the aSPU to pathway-based analysis, one more layer for the genes was introduced as done in the layer of genetic variants.

The pathway-based aSPU model is defined from logistic regression. From the logistic regression model with $k$ SNVs in a pathway, the score vector $U$ for the coefficients and its covariance is derived. Here, the traditional score statistic $T_S$ defined as $T_S = U'V^{-1}U$, but the power of this statistic diminishes as the $k$ gets larger than the number of samples $N$. In order to overcome this issue, the original aSPU suggests the modified statistic of a class of sum of powered score (SPU). The SPU statistic $T_{SPU}$ is essentially a generalization of existing score-based statistics, by defining their statistic as

$$T_{SPU(\gamma)} = \sum_{i=1}^{k} U_i^{\gamma}. \tag{2.3}$$

Pan, et al. (2014) showed that **Equation (2.3)** become any of the three score-based statistics they addressed, according to the value of $\gamma$. Using the property of $T_{SPU}$. They proposed the aSPU statistics as

$$T_{aSPU} = \min_{\gamma \in \Gamma} P_{\gamma}, \tag{2.4}$$

where $\Gamma$ is a set of predetermined values, and $P_{\gamma}$ is the $p$-value that was calculated using $T_{SPU(\gamma)}$. Finally, the permutation procedure is applied to derive the null distribution of $T_{aSPU}$, to yield the $p$-value.

From the original aSPU, the pathway-based aSPU method (referred as "PathSPU" is now defined by extending **Equation (2.4)** and introducing two extra parameters for the gene and the pathway. The PathSPU statistic is defined by the gene-level statistic and the pathway-level statistic which aggregates the gene-level statistic, and those two statistics are defined as

$$SPU(\gamma, w_g; g) = \|U_{g+}\|_{\gamma} = \left( \sum_{j=1}^{k_g} (w_{gj} U_{gj})^{\gamma} / k_g \right)^{\gamma^{-1}}, \tag{2.5}$$

$$PathSPU(\gamma, \gamma_G, \widetilde{w}_1, \dots, \widetilde{w}_{|S|}, \tilde{\beta}; S) = \sum_{g \in S} \left( \beta_g SPU\left(\gamma, w_{g1}, \cdots, w_{gk_g}; g\right) \right)^{\gamma_G}, \tag{2.6}$$

where $w_{gj}$ are gene-specific weights for SNVs in the gene and contain prior information of SNVs (e.g., inverse-MAF weight), and $\tilde{\beta}$ contains gene functional annotations or gene expression data to represent prior likelihoods of their being functional (Pan, et al., 2014).

Similar to the original SPU method, **Equation (2.6)** should be assessed in order to derive the statistical significance. Similar to **Equation (2.2)**, the pathway-based aSPU statistic is defined as

$$T_{aSPUpath} = \min_{\gamma \in \Gamma, \gamma_G \in \Gamma_G} P_{PathSPU(\gamma, \gamma_G, \tilde{w}_1, \ldots, \tilde{w}_{|S|}, \tilde{\beta}; S)}, \tag{2.7}$$

where $\Gamma_G$ is a set of predetermined values for $\gamma_G$. Pan, et al. (2015) suggested that an optimal choice of $\gamma$ and $\gamma_G$ is difficult to determine, since the optimal situation depends on the combination of effect sizes and their directions. In this respect, a grid search over a wide range of possible values for the choice of $\gamma$ and $\gamma_G$ was suggested with the values of the grid, but in empirical and heuristic way. For the calculation of *p*-value, the same permutation technique in the original aSPU was used.

## 2.2.3. Self-contained test: MARV

Despite the efforts for multivariate analysis using rare variant, only a number of pathway-based multivariate analyses have been deemed feasible. Among them, Multi-phenotype Analysis of Rare Variants (MARV) is one of the available multivariate analysis using rare variants. Unlike other existing methods, MARV implements the joint analysis of multiple phenotypes through a special technique called reverse regression.

The MARV method is based on the burden test approach (Li and Leal, 2008), in which rare variants within a predefined region (e.g., genomic region) are collapsed into one variable. In order to construct the model, MARV first calculates the frequency of rare variants, which is defined as $r_i n_i^{-1}$, where $r_i$ is the number of minor alleles at RVs and $n_i$ is the total number of RVs. Next, unlike the other methods, this proportion is modeled as a linear combination of $Q$ phenotypes. In other words, the MARV model uses "reverse regression" as compared to standard GWAS, with the genotype data as the outcome and the phenotypes as the predictors. Thus, the model becomes

$$\mathrm{E}\left(r_i n_i^{-1}\right) = \beta_0 + \sum_{q=1}^{Q} \beta_q y_{iq}, \tag{2.8}$$

where $y_{iq}$ denotes an observation of the $q^{\text{th}}$ phenotype for the $i^{\text{th}}$ individual, and $\beta_0$ and $\beta_q$ represent the slope and the coefficients for the $q^{\text{th}}$ phenotype, respectively. Since the MARV is based on the standard regression model, the estimates of $\beta_0$ and $\beta_q$ are simply obtained via least squares estimation from **Equation (2.8)**. Then, a likelihood ratio test is constructed by comparing the weighted log likelihoods of the fitted model against a null model where $\beta_1 = \cdots = \beta_Q = 0$ . The test statistic has an approximate $\chi^2$ distribution with $K$ degrees of freedom.

The MARV method has several advantages compared to many existing pathway-based approaches, since its basis is the reverse regression, First, its statistical assumption is parametric, hence the calculation of $p$-value does not require the permutation procedure which is computationally expensive. Second, it can accommodate both quantitative and binary phenotypes.

For discovery purposes, the full model, including all the phenotypes is fitted. However, to allow further investigation of the loci reaching genome-wide significance after correction for multiple testing to take into account the number of regions tested within the analysis, an implementation of the method also analyze all phenotype combinations. For model selection purposes, MARV further calculates the Bayesian information criterion.

## 2.3. Generalized structured component analysis

### 2.3.1. The model

In this thesis, the proposed approach is essentially based on Generalized Structured Component Analysis (GSCA) (Hwang and Takane, 2004). The GSCA was originally inspired by Partial Least Squares (PLS) method that has been employed for the analysis with latent variables. Since the PLS lacks a global optimization criterion and a method of evaluating the goodness-of-fit, the GSCA was introduced to overcome the limitations of PLS (Hwang and Takane, 2004). The GSCA model consists of the following three submodels, defined as

$$X = FC + \epsilon, \tag{2.9}$$

$$F = FB + \zeta, \tag{2.10}$$

$$F = XW, \tag{2.11}$$

where $X$ is a $N \times P$ matrix of the observations and assumes that its columns are standardized, $F$ is a $N \times K$ matrix of the latent variables, $W$ is a $P \times K$ matrix of "weighted relation" that are associated with the exogenous variables, $C$ is the so-called "measurement (outer) model" matrix with the dimension of $K \times P$, that relates the components to their observed variables, $B$ is the so-called "structural (inner) model" matrix with the dimension of $K \times K$, that relates the latent variables, and $\epsilon$ and $\zeta$ are the error vectors (Hwang and Takane, 2004). Here, $N$, $K$ and $P$ are the number of observations, the number of latent variables and the number of manifest variables, respectively.

Unlike the PLS, GSCA introduces a unified model to yield an optimizing equation that is consistently minimized to obtain the estimates of model parameters. In this manner, GSCA overcomes the addressed limitations of PLS, while providing all the strengths of PLS, such as a liberal assumption of the underlying distribution or an avoidance of improper solution. With a straightforward manner, the unified GSCA model is given as

$$\begin{aligned} [X \;\; F] &= F[C \;\; B] + [\epsilon \;\; \zeta] \\ X[I \;\; W] &= XW[C \;\; B] + [\epsilon \;\; \zeta] \\ XU &= XWA + e. \end{aligned} \tag{2.12}$$

As shown in **Equation (2.12)**, the model resembles traditional linear model, hence the following single equation for least square that simultaneously estimates $U$ (essentially $W$) and $A$ by minimizing $e_i$, can be derived.

$$\phi_{GSCA} = \sum_{i=1}^{N} (X_i'U - X_i'WA)'(X_i'U - X_i'WA) \tag{2.13}$$

### 2.3.2. Parameter estimation

The issue on the minimization of **Equation (2.13)** is that there are two parameters in the model, and those two parameters are related each other. In order to overcome this issue, a simple algorithm based on Alternating Least Square (ALS) was developed (de Leeuw, et al., 1976). In brief, the extended ALS algorithm for GSCA consists of two steps: updating $A$ for the fixed $W$ and updating $W$ for the fixed $A$. It is trivial that the first step becomes an ordinary least square form. However, the second step requires to consider both part of **Equation (2.13)**, we need to perform column-wise update for each the *k*th column, using the manner that is similar to the first step.

This "alternation" of the estimating parameters is repeated until convergence criterion is satisfied. On the convergence, Hwang and Takane (2004) showed that the ALS algorithm assures its convergence, by showing the proposed ALS algorithm monotonically decreases the value of criterion in **Equation (2.13)**. In addition, the authors also suggested two convenient ways to maximize the possibility of reaching global minimum (Hwang and Takane, 2004).

This chapter was published in *Bioinformatics*
as a partial fulfillment of Sungyoung Lee's Ph.D program.

# Chapter 3

## Pathway-based approach using rare variants

### 3.1. Introduction

In this chapter, we introduce a novel method for analysis of we propose a novel statistical approach for the analysis of rare variants using pathways, named **P**athway-based approach using **H**ier**A**rchical components of collapsed **RA**re variants **O**f **H**igh-throughput sequencing data (***PHARAOH***). The proposed method has several unique distinctive features. First, ***PHARAOH*** can examine associations between a phenotype and entire pathways with a single model, using collapsed rare variants derived from gene information. Using this model, ***PHARAOH*** can evaluate effects of pathways to the phenotype, in addition to effects of genes to the phenotype via the pathway.

Thus, ***PHARAOH*** provides an expansive view of biological processes underlying the trait of interest by examining entire pathways. Second, ***PHARAOH*** can account for potential correlations between pathways by imposing a ridge penalty on the effects of pathways on a phenotype. ***PHARAOH*** also adds another ridge penalty on the weights of genes to their corresponding pathways, allowing consideration of potential correlations between genes. In this regard, ***PHARAOH*** is a doubly ridge-regularized method (Hwang, 2009). Although there is a number of alternative penalization approach such as LASSO (Tibshirani, 1996) or Elastic-Net (Zou and Hastie, 2005), we choose ridge method from its computational efficiency.

The proposed method is capable to analyze a massive genetic dataset which consists of several thousands of samples and tens of millions of genetic variants. In order to make it possible within a reasonable time, ***PHARAOH*** is implemented with C/C++ and is capable to complete an analysis of such large dataset within several hours. Through simulation studies, the proposed method was shown to have higher statistical power than the existing pathway-based methods and gene-based methods. In addition, using large-scale, whole-exome sequencing data from a Korean population study of liver enzyme levels, ***PHARAOH*** was compared to several existing pathway-based analyses of genetic variants, using two well-known pathway databases Biocarta (http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways) and KEGG (Kanehisa, et al., 2004). These comparisons demonstrated that ***PHARAOH*** not only identified associated pathways, with no need for multiple comparisons, but also successfully detected biologically plausible pathways for a phenotype of interest. Furthermore, we developed the ***PHARAOH*** software to provide a graphical display of pathway-based analysis results, thus allowing for easy and detailed interpretations.

With extensive simulations, we compared the proposed methods with existing methods (He, et al., 2014; Zhu and Xiong, 2012), and results showed that the proposed methods were the most efficient in the considered scenarios. Application of the proposed method to schizophrenia and GAW17 illustrated its practical value in real analyses.

## 3.2. Methods

### 3.2.1. Notations and the model

We assume that a dataset has $N$ samples for the interested traits. Then, let us assume that $y_i$ is the $i^{\text{th}}$ observation on a clinical phenotype arising from a distribution in the exponential family $(i = 1, \cdots, N)$. The density function or probability distribution for $y_i$ can be generally expressed as

$$p(y_i; \gamma_i, \delta) = \exp\left(\left(y_i\gamma_i - \xi(\gamma_i)\right)/\zeta(\delta) + \nu(y_i, \delta)\right) \tag{3.1}$$

for some known functions $\xi(\cdot)$, $\zeta(\cdot)$, and $\nu(\cdot)$. If the dispersion parameter $\delta$ is known, **Equation (3.1)** belongs to the exponential family with canonical parameter $\gamma_i$. In **Equation (3.1)**, $y_i$ is independently distributed with mean

$\mu_i$. The dispersion parameter is assumed to be constant over observations (McCullagh and Nelder, 1989).

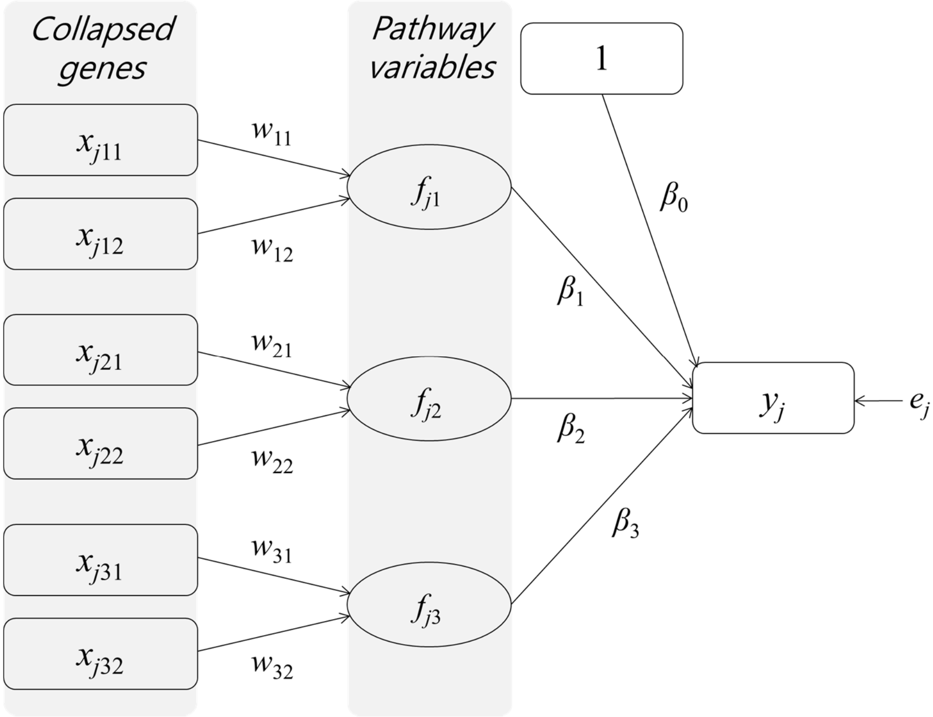
Let the dataset can be mapped onto $K$ pathways. Then, Let $x_{ikt}$ denote the $t^{\text{th}}$ gene in the $k^{\text{th}}$ pathway on the $i^{\text{th}}$ observation $(k = 1, \cdots, K; t = 1, \cdots, T_K)$, where $T_k$ is the number of genes for the $k^{\text{th}}$ pathway. We define each pathway as a weighted composite or component of a set of genes. Let $w_{kt}$ denote a weight assigned to $x_{ikt}$, leading to the $k^{\text{th}}$ pathway. Let $\beta_0$ denote the intercept. Let $\beta_k$ denote the $k^{\text{th}}$ coefficient connecting the $k^{\text{th}}$ pathway to the phenotype $y_i$. Let $\eta_i$ and $g(\cdot)$ denote a linear predictor and a link function, respectively.

Using the above notations, we specify the ***PHARAOH*** model that defines the relationship between a linear predictor and a link function as follows.

$$\eta_i = \beta_0 + \sum_{k=1}^{K} \left[ \sum_{t=1}^{T_k} x_{ikt} w_{kt} \right] \beta_k = \beta_0 + \sum_{k=1}^{K} f_{ik} \beta_k = g(\mu_i), \tag{3.2}$$

where $f_{ik} = \sum_{t=1}^{T_k} x_{ikt} w_{kt}$ indicates the $i^{\text{th}}$ observation's score of the $k^{\text{th}}$ pathway when $k > 0$, and is equal to one when $k = 0$. In case of $\gamma_i = \eta_i$, we have the canonical link; for instance, the identity, logit, log, inverse and squared inverse functions are the canonical links for the normal, binomial, Poisson, gamma, and inverse Gaussian distributions, respectively.

**Figure 3.1. A schematic diagram of the proposed model.** $x_{ikt}$ and $w_{kt}$ denote the $i$th sample's the $t$th collapsed gene within the $k$th pathway and its coefficient, and $f_{ik}$ and $\beta_k$ denotes the $i$th sample's latent variable for the $k$th pathway and its coefficient, respectively. $y_i$ and $e_i$ denote the $i$th sample's phenotype and error, respectively.



### 3.2.2. An exemplary structure

To facilitate an understanding of the proposed model, we provide an example of the model in **Figure (3.1)**. The depicted exemplary model assumes that a phenotype is normally distributed and it involves three pathways ($K = 3$), each of which consists of two genes ($P_1 = P_2 = P_3 = 2$). Each pathway is constructed by adding weights to its genes, featured by straight lines and denoted by $w$; and in turn influences a phenotype, signified by single-headed arrows. When the phenotype is continuous (or normally distributed), this model can be viewed as a special type of structural equation model, called the extended redundancy analysis model (Desarbo, et al., 2013; Hwang, et al., 2013; Takane and Hwang, 2005), in which all latent variables are equivalent to components of observed variables (e.g., genes) and serve as exogenous variables that affect a single endogenous and observed variable (e.g., a phenotype). Nonetheless, the proposed method is built on the framework of generalized linear models (GLM) (McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972), so that it can deal with phenotype data arising from a variety of exponential-family distributions. Furthermore, as will be shown shortly, the proposed method aims to address the issue of multicollinearity in parameter estimation, which is likely to be present among both genes and pathways.

### 3.3.3. Parameter estimation

To estimate parameters, we seek to maximize a penalized log-likelihood function taking the general form as

$$\phi_1 = \sum_{i=1}^{N} \log p(y_i; \gamma_i, \delta) - \frac{1}{2}\lambda_G \sum_{k=1}^{K} \sum_{t=1}^{T_k} w_{kt}^2 - \frac{1}{2}\lambda_P \sum_{k=0}^{K} \beta_k^2, \qquad (3.3)$$

with respect to $w_{kt}$ and $\beta_k$, subject to the conventional scaling constraint $\Sigma_{i=1}^{N} f_{ik}^2 = N$ (Takane and Hwang, 2005), where $\lambda_G$ and $\lambda_P$ are ridge parameters for the genes and the pathways, respectively. This optimization function can be viewed as the $L_2$-norm penalized log-likelihood (e.g., Le cessie and van houwelingen (1992); Lee and Silvapulle (1988)), where the $L_2$-norm or ridge penalty (Hoerl and Kennard, 1970) is imposed on both weights and coefficients. The two ridge penalties are added to address potential multicollinearity in both genes and pathways, which can have adverse effects on the estimation of weights and coefficients.

Let $\widetilde{w}_k = [w_{k1}, \cdots, w_{kT_k}]'$, $\tilde{\beta} = [\beta_0, \beta_1, \cdots, \beta_K]'$, and $F = [\tilde{f}_1, \cdots, \tilde{f}_N]'$ , where $\tilde{f}_i = [1, f_{i1}, \cdots, f_{iK}]'$ . Maximizing **Equation (3.3)** via iteratively reweighted least squares (IRLS) (e.g., Green (1984)) is equivalent to minimizing the following penalized least-squares function

$$\begin{aligned} \phi_2 &= \sum_{i=1}^{N} v_i \left( z_i - \sum_{k=0}^{K} f_{ik}\beta_k \right)^2 + \lambda_G \sum_{k=1}^{K} \sum_{t=1}^{T_k} w_{kt}^2 + \lambda_P \sum_{k=0}^{K} \beta_k^2 \text{,} \\ &= \sum_{i=1}^{N} v_i \left( z_i - \tilde{f}_i \tilde{\beta} \right)^2 + \lambda_G \sum_{k=1}^{K} \tilde{w}_k' \tilde{w}_k + \lambda_P \tilde{\beta}' \tilde{\beta} \qquad (3.4) \\ &= \left( \tilde{z} - F\tilde{\beta} \right)' V \left( \tilde{z} - F\tilde{\beta} \right) + \lambda_G \sum_{k=1}^{K} \tilde{w}_k' \tilde{w}_k + \lambda_P \tilde{\beta}' \tilde{\beta} \end{aligned}$$

with respect to $\tilde{w}_k$ and $\tilde{\beta}$, subject to $\text{diag}(F^T F) = N\text{I}$, where $V$ is an $N$ by $N$ diagonal matrix with elements $v_i = (\partial\mu_i/\partial\eta_i)^2/\tau_i$, where $\tau_i$ is the variance function evaluated at $\mu_i$, $\tilde{z}$ is an $N$ by 1 vector of the so-called adjusted response variable with elements $z_i = \eta_i + (y_i - \mu_i)/v_i$ (McCullagh & Nelder, 1989, Chapter 2).

To minimize **Equation (3.4)**, we develop an iterative algorithm similar to the alternating regularized least-squares algorithm (Hwang, 2009). This algorithm repeats the following steps until no substantial changes in parameter estimates occur.

**Step 1:** We update $\tilde{\beta}$ for fixed $\tilde{w}_k$, $V$, and $\tilde{z}$. This is equivalent to minimizing

$$\phi_3 = \left( \tilde{z} - F\tilde{\beta} \right)' V \left( \tilde{z} - F\tilde{\beta} \right) + \lambda_P \tilde{\beta}^T \tilde{\beta}, \qquad (3.5)$$

with respect to $\tilde{\beta}$. Hence, the estimates of $\tilde{\beta}$ are obtained by

$$\hat{\tilde{\beta}} = (F^T V F + \lambda_P)^{-1} F^T V \tilde{z}. \qquad (3.6)$$

**<u>Step 2</u>:** We update $\widetilde{w}_k$ for fixed $\tilde{\beta}$, $V$, and $\tilde{z}$. Let $W = [\widetilde{w}_1, \cdots, \widetilde{w}_K]$ and $X_k = \left[x_{k1}, \cdots, x_{kP_k}\right]$. This is equivalent to minimizing

$$\begin{aligned}
\phi_4 &= \left(\tilde{z} - F\tilde{\beta}\right)' V \left(\tilde{z} - F\tilde{\beta}\right) + \lambda_G \sum_{k=1}^{K} \widetilde{w}_k^T \widetilde{w} \\
&= \left(\tilde{z} - [1\, X_1 \,\cdots\, X_K] W\tilde{\beta}\right)' V \left( \tilde{z} - X \begin{bmatrix} 1 & & & \\ & \widetilde{w}_1 & & \\ & & \ddots & \\ & & & \widetilde{w}_K \end{bmatrix} \tilde{\beta} \right) + \lambda_G \widetilde{w}^T \widetilde{w} \\
&= \left(\tilde{z} - XW\tilde{\beta}\right)' V \left(\tilde{z} - XW\tilde{\beta}\right) + \lambda_G \widetilde{w}^T \widetilde{w} \\
&= \left(\tilde{z} - \left(\tilde{\beta}^T \otimes X\right)\mathrm{vec}(W)\right)' V \left(\tilde{z} - \left(\tilde{\beta}^T \otimes X\right)\mathrm{vec}(W)\right) + \lambda_G \widetilde{w}^T \widetilde{w} \\
&= (\tilde{z} - \Phi\widetilde{w})' V (\tilde{z} - \Phi\widetilde{w}) + \lambda_G \widetilde{w}^T \widetilde{w}
\end{aligned} \tag{3.7}$$

where $X = [1, X_1, \cdots, X_K]$, $W = \mathrm{diag}(1, \widetilde{w}_1, \cdots, \widetilde{w}_K)$, and $\mathrm{diag}(\cdot)$ denotes diagonalization operator, $\mathrm{vec}(W)$ denotes a supervector formed by stacking all columns of $W$ one below another, $\otimes$ indicates the Kronecker product, $\widetilde{w}$ is equivalent to the vector formed by eliminating fixed elements such as one and zeros from $\mathrm{vec}(W)$, $\Phi$ is the matrix formed by eliminating the columns of $\tilde{\beta}' \otimes X$ corresponding to the fixed elements in $\mathrm{vec}(W)$.

Then, the estimates of $\widetilde{w}$ are obtained by

$$\hat{\widetilde{w}} = (\Phi^T V \Phi + \lambda_G I)^{-1} \Phi^T V \tilde{z}. \tag{3.8}$$

Subsequently, we obtain $\tilde{f}_k$ by $\tilde{f}_k = X_k \widetilde{w}_k$ and standardize it to satisfy the constraint $\tilde{f}_k' \tilde{f}_k = N$.

**Step 3:** We update $V$ and $\tilde{z}$ for fixed $\tilde{\beta}$ and $\widetilde{w}_k$. As previously stated, $\tilde{z}$ is updated based on $z_i = \eta_i + (y_i - \mu_i)/v_i$. The calculation of $V$ depends on which distribution is assumed for responses. For instance, for a normal distribution, $V = \mathrm{I}_N$, and for a binomial distribution, $V$ has elements $v_i = \alpha_i(1 - \alpha_i)$, where $\alpha_i = exp(\eta_i)/(1 + \exp(\eta_i))$. Refer to McCullagh and Nelder (1989) for calculation of $V$ for other exponential-family distributions.

Finally, we should determine the values of $\lambda_G$ and $\lambda_P$ before applying the parameter estimation procedure. We may use *K*-fold cross validation (Hastie et al., 2009, p. 214) to decide the values of $\lambda_G$ and $\lambda_P$. In *K*-fold cross validation, we divide the entire set of data into *K* subsets. Next, we leave one subset as a test set and use the remaining subsets as a training set for estimating parameters. We apply the parameter estimates obtained from the training set to the test set, and calculate its (minus) log-likelihood value. We repeat this procedure *K* times, varying test and training sets systematically. Then, we compute the average of the (minus) log-likelihood values over all *K* test sets. We may consider a number of alternative values of $\lambda_G$ and $\lambda_P$ and reiterate the above procedure for each alternative. The values of $\lambda_G$ and $\lambda_P$ associated with the largest average log-likelihood values (equivalently, the smallest average minus log-likelihood value) may be selected as the final estimates.

The proposed method resulted in ridge estimates of parameters. Thus, the asymptotic approximation to the variances of these parameter estimates cannot be used directly for obtaining their confidence intervals, because their biases should be taken in account (Le cessie and van houwelingen, 1992). Instead, resampling methods can be used to test the statistical significance of the estimated effects of all pathways on the phenotype, as well as the estimated weights assigned to genes. Although other resampling methods such as the bootstrap or jackknife can also be used for examining the statistical significance of the estimates, in the proposed method, we utilize a permutation test to provide $p$-values.

## 3.4. Simulation study

### 3.4.1. The simulation dataset

In order to demonstrate the statistical performance of the proposed model, we conducted a straightforward simulation study. To perform simulation, we used well-established simulation data that was generated under pathway model, Genetic Analysis Workshop (GAW) 17 dataset (Almasy, et al., 2011). In brief, the GAW17 dataset is a simulated dataset consisting of 697 individuals from 1000 Genomes Project and 24,487 SNVs, along with 200 replicates of four simulated traits (Q1, Q2, Q4 and AFFECTED). Among those traits, only Q1 was simulated to be affected by an age factor and 39 SNVs residing in nine genes from the vascular endothelial growth factor (VEGF) pathway defined by Ingenuity Pathway Analysis (http://www.ingenuity.com). Other traits were generated without using pathway information and thus not considered further in our simulation studies. Since Q1 reflects combinatorial effect of multiple genes in a pathway, VEGF, we examined the power of the proposed method by the proportion of identifying the pathway. First, 21,028 SNVs in 3,179 genes from 697 unrelated samples were selected as rare variants by MAF filtering, i.e. less than 5%. Subsequently, all of the rare variants were collapsed into genes. Here, MAFs for all rare variants were computed directly from the data. The names of all the genes were annotated using the HUGO Gene Nomenclature Committee database. Here, each rare variant was assigned to a gene if its location was in the gene or within 10 kilobases 5′ or 3′ to the transcribed region. For pathway-gene mapping, we extracted 217 pathways from Biocarta and 186 pathways from KEGG (Kanehisa, et al., 2004), and mapped the genes to the pathways.

### 3.4.2. Comparison of methods using simulation dataset

For the purpose of power comparison, ***PHARAOH*** and existing pathway-based methods, including aforementioned WKS (WKS-Variant and WKS-MinP) and DRB (Direct-Burden and Direct-SKAT-o) (Wu and Zhi, 2013), were applied to the GAW17 simulation dataset. We did not include the SFPCA method because it was proposed for binary traits (Zhao, et al., 2014). First, the performance of methods was carried out by comparing empirical power which is a proportion of VEGF pathway (true causal pathway in the simulation) $p$-value $< 0.05$ from 200 replicates of Q1. For ***PHARAOH***, the tuning parameters, $\lambda_G$ and $\lambda_P$, were chosen based on five-fold CV using 11 different starting points of ridge parameter ranging from $10^{-2}$ to $10^{8}$ on a logarithmic base 10 scale, and it was fixed to 4,000 across simulation study. An analysis time of ***PHARAOH*** was 15 minutes. As shown in **Figure (3.2)**, the proposed method showed 0.87 of empirical statistical power to detect VEGF pathway, while those of WKS were only 0.105 and 0.055, respectively. Moreover, as shown in **Figure (3.4)**, while the top five pathways of ***PHARAOH*** showed the reasonable detection power (**Figure (3.4A)**), those of WKS-all and WKS-minP (**Figure (3.4B)** and **(3.4C)**) were not far from random chance by the given statistical threshold ($\alpha$ = 0.05). From the result summary, we excluded the results from DRB since it showed substantial inflation of $p$-values (**Figure (3.3)**). Notably, ***PHARAOH*** also identified the focal adhesion pathway in 77% of replicates, since the pathway is a subsequent pathway of VEGF pathway and the pathway contains five of significantly simulated genes (FLT1, FLT4, KDR, VEGFA and VEGFC). Second, we generated and tested another 5,000 replicates of Q1 by permuting the first original replicate, to assess type I error. As shown in **Table (3.1)**, all of the methods controlled their type 1 errors, despite of slight conservative trend of ***PHARAOH***.

**Table 3.1. Type 1 error rates of PHARAOH, WKS and DRB.**

| Method | α=0.05 | α=0.01 |
|---:|:---:|:---:|
| **PHARAOH** | 0.040 (±0.019) | 0.0083 (±0.008) |
| **WKS-Variant** | 0.056 (±0.028) | 0.0156 (±0.017) |
| **WKS-MinP** | 0.049 (±0.025) | 0.0101 (±0.010) |
| **Direct-Burden** | 0.051 (±0.044) | 0.0103 (±0.017) |
| **Direct-SKAT-o** | 0.049 (±0.043) | 0.0105 (±0.017) |

**Figure 3.2. Empirical powers of simulation dataset using KEGG pathway database.** Empirical power indicates the times of identification among 200 replicates. (A) Empirical power of top 5 pathways from ***PHARAOH***. (B) Empirical power of top 5 pathways from WKS-Variant. (C) Empirical power of top 5 pathways from WKS-MinP.

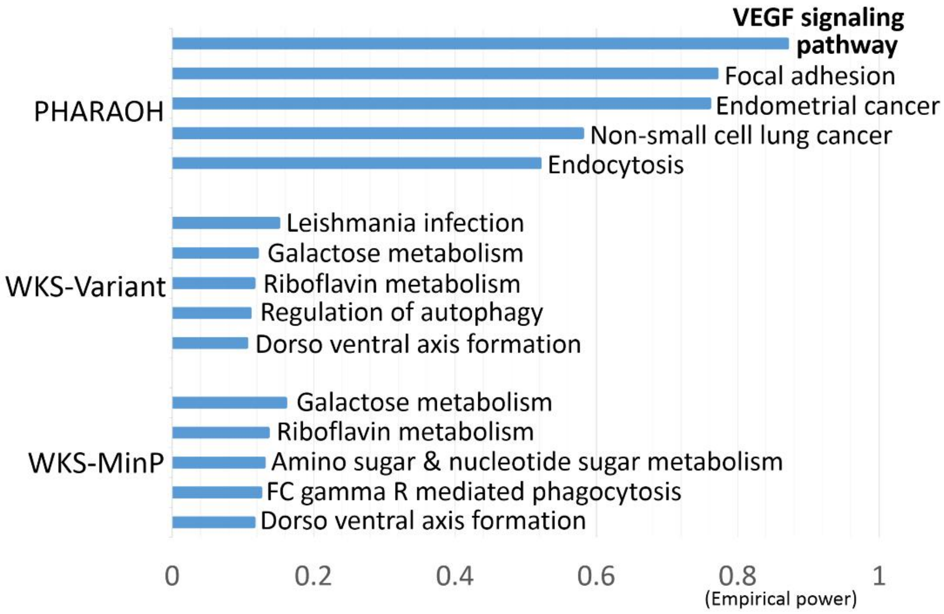


**Figure 3.3. Quantile-quantile plot of GAW17 simulation dataset analysis result using Direct Region-Based (DRB) method.** Three types of DRB tests (Direct-Burden, Direct-SKAT and Direct-SKAT-o) are plotted with green, pink and yellow dots, respectively.

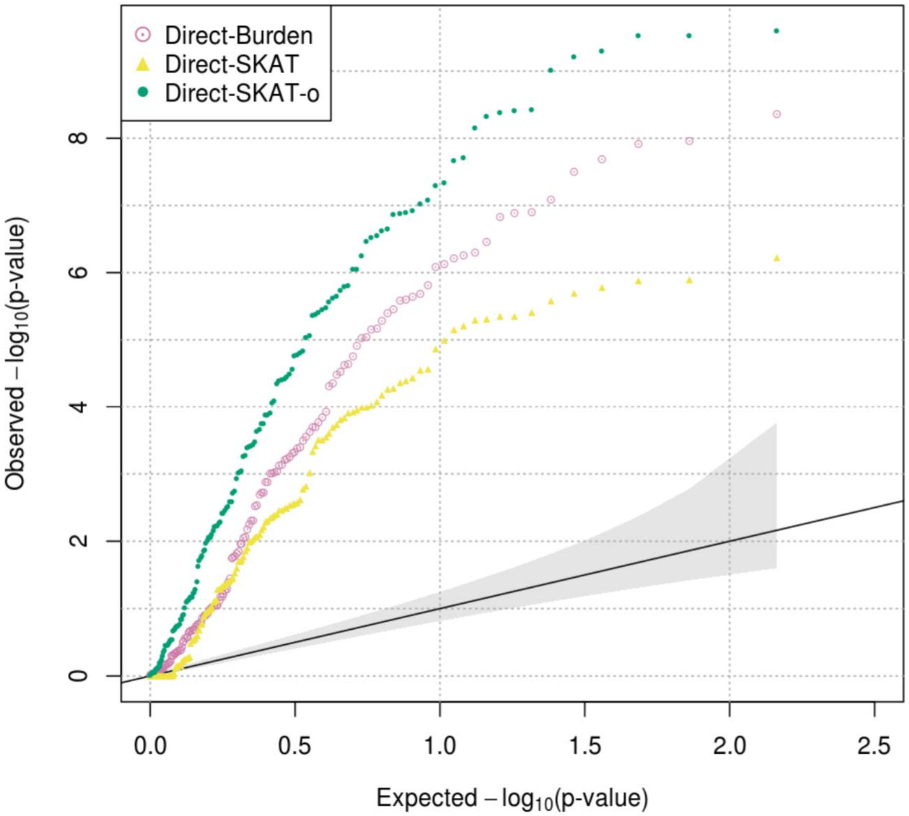
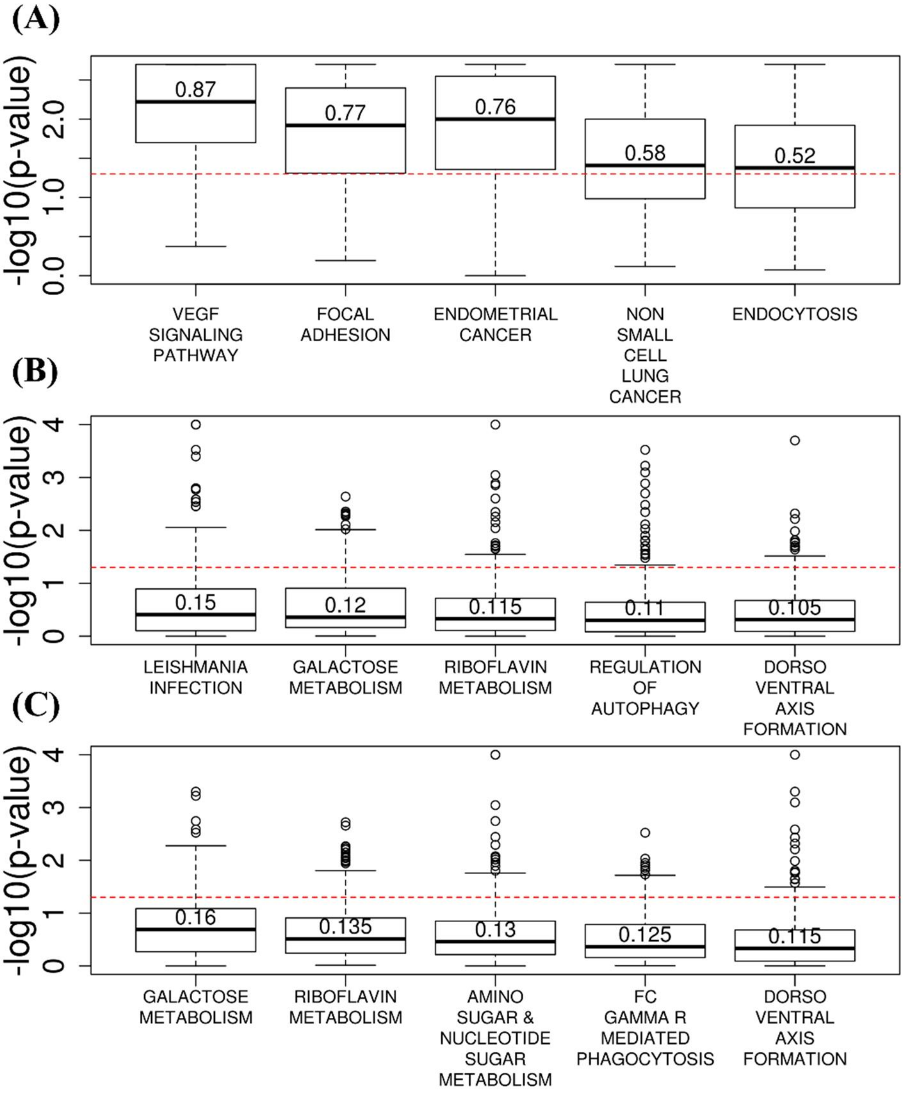


**Figure 3.4. Top five significant pathways from the simulation study.** (A) Top 5 pathways of ***PHARAOH***. (B) Top 5 pathways of WKS-all. (C) Top 5 pathways of WKS-minP.



Finally, we also performed literature search to investigate empirical powers of other methods to detect VEGF pathway in GAW17 dataset. For the methods using both common rare variants, one comparison study that compared four extensions of gene-based methods to pathway-based, using both rare and common variants, showed that the highest empirical power among them was 0.65 (Uh, et al., 2011). In contrast, another comparison demonstrated up to 0.93 of empirical power (Ngwa et al., 2011). However, since they considered all of nine causal genes are belong to VEGF pathway, their assumption contains much more causal genes compared to our KEGG mapping, as shown above. With reflection of this difference, a subsequent analysis using ***PHARAOH*** with modified VEGF pathway contains all of significant genes showed 0.935 of empirical power even without the presence of common variants (data not shown). Among the methods using only rare variants of GAW17, only one method could handle joint effects of multiple rare variants (Hu et al., 2011). Its maximum empirical power for VEGF pathway was only 0.182, which demonstrates superior performance of ***PHARAOH***.

## 3.5. Application to analysis of liver enzymes

### 3.5.1. Whole exome sequencing dataset for pathway discovery

We applied ***PHARAOH*** to perform a pathway analysis of whole-exome sequencing (WES) data from a Korean population study, via our membership in the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium. Specifically, the genomes of 1087 individuals, selected from the Korean Association REsource (KARE) study (Cho, et al., 2009), were sequenced using the Illumina HiSeq2000 platform (Illumina, Inc., San Diego, CA, USA). The levels of aspartate aminotransferase (AST), a liver enzyme, were measured in the morning, before the first meal of the day. Prior to the analysis, 1046 samples were chosen after excluding participants taking medications likely to influence liver enzyme levels. For 1,046 participants, 399,729 variants, mapped to the UCSC hg19 genomic coordination, were retained after a quality control process. Here, the quality control process was an exclusion of variants with genotype call rates $< 95\%$ or Hardy–Weinberg Equilibrium (HWE) test $P < 10^{-5}$. Using 120,807 rare variants with MAF $< 5\%$, rare variant collapsing and pathway-gene mapping were then performed, as in the simulation study. MAFs for all rare variants were computed directly from the data. The final datasets consisted of 1,190 genes, with 55,978 rare variants for Biocarta, and 4,913 genes, with 216,531 rare variants for KEGG, respectively. Note that the numbers of genes and variants per pathway included those shared with other pathways.

AST can be used for determining liver function abnormalities, in addition to other liver enzymes such as alanine aminotransferase (ALT) (Huang, et al., 2006). As WKS and our method require phenotype permutation, we generated 1000 and 10,000 permuted replicates of phenotypes for

***PHARAOH*** and WKS, respectively. Following association tests conducted by Cho, et al. (2009), age, sex and area were included as covariates in the pathway analyses. The chosen $\lambda$ values for AST were 5500 for Biocarta and 9500 for KEGG. The total computing times were 67, 113 and 22 minutes for ***PHARAOH***, WKS and DRB methods, respectively. Quantile–quantile plots of the results showed no explicit inflation or deflation of *P* values (**Figure (3.5)**).

**Figure 3.5. Quantile–quantile (QQ) plots for levels of the liver enzyme AST, with adjustment for covariates.** The QQ-plots are provided for ***PHARAOH***, WKS and DRB, with 95% confidence interval. (A) QQ-plot of ***PHARAOH*** using Biocarta. (B) QQ-plot of ***PHARAOH*** using KEGG. (C) QQ-plot of WKS using Biocarta. (D) QQ-plot of WKS using KEGG. (E) QQ-plot of DRB using Biocarta. (F) QQ-plot of DRB using KEGG.

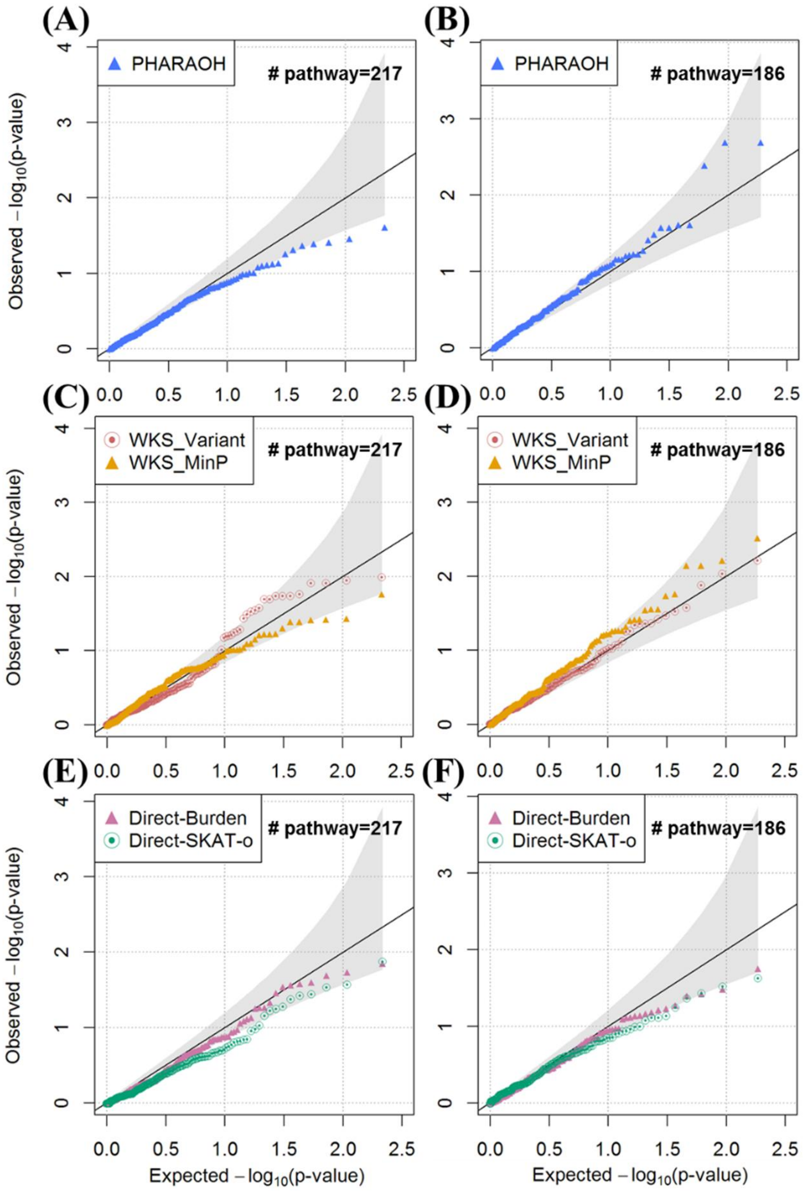


**Table 3.2. Pathways identified by *PHARAOH* in the discovery study with Biocarta pathway database.**

| Pathway | # of mapped SNVs[a] | # of mapped genes[b] | *P* values | | | | |
|---|---|---|---|---|---|---|---|
| | | | PHARAOH | WKS-Variant | WKS-MinP | Direct-Burden | Direct-SKAT-o |
| p38 MAPK Signaling Pathway | 1321 | 47 | **0.024** | 0.259 | 0.735 | 0.175 | 0.316 |
| **Insulin Signaling Pathway** | 806 | 26 | **0.034** | 0.679 | 0.854 | 0.734 | 0.411 |
| Role of Ran in mitotic spindle regulation | 441 | 18 | **0.038** | 0.379 | 0.264 | 0.847 | 1 |
| Hemoglobin's Chaperone | 350 | 12 | **0.040** | 0.176 | 0.857 | 0.099 | 0.17 |
| **Erythrocyte Differentiation Pathway** | 354 | 21 | **0.042** | 0.723 | 0.068 | 0.222 | 0.38 |
| EGF Signaling Pathway | 1173 | 43 | **0.048** | 0.546 | 0.657 | 0.528 | 0.76 |

[a] The number of mapped genetic variants to the pathway. [b] The actual number of genes included in the pathway. Pathway names with bold text and underlined text indicate the replicated pathways in the independent dataset and another independent study (Sookoian and Pirola, 2012), respectively.

**Table 3.3. Pathways identified by *PHARAOH* in the discovery study with KEGG pathway database.**

| Pathway | # of mapped SNVs[a] | # of mapped genes[b] | *P* values | | | | |
|---|---|---|---|---|---|---|---|
| | | | PHARAOH | WKS-Variant | WKS-MinP | Direct-Burden | Direct-SKAT-o |
| Protein export | 575 | 28 | **0.004** | **0.04** | 0.738 | 0.808 | 0.479 |
| **Glycine, serine & threonine metabolism** | 1312 | 39 | **0.024** | 0.127 | 0.467 | 0.893 | 0.589 |
| Other glycan degradation | 879 | 22 | **0.024** | 0.978 | 0.829 | 0.126 | 0.166 |
| **Glycosaminoglycan biosynthesis** (*heparan sulfate*) | 966 | 35 | **0.026** | 0.773 | 0.101 | 0.709 | 0.387 |
| Linoleic acid metabolism | 1180 | 35 | **0.026** | 0.752 | 0.462 | 0.648 | 0.559 |
| Galactose metabolism | 1649 | 43 | **0.032** | 0.677 | **0.033** | 0.207 | 0.164 |
| Sphingolipid metabolism | 1347 | 50 | **0.038** | 0.829 | 0.917 | 0.195 | 0.195 |

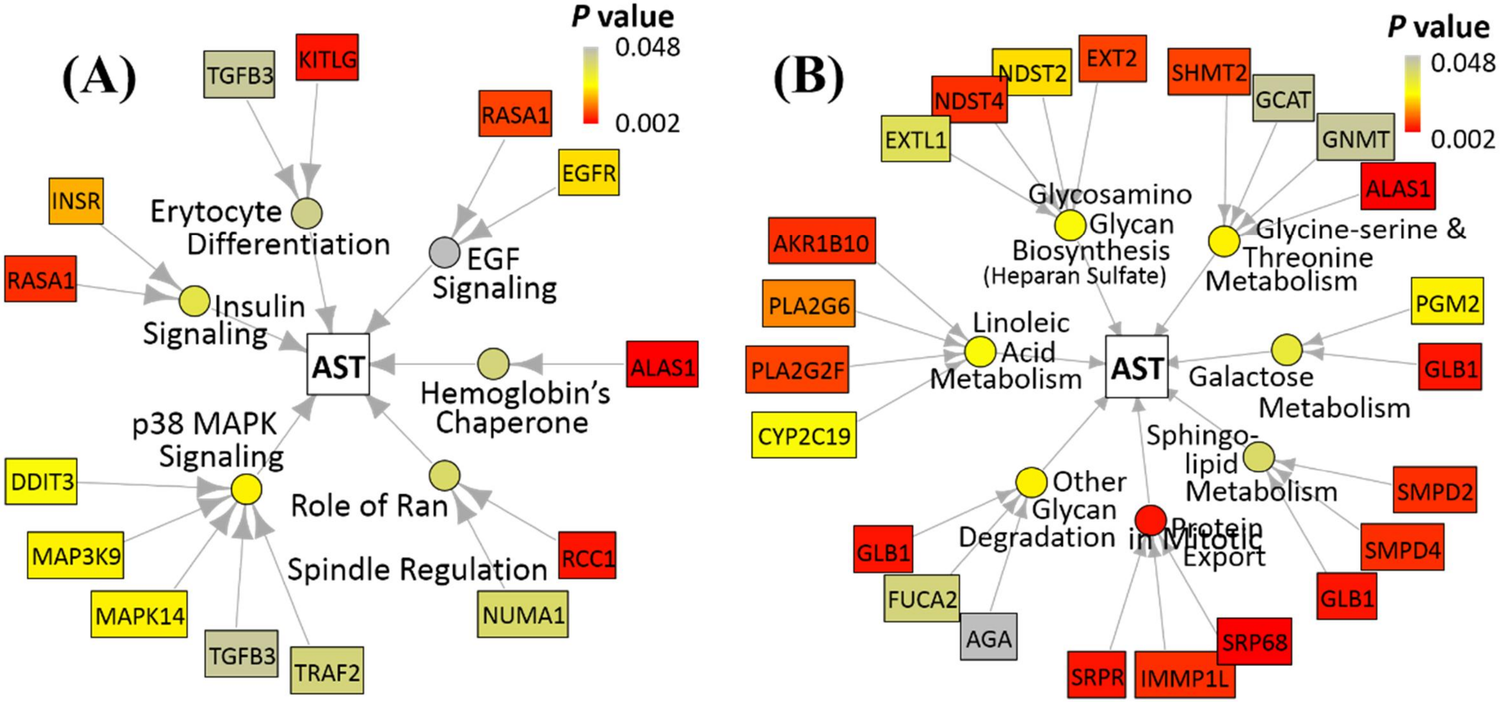
[a]The number of mapped genetic variants to the pathway. [b]The actual number of genes included in the pathway. Pathway names with bold text and underlined text indicate the replicated pathways in the independent dataset and another independent study (Sookoian and Pirola, 2012), respectively.

The discovery study using WES dataset using ***PHARAOH*** identified six pathways for Biocarta (**Table (3.2)**), and seven pathways for KEGG (**Table (3.3)**), at a 5% significance level. Significant pathways and their significant genes for Biocarta and KEGG are depicted in **Figure (3.6A)** and **(3.6B)**, respectively. However, none of the existing methods identified statistically significant pathways after Bonferroni correction at the 5% significance level, as shown in **Table (3.2)** and **(3.3)**. The Bonferroni-corrected *P* value thresholds were $2.3 \times 10^{-4}$ for Biocarta and $2.69 \times 10^{-4}$ for KEGG, and those thresholds were calculated from the number of mapped pathways for each pathway database (217 for Biocarta, and 186 for KEGG).

The pathways identified by ***PHARAOH*** from the discovery study are reported to have strong biological relevance to the liver. The pathways linoleic acid metabolism, galactose metabolism, erythrocyte differentiation and alpha-hemoglobin stabilizing protein all relate to liver function. One previous study showed that dietary conjugated linoleic acid alleviated non-alcoholic fatty liver disease by reducing levels of hepatic injury markers in Zucker (fa/fa) rats (Nagao, et al., 2005). Conjugated linoleic acid supplementation also lowered levels of serum ALT and alkaline phosphatase in Zucker (fa/fa) rats (Noto, et al., 2006). Galactose, a mono saccharide sugar metabolized primarily in the liver, and galactose elimination capacity, have been widely used for estimating quantitative liver function (Lindskov, 1982). Two other pathways, erythrocyte differentiation and alpha-hemoglobin stabilizing protein, were found to be related to red blood cells (**Table (3.2)**). Erythrocyte differentiation pathway and Hemoglobin's Chaperone pathway describe the process of preventing precipitation of hemoglobin alpha-subunits by alpha-hemoglobin-stabilizing protein. The liver is a major hematopoietic organ during fetal life (Cardier and Barbera-Guillem, 1997).

We next compared the list of identified pathways from ***PHARAOH*** with previous pathway-based analyses of AST and ALT results from Sookoian and Pirola (2012) (Hereinafter SP). Despite the use of different pathway databases, we found that the sphingolipid metabolic process was significant in both results ($P = 0.038$ from ***PHARAOH*** and adjusted $P$ value $= 0.018$, where the adjusted $P$ value is derived by Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995)). Notably, ***PHARAOH*** also successfully identified the sphingolipid pathway, well known to relate to liver diseases (Alexaki, et al., 2014; Pralhada Rao, et al., 2013), thus further supporting our approach. Moreover, our detection of the glycine-serine and threonine metabolism pathways of KEGG concurs with SP that identified three of four significant genes in those pathways (*SHMT2*, *GCAT* and *ALAS1*). Moreover, the *ALAS1* gene was also statistically significant in the Hemoglobin's Chaperone pathway of Biocarta ($P = 0.002$) and the glycine serine and threonine metabolism pathways ($P = 0.002$) of KEGG.

**Figure 3.6. Visualizations generated by *PHARAOH* in the discovery study of AST levels.** Outermost rectangles indicate statistically significant genes within significant pathways, circles represent statistically significant pathways, and center square indicate the phenotype of interest. (A) Result using the Biocarta pathway database (B) Result using the KEGG database.



### 3.5.2. Replication study using exome chip dataset

To further confirm our discovered pathways in an independent cohort, we conducted a replication study using an independent Korean cohort from the Health Examinee shared control study (HEXA), a part of the KoGES population based cohort, initiated in 2001 (Kim, et al., 2011). Among these, 3,445 samples were used for the replication study. Samples were genotyped using the HumanExome BeadChip v1.1 (Illumina, Inc., San Diego, CA), which contains approximately 240,000 variants. All samples passed quality control tests using the following exclusion criteria: a genotype call rate < 99%, excessive heterozygosity, and sex inconsistency. The exclusion criterion for variants was as follows: HWE test $P < 10^{-6}$, genotype call rates < 95%, and monomorphic variants. After quality control, 60,628 variants remained for further analysis. For all participants from the cohort, AST was measured identically to the KARE study. Rare variant collapsing and pathway-gene mapping were then performed, as in the discovery study. MAFs for all rare variants were computed directly from the data. Consequently, 517 genes mapped to 210 pathways, and 2,391 genes mapped to 186 pathways, were then used in the replication study for Biocarta and KEGG, respectively.

In the replication study with an independent Korean population dataset using exome array ($n$=3,445), the chosen $\lambda$ values were 245 for Biocarta and 8,000 for KEGG. An execution time was 37 minutes for Biocarta and 38 minutes for KEGG. The replication study identified eight pathways from Biocarta and nine pathways from KEGG (**Table (3.4)**). Despite the limited number of rare variants included on the exome array used in our replication study, we were able to successfully replicate the Erythrocyte Differentiation pathway of Biocarta, and the KEGG pathways glycine-serine and threonine metabolism and glycosaminoglycan biosynthesis. In addition, the insulin signaling pathway was also replicated, despite the differences between pathway databases. Among the replicated pathways, we were able to discover a number of associations between the identified pathways and liver function. The insulin signaling pathway manifests selective insulin resistance in diabetic mice (Li, et al., 2010). Additionally, a study of the AST values of a high protein diet suggested that hepatic utilization of glycine-serine and threonine in the liver varied between fed and starved rats, thus also reflecting the role of the glycine-serine and threonine pathways in the liver (Remesy, et al., 1983).

**Table 3.4. Significant pathways from *PHARAOH* in the replication study.** Pathway names with bold text and underlined text indicate the replicated pathways in the independent dataset using the same pathway database, an independent dataset, and a different pathway database (SP), respectively.

| DB | Pathway | PHARAOH |
|---|---|---|
| Biocarta | HIV-I Nef: negative effector of Fas and TNF | 0.006 |
| | Feeder Pathways for Glycolysis | 0.016 |
| | Human Cytomegalovirus and Map Kinase Pathways | 0.018 |
| | Lck/Fyn tyrosine kinases in initiation of TCR Activation | 0.022 |
| | **Erythrocyte Differentiation Pathway** | 0.026 |
| | NFkB activation by Nontypeable Hemophilus influenza | 0.048 |
| | Growth Hormone Signaling Pathway | 0.049 |
| | Influence of Ras and Rho proteins on G1 to S Transition | 0.049 |
| KEGG | **Glycine, serine and threonine metabolism** | 0.01 |
| | Metabolism of xenobiotics by cytochrome P450 | 0.018 |
| | **Insulin Signaling Pathway** | 0.028 |
| | **Glycosaminoglycan biosynthesis** *(keratan sulfate)* | 0.032 |
| | Phenylalanine metabolism | 0.036 |
| | <u>Tryptophan metabolism</u> | 0.044 |

## 3.6. Discussion

In this study, we developed a novel statistical method for pathway-based analysis of large-scale genetic data. Using GAW17 simulation dataset, we have demonstrated substantial empirical power using ***PHARAOH***, compared to several methods for pathway analysis, with an appropriate control of type 1 error. While other methods require common variants to achieve large empirical power, our method could achieve higher power without common variants. In addition, by applying ***PHARAOH*** to large-scale WES and exome chip data, we identified several pathways biologically associated with levels of AST or overall liver function, in accord with previous findings (Alexaki, et al., 2014; Cardier and Barbera-Guillem, 1997; Li, et al., 2010; Lindskov, 1982; Nagao, et al., 2005; Noto, et al., 2006; Pralhada Rao, et al., 2013; Remesy, et al., 1983; Sookoian and Pirola, 2012). Generally, it is not straightforward to replicate findings of rare variant analysis (Liu and Leal, 2010), because the composition of rare variants can differ in independent datasets. Nonetheless, we successfully replicated four pathways using an independent dataset, representing potential candidates for biological validation.

Compared to other existing pathway-based tests, our method has several advantages. First, the proposed method is not restricted by the multiple testing problem, because ***PHARAOH*** fits only a single model that considers all pathways of interest, testing the statistical significance of all parameter estimates at once. Although the number of tests in a pathway-based analysis is much smaller than that of variant-level or gene-based analysis, its cutoff value of Bonferroni corrected *P*-value at a 5% significance level was $2.3 \times 10^{-4}$ for 217 pathways in Biocarta, making it highly untenable to reject the null hypothesis. Because it is free from the multiple testing problem, ***PHARAOH*** requires substantially smaller numbers of permutations than other existing permutation-based methods. In practice, ***PHARAOH*** requires at most 1,000 permutations at a 5% significance level, whereas other existing permutation-based methods require much larger numbers of permutations (e.g. 10 000 or more) (Kim, et al., 2011; Weng, et al., 2011).

Second, ***PHARAOH*** can accommodate potentially high correlations between pathways, which cannot be efficiently controlled by other existing methods using a series of single pathway analyses. As shown by several studies of pathway-based or gene set-based methods (Alexa, et al., 2006; Jiang and Gentleman, 2007; Skarman, et al., 2012), it is necessary to consider correlations between pathways, because such correlations influence the combined effects of pathways on the phenotype. Whereas other existing methods adopt an additional step to adjust for the effect caused by overlap between pathways, our method seeks to control for correlations between genes in a specific pathway, as well as correlations between pathways, by imposing ridge-type penalties on both gene and pathway coefficient estimates. In addition, the proposed method provides *P* values not only for pathway coefficient estimates, but also for gene estimates per pathway.

Although we identified and addressed a number of issues in this report, several challenges are still remained. Unlike other methods, our proposed approach analyzes all pathways simultaneously in very short time (e.g. several hours). However, the permutation scheme used to obtain *P* values increases the time required for an entire analysis. Thus, it would be desirable to extend the proposed method without heavy permutation, to achieve faster and more accurate computation.

An optimal choice of weight would increase the performance of ***PHARAOH***. The current default weight is the beta-transformed MAF in the collapsing of multiple rare variants of specific genes, as suggested by Wu, et al. (2011) for rare variant analysis. However, recent studies suggest that other weighting approaches, based on the number of informative family members or the predicted functional effects of variants, can reduce false positive rates and increase statistical power (De, et al., 2013; Hu, et al., 2013; Shugart, et al., 2012; Sifrim, et al., 2013). The application of such weighting variants represents one possible extension of our future work.

Cross validation for ***PHARAOH*** can often be time-consuming because it considers large combinations of candidate values for the two penalty parameters for gene and pathway. To reduce computational burden, we applied cross validation to select only a single value for the parameters, constraining them to be equal. This may lead to less optimal values for the parameters. According to our limited experience, if cross validation is applied to decide the two parameters freely without the equality constraint on the parameters, the penalty parameter for gene tends to remain the same as the common penalty parameter obtained under the equality constraint. This may suggest that if we can derive the penalty value for pathway as a function of that for gene in some way (e.g. $\lambda_G = c\lambda_P$, where $c$ is a constant), using cross validation with this constraint could be more computationally efficient. However, a careful investigation into the feasibility of this approach is warranted.

Although there exist other penalization approaches such as LASSO (Tibshirani, 1996) or Elastic-Net (Zou and Hastie, 2005), we choose a ridge method due to its computational efficiency. Hence, our future work can be an extension of the proposed approach to the model using different penalizations. Moreover, ***PHARAOH*** can be flexible by allowing pathways and genes to have their own penalty parameters. We strongly believe that our novel method will enhance the success of pathway-based analysis using genetic datasets, thus addressing, at least in part, the problem of missing heritability.

This chapter was accepted in *BMC Bioinformatics*
as a partial fulfillment of Sungyoung Lee's Ph.D program.

# Chapter 4

## Multivariate pathway-based approach using rare variants

### 4.1. Introduction

In this chapter, we introduce a novel statistical method, "***PHARAOH-multi***" (**P**athway-based approach using **H**ier**A**rchical component of collapsed **RA**re variants **O**f **H**igh-throughput sequencing data), for analyzing **multi**ple phenotypes. In a nutshell, the proposed method is an extension of the ***PHARAOH*** method for multivariate analysis. Here, while keeping the advantages of our previous approach, we extend it to enable analysis of multiple traits using hierarchical components of genetic variants. In addition, the proposed model can identify associations between multiple phenotypes and multiple pathways, with a single model, in the presence of subsequent genes within pathways, as a hierarchy.

Simulation studies successfully demonstrated advantages of multivariate analysis, compared to univariate analysis, and comparison studies showed the proposed approach to outperform existing methods. Moreover, real data analysis of six type 2 diabetes-related traits, using large-scale whole exome sequencing data, identified significant pathways that were not found by univariate analysis. Furthermore, strong relationships between the identified pathways, and their associated metabolic disorder risk factors, were found via literature search, and one of the identified pathway, was successfully replicated by an analysis with an independent dataset.

## 4.2. Methods

### 4.2.1. Notations and the model

Our ultimate goal was to find an association between $Q$ phenotypes and $K$ pathways, each of whose number of genes was $T_1, \cdots, T_K$, under the presence of distinct parameters for ridge penalization. The proposed method is based on Generalized Structural Component Analysis (GSCA). First, we assume that a dataset has $N$ samples for the interested traits. Then, Let $Y = \left[y_{11} \cdots y_{1Q} \, ; \, \cdots ; y_{N1} \cdots y_{NQ}\right]$ be the matrix of phenotypes for $N$ samples, where $y_{iq}$ is the observation of the $i^{\text{th}}$ sample on the $q^{\text{th}}$ phenotype, and let $X$ be the matrix of gene-level collapsed variables generated by summing rare variants according to their gene variant-gene mapping. Let $g_{ij} \in \{0,1,2\}$ be the number of minor alleles for $j^{\text{th}}$ genetic variant of the $i^{\text{th}}$ sample. Regarding the elements of $X$, $x_{ikt}$ is a weighted sum of the $i^{\text{th}}$ sample's rare variants in the $t^{\text{th}}$ gene of the $k^{\text{th}}$ pathway, denoted by $x_{ikt} = \Sigma_{j\in M_{kt}}\omega_j g_{ij}$, where $M_{kt}$ represents indices of the rare variants for the $t^{\text{th}}$ gene in the $k^{\text{th}}$ pathway. Several weighting parameters, $\omega_i$, can be used, as previously described in (Lee, et al., 2016). By imposing two penalty parameters $\lambda_G$ and $\lambda_P$ on the genes-pathway and pathways-phenotype, we sought to address potential multicollinearity problems, in both genes and pathways, in the proposed method. Such problems may adversely affect the estimation of weights and coefficients. The proposed model assumes that the phenotype, $y_{iq}$, arises from the multivariate normal distribution with mean $\mu$ and covariance $\Sigma$ ($q = 1,\cdots,Q$ and $i = 1,\cdots,N$). Then we define the proposed ***PHARAOH-multi*** model as

$$y_{iq} = \beta_{0q} + \sum_{k=1}^{K}\left(\sum_{t=1}^{T_k} x_{tki}w_{tk}\right)\beta_{kq} + \tilde{\epsilon}_i = \beta_{0q} + \sum_{k=1}^{K} f_{ik}\beta_{kq} + \tilde{\epsilon}_i \tag{4.1}$$

$$= F_i\tilde{\beta}_q + \tilde{\epsilon}_i.$$

Here, $f_{ik} = \sum_{t=1}^{T_k} x_{ikt}w_{kt}$ and $F_i$ indicate the $i^{\text{th}}$ observation's score of the $k^{\text{th}}$ pathway, and its matrix form across $Q$ phenotypes, respectively. Moreover, $\tilde{\beta}_q = [\beta_{0q}\,\beta_{1q}\cdots\,\beta_{Kq}]$ is a vector of coefficients for the $q^{\text{th}}$ phenotype, and $\tilde{\epsilon}_i = [\epsilon_{i1}\cdots\epsilon_{iQ}]$ is a vector of residuals for the $i^{\text{th}}$ sample.

### 4.2.2. An exemplary structure

The proposed method is based on Generalized Structural Component Analysis (GSCA) (Hwang and Takane, 2004), therefore the model has a hierarchy that consists of manifest variables, latent variables and phenotypes. In order to provide an intuitive explanation about the proposed method, an exemplary structure of the model is shown in **Figure (4.1)**. In the example, three pathways ($K = 3$) are defined as the latent variables. For the first to the third pathways, two, three and two genes are mapped, respectively. Since we calculate the value of each gene as a weighted sum of included SNVs, we define the genes as the manifest variables. By relating the collapsed genes onto the pathways according to their mapping, we construct the outer model. Next, we define the inner model. Since there are three phenotypes ($Q = 3$) are assigned, we need to related all the latent variables onto the phenotypes. Finally, the two covariates in the example are added on the inner model, to complete the construction.

**Figure 4.1. A graphical representation of *PHARAOH-multi*.** The exemplary model is described with the number of pathways $K$=3, the number of phenotypes $Q$=3, the number of covariates $L$=2, and the number of genes for each pathway $T_1$, $T_2$ and $T_3$ are 2, 3 and 2, respectively. Variable $w_{kt}$ denote the weights assigned to the collapsed genes, and $\beta_{ik}$ are coefficients on the pathway latent variables. Residual terms were omitted.

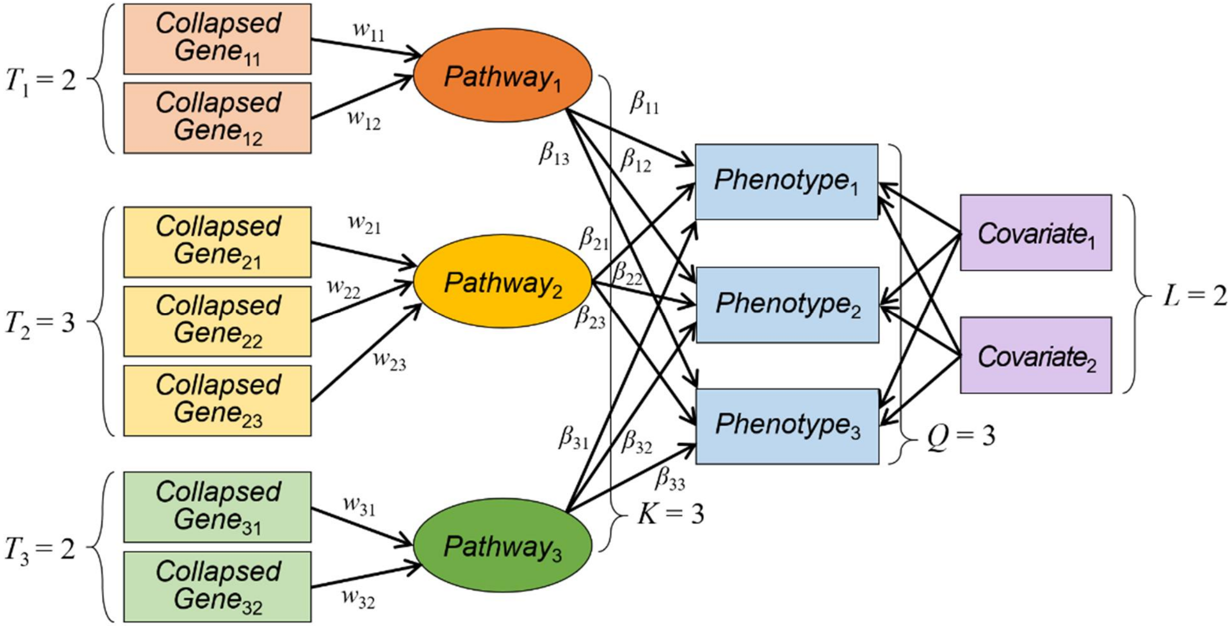


**Table 4.1. Parameters related to specific relationships for the proposed model.** $\boldsymbol{P}_k$ indicates the $k^{\text{th}}$ pathway, $\boldsymbol{Y}_q$ is the $q^{\text{th}}$ phenotype, $\boldsymbol{Y}_*$ indicates all phenotypes, and $\boldsymbol{G}_{tk}$ indicates the $t^{\text{th}}$ gene in the $k^{\text{th}}$ pathway.

| | | **Coefficients** |
|---|---|---|
| **Relationship** | $\boldsymbol{P}_k \rightarrow \boldsymbol{Y}_*$ | $\beta_{k1}, \ldots, \beta_{kQ}$ |
| | $\boldsymbol{P}_k \rightarrow \boldsymbol{Y}_q$ | $\beta_{kq}$ |
| | $\boldsymbol{G}_{tk} \rightarrow \boldsymbol{Y}_*$ | $w_{tk}\beta_{k1}, \ldots, w_{tk}\beta_{kQ}$ |
| | $\boldsymbol{G}_{tk} \rightarrow \boldsymbol{Y}_q$ | $w_{tk}\beta_{kq}$ |

### 4.2.3. Parameter estimation

The proposed model seeks to associate pathways and phenotypes. The effect of the $k^{\text{th}}$ pathway, on multiple phenotypes, can be determined by testing all coefficients of the pathways simultaneously ($\mathrm{H}_0 : \beta_{k1} = \cdots = \beta_{kQ} = 0$).

Moreover, by its nature, the proposed method can further assess three more associations: (1) the effect of a gene on multiple phenotypes conditioned by a given pathway; (2) the effect of a gene on a phenotype conditioned by the pathway; and (3) the effect of a pathway on a phenotype. Detailed characteristics including relationships and coefficients of the proposed model (***PHARAOH-multi***) are shown in **Table (4.1)**.

In order to estimate the parameters, let $B$ is a matrix of $\tilde{\beta}_1, \cdots, \tilde{\beta}_Q$. From the above model, we seek to maximize the penalized log-likelihood function, to estimate the parameters $w_{kt}$ and $\beta_{kq}$, subject to the conventional scaling constraint $\sum_{j=1}^{N} f_{jk}^2 = N$ (Takane and Hwang, 2005). The penalized log-likelihood function is expressed to

$$\begin{aligned}\ell(B, W, \Sigma | Y_i, X) = & -\frac{NQ}{2}\log\pi - \frac{N}{2}\log\det\Sigma \\ & -\frac{1}{2}\sum_{i=1}^{N}(Y_i - B'F_i)'\Sigma^{-1}(Y_i - B'F_i) - \frac{1}{2}\lambda_G\sum_{k=1}^{K}\sum_{t=1}^{T_k}\|w_{kt}\|_2 \\ & -\frac{1}{2}\lambda_G\sum_{q=1}^{Q}\sum_{k=1}^{K}\|\beta_{kq}\|_2,\end{aligned} \tag{4.2}$$

where $\lambda_G$ and $\lambda_P$ are the penalty parameters for each specific gene and pathway, respectively, and $\|w_{tk}\|_2$ and $\|\beta_{kq}\|_2$ are the ridge penalties.

In the method section of ***PHARAOH***, we introduced an iteratively reweighted least square (IRLS) method to minimize an univariate version of **Equation (4.2)** under the presence of ridge penalties, which is similar to the alternating regularized least-squares algorithm which was proposed by (Hwang, 2009). Here we extend the previous algorithm to multivariate analysis. Let $R_i$ be a “column-trimmed” matrix of GSCA (Hwang, 2009), defined by $F_i \otimes I_K$, where $\otimes$ is Kronecker product, and $I_K$ is $K \times K$ identity matrix. Maximization of **Equation (4.2)** in respect of $B$ and $W$ is equivalent to minimizing the following least-square functions:

$$
\begin{aligned}
\phi_B &= \sum_{i=1}^{N}(Y_i - B'F_i)'\Sigma^{-1}(Y_i - B'F_i) + \lambda_P \sum_{q=1}^{Q}\sum_{k=1}^{K}\|\beta_{kq}\|_2 \\
&= \sum_{i=1}^{N}\big(Y_i - R_i\mathrm{vec}(B)\big)'\Sigma^{-1}\big(Y_i - R_i\mathrm{vec}(B)\big) + \lambda_P\mathrm{vec}(B)'\mathrm{vec}(B) \\
&= \big(\mathrm{vec}(Y) - R\mathrm{vec}(B)\big)'\big(\mathrm{vec}(Y) - R\mathrm{vec}(B)\big) + \lambda_P\mathrm{vec}(B)'\mathrm{vec}(B)
\end{aligned}
\tag{4.3}
$$

$$
\begin{aligned}
\phi_W &= \sum_{i=1}^{N}(Y_i - B'X_iW)'\Sigma^{-1}(Y_i - B'X_iW) + \lambda_G \sum_{k=1}^{K}\sum_{t=1}^{T_k}\|w_{tk}\|_2 \\
&= \sum_{i=1}^{N}(Y_i - \Phi_iW)'\Sigma^{-1}(Y_i - \Phi\,_iW) + \lambda_G \sum_{k=1}^{K} w_k'w_k \\
&= (\mathrm{vec}(Y) - \Phi W)'(\mathrm{vec}(Y) - \Phi W) + \lambda_G \sum_{k=1}^{K} w_k'w_k
\end{aligned}
\tag{4.4}
$$

These least-square functions are subject to $\mathrm{diag}(R'R) = N\mathrm{I}_{NQ}$, where $\Phi_i$ is a column-trimmed matrix of $B' \otimes X_i$ (Hwang, 2009), and $\mathrm{vec}(\cdot)$ is a vectorization operator. Then, it can be easily shown that the covariance matrix Σ is not related to the above equations, since the ***PHARAOH-multi*** model uses multivariate linear model. In this respect, an estimation of Σ can be done after convergence of $B$ and $W$, by minimizing the first derivate of **Equation (4.2)** with respect to Σ, as:

$$\hat{\Sigma} = \frac{1}{N}\big(Y - Rvec(B)\big)'\big(Y - Rvec(B)\big) \tag{4.5}$$

Similarly, $B$ and $W$ can be updated by equating **Equations (4.3)** and **(4.4)** to zero. This then gives the estimating equation of $B$ and $W$ as:

$$vec(\hat{B}) = (R'R)^{-1}R'vec(Y) \tag{4.6}$$

$$vec(\widehat{w}) = (\Phi'\Phi)^{-1}\Phi'vec(Y) \tag{4.7}$$

Taken together, the overall procedure of the proposed algorithm is as follows:

1. Let $t = 1$.

2. Assign random initial values to $W$, which are then represented by $W_{(0)}$.

3. Calculate $F_{(t)}$, using $W_{(t-1)}$.

4. Update $B_{(t)}$, using $F_{(t)}$.

5. Update $W_{(t)}$, using $F_{(t)}$ and $B_{(t)}$.

6. Repeat until the sum of the differences $\left|W_{(t)} - W_{(t-1)}\right| + \left|B_{(t)} - B_{(t-1)}\right|$ converges the threshold.

Finally, we determine the values of $\lambda_G$ and $\lambda_P$, before applying the parameter estimation procedure. To that end, we can implement $k$-fold cross-validation (CV) to determine the values of $\lambda_G$ and $\lambda_P$. First, we construct a two-dimensional grid of different $\lambda_G$ and $\lambda_P$ values. Then we compute the deviance of each model with the given $\lambda_G$ and $\lambda_P$, for all CV fold values. Finally, $\lambda_G$ and $\lambda_P$ are selected by their average deviance, which is minimized.

### 4.2.4. Significance testing

In order to assess the significance of association between phenotypes and genes or phenotypes and pathways, resampling methods can be used to test the statistical significance of the estimated effects of all pathways on the phenotype. In the proposed method, we utilize a permutation test to obtain $p$-values. By permuting the given phenotype, our method first generates null distributions for both pathways and gene coefficients. By computing the quantile of estimated pathway and gene coefficients from the non-permuted dataset in each empirical null distribution, we can obtain an empirical $p$-value for any specific pathway and gene. Unlike the ***PHARAOH*** method, ***PHARAOH-multi*** has two types of $p$-value: phenotype-wise type and joint type.

The testing of joint effects across multiple phenotypes is crucial to ***PHARAOH-multi***. As shown in **Table (4.1)**, ***PHARAOH-multi*** provides the individual effects of a pathway on each phenotype through $\beta_{1k}, \ldots, \beta_{Qk}$. The global effect of a pathway, on all phenotypes, can be evaluated by jointly testing $\beta_{1k}, \ldots, \beta_{Qk}$. Here, we introduce two different schema for determining a joint $p$-value for the $k^{\text{th}}$ pathway, from multiple phenotypes.

Our first approach was to combine the individual $p$-values (referred as "P_K"). Since there are considerations among the estimated coefficients $\beta_{1k}, \ldots, \beta_{Qk}$, these correlations should be accounted for combining multiple $p$-values. Let the $p$-values from the $k^{\text{th}}$ pathway be denoted by $P_{1k}, \ldots, P_{Qk}$. The simplest way to combine those $p$-values is to use Fisher's method, which is denoted by $\Psi_k = -2\sum_{i=1}^{Q} \log P_{ik}$ under the independence assumption. Then, it is known that the test statistic $\Psi_k$ follows the $\chi^2$ distribution with the degrees of freedom $2Q$, under the null hypothesis. In the other hand, an extended version of Fisher's method, Brown's method, can combine dependent $p$-values using a rescaled $\chi^2$ distribution and covariance of $p$-values (Brown, 1975). However, an analytical computation of the covariance is not feasible for large-scale datasets, due to their computational complexity. A solution for this problem (Kost and McDermott, 2002) introduced an approximation using a third-order polynomial for the covariance, denoted by $\text{cov}(-2\log P_i, -2\log P_j) \approx 3.263\rho_{ij} + 0.71\rho_{ij}^2 + 0.027\rho_{ij}^3$. To that end, Kost's approach has been shown to be one of the best working methods for combining $p$-values (Alves and Yu, 2014). Here, we adopt Kost's method by substituting $\rho$ to the empirical correlation of estimated coefficients, $\beta_{1k}$, ... , $\beta_{Qk}$,, and derive the statistic for joint effect between the $k^{\text{th}}$ pathway and multiple phenotypes, as follows:

$$P_{Kost,k} = 1 - \Phi_{2d_k}(\Psi_k \,/\, c_k), \tag{4.8}$$

where $c_k$, $d_k$ and $\Phi_{2d_k}$ are the scale parameter, the re-scaled degree of freedom, and the cumulative distribution function of $\chi^2$, with the degree of freedom $2d_k$ for the $k^{\text{th}}$ pathway, respectively (Kost and McDermott, 2002). Unlike other methods for combining multiple $p$-values such as Fisher's method, the Kost's method has an advantages from considering the correlation across $p$-values.

Our second approach was to construct a single statistic that combines all $Q$ coefficients (referred as "P_M"). Here, we define a simple Wald-type statistic, $\boldsymbol{T}$, as below, and utilize $\boldsymbol{T}$ for the following permutation testing scheme:

$$\boldsymbol{T} = \tilde{\beta}_k' cov^{-1}(\tilde{\beta}_k)\tilde{\beta}_k \tag{4.9}$$

Then, the estimated covariance $cov\left(\hat{\tilde{\beta}}_k\right)$ can be directly estimated using **Equation (4.6)** with equation $cov\left(\hat{\tilde{\beta}}\right) = (F'F + \lambda_P I)^{-1}F'F(F'F + \lambda_P I)^{-1} \otimes \hat{\Sigma}$ (Hoerl and Kennard, 1970), or can be altered by calculating sample covariance of $\tilde{\beta}_k$, from permutations.

### 4.2.5. Multiple testing correction

An association study with large number of variables suffers from the "multiple testing problem". This problem occurs when we perform a set of statistical tests simultaneously, and becomes a critical barrier especially for the sequencing dataset. For example, assume that we want to perform an association test for each genetic variant at the 5% significance level. For each test, if the null hypothesis is true, there is 5% probability of incorrect rejection. However, if our dataset has 10,000,000 variants and the null hypothesis is true for all the tested variants, there will be 500,000 of the incorrectly rejected tests, which is unacceptably large number. In order to overcome the multiple testing problem, Bonferroni correction can be a straightforward approach. However, it may impose an adjustment that is too stringent. For instance, the adjusted cutoff at 5% using Bonferroni correction is $5 \times 10^{-9}$, from the above example.

While the association tests using gene- or pathway-level information can substantially relax the adjusted cutoff of Bonferroni correction by reducing the number of tests, the "multiple testing problem" still remains. Especially, since the Bonferroni correction assumes that the tests are independent, it gives more stringent correction for correlated results (Meinshausen, et al., 2011). To overcome this issue, we applied two types of multiple testing corrections.

First, ***PHARAOH-multi*** corrects $p$-values using the Westfall & Young permutation procedure (Westfall and Young, 1993), which can be easily adopted, since ***PHARAOH-multi*** already uses a permutation scheme. Let $T_{(0)}$

be a vector of the statistics calculated using observed, non-permuted phenotypes, and let $T_{(j)}$ be those from the $j^{\text{th}}$ permutation. First, we rank the values of $T_{(0)}$ in ascending order, and let the rank of the $k^{\text{th}}$ pathway, and the $k^{\text{th}}$ index, be $r_k$ and $r_{(k)}$ , respectfully. Then, for each permutation $j = 0,1,\cdots,J$, let $T'_{(j)}$ be $T_{(j)r_{(1)}},\cdots,T_{(j)r_{(K)}}$, to define $T^M_{(j)}$ as a cumulative maximum of $T'_{(j)}$. Let $I_{j,k}$ be an indicator function that resolves to 1.0, if $T'_{(0)r_k} < T^M_{(j)r_k}$, or 0.0, if that condition does not hold. The adjusted $p$-value for the $k^{\text{th}}$ pathway, by the Westfall & Young procedure, is then defined as:

$$P_k^{adj} = \frac{1 + \sum_{j=1}^{J} I_{j,k}}{1 + J} \tag{4.10}$$

Second, ***PHARAOH-multi*** provides multiple testing adjustment, by calculating adjusted $p$-values by Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Here, we first obtain $K$ as the number of permutation $p$-values, and from those, we can derive the adjusted $p$-values, using the Benjamini-Hochberg step-up procedure which is expressed by the below procedure.

1) For a given significance level $\alpha$, find the largest $k$ such that $P_{(k)} \leq \frac{k}{m}\alpha$.

2) Reject the null hypothesis for all $H_{(i)}$ for $i = 1,\cdots,k$.

## 4.3. Simulation study

### 4.3.1. The simulation model

To evaluate the performance of the proposed method, we conducted simulation studies, under the two interested scenarios. First, we compared the statistical powers of the ***PHARAOH-multi*** and the compared methods to investigate whether the proposed method exhibits better statistical power or not. Second, we investigated the statistical powers of ***PHARAOH-multi*** by individually varying the simulation parameters, to assess the effect of each simulation parameter in detail.

For generating rare variants, we first produced a pool of genetic variants, using SimRare (Li, et al., 2012), a rare variant simulator with well-established genetic assumptions. A pool was then generated, with default settings and gene lengths of 1Kbp. Next, we generated a simulation dataset of 10 pathways, with 1,000 samples, for each replicate. All simulation scenarios were evaluated, using 1,000 replicates. Based on the genotypes, the simulated phenotypes were generated by the following model, with an assumption that only the first pathway is causal to the phenotypes:

$$y_{iq} = \beta_{1q}\tilde{f}_{i1} + \epsilon_{iq} = \beta_{1q}\sum_{t=1}^{H_1} w_{1t}x_{i1t} + \epsilon_{iq} = \beta_{1q}\sum_{t=1}^{H_1}\left(w_{1t}\sum_{j=1}^{M_{1t}}\gamma_{1tj}g_{i1tj}\right) + \epsilon_{iq} \quad (4.11)$$

This is then subject to $\mathrm{diag}(F'F) = N\mathrm{I}_K$, where $H_1$ is the number of causal genes in the first pathway, and $M_{1t}$ is the number of rare variants in the $t^{\mathrm{th}}$ gene of the first pathway (i.e., causal pathway).

In the above model using **Equation (4.11)**, $\gamma_{1tj}$ denotes the effect of the $j^{\text{th}}$ genetic variant, of the $t^{\text{th}}$ gene, set to $|\log_{10} MAF_{tj}|$, such that $\epsilon_{iq}$ denotes the residual and follows $\mathrm{MVN}(0, \Sigma)$. In our simulation, the settings $q$ = 1,2, and $H_1$ = 1, 2, 5, 10, were used. For each replicate, all rare variants were collapsed into genes.

For the simulation and the analysis of the real dataset, a workstation system with two Intel Xeon E5-2620 CPUs with a combined RAM of 128GiB, were used. Note that the aSPU and MARV analyses were performed using the default settings, except that aSPU was performed without "genetic variant pruning" capability, as we observed that aSPU raises "unrecoverable error" with that capability. For our proposed method and aSPU, the number of permutations was 5,000, to prevent possible lower bound limitation. For ***PHARAOH-multi***, we selected the tuning parameters $\lambda_G$ and $\lambda_P$,, based on three-fold CV for each replicate, using two-dimensional grids of $\lambda_G$ and $\lambda_P$,, with six different starting points of ridge parameters, ranging from $10^1$ to $10^6$ (on a logarithmic base 10 scale).

**Figure 4.2. Type 1 error simulation result.** Plots in the first and second row represent the type 1 error at $\alpha$=0.05 and $\alpha$=0.01, respectively. Each bar is the mean and the error bars represent SD.

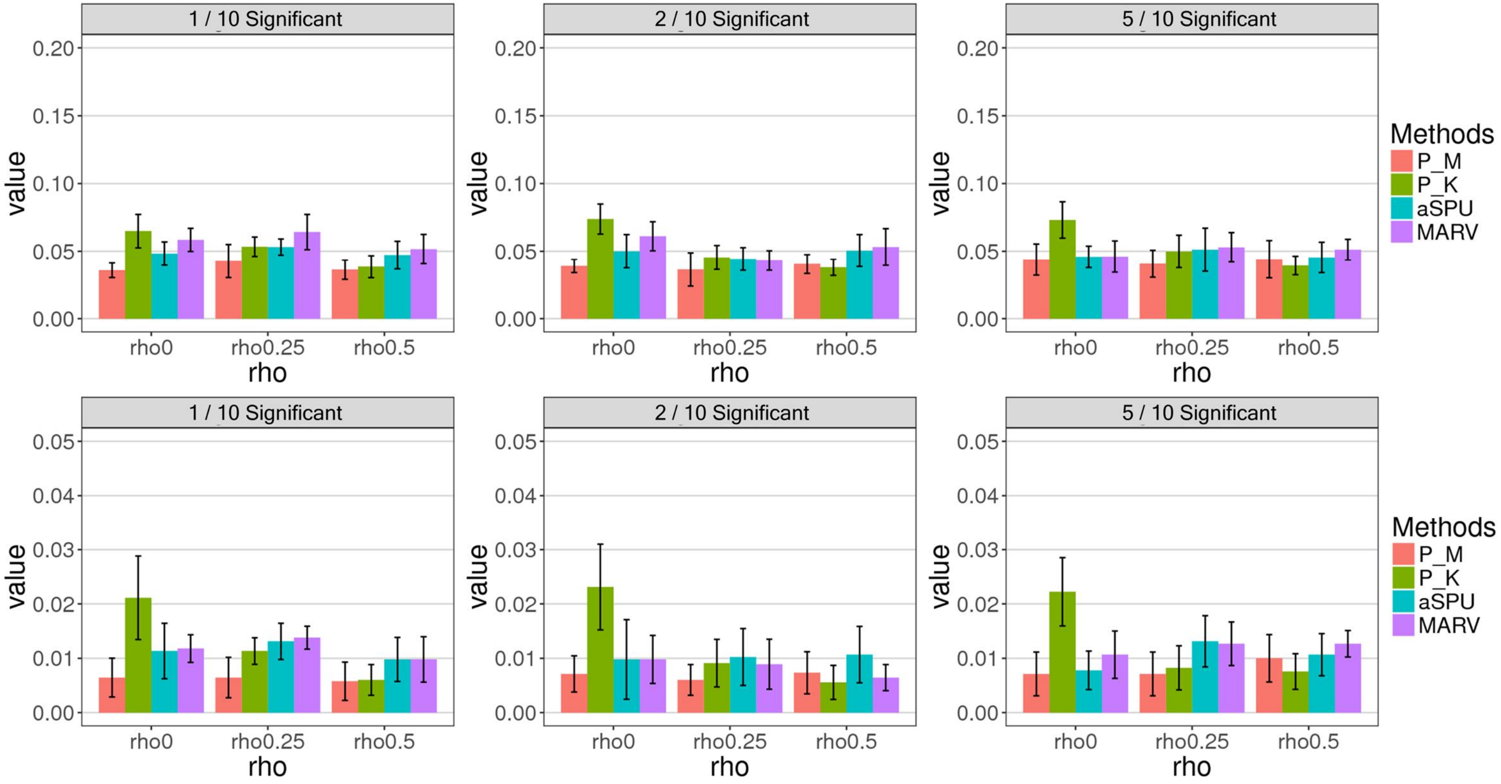
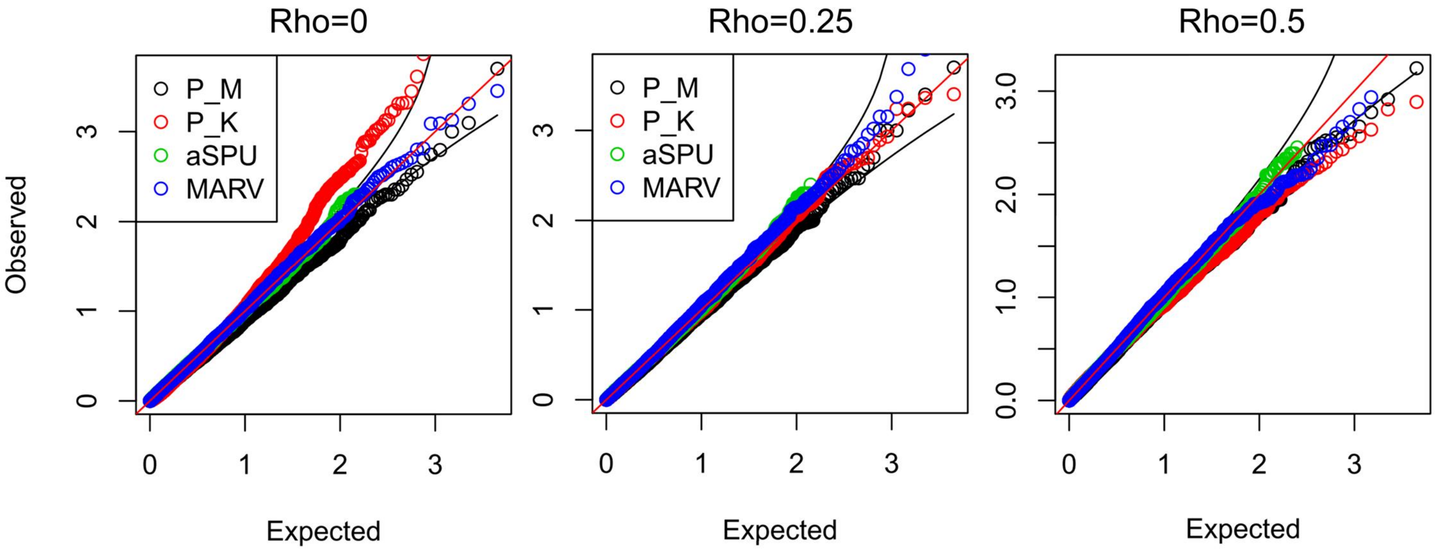


**Figure 4.3. Quantile-Quantile plots of type 1 error evaluation, without multiple testing adjustment.**



### 4.3.2. Evaluation with simulated data

In our simulation, $H_1$, $w_{1t}$, $\beta$, and $\rho$ were assumed to be 1, 2, and 5; 0.1 and 0.2; 0.1, 0.15, and 0.2; and 0, 0.25, 0.5, respectively, with evaluation of their exhaustive combinations. Other parameters, $Q$, $K$ and $T_K$, were fixed to 2, 10, and 10, respectively. To calculate the type 1 error rate and the statistical power, 1,000 and 500 replicates of the independently generated phenotype sets were respectively used.

We first compared the type 1 error rates of the proposed method versus the exiting methods. Here, type 1 error rate was computed as a proportion of $p$-values for the pathways with no effect, and was less than the significance level, across 1,000 replicates of permuted phenotypes. As shown in **Figure (4.2)**, we evaluated the type 1 errors using two significance levels, 0.01 and 0.05. As a result, type 1 errors were controlled well in the traditional methods, but ***PHARAOH-multi*** showed a moderately deflated type 1 error rate (P_M), while the inflated rate is P_K, when $\rho$=0. In contrast, the quantile-quantile (Q-Q) plots in **Figure (4.3)** show no inflation or deflation pattern, in all the methods, except for P_K, with no correlation between phenotypes.

It was also worthwhile to assess the gain of power in the multivariate analysis, as compared to univariate analysis. In this respect, our simulation study was conducted to compare the power gain from multivariate methods, and between multivariate and univariate analyses.

First, we checked whether ***PHARAOH-multi*** with multiple phenotypes boosts power compared to ***PHARAOH*** with a single phenotype, under the same scenarios of the power simulation. As a result, we observed that the power of PHAROH-multi was at least 2.52 times larger than ***PHARAOH***, and this difference becomes even larger, as $w$ and $\beta$ increase (data not shown).

Second, we assessed the statistical power of ***PHARAOH-multi*** and the compared methods, defined as the proportion of the adjusted $p$-value of the simulated causal pathway (the first pathway) being less than the significance threshold, e.g., 0.05. Despite the proposed method supporting the Westfall-Young permutation procedure, it was not considered in the simulation study, due to the absence of corresponding adjustments in the compared methods.

**Figure (4.4)** to **(4.9)** show comparison results of statistical power simulation from 1,000 replications. Each figure represents the same settings of $w$ and $\beta$, with different numbers of causal genes in the causal pathway, and each column represents the same number of causal genes, with different effect settings across the figures.

In most scenario comparisons, the two proposed statistics obtained by ***PHARAOH-multi*** (P_M) and $p$-value aggregation (P_K) showed greater power than the other two approaches, aSPU and MARV. However, this did not hold when 50% of the genes were causal for a specific pathway, with effect sizes of $w$=0.1 and $\beta$=0.1. In order to investigate whether or not there are significant differences among powers, we performed paired $t$-tests between a pair of methods. In **Figure (4.4)** to **(4.6)** for the case of $w$ = 0.1, the $p$-values were $3\times10^{-7}$ for comparing powers of P_M and aSPU, and $4.2\times10^{-6}$ for comparing those of P_M and MARV. In **Figure (4.7)** to **(4.9)** for the case of $w$ = 0.2, the same pairwise comparison for the $p$-values were $7.3\times10^{-7}$ and $4.2\times10^{-6}$, respectively. In overall scenarios, powers of P_M were larger up to 18%p compared to aSPU in $H_1$=5, $w$=0.2 and $\beta$=0.2, and were larger up to 83%p compared to MARV. Generally, P_K exhibited smaller power than P_M, and showed comparable or slightly smaller power, than aSPU.

Here, we observed three interesting patterns in the results. First, the proposed P_M and aSPU methods showed lower power, when the proportion of causal genes increase, compared to MARV. Second, the power rapidly increased, as $\beta$ increased, as shown in **Figure (4.4)** to **(4.9)**. Third, the contribution of $w$ to the power was relatively moderate, compared to $\beta$, as shown in corresponding scenarios of **Figure (4.4)** to **(4.9)**. The reason for the occurrence of those two patterns is that the model(s) generate phenotypes for power simulation, and **Equation (4.11)** requires the constraint of the so-called "latent variable," in GSCA (see Methods). While both ***PHARAOH-multi*** and aSPU construct hierarchies of genes and pathways, MARV essentially treats a pathway as a large set of SNVs, since the motivation of MARV is for region- vs. pathway-based tests. The simulation setting and its overall effect on phenotypes is summarized, first at the gene-level, and then by the expression of a linear combination of those genes. In this respect, the results of ***PHARAOH-multi*** and aSPU were more plausible than those of MARV, because those two methods more properly reflected the simulation settings.

**Figure 4.4. Comparison of simulation results of statistical power from various methods of multiple testing adjustment (*w* = 0.1 and *β* = 0.1).** Red and green bars represent results obtained by the ***PHARAOH-multi*** ("P_M", using joint testing) and *p*-value aggregation ("P_K") methods, respectively. Teal and purple bars represent the aSPU and MARV methods, respectively. Powers were calculated by the proportion of the causal pathway's adjusted *p*-value, obtained by the Benjamini-Hochberg procedure (< 0.05). Paired *t*-test *p*-values for *w* = 0.1 are $7.3\times10^{-7}$ and $4.2\times10^{-6}$ for P_M vs. aSPU and P_M vs. MARV, respectively.

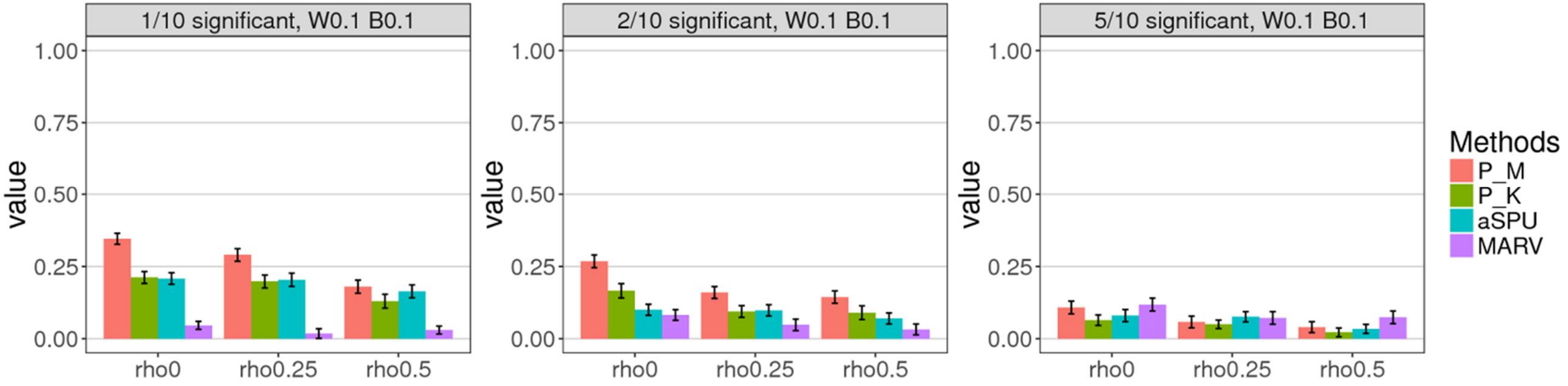


**Figure 4.5. Comparison of simulation results of statistical power from various methods of multiple testing adjustment ($w$ = 0.1 and $\beta$ = 0.15).** Red and green bars represent results obtained by the ***PHARAOH-multi*** ("P_M", using joint testing) and $p$-value aggregation ("P_K") methods, respectively. Teal and purple bars represent the aSPU and MARV methods, respectively. Powers were calculated by the proportion of the causal pathway's adjusted $p$-value, obtained by the Benjamini-Hochberg procedure (< 0.05). Paired $t$-test $p$-values for $w$ = 0.1 are $7.3\times10^{-7}$ and $4.2\times10^{-6}$ for P_M vs. aSPU and P_M vs. MARV, respectively.

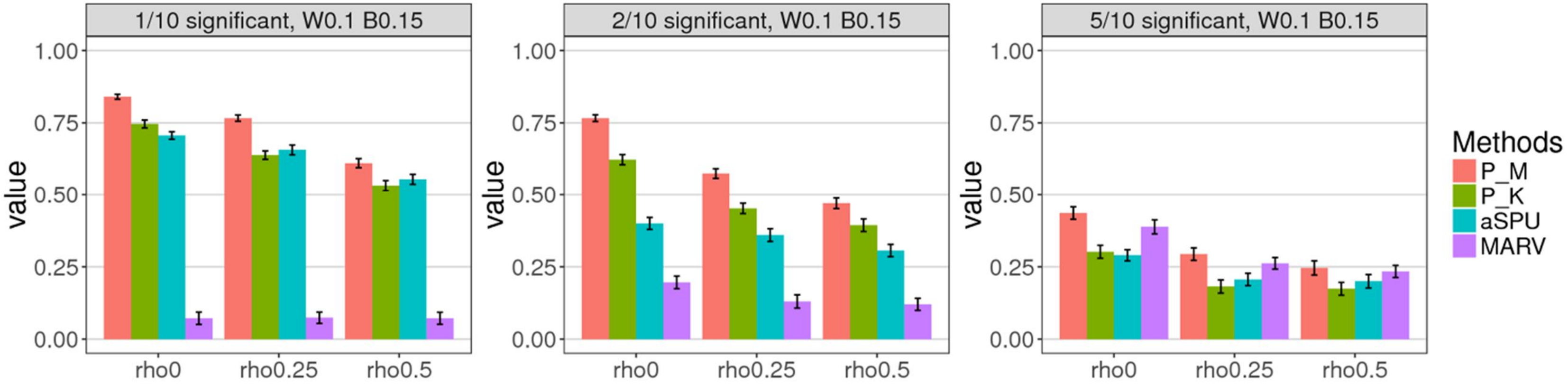


**Figure 4.6. Comparison of simulation results of statistical power from various methods of multiple testing adjustment (*w* = 0.1 and *β* = 0.2).** Red and green bars represent results obtained by the ***PHARAOH-multi*** (“P_M”, using joint testing) and *p*-value aggregation (“P_K”) methods, respectively. Teal and purple bars represent the aSPU and MARV methods, respectively. Powers were calculated by the proportion of the causal pathway’s adjusted *p*-value, obtained by the Benjamini-Hochberg procedure (< 0.05). Paired *t*-test *p*-values for *w* = 0.1 are $7.3\times10^{-7}$ and $4.2\times10^{-6}$ for P_M vs. aSPU and P_M vs. MARV, respectively.

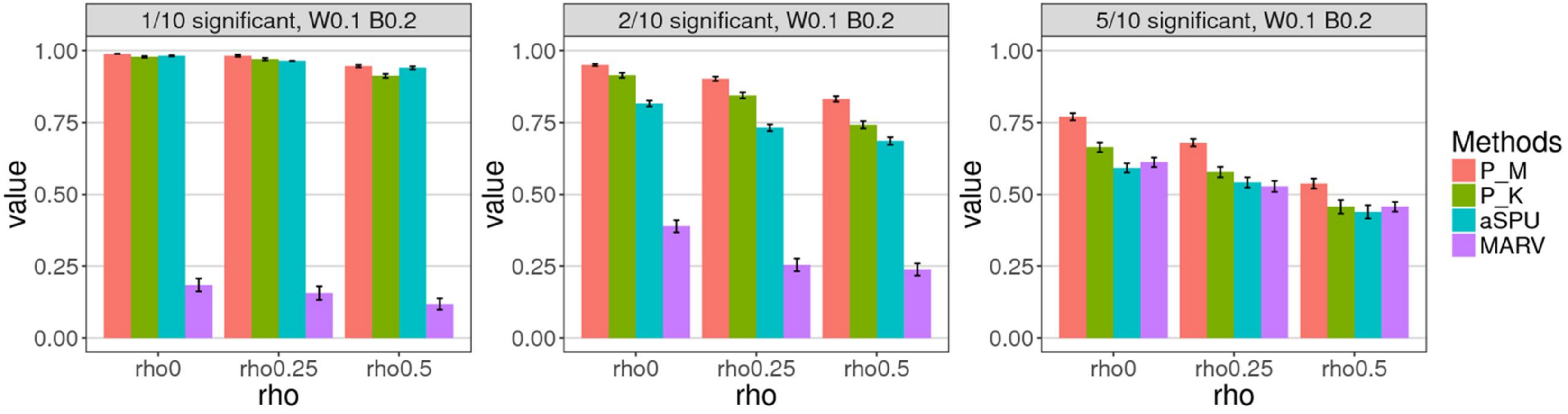


**Figure 4.7. Comparison of simulation results of statistical power from various methods of multiple testing adjustment ($w$ = 0.2 and $\beta$ = 0.1).** Red and green bars represent results obtained by the ***PHARAOH-multi*** ("P_M", using joint testing) and $p$-value aggregation ("P_K") methods, respectively. Teal and purple bars represent the aSPU and MARV methods, respectively. Powers were calculated by the proportion of the causal pathway's adjusted $p$-value, obtained by the Benjamini-Hochberg procedure (< 0.05). Paired $t$-test $p$-values for $w$ = 0.1 are 7.3×$10^{-7}$ and 4.2×$10^{-6}$ for P_M vs. aSPU and P_M vs. MARV, respectively.

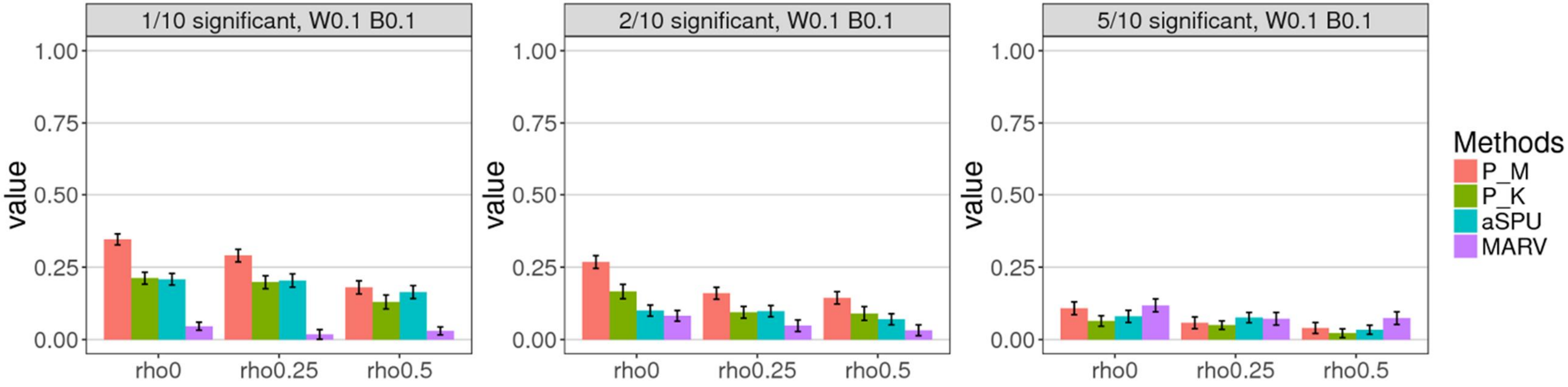


**Figure 4.8. Comparison of simulation results of statistical power from various methods of multiple testing adjustment (*w* = 0.2 and *β* = 0.15).** Red and green bars represent results obtained by the ***PHARAOH-multi*** ("P_M", using joint testing) and *p*-value aggregation ("P_K") methods, respectively. Teal and purple bars represent the aSPU and MARV methods, respectively. Powers were calculated by the proportion of the causal pathway's adjusted *p*-value, obtained by the Benjamini-Hochberg procedure (< 0.05). Paired *t*-test *p*-values for *w* = 0.1 are 7.3×$10^{-7}$ and 4.2×$10^{-6}$ for P_M vs. aSPU and P_M vs. MARV, respectively.

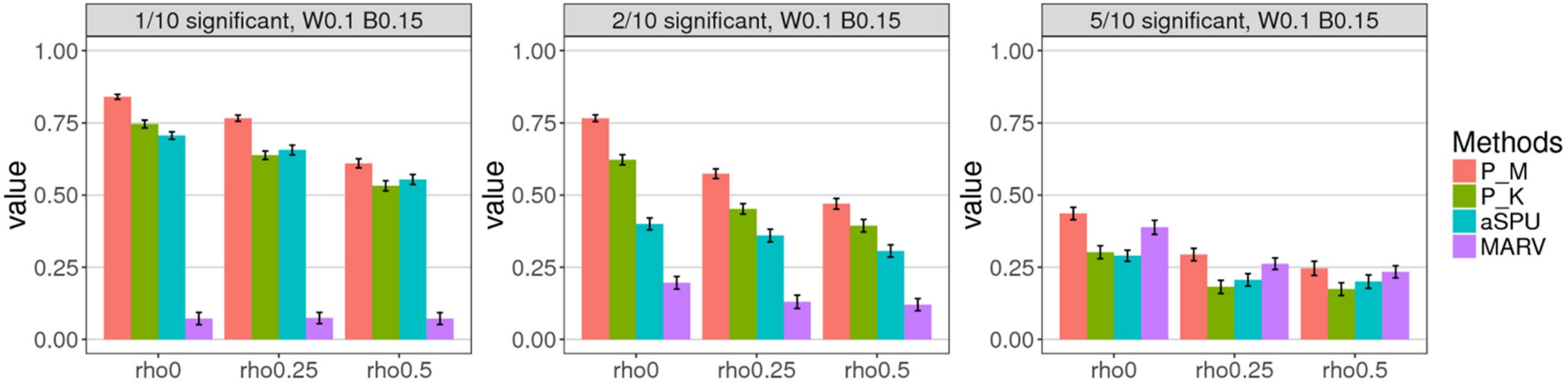


**Figure 4.9. Comparison of simulation results of statistical power from various methods of multiple testing adjustment (*w* = 0.2 and *β* = 0.2).** Red and green bars represent results obtained by the ***PHARAOH-multi*** ("P_M", using joint testing) and *p*-value aggregation ("P_K") methods, respectively. Teal and purple bars represent the aSPU and MARV methods, respectively. Powers were calculated by the proportion of the causal pathway's adjusted *p*-value, obtained by the Benjamini-Hochberg procedure (< 0.05). Paired *t*-test *p*-values for *w* = 0.1 are $7.3\times10^{-7}$ and $4.2\times10^{-6}$ for P_M vs. aSPU and P_M vs. MARV, respectively.

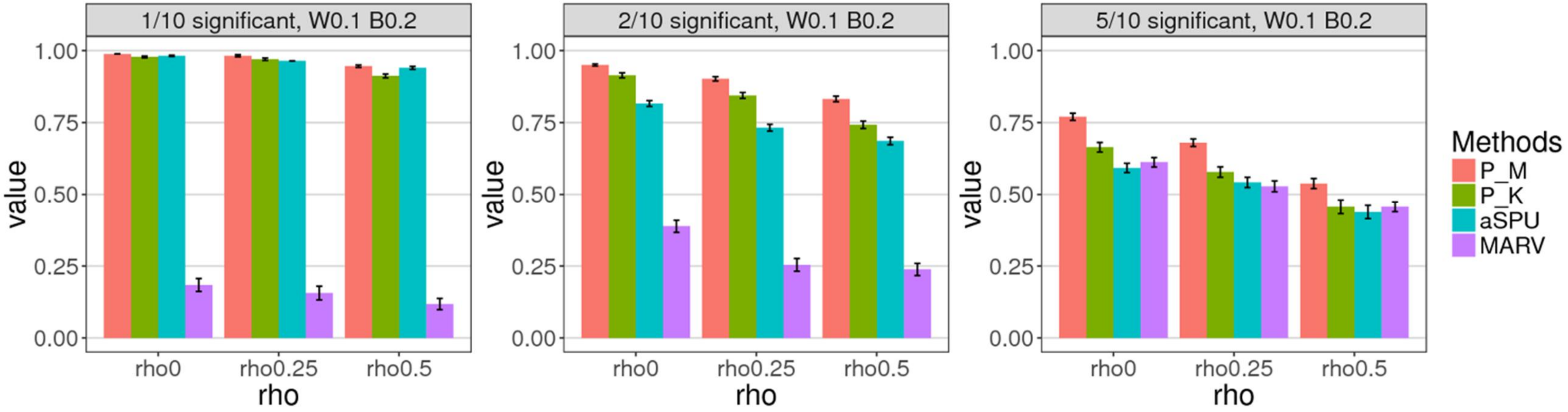
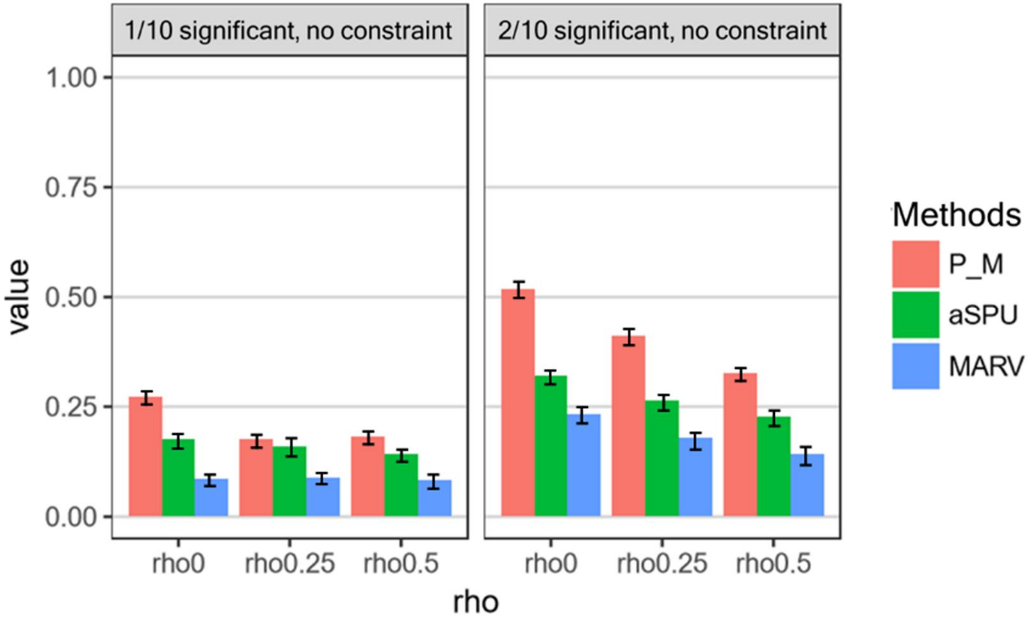


**Figure 4.10. Result of additional power simulation, without constraint of manifest variables.**



To confirm the above hypothesis, we performed an additional comparison using the same dataset, except that the phenotypes were generated without the constraint. As shown in **Figure (4.10)**, ***PHARAOH-multi*** ("P_M") showed larger power than in the previous simulation, while the power of MARV increased as the number of causal genes increased. However, in contrast to the previous results, the powers of ***PHARAOH-multi*** and aSPU also increased.

Finally, we investigated whether or not the statistical power changes by $M_{kt}$. For simplicity, we split 1,000 simulation datasets into two groups: the first group where the number of variants is small and the second where it is large. Then, we compared the power of each method between two groups using *t*-test. As a result, the *p*-values were 0.097 for aSPU, 0.684 for MARV, and 0.825 for ***PHARAOH-multi***. Thus, we concluded that $M_{kt}$ is unlikely to affect the simulation result regardless of the methods.

## 4.4. Application to the real datasets

To demonstrate the validity of the proposed method for examining large-scale datasets with multiple phenotypes, in real (biological) data analysis, we analyzed whole-exome sequencing (WES) data from a Korean population study.

**Figure 4.11. Q-Q plots of univariate *PHARAOH* for each pathway database.**

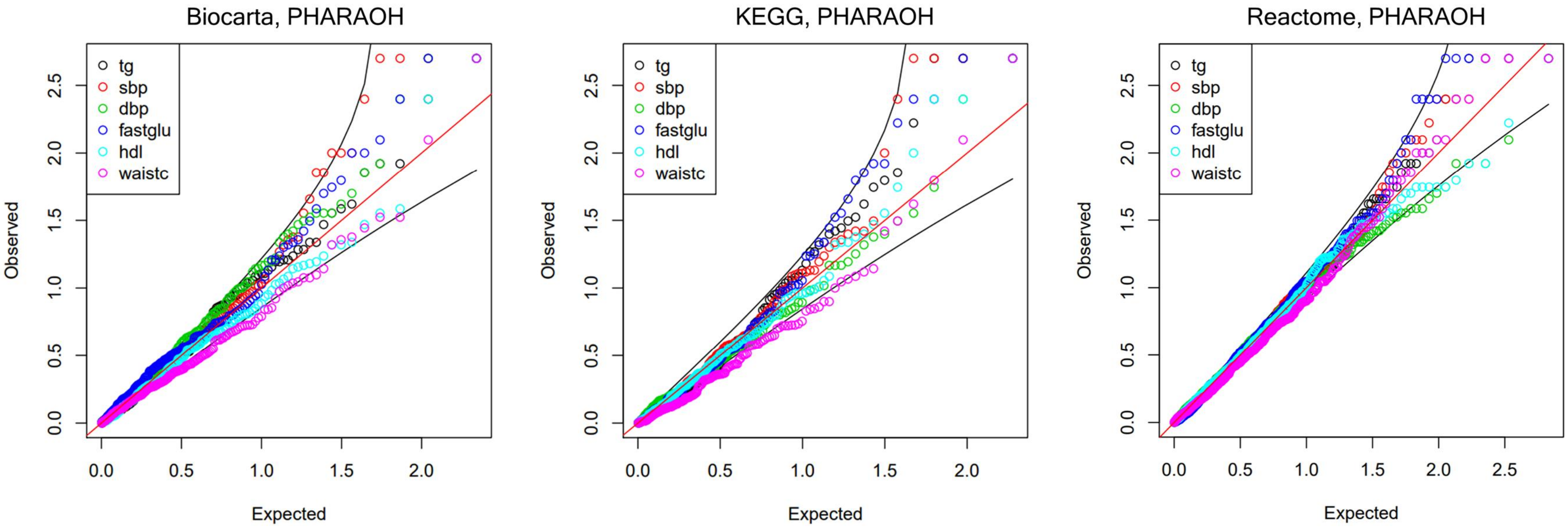


**Figure 4.12. Q-Q plots of discovery study using multivariate methods (*PHARAOH-multi*).** The plots were drawn by unadjusted *p*-values. Each Q-Q plot sequentially represents the results using three pathway databases (Biocarta, KEGG and Reactome).

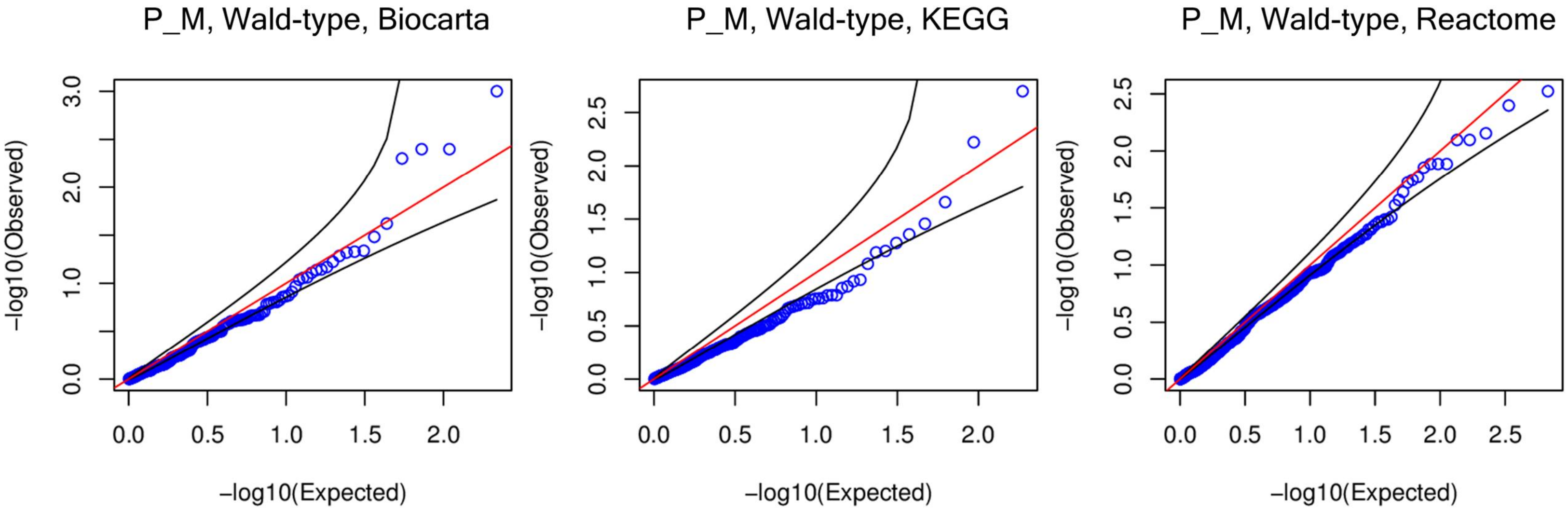


**Figure 4.13. Q-Q plots of discovery study using multivariate methods (aSPU).** The plots were drawn by unadjusted *p*-values. Each Q-Q plot sequentially represents the results using three pathway databases (Biocarta, KEGG and Reactome). Many *p*-values of aSPU are not appear in the Q-Q plots since aSPU reports many zero *p*-value that cannot be drawn in log scale.

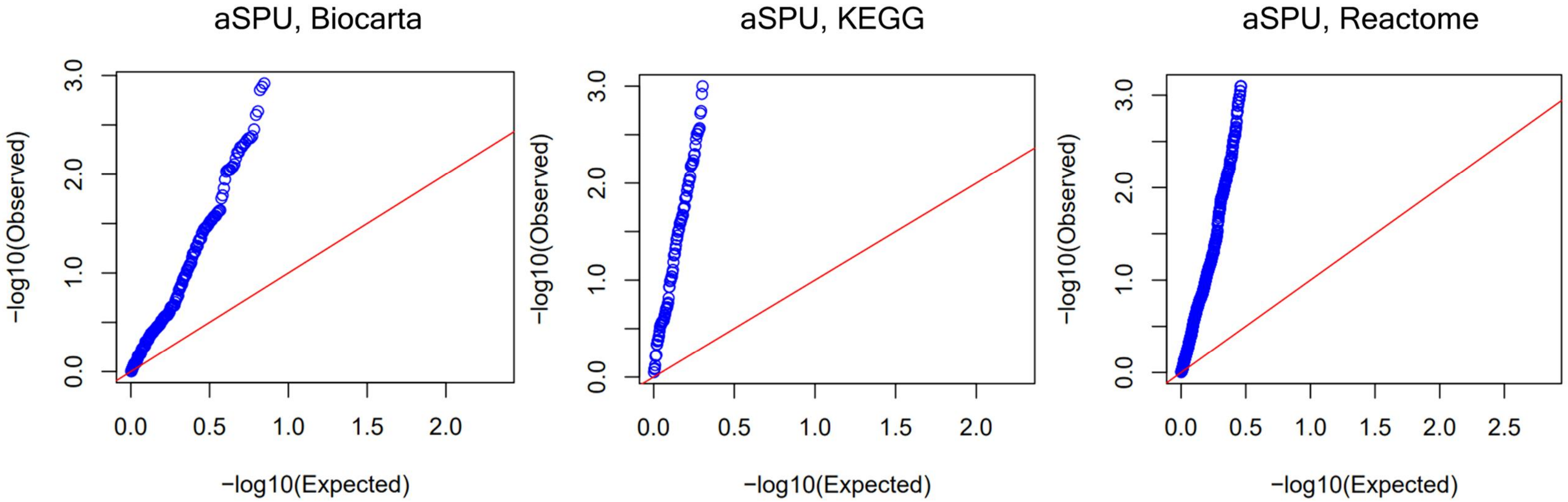
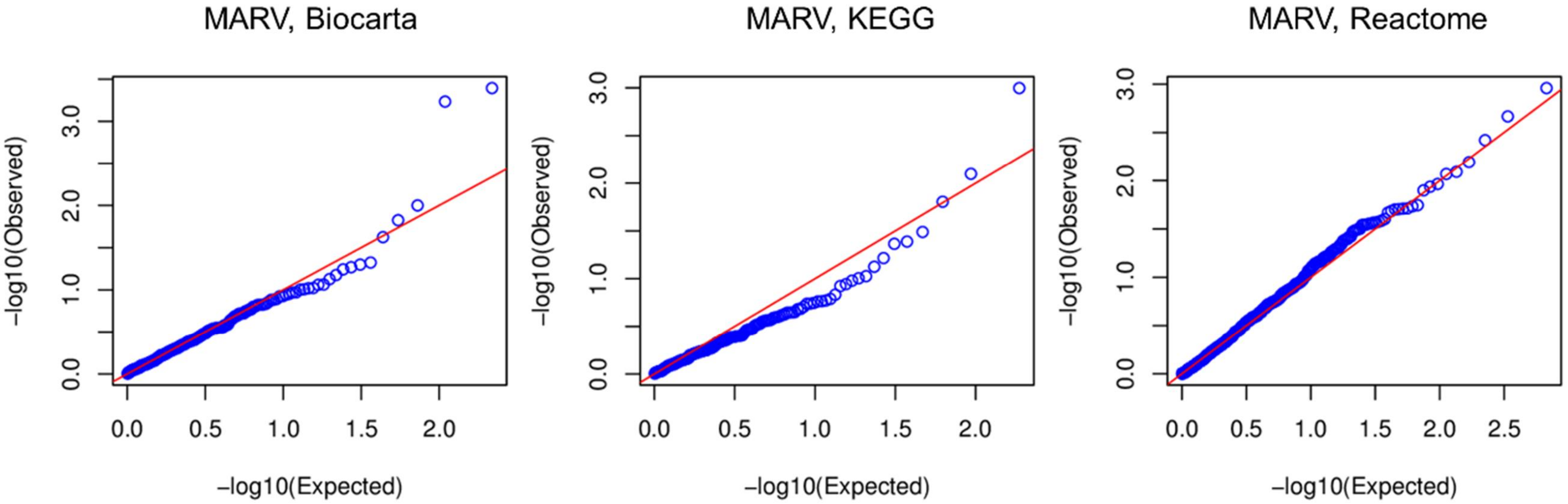


**Figure 4.14. Q-Q plots of discovery study using multivariate methods (MARV).** The plots were drawn by unadjusted *p*-values. Each Q-Q plot sequentially represents the results using three pathway databases (Biocarta, KEGG and Reactome).



MARV, Biocarta

MARV, KEGG

MARV, Reactome

**Table 4.2. Significant Biocarta pathways of *PHARAOH-multi* and MARV, and their adjusted *p*-values of multivariate and univariate analyses.** Bold numbers are the adjusted *p*-values below the significance threshold 0.1. P_M, aSPU and MARV indicate adjsuted *p*-values from the joint testing method of multiple phenotypes, and univariate adjusted *p*-values indicate the adjusted *p*-values of ***PHARAOH*** analysis for each phenotype.

| DB | Pathway | # variants | Multivariate adjusted *p*-value | | | Univariate adjusted *p*-value (*PHARAOH*) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **P_M** | **aSPU** | **MARV** | **tg** | **sbp** | **dbp** | **fastglu** | **hdl** | **waistc** |
| **Biocarta** | | | | | | | | | | | |
| | CDMAC pathway | 63 | **0.0858** | 0.5739 | 0.9638 | 0.9817 | 0.1094 | 0.5743 | 0.953 | 0.9967 | 0.9962 |
| | Cell2cell pathway | 112 | **0.0208** | 0 | 0.9638 | 0.7293 | 0.3063 | 0.5722 | 0.8234 | 0.9967 | 0.9962 |
| | GABA pathway | 46 | **0.0497** | 0.0085 | 0.8134 | 0.9783 | 0.1094 | 0.5722 | 0.8234 | 0.9967 | 0.9962 |
| | MPR pathway | 179 | **0.0208** | 0 | 0.9638 | 0.9783 | 0.1094 | 0.2188 | 0.8234 | 0.9845 | 0.9962 |
| | Caspase pathway | 649 | 0.8358 | 0.1741 | **0.0634** | 0.997 | 0.8863 | 0.6418 | 0.7584 | 0.999 | 0.9928 |
| | D4GDI pathway | 422 | 0.7626 | 0.3727 | **0.0634** | 0.997 | 0.9867 | 0.6418 | 0.4377 | 0.999 | 0.995 |

**Table 4.3. Significant KEGG and Reactome pathways of *PHARAOH-multi* and MARV, and their adjusted *p*-values of multivariate and univariate analyses.** Bold numbers are the adjusted *p*-values below the significance threshold 0.1. P_M, aSPU and MARV indicate adjusted *p*-values from the joint testing method of multiple phenotypes, and univariate adjusted *p*-values indicate the adjusted *p*-values of ***PHARAOH*** analysis for each phenotype.

| DB | Pathway | # variants | **Multivariate adjusted *p*-value** | | | **Univariate adjusted *p*-value (*PHARAOH*)** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **P_M** | **aSPU** | **MARV** | **tg** | **sbp** | **dbp** | **fastglu** | **hdl** | **waistc** |
| **KEGG** | | | | | | | | | | | |
| | Peroxisome | 421 | **0.0396** | 0 | 0.9826 | 0.6886 | 0.7 | 0.9138 | 0.9899 | 0.9942 | 1 |
| | Glutathione metabolism | 187 | **0.044** | 0.0076 | 0.9826 | 0.999 | **0.0939** | 0.9138 | 0.993 | 0.9942 | 1 |
| **Reactome** | | | | | | | | | | | |
| | Glutathione conjugation | 99 | **0.0979** | 0.0567 | 0.9979 | 0.9813 | 0.3859 | 0.9254 | 0.999 | 0.9793 | 0.979 |
| | Phase II conjugation | 270 | **0.0571** | 0 | 0.9979 | 0.9813 | 0.3859 | 0.9254 | 0.999 | 0.9793 | 0.99 |

### 4.4.1. Real data discovery from whole-exome sequencing dataset

As a discovery study, we conducted an analysis using a large-scale sequencing dataset. Many studies suggest that the major underlying risk factors for metabolic disorders include high density lipoprotein (HDL), blood pressure (SBP, DBP), waist circumference (WAISTC), fasting glucose (FAST_GLU), and triglycerides (TG). In this regard, we conducted a multivariate analysis of metabolism-related traits, using a large-scale sequencing dataset, obtained from the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium, comparing our proposed (***PHARAOH-multi***) and other common methods. In detail, we analyzed a dataset consisting of 1,086 samples selected from the Korean Association REsource (KARE) study (Cho, et al., 2009). The original dataset consists of next generation sequencing of the individuals' genomes, using the Illumina HiSeq2000 platform (Illumina, Inc., San Diego, CA). For pathway-gene mapping, we retrieved pathway information from MSigDB (Liberzon, et al., 2011), and mapped the genes to 217, 186 and 674 pathways extracted from the Biocarta, KEGG (Kanehisa, et al., 2012) and Reactome (Fabregat, et al., 2016), respectively.

After removal of samples with any missing observations of the aforementioned six phenotypes, we included 1,085 samples for analysis. The quality controls with genotype call rates were < 95%, or for Hardy-Weinberg Equilibrium (HWE) test $P < 10^{-5}$, the minor allele frequency was < 5% and the minor allele count was > 2, resulted in 198,761 variants. The final dataset was then mapped to genes, using the human genome-19 (hg19) reference genome coordinates, with 10Kbp flanking regions. The gene range of hg19 reference, was extracted from RefSeq track of UCSC Table Browser, as of October 2014. Finally, the gene-level collapsed variable was generated using Workbench for Integrated Superfast Association study with Related Data (WISARD), with beta-transformation weighting, as suggested in (Wu, et al., 2011), with the number of genes being 4,388.

Next, we compared our multivariate and univariate analysis results, using ***PHARAOH-multi*** and ***PHARAOH***. As shown in **Figure (4.11)** and **(4.12)**, Q-Q plots of the results showed no substantial inflation or deflation pattern for either the multivariate or univariate results. However, with regard to pathway discovery, the results did show significant differences.

These comparisons clearly support the one advantage of multivariate analysis that we discussed above: elevation of statistical power. As with multivariate analysis, we calculated adjusted *p*-values for each univariate result. Interestingly, no univariate analysis identified significant pathways, except for SBP with KEGG, which identified three pathways (drug metabolism cytochrome P450, glutathione metabolism and progesterone-mediated oocyte maturation). As shown in **Table (4.2)**, only one pathway, glutathione metabolism, was identified in the univariate analysis, and the adjusted *p*-values of univariate analyses for pathways identified by multivariate analysis, were not significant.

Second, we compared the result of multivariate analyses, using ***PHARAOH-multi***, aSPU and MARV. As shown in **Figure (4.12)**, ***PHARAOH-multi*** exhibited generally acceptable *p*-value trends, despite the result from KEGG being modestly deflated, due to the optimization of lambda. The Q-Q plots of MARV look similar to ***PHARAOH-multi*** (**Figure (4.14)**). In contrast, as shown in **Figure (4.13)**, aSPU showed unacceptably inflated patterns of Q-Q plots, regardless of the pathway databases, which were not used in the simulation study. This could possibly be due to substantial overlap of existing pathway databases.

As shown in **Table (4.2)** and **(4.3)**, the multivariate analysis successfully identified eight pathways from three pathway databases, with Benjamini-Hochberg adjusted *p*-value < 0.1. Interestingly, ***PHARAOH-multi*** identified glutathione-related pathways in both KEGG and Reactome pathway databases, which supports the result of PHRAOH-multi. As shown in **Figure (4.13)**, the quantile-quantile plots of aSPU for the real dataset are highly inflated (i.e., their *p*-values are very small). As a result, 57.7% (Reactome), 29.5% (Biocarta) and 71.5% (KEGG) of the tested pathways by aSPU were statistically significant (adjusted *p*-value < 0.1). Unfortunately, these pathways are highly false positives. In this respect, we included the results of significant pathways identified by either ***PHARAOH-multi*** or MARV.

The identified pathways suggested evident relationships with metabolic syndrome. Since the peroxisome pathway elucidates peroxisome biogenesis, which contributes to fatty acid oxidation and biosynthesis of ether lipids, many studies have discussed interrelationship between peroxisomes and metabolic processes (Azhar, 2010; Hall, et al., 2010). Likewise, identification of the GABA pathway can also be explained by the relationship between GABA and peroxidation, and putative relationship of obesity (Deng, et al., 2010; Ma, et al., 2000). Moreover, identification of glutathione metabolism, and its conjugation, explain that ***PHARAOH-multi*** successfully captured a key process of metabolic disorders (Wu, et al., 2004). Finally, another report suggested a putative role of adhesion molecules in metabolic diseases, as explained by "cell2cell" pathway (Wagner and Jilma, 1997). For the two pathways identified by MARV, we found that Caspase pathway has been known to be related to metabolic stress or perturbation (McIlwain, et al., 2015), but no evidence for D4GDI pathway was found.

### 4.4.2. Replication study using independent exome chip dataset

For replication of the identified pathways from the discover study, an independent cohort from Koreans, the Health Examinee shared control study (HEXA), was used. HEXA is a part of the KoGES population based cohort, initiated in 2001 (Kim, et al., 2011). In total, genotypes of 3,445 individuals were acquired using the HumanExome BeadChip v1.1 (Illumina, Inc., San Diego, CA). With same quality control criteria, 24,474 rare variants were used in the analysis.

Here, we conducted a replication study using exome chip dataset from an independent cohort, using the identified pathways in the discovery study. Despite the insufficiency of detected variants in the exome chip dataset, as a result, we successfully replicated two pathways with $p$-value < 0.1, the peroxisome pathway in KEGG ($p$=0.059) and cell2cell pathway in Biocarta ($p$=0.093). As shown in the literature search, the two pathways we replicated have strong relationships with metabolic disorders.

## 4.5. Discussion

In this study, we proposed a novel statistical approach for multivariate pathway-based analysis of rare variants, from large-scale sequencing datasets. Analyses of multiple phenotypes have been successful in analyzing various complex diseases, including type-2 diabetes (T2D) or hypertension. In general, curated guidelines suggest diagnosing T2D according to traits observed in the individual. Consequently, incorporating multiple correlated traits, to be investigated for association with specific diseases, via multivariate analysis, elevates the statistical power. In this respect, our simulation study reflects the relationship between diseases and their related traits. Throughout the simulation study, ***PHARAOH-multi*** outperformed existing multivariate methods. In addition, our proposed method successfully demonstrated several advantages of multivariate analysis, including significantly improving the detection power of causal pathways, as compared to univariate analysis, while also retaining detection power for the individual phenotype. Moreover, we successfully demonstrated that the proposed method is capable of identifying plausible pathways in the real dataset, by identifying eight pathways in the discovery study, and replicating two pathways in the replication study.

For analysis times of both simulation and real datasets, MARV was the fastest among all the methods, while ***PHARAOH-multi*** ran slightly faster than aSPU. For example, in the analysis of simulation dataset of 100 genes with 1,000 samples, the running times of MARV, ***PHARAOH-multi***, and SPU were 13, 67 and 235 seconds, respectively. The trends of execution time were consistent regardless of simulation parameters or datasets. However, ***PHARAOH-multi*** can be further accelerated with multithreading which is not supported by MARV and aSPU. With multithreading of 8 threads, the analysis time of ***PHARAOH-multi*** was reduced to 12 seconds.

Compared to univariate approaches, which analyze each phenotype individually, our real data analysis successfully demonstrated that the multivariate approach could identify pathways commonly associated with specific traits. It is important to construct a systematic analysis that considers the correlation between complex diseases and their underlying biological traits. In addition, our results from two well-established pathway databases were strongly supported by many existing publications, thus demonstrating the advantage of our proposed approach.

Compared to existing multivariate analysis methods, ***PHARAOH-multi*** features several advantages. Firstly, by constructing a hierarchical structure of genes-pathways-phenotypes, four types of associations (gene-single phenotype, gene-multiple phenotypes, pathway-single phenotype, and pathway-multiple phenotypes) can be estimated simultaneously. Moreover, the proposed method can address the pleiotrophy by testing the joint hypothesis for multiple phenotypes.

Compared to our proposed method, existing methods of multivariate analysis were limited to gene-level analysis, and hence, the combinatorial effect of multiple genes, via biological pathways, was impossible to estimate. In addition, the proposed method considers the correlation between genes, pathways, and phenotypes, by imposing penalty parameters on the estimation procedure.

Secondly, ***PHARAOH-multi*** provides multiple options for correcting for the multiple testing issue. Although Bonferroni correction is simple, and powerfully controls type 1 error, it is a well-known fact that the Bonferroni correction often results in controls that are too stringent, when the tests are correlated. Under such conditions, application of the Westfall-Young permutation procedure can be an appropriate alternative, since its asymptotic optimality under dependence is known (Meinshausen, et al., 2011). In this respect, the proposed method has the advantage of identifying causal pathways, by considering correlation among pathways.

At this point, there are a number of subjects we can consider for future research. Our current analysis is limited only to Korean population. In our future study, we apply our method to the whole data of 13,000 WES dataset of

T2D-GENES consortium (Fuchsberger, et al., 2016) which contains our KARE samples. It would be a challenging work to identify novel pathways across multiple populations. For the methodological aspect, our approach uses gene-level collapsing of multiple rare variants. Although the collapsing method has the advantage that the analysis of very rare variants is possible, it cancels out the effects of variants with opposite direction (e.g, gene upregulation vs. downregulation). Despite such limitations, our method showed great potential in identifying causal genetic structure in the real data analysis. However, further research, on a more sophisticated approach that can consider the effect direction of variants, is needed. Moreover, we plan to improve our proposed multivariate analysis by applying Generalized Estimating Equations (GEE) or Linear Mixed Model (LMM). Our method can be extended to prediction models, rather than association tests, using other types of penalization, such as LASSO or SCAD (Fan and Li, 2001). However, LASSO imposes a potential problem that is selecting one of highly correlated pathways or genes, which leads to the elimination of true causal variables. An application of group LASSO (Yuan and Lin, 2006) might be a good choice in order to overcome this problem.

Lastly, our method can also be extended to pathway interaction analysis that has been commonly performed in gene expression data analysis (Liu, et al., 2012). We firmly believe that the proposed method will assist researchers in understanding the genetic structures that underlie many complex diseases.

# Chapter 5

## Summary & Conclusions

In this thesis, we addressed the efforts on the genetic association studies for the last two decades. In its turn, we found that the application of pathway information has several advantages, such as reduction of computational complexity, aid of the biological interpretation, and boost of the statistical power. Despite the advantages of the pathway-based approach, it has been suffered by many challenges (e.g., substantial correlation among pathways, absence of biological hierarchy).

Here we focused on developing a novel statistical method by extending an existing methodology with the flexible statistical framework. With the first study and an additional study to make the proposed method applicable to multivariate analysis, we successfully demonstrated that the proposed method is capable to analyze the large-scale sequencing data in computationally efficient manner and outperforms the existing methods in terms of the statistical power and the reproducibility. In this thesis, we suggested the ways to identify statistically associated genes or pathways from the large-scale sequencing dataset.

The major contributions of this thesis are as the follows:

1) We proposed a method that models the underlying biological process from the large-scale sequencing dataset. With an integration of the GLM, the proposed method can handle a variety of phenotypes under the concrete statistical framework. The proposed integrated model of all pathways and all genes considers the complexity of the underlying biological relationship. By imposing the ridge penalty on both genes and pathways, this so-called doubly ridge model can efficiently dissolve the correlation among the genes and the pathways. As shown in Chapter 3, from the analyses of the large-scale WES and exome chip datasets, we successfully demonstrated the two main advantages of the proposed method: interpretability and reproducibility.

2) We further extended the proposed model by considering multiple phenotypes at once, to substantiate the flexibility of ***PHARAOH*** and make the methodology more practical to use in the real data analysis. As shown in Chapter 4, our method for the multivariate extension was very straightforward. Despite its simplicity, the simulation study successfully showed that the ***PHARAOH-multi*** outperforms the existing multivariate pathway-based methods, and the real data analysis successfully replicated multiple pathways.

Despite the history of twenty years of large-scale genetic association study, the current knowledge to understanding the complex biological behavior remains many shades to unveil. For the tremendous expectation of the forthcoming "multi-omics big data" era, we want to conclude with some suggestions for further extending the proposed approach for the multi-omics dataset.

First, the proposed method can gain more advantages by considering the "topology" of the pathways. One successful method in cancer genomics is HotNet (Vandin, et al., 2011), an algorithm to discover mutated pathways by investigating patterns of somatic mutations. The HotNet algorithm seeks the subnetworks whose genes have more mutations than expected by chance. Since such approach is suffered by extensive searching space and large number of null hypotheses, the algorithm solved the problems by shrinking the searching space with their novel algorithm called "influence graph" and by reducing the number of null hypotheses with two-stage multiple hypothesis test strategy (Vandin, et al., 2011). By considering the topology of pathways, the HotNet algorithm become the most promising algorithm in cancer genomics, with many successful discoveries, such as ovarian cancer or prostate cancer (Cancer Genome Atlas Research, 2011; Grasso, et al., 2012). Since the proposed method can reflect the topology of pathways by extending the "inner" model, we expect that such extension would be useful to an analysis of cancer datasets.

Second, an extension to consider "exclusivity" between samples is possible. There are two well-known exclusivities: population-specific variants and Mutual Exclusivity Modules in cancer (MEMo). A large-scale sequencing project usually consists of multiple ethnicities, which leads to large number of population-specific variants, and this situation becomes severe as the MAF becomes rarer (van Rooij, et al., 2017). Moreover, the similar pattern is observed in the cancer samples and is called MEMo. A study on the dataset of The Cancer Genome Atlas (TCGA) showed that the MEMo is actually observed in many cancer types and is resulted from the different pathway behaviors (Ciriello, et al., 2012). In this respect, a novel strategy to consider such exclusive pattern across the samples is demanded to improve the proposed methods.

Finally, since our method emulates the natural hierarchy of the biological behavior, it would be a good next step if the method is extended to consider an integration of multi-omics dataset. Moreover, a concrete investigation on the assessment of statistical significance of the proposed model might lead to a substantial improvement.

# Bibliography

Ahituv, N., *et al.* Medical sequencing at the extremes of human body mass. *American Journal of Human Genetics* 2007;80(4):779-791.

Alexa, A., Rahnenfuhrer, J. and Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 2006;22(13):1600-1607.

Alexaki, A., *et al.* Autophagy regulates sphingolipid levels in the liver. *Journal of lipid research* 2014.

Almasy, L., *et al.* Genetic Analysis Workshop 17 mini-exome simulation. *BMC Proc* 2011;5 Suppl 9:S2.

Altshuler, D., Daly, M.J. and Lander, E.S. Genetic mapping in human disease. *Science* 2008;322(5903):881-888.

Alves, G. and Yu, Y.K. Accuracy evaluation of the unified P-value from combining correlated P-values. *PloS one* 2014;9(3):e91225.

American Diabetes, A. Diagnosis and classification of diabetes mellitus. *Diabetes Care* 2014;37 Suppl 1:S81-90.

Askland, K., Read, C. and Moore, J. Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Human genetics* 2009;125(1):63-79.

Azhar, S. Peroxisome proliferator-activated receptors, metabolic syndrome and cardiovascular disease. *Future Cardiol* 2010;6(5):657-691.

Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995;57(1):289-300.

Brown, M.B. 400: A Method for Combining Non-Independent, One-Sided Tests of Significance. *Biometrics* 1975;31(4):987-992.

Brunham, L.R., Singaraja, R.R. and Hayden, M.R. Variations on a gene: rare and common variants in ABCA1 and their impact on HDL cholesterol levels and atherosclerosis. *Annual review of nutrition* 2006;26:105-129.

Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011;474(7353):609-615.

Cardier, J.E. and Barbera-Guillem, E. Extramedullary hematopoiesis in the adult mouse liver is associated with specific hepatic sinusoidal endothelial cells. *Hepatology* 1997;26(1):165-175.

Cho, Y.S.*, et al.* A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* 2009;41(5):527-534.

Ciriello, G.*, et al.* Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res* 2012;22(2):398-406.

Cohen, J.C.*, et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 2004;305(5685):869-872.

Costanzo, M.*, et al.* The genetic landscape of a cell. *Science* 2010;327(5964):425-431.

De, G.*, et al.* Rare variant analysis for family-based design. *PloS one* 2013;8(1):e48495.

de Leeuw, C.A.*, et al.* The statistical properties of gene-set analysis. *Nature reviews. Genetics* 2016;17(6):353-364.

de Leeuw, J., Young, F.W. and Takane, Y. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika* 1976;41(4):471-503.

Deng, Y.*, et al.* New perspective of GABA as an inhibitor of formation of advanced lipoxidation end-products: it's interaction with malondiadehyde. *J Biomed Nanotechnol* 2010;6(4):318-324.

Desarbo, W.S.*, et al.* Constrained Stochastic Extended Redundancy Analysis. *Psychometrika* 2013.

Duncan, L.E. and Keller, M.C. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *Am J Psychiatry* 2011;168(10):1041-1049.

Fabregat, A.*, et al.* The Reactome pathway Knowledgebase. *Nucleic acids research* 2016;44(D1):D481-487.

Fan, J. and Li, R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association* 2001;96(456):1348-1360.

Fuchsberger, C.*, et al.* The genetic architecture of type 2 diabetes. *Nature* 2016;536(7614):41-47.

Grasso, C.S.*, et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* 2012;487(7406):239-243.

Green, P.J. Iteratively Reweighted Least-Squares for Maximum-Likelihood Estimation, and Some Robust and Resistant Alternatives. *J Roy Stat Soc B Met* 1984;46(2):149-192.

Hall, D.*, et al.* Peroxisomal and microsomal lipid pathways associated with resistance to hepatic steatosis and reduced pro-inflammatory state. *J Biol Chem* 2010;285(40):31011-31023.

He, Z.*, et al.* Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *Am J Hum Genet* 2014;94(1):33-46.

Hirschhorn, J.N. Genomewide association studies--illuminating biologic pathways. *The New England journal of medicine* 2009;360(17):1699-1701.

Hoerl, A.E. and Kennard, R.W. Ridge Regression - Biased Estimation for Nonorthogonal Problems. *Technometrics* 1970;12(1):55-&.

Hu, H.*, et al.* VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet Epidemiol* 2013;37(6):622-634.

Huang, X.J.*, et al.* Aspartate aminotransferase (AST/GOT) and alanine aminotransferase (ALT/GPT) detection techniques. *Sensors-Basel* 2006;6(7):756-782.

Hwang, H. Regularized Generalized Structured Component Analysis. *Psychometrika* 2009;74(3):517-530.

Hwang, H.*, et al.* Generalized Functional Extended Redundancy Analysis. *Psychometrika* 2013.

Hwang, H. and Takane, Y. Generalized structured component analysis. *Psychometrika* 2004;69(1):81-99.

International Multiple Sclerosis Genetics, C. Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls. *Am J Hum Genet* 2013;92(6):854-865.

Jiang, Z. and Gentleman, R. Extensions to gene set enrichment. *Bioinformatics* 2007;23(3):306-313.

Kaakinen, M.*, et al.* A rare-variant test for high-dimensional data. *Eur J Hum Genet* 2017.

Kanehisa, M.*, et al.* The KEGG resource for deciphering the genome. *Nucleic acids research* 2004;32(Database issue):D277-280.

Kanehisa, M.*, et al.* KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* 2012;40(Database issue):D109-114.

Khatri, P., Sirota, M. and Butte, A.J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology* 2012;8(2):e1002375.

Kim, K.*, et al.* Urine metabolomic analysis identifies potential biomarkers and pathogenic pathways in kidney cancer. *Omics : a journal of integrative biology* 2011;15(5):293-303.

Kim, Y.J.*, et al.* Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. *Nat Genet* 2011;43(10):990-995.

Kost, J.T. and McDermott, M.P. Combining dependent P-values. *Stat Probabil Lett* 2002;60(2):183-190.

Kwak, I.Y. and Pan, W. Adaptive gene- and pathway-trait association testing with GWAS summary statistics. *Bioinformatics* 2016;32(8):1178-1184.

Lamparter, D.*, et al.* Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput Biol* 2016;12(1):e1004714.

Le cessie, S. and van houwelingen, J.C. Ridge Estimators in Logistic-Regression. *Applied Statistics-Journal of the Royal Statistical Society Series C* 1992;41(1):191-201.

Lee, A.H. and Silvapulle, M.J. Ridge Estimation in Logistic-Regression. *Communications in Statistics-Simulation and Computation* 1988;17(4):1231-1257.

Lee, S.*, et al.* Pathway-based approach using hierarchical components of collapsed rare variants. *Bioinformatics* 2016;32(17):i586-i594.

Lee, S., Wu, M.C. and Lin, X.H. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012;13(4):762-775.

Lesnick, T.G.*, et al.* A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. *PLoS Genet* 2007;3(6):e98.

Li, B., Wang, G. and Leal, S.M. SimRare: a program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits. *Bioinformatics* 2012;28(20):2703-2704.

Li, B.S. and Leal, S.M. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics* 2008;83(3):311-321.

Li, S., Brown, M.S. and Goldstein, J.L. Bifurcation of insulin signaling pathway in rat liver: mTORC1 required for stimulation of lipogenesis, but not inhibition of gluconeogenesis. *Proceedings of the National Academy of Sciences of the United States of America* 2010;107(8):3441-3446.

Liberzon, A.*, et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;27(12):1739-1740.

Lindskov, J. The Quantitative Liver-Function as Measured by the Galactose Elimination Capacity .1. Diagnostic-Value and Relations to Clinical, Biochemical, and Histological-Findings in Patients with Steatosis and Patients with Cirrhosis. *Acta Med Scand* 1982;212(5):295-302.

Liu, D.J. and Leal, S.M. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet* 2010;6(10):e1001156.

Liu, K.Q., *et al.* Identifying dysregulated pathways in cancers from pathway interaction networks. *BMC Bioinformatics* 2012;13:126.

Ma, Y.H., *et al.* Transgenic mice overexpressing gamma-aminobutyric acid transporter subtype I develop obesity. *Cell Res* 2000;10(4):303-310.

Maher, B. Personal genomes: The case of the missing heritability. *Nature* 2008;456(7218):18-21.

Manolio, T.A., Brooks, L.D. and Collins, F.S. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008;118(5):1590-1605.

Manolio, T.A., *et al.* Finding the missing heritability of complex diseases. *Nature* 2009;461(7265):747-753.

McCarthy, M.I., *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics* 2008;9(5):356-369.

McCullagh, P. and Nelder, J.A. Generalized linear models. London ; New York: Chapman and Hall; 1989.

McIlwain, D.R., Berger, T. and Mak, T.W. Caspase functions in cell death and disease. *Cold Spring Harb Perspect Biol* 2015;7(4).

Meinshausen, N., Maathuis, M.H. and Bühlmann, P. Asymptotic optimality of the Westfall–Young permutation procedure for multiple testing under dependence. *The Annals of Statistics* 2011;39(6):3369-3391.

Nagao, K., *et al.* Dietary conjugated linoleic acid alleviates nonalcoholic fatty liver disease in Zucker (fa/fa) rats. *J Nutr* 2005;135(1):9-13.

Neale, B.M., *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet* 2011;7(3):e1001322.

Nelder, J.A. and Wedderburn, R.W.M. Generalized Linear Models. *J R Stat Soc Ser a-G* 1972;135(3):370-&.

Noto, A.*, et al.* Conjugated linoleic acid reduces hepatic steatosis, improves liver function, and favorably modifies lipid metabolism in obese insulin-resistant rats. *Lipids* 2006;41(2):179-188.

O'Dushlaine, C.*, et al.* The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics* 2009;25(20):2762-2763.

O'Reilly, P.F.*, et al.* MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PloS one* 2012;7(5):e34861.

Pan, W.*, et al.* A powerful and adaptive association test for rare variants. *Genetics* 2014;197(4):1081-1095.

Pan, W., Kwak, I.Y. and Wei, P. A Powerful Pathway-Based Adaptive Test for Genetic Association with Common or Rare Variants. *Am J Hum Genet* 2015;97(1):86-98.

Pralhada Rao, R.*, et al.* Sphingolipid metabolic pathway: an overview of major roles played in human diseases. *Journal of lipids* 2013;2013:178910.

Price, A.L.*, et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010;86(6):832-838.

Qian, D.C.*, et al.* A novel pathway-based approach improves lung cancer risk prediction using germline genetic variations. *Cancer Epidemiol Biomarkers Prev* 2016.

Remesy, C., Fafournoux, P. and Demigne, C. Control of hepatic utilization of serine, glycine and threonine in fed and starved rats. *J Nutr* 1983;113(1):28-39.

Risch, N. and Merikangas, K. The future of genetic studies of complex human diseases. *Science* 1996;273(5281):1516-1517.

Shugart, Y.Y.*, et al.* Weighted pedigree-based statistics for testing the association of rare variants. *BMC genomics* 2012;13:667.

Sifrim, A.*, et al.* eXtasy: variant prioritization by genomic data fusion. *Nature methods* 2013;10(11):1083-1084.

Skarman, A.*, et al.* A Bayesian variable selection procedure to rank overlapping gene sets. *BMC Bioinformatics* 2012;13:73.

Sladek, R., *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007;445(7130):881-885.

Slatter, T.L., *et al.* Novel rare mutations and promoter haplotypes in ABCA1 contribute to low-HDL-C levels. *Clinical genetics* 2008;73(2):179-184.

Sookoian, S. and Pirola, C.J. Alanine and aspartate aminotransferase and glutamine-cycling pathway: their roles in pathogenesis of metabolic syndrome. *World journal of gastroenterology : WJG* 2012;18(29):3775-3781.

Sun, J., *et al.* A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects. *Eur J Hum Genet* 2016;24(9):1344-1351.

Takane, Y. and Hwang, H. An extended redundancy analysis and its applications to two practical examples. *Computational Statistics & Data Analysis* 2005;49(3):785-808.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met* 1996;58(1):267-288.

Uh, H.W., Tsonaka, R. and Houwing-Duistermaat, J.J. Does pathway analysis make it easier for common variants to tag rare ones? *BMC Proc* 2011;5 Suppl 9:S90.

van Rooij, J.G.J., *et al.* Population-specific genetic variation in large sequencing data sets: why more data is still better. *European Journal Of Human Genetics* 2017;25:1173.

Vandin, F., Upfal, E. and Raphael, B.J. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* 2011;18(3):507-522.

Visscher, P.M., *et al.* Five years of GWAS discovery. *Am J Hum Genet* 2012;90(1):7-24.

Wagner, O.F. and Jilma, B. Putative role of adhesion molecules in metabolic disorders. *Horm Metab Res* 1997;29(12):627-630.

Walsh, T., *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 2008;320(5875):539-543.

Wang, K., Li, M. and Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* 2007;81(6):1278-1283.

Weng, L.*, et al.* SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics* 2011;12:99.

Westfall, P.H. and Young, S.S. Resampling-based multiple testing : examples and methods for P-value adjustment. New York: Wiley; 1993.

Wu, G.*, et al.* Glutathione metabolism and its implications for health. *J Nutr* 2004;134(3):489-492.

Wu, G. and Zhi, D. Pathway-based approaches for sequencing-based genome-wide association studies. *Genet Epidemiol* 2013;37(5):478-494.

Wu, M.C.*, et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011;89(1):82-93.

Yan, Q.*, et al.* Kernel-machine testing coupled with a rank-truncation method for genetic pathway analysis. *Genet Epidemiol* 2014;38(5):447-456.

Yang, J.*, et al.* GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;88(1):76-82.

Yang, Q. and Wang, Y. Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies. *J Probab Stat* 2012;2012:652569.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2006;68(1):49-67.

Zhao, J.*, et al.* Pathway analysis with next-generation sequencing data. *Eur J Hum Genet* 2014.

Zhou, X. and Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods* 2014;11(4):407-409.

Zhu, Y. and Xiong, M. Family-Based Association Studies for Next-Generation Sequencing. *Am J Hum Genet* 2012;90(6):1028-1045.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *J Roy Stat Soc B* 2005;67:301-320.

Zuk, O., *et al.* The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America* 2012;109(4):1193-1198.

# 초 록

지난 20 년에 걸쳐, 유전체 서열을 판독하기 위한 생물학적 기술의 급격한 발전으로 말미암아 인간의 유전적 구조를 밝혀내기 위한 폭넓은 연구가 진행되어 왔으며, 특히 복잡형질(complex trait)과 연관된 유전적 변이를 밝혀내기 위한 활발한 연구가 이루어졌다. 전장 유전체 연관성 분석(GWAS)은 이러한 연구의 가장 대표적인 예 중 하나로써, 다양한 복잡형질과 관련된 단일 염기서열 변이(SNV)를 밝혀내는 데에 주요한 역할을 수행하였다. 그러나 이러한 수많은 GWAS 의 성공으로 밝혀진 유전변이들은 대부분의 표현형에 대해 잠재적인 유전율(heritability) 중 일구밖에 설명하지 못하는 현상이 나타났으며, 소위 "잃어버린 유전율(missing heritability)"으로 불리는 이러한 현상을 설명하기 위해 희귀 유전변이(rare variant), 유전자-유전자 상호작용(gene-gene interaction), 다중 오믹스(multi-omics) 통합 분석 등과 같은 많은 가설이 제시되었다.

이러한 가설 중 희귀 유전변이를 분석하기 위한 방법들은 각 유전변이 별로 분석하는 형태에서 유전자 단위로 모아 분석하고, 나아가 다수의 표현형을 모아 분석하는 형태로 발전하였다. 이러한 흐름과 함께 공개적으로 사용할 수 있는 생물학적 데이터베이스의 수가 증가함에 따라 패스웨이 정보를 사전 정보로써 활용하여 희귀 유전변이를 분석하는 형태가 최근의 방법론에서 많이 시도되고 있다. 결론적으로 희귀 유전변이를 패스웨이 단위로 모아 개별적으로 분석하는 다양한 방법론이 등장하였다. 비록 이러한 형태의 방법론은 패스웨이간의 상관성(correlation)을 무시함으로써 잘못된 결론을 도출할 수 있으나, 모든 패스웨이를 상관성을

고려하여 한 번에 분석하는 시도는 막대한 계산량에 가로막혀 있다. 이러한 이유로 다수의 표현형을 고려하는 방법론에서는 개별 패스웨이 단위로 분석하는 소수의 방법론만이 존재할 뿐, 모든 패스웨이를 한 번에 분석하는 통합된 통계모형은 아직 공개되지 않았다.

본 연구에서는 패스웨이 정보를 이용하여 대용량 유전체 자료를 분석할 수 있는 새로운 통계적 방법론인 PHARAOH 와 해당 방법론의 다변량 분석에 대한 확장인 PHARAOH−multi 를 제시한다. PHARAOH 방법론은 일반화 구조적 요소 분석(generalized structural component analysis)을 확장시킨 형태이며, 일반화 선형 모형(generalized linear models)을 접목함으로써 하여금 지수족 분포(exponential family distribution)로부터 유도될 수 있는 다양한 형태의 표현형을 분석할 수 있는 방법론이다. PHARAOH 는 통합된 계층적 모형(hierarchical model)을 구축하며, 다수의 패스웨이와 해당 패스웨이에 속한 유전자를 희귀 유전변이의 가중치 합(weighted sum)으로 나타냄으로써 계층적으로 모형화한다. 전체 패스웨이의 분석에는 유전자−패스웨이 및 패스웨이−표현형 사이에 릿지 벌점화(ridge−type penalty) 방법을 적용함으로써 모든 패스웨이 간의 상관성을 고려한 통합적 분석이 가능하다.

PHARAOH−multi 는 PHARAOH 모형을 다변량 자료 분석에 대해 확장하여, 다수의 표현형에 대한 유전변이의 계층적 요소에 기반한 분석을 가능하게 하였다. PHARAOH−multi 는 다수의 표현형과 다수의 패스웨이 간의 연관성을 패스웨이와 이에 속한 유전자들의 계층 구조에 기반하여 하나의 모형으로 발굴해 낼 수 있다.

**주요어:** 전장 유전체 연관성 분석, 시퀀싱 자료 분석, 희귀 유전변이 연관성 검정, 패스웨이 분석, 다변량 분석

**학 번:** 2010-23162