A BAYESIAN HIERARCHICAL FRAMEWORK FOR PATHWAY ANALYSIS IN GENOME-WIDE ASSOCIATION STUDIES

by

Lei Zhang

APPROVED BY SUPERVISORY COMMITTEE:

Dr. Swati Biswas, Chair

Dr. Sam Efromovich

Dr. Min Chen

Dr. Pankaj K. Choudhary

Copyright © 2018

Lei Zhang

All rights reserved

This dissertation is dedicated to my advisor and my family.

A BAYESIAN HIERARCHICAL FRAMEWORK FOR PATHWAY ANALYSIS IN GENOME-WIDE ASSOCIATION STUDIES

by

LEI ZHANG, BS, MS

DISSERTATION
Presented to the Faculty of
The University of Texas at Dallas
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY IN
STATISTICS

THE UNIVERSITY OF TEXAS AT DALLAS
December 2018

ACKNOWLEDGMENTS

I would like to express my sincerest gratitude to my advisor, Dr. Swati Biswas, for being patient with me, keeping me motivated, and sharing her immense knowledge during my PhD study and research. To me, she is not only my advisor offering me continuous support and guidance in research and dissertation writing, but also a role model who is dedicated to her work and meanwhile taking great care of her family.

Besides my advisor, I would like to thank the rest of my supervisory committee: Dr. Sam Efromovich, Dr. Min Chen, and Dr. Pankaj K. Choudhary for serving on my committee and providing their support in completing this work. I have always enjoyed taking your classes. Your profound knowledge of statistics and the enthusiasm for teaching and sharing have opened a door for me to a world full of fun in studying and learning.

Words can not express how grateful I am to my family members for your love and support.

This research was partially supported by allocations of computing times from the Texas Advanced Computing Center at The University of Texas at Austin. The breast and renal cancer datasets were obtained from NIH's database of Genotypes and Phenotypes (dbGaP).

July 2018

A BAYESIAN HIERARCHICAL FRAMEWORK FOR PATHWAY ANALYSIS IN GENOME-WIDE ASSOCIATION STUDIES

Lei Zhang, PhD
The University of Texas at Dallas, 2018

Supervising Professor: Dr. Swati Biswas, Chair

The genome-wide association studies (GWAS) aim to identify genetic variants, typically single nucleotide polymorphisms (SNPs), associated with a disease/trait. A commonly used analytic strategy in GWAS is to test for association with one single SNP at a time. However, such a strategy lacks power to detect associations that are caused by joint effects of multiple variants, each with a modest effect of its own. Pathway analysis jointly tests the combined effects of all SNPs in all genes belonging to a molecular pathway. This analysis is usually more powerful than single-SNP analyses for detecting joint effects of variants in a pathway. Moreover, due to biological functionality of pathways, a significant result lends itself more easily to interpretation.

In this dissertation, we develop a Bayesian hierarchical model that fully models the natural three-level hierarchy inherent in pathway structure, namely SNP—gene—pathway, unlike most other methods that use ad hoc ways of combining such information. We model the effects at each level conditional on the effects of the levels preceding them within the generalized linear model framework. This joint modeling allows detection of not only the associated pathways but also testing for association with genes and SNPs within significant pathways and significant genes in a hierarchical manner, which can be useful for follow-up studies. To deal with the high dimensionality of such a unified model, we regularize the regression coefficients through an appropriate choice of priors. We fit the model using a combination of Iteratively Weighted Least Squares and Expectation-Maximization algorithms to estimate the posterior modes and their standard errors. The inference is carried out in a hierarchical manner from pathways to genes to SNPs. Hierarchical false discovery rate (FDR) is used for multiplicity adjustment of the entire inference procedure. We also explore the utility of effective number of parameters proposed in the Bayesian literature in our context of multiplicity adjustment using the hierarchical FDR.

To study the proposed approach, we conduct simulations with samples generated under realistic linkage disequilibrium patterns obtained from the HapMap project. We find that our method has higher power than some standard approaches in several settings for identifying pathways that have multiple modest-sized variants. Moreover, it can also pinpoint associated genes once a pathway is implicated, a feature unavailable in other methods. We also find that the use of the effective number of parameters can boost the power to detect associated genes and helps in distinguishing them from the null genes. We apply the proposed method to two GWAS datasets on breast and renal cancer.

TABLE OF CONTENTS

ACKNOWLEDGMENTS v

ABSTRACT vi

LIST OF FIGURES x

LIST OF TABLES xii

CHAPTER 1 INTRODUCTION 1

- 1.1 Background and Motivation 1
- 1.2 Organization of the Dissertation 5
- 1.3 Existing Pathway Association Methods 5
  - 1.3.1 Gene set Ridge regression in ASsociation Studies (GRASS) 5
  - 1.3.2 PLINK set-based test 7
  - 1.3.3 Association LIst Go AnnoTatOR (ALIGATOR) 7
  - 1.3.4 Hierarchical Generalized Linear Mixed Model 8
  - 1.3.5 Adaptive Rank Truncated Product (ARTP) 10
  - 1.3.6 Sequence-Kernel-Association Test — ARTP (SKAT—ARTP) 11
  - 1.3.7 Bayesian Gene-Set Analysis using SNP data (BGSAsnp) 12
- 1.4 Overview of the Proposed Method 14
- 1.5 Preliminaries 15
  - 1.5.1 Generalized Linear Model Fitted with IWLS Algorithm 15
  - 1.5.2 Expectation-Maximization (EM) Algorithm 17
  - 1.5.3 Bayesian Hierarchical Priors as Additional Data Points 18
  - 1.5.4 Hierarchical False Discovery Rate (FDR) 20
  - 1.5.5 Effective Number of Parameters 22

CHAPTER 2 METHODS 23

- 2.1 Hierarchical Prior Structure 23
- 2.2 Posterior Mode Estimation 27
- 2.3 EM Algorithm for Estimating $\boldsymbol{\phi}$ 30
- 2.4 IWLS-EM Algorithm for Model Fitting 35
- 2.5 Inference and Hierarchical FDR 36

2.6 Effective Number of Parameters 36

2.6.1 Illustration of Calculation of V_{ψ}^{-1} 37

2.7 Software 45

CHAPTER 3 SIMULATION STUDY 46

3.1 Setting 1 47

3.2 Setting 2 47

3.3 Setting 3 51

3.4 Setting 4 52

3.5 Setting 5 54

CHAPTER 4 REAL DATA APPLICATIONS 62

4.1 Application to Breast Cancer Data 62

4.2 Application to Renal Cancer Data 65

CHAPTER 5 DISCUSSION AND FUTURE WORK 74

5.1 Discussion 74

5.1.1 Sensitivity Analysis of Hyper-Parameters 75

5.2 Future Work 76

5.2.1 Overlapping SNPs/Genes in Multiple Genes/Pathways 76

5.2.2 Extension to Quantitative Trait 79

5.2.3 Extension to Multiple Dependent Phenotypes 79

REFERENCES 82

BIOGRAPHICAL SKETCH 89

CURRICULUM VITAE

LIST OF FIGURES

1.1 Example of SDC1 pathway structure (Menashe et al., 2010). Ellipses represent genes and blue diamonds represent the cellular processes induced by the pathway. 2

1.2 Example of hierarchical testing procedure using a two-pathway structure. Nodes in green are found to be significant and thus the nodes nested below them (if any) are tested subsequently. 21

3.1 Setting 1: Structure of the three non-null pathways. There are three additional null pathways of the same structure with all SNPs being null (not shown). The components in green (dark shade) are non-null while the ones in gray (light shade) are null. 48

3.2 Setting 1 Results: Power for detecting the three non-null pathways. 49

3.3 Setting 2: Structure of the non-null pathway. There is also a null pathway of the same structure with all SNPs being null (not shown). The components in green are non-null. 50

3.4 Setting 2, Sub-Setting 1 Results: Power for detecting the non-null pathway. . . . 51

3.5 Setting 2, Sub-Setting 2 Results: Power for detecting the non-null pathway. . . . 52

3.6 Setting 2, Sub-Setting 3 Results: Power for detecting the non-null pathway. . . . 53

3.7 Setting 2, Sub-Setting 1 Results: The number of times each gene is detected once the pathway is detected (a) without and (b) with using the effective number of parameters. The numerator in the top row of each table is the number of times the pathway is found to be significant (out of 1000) and the numbers in the last row are the numbers of times the genes are found to be significant (out of the number of times the pathway is found to be significant). The hierarchical FDR of the entire inference procedure is about 0.1. 54

3.8 Setting 2, Sub-Setting 2 Results: The number of times each gene is detected once the pathway is detected (a) without and (b) with using the effective number of parameters. The numerator in the top row of each table is the number of times the pathway is found to be significant (out of 1000) and the numbers in the last row are the numbers of times the genes are found to be significant (out of the number of times the pathway is found to be significant). The hierarchical FDR of the entire inference procedure is about 0.1. 55

3.9 Setting 2, Sub-Setting 3 Results: The number of times each gene is detected once the pathway is detected (a) without and (b) with using the effective number of parameters. The numerator in the top row of each table is the number of times the pathway is found to be significant (out of 1000) and the numbers in the last row are the numbers of times the genes are found to be significant (out of the number of times the pathway is found to be significant). The hierarchical FDR of the entire inference procedure is about 0.1. 56

3.10 Setting 3: Structure of the three non-null pathways. There are also three null pathways of the same structure with all SNPs being null (not shown). The components in green are non-null. 57

3.11 Setting 3 Results: Power for detecting the three non-null pathways. 58

3.12 Setting 4: Structure of the three non-null pathways. There are also three null pathways of the same structure with all SNPs being null (not shown). The components in green are non-null. 59

3.13 Setting 4 Results: Power for detecting the three non-null pathways. 60

3.14 Setting 5 Results: Power for detecting the non-null pathway. 61

5.1 ROC curves with varying s_ξ values. 78

LIST OF TABLES

4.1 Summary on gene counts, SNP counts, and subjects across 298 KEGG pathways for the Breast Cancer data. 63

4.2 Pathways with p-values not exceeding 0.05 using BHPathway applied on the Breast Cancer data. 65

4.3 Pathways with p-values not exceeding 0.05 using GRASS or PLINK applied on the Breast Cancer data. 66

4.4 Pathways with p-values between 0.05 and 0.1 using BHPathway, GRASS, or PLINK applied to the Breast Cancer data. 69

4.5 Summary on gene and SNP counts across 298 KEGG pathways for Renal Cancer data. 70

4.6 Pathways with p-values not exceeding 0.05 using BHPathway applied to the Renal Cancer data. 70

4.7 Genes in pathway hsa03050 found to be significant using the effective number of parameters at hierarchical FDR boundary of 0.2 in the Renal Cancer data. . . . 71

4.8 Pathways with p-values not exceeding 0.05 using GRASS, or PLINK applied to the Renal Cancer data. 72

4.9 Pathways with p-values between 0.05 and 0.1 using BHPathway, GRASS, or PLINK applied to the Renal Cancer data. 73

5.1 Pathway p-value for hsa00760 in the Renal Cancer data under varying s_ξ values. 77

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

In the last decade, the genome-wide association studies (GWAS) have identified many SNPs associated with complex common diseases. However, for most diseases, the implicated SNPs explain only a small fraction of the genetic risk (Manolio et al., 2009). One of the reasons for limited success in unraveling the genetic basis stems from the fact that the commonly used analytic approach in GWAS tests for association with each SNP individually. Such an approach has little power for uncovering associations that are caused by joint effect of multiple variants, each with a modest effect of its own. There is growing evidence that many SNPs and genes within a biologically defined pathway or gene-sets (defined based on criteria not using the data under study) function in tandem to produce a larger effect on a complex trait compared to their own individual effects (Cantor et al., 2010; Fridley and Biernacka, 2011). In such situations, jointly analyzing the combined effects of all variants/genes in a pathway will typically lead to more power (Menashe et al., 2010; Torkamani et al., 2008; Ramanan et al., 2012). Further, due to biological relevance and functionality of pathways, significant results are much more amenable to interpretation and further validation. An example of SDC1 pathway is shown in Figure 1.1.

Recognizing this potential of pathway analysis, several statistical methods have been proposed in the last several years. Some early ones are adaptations of the methods used in gene expression studies. Adaptations are needed because there are many SNPs within a gene

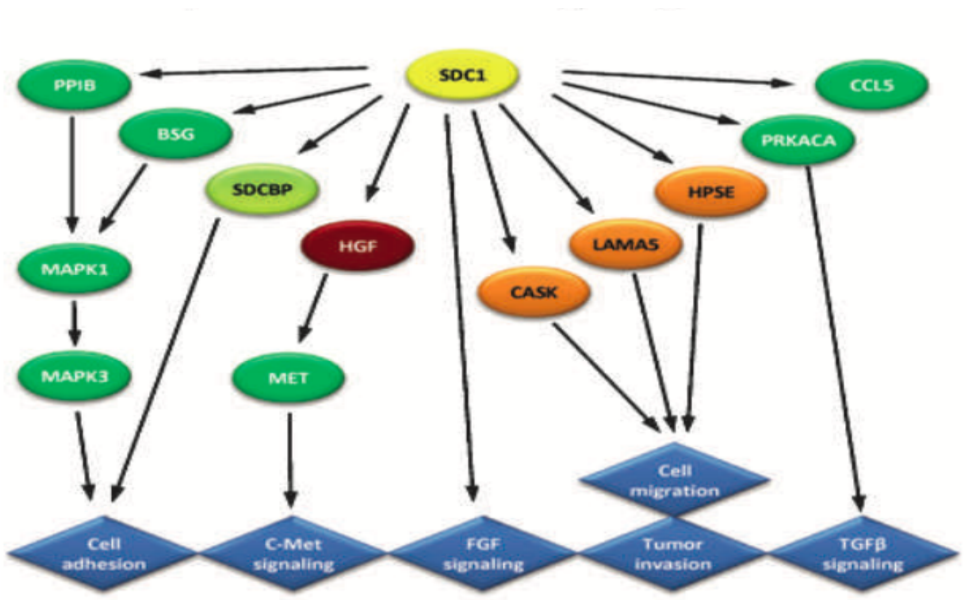


Figure 1.1. Example of SDC1 pathway structure (Menashe et al., 2010). Ellipses represent genes and blue diamonds represent the cellular processes induced by the pathway.

and information over all the SNPs needs to be combined to get a gene-level measure for each gene. The different adaptations combine SNP- and gene-level information in different ways (Wang et al., 2007; Holmans et al., 2009; Purcell et al., 2007; Holden et al., 2008; O'Dushlaine et al., 2009; Zhang et al., 2010; Weng et al., 2011; Yu et al., 2009; Chen et al., 2010). In general, the pathway analysis approaches can be broadly classified based on the type of null hypothesis — *competitive* or *self-contained* — depending on whether the pathway of interest is compared with other pathways in the genome (competitive) or a non-associated, i.e., null pathway (self-contained). One issue with competitive tests is that any significance has to be interpreted with respect to the comparison (reference) sets. Also, the resulting p-values across different pathways are dependent, complicating the multiplicity adjustment issue. Moreover, a non-significant result does not provide direct evidence that the SNP/genes in the pathway are not associated with the trait (Cantor et al., 2010; Fridley and Biernacka, 2011; Schaid et al., 2012). For these reasons, self-contained pathway analysis may be preferable and is also the focus of this dissertation.

Pathway analysis methods can also be classified based on whether they take the input data of SNP p-values or raw genotype data. An obvious advantage of the former is their simplicity and portability across studies, nonetheless, there are some limitations as well (Cantor et al., 2010; Fridley and Biernacka, 2011; Schaid et al., 2012). For example, they cannot appropriately account for dependency among SNPs and genes, and use ad hoc corrections. Further, there is the issue of determining an appropriate cutoff for declaring SNPs to be significant. Besides, the multiplicity adjustment for different genes/pathways of different sizes is not straightforward. Some methods that use p-values for calculating the test statistics, in fact, require raw genotype data to assess significance using permutations (Wang et al., 2007; Purcell et al., 2007). The permutation method is also used by several methods that directly use the raw genotype data for modeling the joint effects of SNPs and/or genes (Schaid et al., 2012; Chen et al., 2010; Shahbaba et al., 2012).

In our view, a formal modeling approach using raw genotype data is a much more cohesive way of synthesizing information from various SNPs and genes in a pathway compared to ad hoc ways of combining these pieces of information to define a pathway level measure. Some methods have been proposed that utilize a formal model such as a linear model or a generalized linear model (GLM) at the SNP or gene level (Chen et al., 2010; Schaid et al., 2012; Shahbaba et al., 2012; Silver et al., 2012). However, to define a pathway level measure, the different pieces of information are again somehow combined in a piecemeal manner. In particular, what is lacking is a formal modeling approach to cohesively unify and synthesize the information across the three levels of hierarchy, namely, SNP–gene–pathway. In this regard, a hierarchical modeling approach can exploit the dependency within and between the different levels, and has the potential to fully utilize all the information, which would lead to greater power (Wang et al., 2010). Although few hierarchical approaches have been proposed, they are either somewhat piecemeal in nature, e.g., involve separate and independent models for different levels of hierarchy, or integrate out gene/SNP effects, implying that they do not fully utilize the joint distribution of all components in a pathway (Wang et al., 2011; Evangelou et al., 2014; Shahbaba et al., 2012).

Further, these approaches typically model a test statistic (such as Cochran-Armitage trend statistic) at the SNP or gene level rather than directly modeling the genotype data. Another approach models the mean of SNP-level test statistics such as Cochran-Armitage trend statistics as a linear mixed effects model with fixed pathway effects and random genes effects (Wang et al., 2011), and tests for significance of the pathway effect. There is a variation of this idea in the Bayesian framework that assumes a prior distribution for the SNP-level statistics and obtains a similar gene-level summary measure, which is then further assumed to follow a prior distribution and a pathway-level measure is obtained (Shahbaba et al., 2012). That is, two separate models are used for SNP- and gene-level measures that are not directly connected through a joint hierarchical model.

Also, most pathway analysis methods test for one pathway at a time, and require multiplicity adjustment correction when multiple pathways are tested. Joint consideration of multiple pathways through a single model and analysis can be potentially more powerful and efficient. Further, many methods are applicable to either a quantitative trait or a binary (case/control) response, and cannot handle both (or more) types of responses in a unified manner. Finally, there is no pathway analysis approach that allows for a formal inference on component genes and SNPs once a pathway is found to be significant.

1.2 Organization of the Dissertation

In the remaining of this chapter, we first describe some commonly used standard pathway analysis methods and then provide an overview of our proposed method. Then we discuss some key statistical concepts and tools to be used in this dissertation.

In Chapter 2, we describe the proposed method in details. In Chapter 3, we carry out simulation studies to study the properties of the proposed methodology and compare it with three commonly used methods for pathway analysis, namely, ALIGATOR, PLINK, and GRASS (Purcell et al., 2007; Holmans et al., 2009; Chen et al., 2010). In Chapter 4, we analyze two GWAS datasets on breast and renal cancers obtained from dbGaP (database of Genotypes and Phenotypes, 2018). The dissertation ends with Chapter 5 providing a discussion and directions for future work.

1.3 Existing Pathway Association Methods

1.3.1 Gene set Ridge regression in ASsociation Studies (GRASS)

GRASS (Chen et al., 2010) is a self-contained pathway analysis method and utilizes raw SNP genotype data. It is based on penalized logistic regression and the covariates are eigenSNPs extracted using principle component analysis as follows. The genotype matrix of SNPs in a gene is first standardized, denoted as Z. The principle components of Z are obtained and referred to as eigenSNPs. Suppose there are a total of m SNPs in a gene. Each eigenSNP that explains at least $1/m$ proportion of variation and together with similar eigenSNPs explain around 95% of the total variance are chosen as covariates for model fitting.

GRASS adds group ridge penalty term to the likelihood function for estimation of β coefficients of eigenSNPs with the object function to be minimized being $\boldsymbol{S}_\lambda(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \lambda \sum_{g=1}^{G} w_g(\|\hat{\boldsymbol{\beta}}_g\|_1)^2$, where G is the total number of genes, w_g is weight for g^{th} gene, l is the log-likelihood, and $\|\hat{\boldsymbol{\beta}}_g\|$ is as defined below. This penalty term is a combination of L2 norm at the gene level and L1 norm at the SNP level. In particular, it is a weighted sum of squares to combine summary statistics at the gene level and sum of absolute values of estimated β coefficients of eigenSNPs within each gene.

Summary statistic for gene g is calculated as $\|\hat{\boldsymbol{\beta}}_g\| = \sqrt{\hat{\beta}_{g1}^2 + \hat{\beta}_{g1}^2 + \cdots + \hat{\beta}_{gk_g}^2}$, where $\hat{\beta}_{g1}, \cdots, \hat{\beta}_{gk_g}$ represent estimated coefficients for eigenSNPs in g^{th} gene. By standardizing $\|\hat{\boldsymbol{\beta}}_g\|$ to $\boldsymbol{\beta}_g = \frac{\|\hat{\boldsymbol{\beta}}_g\| - \hat{\boldsymbol{\mu}}_g}{\hat{\boldsymbol{\sigma}}_g}$ using mean $\hat{\boldsymbol{\mu}}_g$ and standard deviation $\hat{\boldsymbol{\sigma}}_g$ of the null distribution, whose estimation is explained later, pathway level statistic is calculated as $T^{obs} = \sqrt{(\boldsymbol{\beta}_1)^2 + (\boldsymbol{\beta}_2)^2 + \cdots + (\boldsymbol{\beta}_G)^2}$ using standardized gene level statistics $(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_G)$ for the G genes in this pathway.

The null distribution is estimated using a permutation method. In particular, by permuting case/control status once, a $\|\hat{\boldsymbol{\beta}}_g^b\|$ value for gene g and permutation b is calculated and this permutation procedure is repeated B times. Then $\hat{\boldsymbol{\mu}}_g$ and $\hat{\boldsymbol{\sigma}}_g$ are calculated as the mean and standard deviation of $\|\hat{\boldsymbol{\beta}}_g^b\|, b = 1, \cdots, B$ values obtained for gene g. These are used in previously mentioned formula of $\boldsymbol{\beta}_g$ to standardize $\|\hat{\boldsymbol{\beta}}_g\|$ (calculated from observed data) as well as $\|\hat{\boldsymbol{\beta}}_g^b\|$ (calculated from b^{th} permuted sample) for $b = 1, \cdots, B$. At b^{th} permutation, the pathway test statistic is calculated as $T^b = \sqrt{(\hat{\boldsymbol{\beta}}_1^b)^2 + (\hat{\boldsymbol{\beta}}_2^b)^2 + \cdots + (\hat{\boldsymbol{\beta}}_G^b)^2}$ for $b = 1, \cdots, B$. The percentage of times that T^b, $b = 1, \cdots, B$, exceeds T^{obs} serves as empirical p-value for testing statistical significance of the pathway.

1.3.2 PLINK set-based test

PLINK (Purcell et al., 2007) set-based test can be used for any set of SNPs, e.g., for SNPs in a given pathway. It tests for the self-contained null hypothesis and utilizes raw SNP genotype data. It applies two levels of filtering on the group of SNPs under study. The first level of filtering is applied to get an independent set of SNPs whose pairwise r^2 statistics are below a certain threshold (default is 0.5). The second level of filtering retains only those SNPs from the independent set whose p-values of single-SNP analysis (e.g, obtained using χ^2 test statistic) are less than a certain value (default is 0.05). With the final set of selected SNPs, the average value of the single-SNP test statistics is used as the set-based test statistic.

By permuting the phenotypes, the null distribution of the set-based test statistic is estimated. The percentage of times that the set-based test statistics calculated from the permuted samples exceeds the observed set-based test statistic is the empirical p-value for statistical inference.

1.3.3 Association LIst Go AnnoTatOR (ALIGATOR)

ALIGATOR (Holmans et al., 2009) tests for competitive null hypothesis, unlike PLINK or GRASS. ALIGATOR considers a SNP to be significant if its p-value from the chi-square test is less than a certain threshold (default is 0.05) and a gene is considered as significant if it contains at least one significant SNP. Thus, ALIGATOR only requires summary statistics (or p-values) of SNPs for testing association. The original ALIGATOR method was proposed based on Gene Ontology (GO) annotation database although the method can be applied to any database of pathways. GO classifies genes into different categories. Once ALIGATOR identifies all significant genes, the corresponding GO categories are noted. The total number of significant genes found in a particular GO category is the test statistic for this category. A gene is counted only once regardless of how many significant SNPs it contains.

The significance of the test statistic is obtained in the following manner. A replicate sample is created by randomly selecting with replacement the same number of SNPs as in the original sample. As the p-values of all SNPs are available from the original sample, the total number of significant genes in each GO category can be found for the replicated sample. By repeating this process 5000 times, a category-specific p-value is assessed by the number of times among all the replicate samples that this category contains a larger number of significant genes than the observed test statistic for the category.

To perform pathway analysis using database other than GO, a pathway can be viewed as a GO category within which genes are grouped by their biological functionality and statistical analysis can be carried out using the procedure described above.

1.3.4 Hierarchical Generalized Linear Mixed Model

Similar to ALIGATOR, this method by (Wang et al., 2011) only requires summary statistics on SNPs as input data. It models log of the Cochran-Armitage trend test statistic y_{ij}

for SNP j in gene i using a mixed model. In particular, $y_{ij} \sim \chi^2_{\lambda_{ij}}$. The model has a fixed effect β and a random gene effect u_i. Then for each SNP j in gene i, the mixed model is $\log(\lambda_{ij}) = \eta_{ij} = \beta + u_i \quad i = 1, \cdots, g;\ j = 1, \cdots, s_i$ for all g genes. Further, $\boldsymbol{u} = (u_1, \cdots, u_g) \sim N(\mathbf{0}, G)$, where G is gene by gene covariance matrix and β is the mean of pathway effect. Under the null hypothesis of no association, $\beta = 0$ and $\log(\lambda_{ij})$ is 0 as $E(u_i) = 0$. Under H_0, a standardized estimate of β, i.e., $\hat{\beta}/\mathrm{SE}(\hat{\beta})$ follows t distribution with d.f. $= 1$ and its asymptotic p-value indicates the strength of the pathway association.

As an alternative way of calculating pathway significance instead of relying on its asymptotic p-value, the authors suggest converting the t-statistic t_p of each pathway to a standardized z-score s_p in the following way: For p^{th} pathway, its z-score z_p corresponding to the p-value for t-statistic t_p can be obtained as $\Phi^{-1}(F_d(t_p))$, where F_d and Φ are the cumulative density functions of t_p and $N(0,1)$ distributions. Then subtracting both the median of all z_p's (denoted as m) as well the location parameter (δ) from each of the z-score and then dividing by the scale parameter (σ) would give the standardized z-score s_p as $s_p = \frac{z_p - m - \delta}{\sigma}$. The p-value of s_p is calculated as $p_p = 1 - \Phi(s_p)$ and it can then be used as an alternative p-value for testing statistical significance.

The authors also consider the issue of overlapping SNPs, i.e., when a SNP is mapped to multiple genes. A value of 1 is assigned in the design matrix whenever a SNP under study is mapped to a certain gene and 0 otherwise. To be more specific, let us suppose the SNP j is mapped to both gene m and gene n, then the model for SNP j is written as $\log(\lambda_j) = \beta + \mu_m + \mu_n$. Reflecting this in the design matrix, the cells corresponding to (SNP

j, gene m) and (SNP j, gene n) will both be set equal to 1 to indicate membership of SNP j to both genes.

1.3.5 Adaptive Rank Truncated Product (ARTP)

ARTP (Yu et al., 2009) is a gene-based pathway analysis approach that combines gene-level association evidence through an adaptive extension of rank truncated product (RTP) method (Dudbridge and Koeleman, 2003). It tests for competitive null hypothesis and utilizes SNP p-values, similar to ALIGATOR. The p-value for each SNP is obtained through the Cochran-Armitage trend test and then ordered from the smallest to the largest. The RTP method uses the first K (a fixed value) smallest p-values at a given truncation point. However, in the ARTP method, a truncation point-specific K value for multiple candidate truncation points are considered.

More specifically, in RTP, single SNP p-values using Cochran-Armitage trend test are first obtained as $p_1, \cdots, p_L$ for a total of L SNPs and ordered as $p_{(1)}, \cdots, p_{(L)}$. With a given truncation point K, the RTP test statistic is $W(K) = \prod_{i=1}^{K} p_{(i)}$ as the product of the first K smallest single-SNP p-values. Pathway association is tested using p-value of $W(K)$ computed using permutation method and is denoted by $\hat{s}_{(K)}$. ARTP differs from RTP in the way that it proposes J different truncation points $K_1, \cdots, K_J$ and uses $W(K_1), \cdots, W(K_J)$ for all J truncations to formulate a test statistic. Suppose $\hat{s}_{(K_j)}$ is the estimated p-value for $W(K_j), 1 \leq j \leq J$. The pathway test statistic used in ARTP method is $MinP = \min\ \hat{s}_{(K_j)}$ for $1 \leq j \leq J$.

To estimate the null distribution of the pathway test statistic $MinP$, case/control status are permuted in the following way. For b^{th} permuted sample at j^{th} truncation point, $W_j^{(b)} = \prod_{i=1}^{K_j} p_{(i)}^{(b)}$ is obtained and its p-value is defined as $\hat{s}_j^{(b)} = \frac{\sum_{b*=0}^{B} \boldsymbol{I}(W_j^{(b*)} \leq W_j^{(b)})}{B+1}$, i.e., the percentage of times the W_j values in B permuted samples and the original sample (i.e., in $B+1$ samples) do not exceed the current $W_j^{(b)}$. Then pathway test statistic $MinP^{(b)}$ at b^{th} permutation is $MinP^{(b)} = \min\{\hat{s}_1^{(b)} \cdots, \hat{s}_J^{(b)}\}$. The adjusted empirical p-value of the pathway is then defined by $\frac{\sum_{b=0}^{B} \boldsymbol{I}(MinP^{(b)} \leq MinP^{(0)})}{B+1}$, where $MinP^{(0)}$ is the value of $MinP$ calculated using the observed data.

1.3.6 Sequence-Kernel-Association Test — ARTP (SKAT—ARTP)

In the ARTP method mentioned previously, pathway significance is summarized using SNP level p-values. SKAT-ARTP (Yan et al., 2014) is an approach for two-stage pathway analysis that combines SNP information using a popular rare variant test statistic used in SKAT (Wu et al., 2011) to get gene level summary statistics and then uses an extended ARTP statistic to obtain the pathway p-value. It tests self-contained null hypothesis and the model is flexible to deal with both binary and continuous traits.

In the first stage, a weighted GLM is fitted using phenotype y as response and genotypes of SNPs in a gene as input data to obtain gene-level test statistics as follows. SNP j is assigned a random effect γ_j, which follows $N(0, \tau w_j)$. The choice of w_j will be discussed later. Testing the null hypothesis of $\gamma_j = 0$ is the same as testing $H_0 : \tau = 0$ using a variance-component test statistic $Q = (y - \hat{\mu})' GWG' (y - \hat{\mu})$ for each gene. Here $\hat{\mu}$ is the average of estimated response, G is genotype matrix, and W is the diagonal weight matrix for the set of SNPs mapped to the gene to be tested. P-value of the gene-level statistic Q can be calculated using Davies method (Davies, 1980) and the above process can be applied to all genes in a pathway of interest to obtain the gene-level p-values.

Several ways of assigning SNP-specific weights are proposed, e.g., $\sqrt{w_j}$ = Beta(MAF_j; a_1; a_2), where a_1 and a_2 are constants and MAF_j is the minor allele frequency of SNP j, or $\sqrt{w_j}$ = $\frac{1}{0.1+pvalue\ of\ SNPj}$ to assign higher weight to variants with strong signals, or as proposed in this paper: $\sqrt{w_j}$ = 0.5*Beta(MAF_j; 1; 25)+$\frac{1.25}{0.1+pvalue\ of\ SNPj}$.

In the second stage, the ARTP method (Yu et al., 2009) is employed to test pathway significance. Let $p_1^{(0)}, \cdots, p_L^{(0)}$ be p-values for the L genes in a pathway using observed data and $p_1^{(b)}, \cdots, p_L^{(b)}$ be the ones obtained from b^{th} permuted sample for $b = 1, \cdots, B$ by permuting case/control status. Given each truncation point $1 \leq j \leq J$, $V_j^{(b)} = \prod_{i=1}^{K_j} p_{(i)}^{(b)}$ is calculated as the product of first K_j smallest gene p-values from b^{th} permuted sample and $\hat{s}_j^{(b)} = \frac{\sum_{b*=0}^{B} \boldsymbol{I}(V_j^{(b*)} \leq V_j^{(b)})}{B+1}$ is the empirical p-value for $V_j^{(b)}$ by comparing it to V_j's calculated from both the observed data and B permuted samples.

Pathway-level test statistic for b^{th} permuted sample is defined as $MinP^{(b)} = \min\ \hat{s}_j^{(b)}$ for $1 \leq j \leq J$ and the adjusted empirical pvalue of pathway is then defined by $\frac{\sum_{b=0}^{B} \boldsymbol{I}(MinP^{(b)} \leq MinP^{(0)})}{B+1}$.

1.3.7 Bayesian Gene-Set Analysis using SNP data (BGSAsnp)

BGSAsnp (Shahbaba et al., 2012) only requires SNP level test statistics, in particular, Cochran-Armitage trend test statistics. Let $T_{j1}, \cdots, T_{jm_j}$ be the unsquared Cochran-Armitage trend test statistics for the m_j SNPs in gene j. A hierarchical Bayesian structure is assigned as $T_{ji} \sim N(0, \eta_j^2)$ for $i = 1, \cdots, m_j$, $\eta_j^2 \sim$ Inv-$\chi^2(\kappa, \psi^2)$. With $\boldsymbol{T} = (T_{j1}, \cdots, T_{jm_j})$ given,

$\eta_j^2|\boldsymbol{T},\kappa,\psi^2 \sim \text{Inv-}\chi^2(\kappa+m_j, \frac{\kappa\psi^2+\sum_{i=1}^{i=m_j}T_{ji}^2}{\kappa+m_j})$ with mean $E(\eta_j^2|\boldsymbol{T},\kappa,\psi^2) = \frac{\kappa\psi^2+\sum_{i=1}^{m_j}T_{ji}^2}{\kappa+m_j-2}$ for $\kappa + m_j > 2$. Then the summary statistic for gene j is $Z_j = \text{sign}(T_{j(1)})E(\eta_j^2|\boldsymbol{T},\kappa,\psi^2)$, where $T_{j(1)} = \min(T_{j1},\cdots,T_{jm_j})$.

Borrowing the same idea of constructing hierarchical priors to gene level statistic $\boldsymbol{Z} = (Z_1,\cdots,Z_g)$, it is assumed that gene level test statistic Z_{sj} for gene j in pathway s follows $N(0,\tau_s^2)$. Rather than using only a single scaled-inv-χ^2 as prior for τ_s^2, the authors propose a mixture distribution to distinguish between relevant and irrelavant pathways. In particular, a mixture of two scaled-inv-χ^2 distributions, F_0 and F_1, is assigned as prior for τ_s^2. $F_0 = \text{Inv}-\chi^2(\nu,\phi_0^2)$ is assumed to be the distribution of τ_s^2 for the irrelevant group of pathways and $F_1 = \text{Inv}-\chi^2(\nu,\phi_0^2+\phi_1^2)$ is assumed for the relevant group. More specifically,

$$\begin{aligned}
\tau_s^2|\lambda,\phi_0,\phi_1 &\sim (1-\lambda)\text{Inv}-\chi^2(\nu,\phi_0^2)+\lambda\text{Inv}-\chi^2(\nu,\phi_0^2+\phi_1^2),\\
\phi_0^2,\phi_1^2 &\sim \text{Gamma}(a_\phi,b_\phi),\\
\nu &\sim \text{Gamma}(a_\phi,b_\phi),\ \text{and}\\
\lambda &\sim \text{Beta}(a_\lambda,b_\lambda).
\end{aligned}$$

For inference on pathway level significance, a measurement similar to p-value is proposed. It is given by $p_s = E[P(T \geq \nu\phi_0/\tau_s^2|\boldsymbol{Z})]$ with $T \sim \chi_\nu^2$ and $\nu\phi_0/\tau_s^2 \sim \chi_\nu^2$. A smaller value of p_s indicates stronger statistical significance against the null hypothesis based on the following justification. For an irrelevant pathway, $P(T \geq \nu\phi_0/\tau_s^2|\boldsymbol{Z})$ would have a uniform distribution and p_s would be close to 0.5 while for a relevant pathway p_s would be small. This is because for a relevent pathway, τ_s^2 will tend to be large as it is sampled from Inv-$\chi^2(\nu,\phi_0^2+\phi_1^2)$ with a larger scale parameter $\phi_0^2 + \phi_1^2$ compared to τ_s^2 sampled from Inv-$\chi^2(\nu, \phi_0^2)$ for an irrelevant pathway.

1.4 Overview of the Proposed Method

Our goal is to develop a hierarchical model for pathway analysis that addresses the limitations of the existing approaches described above. The proposed model fully utilizes the SNP—gene—pathway hierarchy by modeling the effects at each level conditional on the effects of the preceding level within the generalized linear model (GLM) framework. For such a hierarchical modeling framework to work in practice, two challenges must be overcome. The first is dealing with the high dimensionality of the data. For example, the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (KEGG database, 2018) has 298 pathways that have genes ranging in number from 10 to 300 (a commonly applied filter for choosing pathways from a database). These contain a total of about 6,689 unique genes (with ensemble gene IDs). In a typical GWAS data, these genes may further contain a total of about 124,000 SNPs within ±10 KB upstream and downstream of each gene. To handle this, we work within a Bayesian framework as it naturally allows modeling of a multi-level hierarchy as well as regularization of effects through an appropriate choice of priors. Also, the multiple testing burden can be less severe in a Bayesian hierarchical modeling compared to classical setting (Gelman et al., 2012).

The second challenge is controlling the computational burden. To this end, we forgo the full posterior simulation, which requires computationally intensive methods such as Markov chain Monte Carlo (MCMC) algorithms in favor of estimation of appropriate conditional posterior modes. This involves a combination of Iteratively Weighted Least Squares (IWLS) and Expectation Maximization (EM) algorithms that are computationally efficient and scalable.

The inference based on the proposed unified model is carried out in a hierarchical manner from pathways to genes to SNPs. We use hierarchical false discovery rate (FDR) for multiplicity adjustment of the entire inference procedure. We also explore hierarchical FDR in conjunction with *effective number of parameters* (Spiegelhalter et al., 2002). This idea has not been used in pathway analysis or in genetic association studies in general (to the best of our knowledge) but can potentially increase the power of detecting associations.

1.5 Preliminaries

This section will briefly go over the statistical concepts, algorithms, and models used in this dissertation.

1.5.1 Generalized Linear Model Fitted with IWLS Algorithm

In many regression problems, the distribution of the dependent variable Y is not necessarily normal, but a member of the exponential family of distributions. The generalized linear model (GLM) is a family of models, which does not require normality and constant variance assumptions (McCullagh and Nelder, 1989).

Suppose Y_i follows a probability distribution from the exponential family with $E(Y_i) = \mu_i$ for $i = 1, \cdots, n$. The mean response μ_i links Y_i to the linear form $X_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \cdots +$

$\beta_{p-1}x_{p-1,i}$ through a link function $\eta_i = g(\mu_i) = X_i'\boldsymbol{\beta}$, where g can be a linear or non-linear function. Variance of Y_i must be a function of x_i's through the mean response μ_i only.

For fitting a GLM, an iteratively weighted least squares method can be used. It involves constructing pseudo response z_i and pseudo variance $\sigma_{z_i}^2$ so that the likelihood function $p(y_i|X_i'\boldsymbol{\beta})$ can be approximated by $N(z_i|X_i'\boldsymbol{\beta}, \sigma_{z_i}^2)$ (Gelman et al., 2014), and weighted least squares can be applied in a iterative fashion. Then, in the regression model, the dependent variable is z instead of Y and the weights are functions of the fitted values $\hat{\mu}$. Pseudo response z and pseudo weight W can be calculated in the following manner:

$$z = \eta + (y - \mu)\frac{d\eta}{d\mu}$$

and weights

$$W = V^{-1}\left(\frac{d\mu}{d\eta}\right)^2,$$

where V is variance function. Here z is a linearized form of the link function applied to the data using the Taylor series expansion to the first order, i.e.,

$$\begin{aligned} g(y) &\approx g(\mu) + (y - \mu)g'(\mu) \\ &= \eta + (y - \mu)\frac{d\eta}{d\mu}. \end{aligned}$$

To estimate the $\boldsymbol{\beta}$ coefficients, weighted least squares method is used iteratively using the pseudo data as the response and the pseudo weights as the weights.

- Step 0: Assign initial values for $\boldsymbol{\beta}$.

- Step 1: Use current estimates of $\boldsymbol{\beta}$ to generate z and W.

- Step 2: Regress z on X with the diagonal weight matrix W using weighted least squares to get the current estimates of $\hat{\boldsymbol{\beta}} = (X^T W X)^{-1} X^T W z$ and $\hat{V}_{\boldsymbol{\beta}} = (X^T W X)^{-1}$.

- Repeat Steps 1-2 until a convergence criteria is met, which can be the absolute or relative change in the parameter estimates being less than a certain value.

1.5.2 Expectation-Maximization (EM) Algorithm

EM algorithm is an iterative procedure for finding maximum-likelihood estimators of parameters when the observed data likelihood is difficult to maximize, however, with some "missing data" filled in, the complete data likelihood is easy to maximize. It makes a clear distinction between the observed, incomplete data Y and the unobserved, complete data X consisting of Y and missing data, say Z, i.e., $X = (Y, Z)$ (Lange, 2003). X (and hence Z) should be chosen in such a manner that the complete data likelihood is trivial to maximize.

The algorithm starts with an initial guess of the parameters and iterates between E-step and M-step until convergence. The E-step involves taking expectation of the log-likelihood function of the complete data with respect to the missing data with the missing data replaced by its current estimate. The M-step finds the parameter estimates that maximize the expected log-likelihood obtained in the E-step (Casella and Berger, 2002).

When adopted in Bayesian analysis, let ϕ be the unknown parameter and γ be the missing data. The EM algorithm can be used to find the mode of the marginal posterior distribution $p(\phi|y)$, averaged over the parameter γ. It is helpful in the situation when it is hard to maximize $p(\phi|y)$ directly but easy to work with $p(\gamma|\phi, y)$ and $p(\phi|\gamma, y)$. The steps in the EM algorithm to find estimate of ϕ are as follows (Gelman et al., 2014):

1. Start with an initial guess, ϕ^0.

2. At t^{th} iteration:

In E-step:

Take expectation of the log posterior density function with respect to γ using the conditional posterior distribution of γ as

$$E_{old}(\log p(\gamma, \phi|y)) = \int (\log p(\gamma, \phi|y))p(\gamma|\phi, y)d\gamma,$$

where $\phi^{\text{old}} = \phi^{t-1}$, the estimate of ϕ at $(t-1)^{th}$ iteration.

In M-step:

Maximize $E_{old}(\log p(\gamma, \phi|y)$ obtained in the E-step to find ϕ^t.

3. Iterate between the E- and M-steps to get estimated value of ϕ at convergence.

1.5.3 Bayesian Hierarchical Priors as Additional Data Points

Consider the GLM described in Section 1.5.1. Under the Bayesian set-up, t prior distributions can be assigned to the regression coefficients to provide minimal prior information to constrain the coefficients in a reasonable range (Gelman et al., 2008). The t prior can be represented hierarchically as a mixture of normal distribution and scaled inverse chi-square distribution. More specifically, J independent normal priors are assigned to $\beta = (\beta_1, \cdots, \beta_J)$ as $\beta_j \sim \text{N}(\mu_j, \sigma_j^2)$ and $\sigma_j^2 \sim \text{Inv-}\chi^2(\nu, s^2)$ for $j = 1, \cdots, J$.

With β coefficients following normal distribution given σ^2, the prior information can be added to classical linear regression model as “additional data” and correspondingly an augmented $\boldsymbol{X}^*$ can be formed (Gelman et al., 2008, 2014). With the augmented $\boldsymbol{X}^*$, the parameters are identifiable and thus the resulting β estimate is well defined and has finite variance, even if the original data are high-dimensional and have collinearity or separation that would result in nonidentifiability of the classical maximum likelihood estimate (Gelman et al., 2014).

The vector $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_J)$ is added to pseudo response vector $\boldsymbol{z} = (z_1, \cdots, z_n)$ introduced in Section 1.5.1 for n observations to define an augmented response vector $\boldsymbol{z}_*$ as

$$\boldsymbol{z}^* = \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{\mu} \end{pmatrix}.$$

Similarly, a $J \times J$ identity matrix $\boldsymbol{I_J}$ is appended to the design matrix $\boldsymbol{X}$ to obtain the augmented design matrix $\boldsymbol{X}^*$ as

$$\boldsymbol{X}^*_{(n+J)\times J} = \begin{pmatrix} \boldsymbol{X}_{n\times J} \\ \boldsymbol{I_{J\times J}} \end{pmatrix}.$$

and an augmented diagonal weight matrix $\boldsymbol{w}^*$ is formed as $\boldsymbol{w}^* = diag(\sigma^2_{z_1}, \cdots, \sigma^2_{z_n}, \sigma^2_1, \cdots, \sigma^2_J)$. With $\boldsymbol{z}^*, \boldsymbol{X}^*$ and $\boldsymbol{w}^*$, β coefficients can be estimated using IWLS as described in Section 1.5.1.

For the case when $\sigma^2_1, \cdots, \sigma^2_J$ are unknown variables with given prior distributions, β coefficients could be estimated by using a combination of IWLS and EM algorithms with the latter one used for estimating $\sigma^2_1, \cdots, \sigma^2_J$ by treating β as the missing data (Gelman et al., 2014).

1.5.4 Hierarchical False Discovery Rate (FDR)

Hierarchical FDR (Yekutieli, 2008) has been proposed for multiplicity adjustment in situations where the hypotheses are tested in a hierarchical manner. It involves adopting the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) to each family of hypotheses under each node in the hierarchical tree. As an example in the context of pathway analysis, consider two pathways and genes and SNPs nested hierarchically in Figure 1.2. The hierarchical testing involves starting with the family of two pathways at the top level and testing them. Suppose pathway 1 is found to be significant (colored green) while pathway 2 is not (colored red). For the latter, the testing stops at that point. For pathway 1, we continue to testing genes 1 to 3 as a family of hypotheses and then go further down to a family of SNPs if the gene that they belong to is found to be significant. The SNPs at the bottom level are considered outer-nodes in the tree.

The following steps describe how BH procedure is applied to a family of m hypotheses:

- Step 1: Let $P_{(1)} \leq \ldots P_{(m)}$ denote the set of ordered p-values corresponding to the m hypotheses being tested.

- Step 2: Let $r = \max_i \{P_{(i)} \leq i \cdot q/m\}$ for a given q-value.

- Step 3: If $r > 0$, then reject the r hypotheses corresponding to $P_{(1)}, \ldots, P_{(r)}$.

The q-value is used for inference to control the FDR.

Three versions of hierarchical FDR boundaries are proposed depending on which levels of the hierarchy are of interest for testing. These are full-tree FDR, level-restricted FDR,

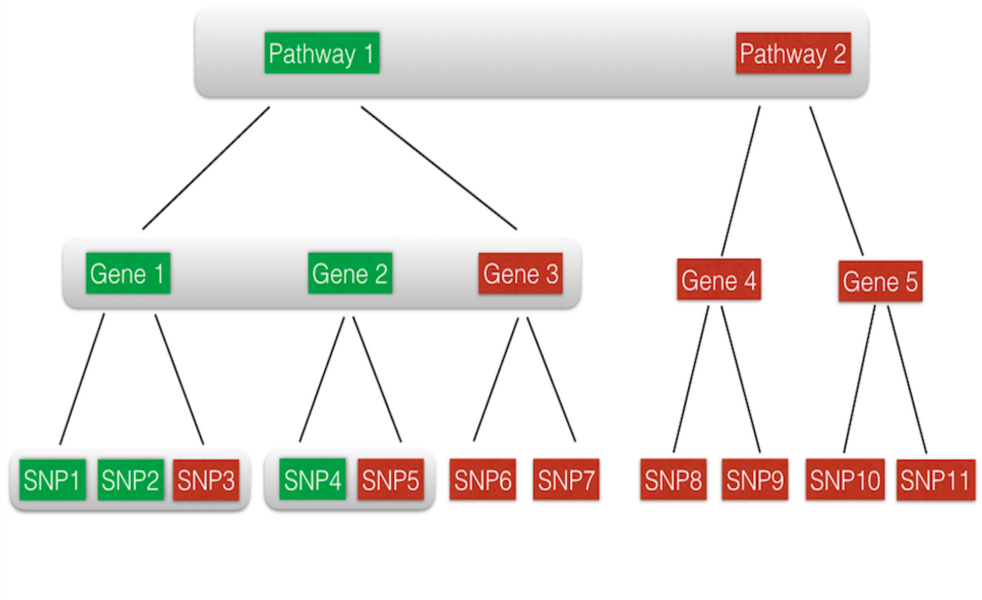


Figure 1.2. Example of hierarchical testing procedure using a two-pathway structure. Nodes in green are found to be significant and thus the nodes nested below them (if any) are tested subsequently.

and outer-nodes FDR. Full-tree FDR focuses on the entire set of discoveries in the tree at all levels. Level restricted FDR is helpful when researchers are interested in discoveries at a particular level of the tree. Outer-nodes FDR is applicable when discoveries found at the outer most leaves of the tree are of interest only.

The upper bound for the full-tree FDR is $q \cdot \delta^* \cdot 2$ and for the outer-nodes FDR, it is $L \cdot q \cdot \delta^* \cdot 2$ for a tree with L levels. There is no theoretical boundary available for level-restricted FDR. However, (Yekutieli, 2008) provides a universal approximate FDR boundary that is applicable to all three types of FDR and guarantees that the actual FDR of the entire hierarchical inference procedure is less than:

$$q \cdot \delta^* \cdot \frac{observed\ no.\ of\ discoveries + observed\ no.\ of\ families\ tested}{observed\ no.\ of\ discoveries + 1},$$

where δ^* can be usually chosen as 1.

1.5.5 Effective Number of Parameters

When testing a family of p hypotheses simultaneously, the total number of hypotheses, i.e., p, is commonly used in many multiplicity adjustment methods (e.g. Bonferroni method, FDR). However, under the Bayesian setting, the number of *independent* hypotheses could be less than p because of the possible dependency across the parameters, especially in hierarchical models (Gelman et al., 2012). Using effective number of parameters (Spiegelhalter et al., 2002) instead of total number of hypotheses, it could potentially account for this dependency and increase power.

The effective number of parameters (p_D) is defined as the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest. To be more specific, $p_D = E_{\psi|y}[-2\log\{\pi(y|\psi)\}] + 2\log[\pi\{y|\bar{\psi}\}]$ for a family of hypotheses where ψ is the set of parameters under study, $\bar{\psi}$ is the posterior mean of ψ, and $\pi(y|\psi)$ is the distribution of y at ψ.

For the case of general hierarchical normal model (Lindley and Smith, 1972)

$$y \sim N(A_1\theta, C_1),\ \theta \sim N(A_2\phi, C_2),$$

it is shown that $p_D = p - \mathrm{tr}(C_2^{-1}V)$, where $V^{-1} = A_1^T C_1^{-1} A_1 + C_2^{-1}$. Here, $0 \leq p_D \leq p$ and $\mathrm{tr}(C_2^{-1}V)$ measures the “shrinkage” of the posterior estimates towards the prior means.

CHAPTER 2
METHODS

Let y_i represent the phenotype response and $\mathbf{x}_i$ represent the vector of genotypes at the SNPs belonging to one or more pathways under consideration for subject $i = 1, \ldots, n$. We assume a GLM for the responses $\mathbf{y} = (y_1, \ldots, y_n)$ from n independent subjects. We focus on $\mathbf{y}$ being binary responses, e.g., case-control status. However, being in the GLM framework, the proposed methodology can also be applied to $\mathbf{y}$ belonging to an exponential family. We model $\mu_i = E(y_i)$ through a link function $\eta_i = g(\mu_i) = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta}$, where β_0 is the intercept and $\boldsymbol{\beta}$ is the vector of the SNP effects. For fitting a classical GLM, an IWLS algorithm is employed to find MLEs of the coefficient vector $\boldsymbol{\beta}$ by fitting $\tilde{\mathbf{y}} = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta}$ with diagonal weight matrix $\mathbf{W}_{\mathbf{y}}$, where $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)$ is a vector of pseudo responses. We will use a similar IWLS approach to estimate the posterior mode of $\boldsymbol{\beta}$ and other model parameters to be introduced in the following.

2.1 Hierarchical Prior Structure

Suppose there are P pathways under investigation and the pth pathway has J_p genes and the jth gene in the pth pathway has S_{jp} SNPs, $j = 1, \ldots, J_p$, $p = 1, \ldots, P$. Thus, there are a total of $J = \sum_{p=1}^{P} J_p$ genes and $S = \sum_{p=1}^{P} \sum_{j=1}^{J_p} S_{jp}$ SNPs. Denote the effects of SNPs by β_{sjp} (referred as β in above), the effects of genes by ξ_{jp}, and the effects of pathways by θ_p, $s = 1, \ldots, S_{jp}$, $j = 1, \ldots, J_p$, $p = 1, \ldots, P$. The vectors of the effects of SNPs, genes, and pathways are denoted by $\boldsymbol{\beta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\theta}$, respectively. To build the three-level SNP–gene–pathway hierarchy, we model the SNP effects β_{sjp} conditional on the gene effects ξ_{jp}, the gene effects are modeled conditional on the pathway effects θ_p, and the effects are assumed to be *conditionally* independent. Specifically, the hierarchical structure of the model is as follows: for $s = 1, \ldots, S_{jp}$, $j = 1, \ldots, J_p$, $p = 1, \ldots, P$,

$$\begin{aligned}
&\text{SNP level (conditional on gene): } \beta_{sjp} | \xi_{jp}, \sigma^2_{\beta jp} \overset{\text{ind}}{\sim} N(\xi_{jp}, \sigma^2_{\beta jp}),\\
&\text{Gene level (conditional on pathway): } \xi_{jp} | \theta_p, \sigma^2_{\xi p} \overset{\text{ind}}{\sim} N(\theta_p, \sigma^2_{\xi p}), \sigma^2_{\beta jp} | s^2_{\beta p} \overset{\text{ind}}{\sim} \text{Inv-}\chi^2(\nu_\beta, s^2_{\beta p}),\\
&\text{Pathway level: } \theta_p | \sigma^2_\theta \overset{\text{ind}}{\sim} N(0, \sigma^2_\theta), \sigma^2_{\xi p} \overset{\text{ind}}{\sim} \text{Inv-}\chi^2(\nu_\xi, s^2_\xi), s^2_{\beta p} \overset{\text{ind}}{\sim} \text{Gamma}(a, b), \sigma^2_\theta \sim \text{Inv-}\chi^2(\nu_\theta, s^2_\theta).
\end{aligned} \tag{2.1}$$

Thus, the SNPs within the jth gene of the pth pathway share a common mean ξ_{jp} and variance $\sigma^2_{\beta jp}$. These parameters are assigned priors using hyper-parameters θ_p, $\sigma^2_{\xi p}$, and s^2_p to capture the shared pathway effect. The hyper-parameters are further assigned priors. This hierarchy allows accounting for dependence among SNPs within a gene and among genes within a pathway. Note that SNP (gene) effects are conditionally independent given the mean and variance parameters that they share. Unconditionally, they are dependent due to the hierarchical structure.

For β_0, we assign a $N(0, \sigma^2_0)$ prior with $\sigma^2_0 \sim \text{Inv-}\chi^2(\nu_0, s^2_0)$, which corresponds to a t marginal distribution with center zero, scale s_0, and degrees of freedom ν_0. By the same token, the above hierarchical specification for the effects can be viewed as a marginal t distribution with mean at each level being zero. This feature serves to regularize the effects of SNPs, genes, and pathways. The amount of shrinkage at each level is controlled by the scale parameters and their hyper-priors. Let $\boldsymbol{\beta}^* = (\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\theta})$ and $\boldsymbol{\phi} = (\sigma^2_0, \boldsymbol{\sigma}^2_\beta, \boldsymbol{\sigma}^2_\xi, \sigma^2_\theta, \mathbf{s}^2_\beta)$ respectively denote the vectors of mean related and scale related parameters in model (2.1).

Here $\boldsymbol{\beta}$, $\boldsymbol{\xi}$, $\boldsymbol{\theta}$, $\boldsymbol{\sigma}^2_\beta$, $\boldsymbol{\sigma}^2_\xi$, and $\mathbf{s}^2_\beta$ are themselves vectors consisting of relevant elements given in model (2.1). For example, $\boldsymbol{\sigma}^2_\beta$ is the vector of prior variances $\sigma^2_{\beta jp}$ of the SNP effects. The hyper-parameters $(s_0^2, s_\xi^2, s_\theta^2, \nu_0, \nu_\beta, \nu_\xi, \nu_\theta, a, b)$ will be assigned known values.

Next, in Bayesian analysis, the prior information on regression coefficients can be treated as additional "data" points, and is combined with the likelihood by augmenting the response vector and the design matrix by adding extra rows and columns (Gelman et al., 2014). The model for the augmented data can be written as a linear model $\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}^*$, where $\mathbf{Y}^*$ consists of the pseudo data vector $\tilde{\mathbf{y}}$ (the one used in fitting a classical GLM) augmented with a vector $\mathbf{0}$ of length $1 + S + J + P$. That is, $\mathbf{Y}^* = (\tilde{y}_1, \ldots, \tilde{y}_n, 0, \cdots, 0)$. The matrix $\boldsymbol{X}^*$ has $n + 1 + S + J + P$ rows and $1 + S + J + P$ columns consisting of the usual design matrix $\boldsymbol{X}$ and additional rows and columns appended to reflect the prior mean at each level. In particular,

$$\boldsymbol{X}^*\boldsymbol{\beta}^* = \begin{pmatrix} \mathbf{1}_{n\times 1} & \boldsymbol{X}_{n\times S} & \mathbf{0}_{n\times J} & \mathbf{0}_{n\times P} \\ 1_{1\times 1} & \mathbf{0}_{1\times S} & \mathbf{0}_{1\times J} & \mathbf{0}_{1\times P} \\ \mathbf{0}_{S\times 1} & \boldsymbol{I}_{S\times S} & -\boldsymbol{X}^\beta_{S\times J} & \mathbf{0}_{S\times P} \\ \mathbf{0}_{J\times 1} & \mathbf{0}_{J\times S} & \boldsymbol{I}_{J\times J} & -\boldsymbol{X}^\xi_{J\times P} \\ \mathbf{0}_{P\times 1} & \mathbf{0}_{P\times S} & \mathbf{0}_{P\times J} & \boldsymbol{I}_{P\times P} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_{S\times 1} \\ \boldsymbol{\xi}_{J\times 1} \\ \boldsymbol{\theta}_{P\times 1} \end{pmatrix}, \tag{2.2}$$

where the matrices $\boldsymbol{X}^\beta$ and $\boldsymbol{X}^\xi$ consist of 0 and 1's arranged in such a way that when combined with the elements of the identity matrices in the same row reflects the membership of each SNP to its gene and each gene to its pathway. The elements of $\boldsymbol{X}^\beta$ and $\boldsymbol{X}^\xi$ are derived as follows.

Assigning normal prior to $\beta_{sjp} \sim N(\xi_{jp}, \sigma^2_{\beta_{jp}})$ for SNP effect is equivalent to the following model:

$$0 = \beta_{sjp} - \xi_{jp} + e_{sjp},\ e_{sjp} \sim N(0, \sigma^2_{\beta_{jp}}),\ i = 1, \dots, S_{jp},\ j = 1, \dots, J_p,\ p = 1, \dots, P.$$

The value 0 on the left side is present in the regression model as response. Meanwhile, the coefficients 1 for β_{sjp} and -1 for ξ_{jp} will show up in the augemented design matrix corresponding to the gene that the SNP with index sjp belongs to. Similarly, for the gene effect $\xi_{jp} \sim N(\theta_p, \sigma^2_{\xi_p})$ and pathway effect $\theta_p \sim N(0, \sigma^2_\theta)$, the equivalent models are:

$$0 = \xi_{jp} - \theta_p + e_{jp},\ e_{jp} \sim N(0, \sigma^2_{\xi_p}),\ j = 1, \dots, J_p,\ p = 1, \dots, P.$$

$$0 = \theta_p + e_p,\ e_p \sim N(0, \sigma^2_\theta),\ p = 1, \dots, P.$$

As a result,

$$X^{\beta}_{S\times J} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\ \vdots\\ 1\end{matrix}}\right\}\text{SNPs in Gene 1} \\ \left.\vphantom{\begin{matrix}1\\ \vdots\\ 1\end{matrix}}\right\}\text{SNPs in Gene 2} \\ \vdots \\ \left.\vphantom{\begin{matrix}1\\ \vdots\\ 1\end{matrix}}\right\}\text{SNPs in Gene } J \end{matrix}$$

$$\text{and} \qquad X^{\xi}_{J\times P} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}}\right\} J_1 \text{ Genes in Pathway 1} \\ \left.\vphantom{\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}}\right\} J_2 \text{ Genes in Pathway 2} \\ \vdots \\ \left.\vphantom{\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}}\right\} J_P \text{ Genes in Pathway } P \end{matrix}$$

2.2 Posterior Mode Estimation

The posterior of $\boldsymbol{\beta}^*$ conditional on $\boldsymbol{\phi}$ can be written as

$$\begin{aligned} \pi(\boldsymbol{\beta}^*|\boldsymbol{\phi},\mathbf{y}^*) &\propto \pi(\mathbf{y}^*|\beta_0,\boldsymbol{\beta})\pi(\beta_0|\sigma_0^2)\pi(\boldsymbol{\beta}|\boldsymbol{\xi},\boldsymbol{\sigma}_\beta^2)\pi(\boldsymbol{\xi}|\boldsymbol{\theta},\boldsymbol{\sigma}_\xi^2)\pi(\boldsymbol{\theta}|\sigma_\theta^2) \\ &= \text{constant} \cdot \left[\prod_{i=1}^{n}\left(\mu_i^{y_i}(1-\mu_i)^{1-y_i}\right)\right]\left[(2\pi)^{-\frac{1}{2}}(\sigma_0^2)^{-\frac{1}{2}}\exp(-\frac{\beta_0^2}{2\sigma_0^2})\right] \\ &\cdot \prod_p\prod_j\prod_s\left\{(2\pi)^{-\frac{1}{2}}(\sigma_{\beta_{jp}}^2)^{-\frac{1}{2}}\exp(-\frac{(\beta_{sjp}-\xi_{jp})^2}{2\sigma_{\beta_{jp}}^2})\right\} \\ &\cdot \prod_p\prod_j\left\{(2\pi)^{-\frac{1}{2}}(\sigma_{\xi_p}^2)^{-\frac{1}{2}}\exp(-\frac{(\xi_{jp}-\theta_p)^2}{2\sigma_{\xi_p}^2})\right\} \\ &\cdot \prod_p\left\{(2\pi)^{-\frac{1}{2}}(\sigma_\theta^2)^{-\frac{1}{2}}\exp(-\frac{(\theta_p-0)^2}{2\sigma_\theta^2})\right\}. \end{aligned} \tag{2.3}$$

As mentioned previously, to maintain scalability of the method while keeping the computational burden manageable, we forgo a full posterior simulation in favor of finding mode of the conditional posterior distribution of $\boldsymbol{\beta}^*$ given appropriate values for $\boldsymbol{\phi}$, and use a normal approximation for this mode to perform inference.

The vector $\boldsymbol{\epsilon}^*$ is distributed as $N(0, \mathbf{W}^{-1})$, where $\mathbf{W} = \text{diag}\{\mathbf{W}_y, W_0, \mathbf{W}_\beta, \mathbf{W}_\xi, \mathbf{W}_\theta\}$ is a weight matrix of order $n + 1 + S + J + P$. It is derived by equating the Fisher information matrix to $\boldsymbol{X}^{*\prime}\boldsymbol{W}\boldsymbol{X}^*$ as shown in the following paragraph. $\mathbf{W}_y$ is the usual weight matrix used in fitting classical GLM, $W_0 = \sigma_0^{-2}$, and $\mathbf{W}_\beta$, $\mathbf{W}_\xi$, and $\mathbf{W}_\theta$ are diagonal matrices with elements consisting of $\sigma_{\beta jp}^{-2}$, $\sigma_{\xi p}^{-2}$, and σ_θ^{-2}, respectively. This linear model representation allows obtaining the approximate mode of $\pi(\boldsymbol{\beta}^*|\boldsymbol{\phi}, \mathbf{y}^*)$ via an IWLS algorithm by fitting a classical GLM for which the algorithm available in `R` is state-of-the-art in terms of efficiency and scalability. The mode is $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\phi}) = (\boldsymbol{X}^{*\prime}\boldsymbol{W}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{W}\boldsymbol{Y}^*$ with approximate covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\phi}) = (\boldsymbol{X}^{*\prime}\boldsymbol{W}\boldsymbol{X}^*)^{-1}$. The mode is "approximate" due to the normal approximation of the likelihood. It is exact for a normal response.

Denote $l = \log \pi(\boldsymbol{\beta}^*|\boldsymbol{\phi}, \mathbf{y}^*)$ where $\pi(\boldsymbol{\beta}^*|\boldsymbol{\phi}, \mathbf{y}^*)$ is defined in equation (2.3). For simplicity of mathematical exposition, let $\xi_{(r)}$ and $\sigma_{(r)}^2$ denote the effects of the gene and pathway that the SNP r belongs to. We could obtain $\boldsymbol{W}$ by equating Fisher information matrix $I_{\boldsymbol{\beta}^*} = \{-E\frac{\partial^2 \log l}{\partial \beta_r \beta_s}\}$ with $\boldsymbol{X}^{*\prime}\boldsymbol{W}\boldsymbol{X}^*$ in the following calculation:

$$
\begin{aligned}
\frac{\partial l}{\partial \beta_r} &= \sum_{i=1}^{n} \left[\left(\frac{y_i}{\mu_i} + \frac{y_i - 1}{1 - \mu_i} \right) \frac{\partial \mu_i}{\partial \beta_r} \right] - \frac{2(\beta_r - \xi_{(r)})}{2\sigma^2_{(r)}} \\
\frac{\partial^2 l}{\partial \beta_r \partial \beta_s} &= 0 \\
\frac{\partial^2 l}{\partial \beta_r^2} &= \sum_{i=1}^{n} \left[\left(-\frac{y_i}{\mu_i^2} \frac{\partial \mu_i}{\partial \beta_r} - \frac{y_i - 1}{(1 - \mu_i)^2} \left(-\frac{\partial \mu_i}{\partial \beta_r} \right) \right) \frac{\partial \mu_i}{\partial \beta_r} + \left(\frac{y_i}{\mu_i} + \frac{y_i - 1}{1 - \mu_i} \right) \frac{\partial^2 \mu_i}{\partial \beta_r^2} \right] - \frac{1}{\sigma^2_{(r)}} \\
&= \sum_{i=1}^{n} \left[\left(-\frac{y_i}{\mu_i^2} + \frac{y_i - 1}{(1 - \mu_i)^2} \right) \left(\frac{\partial \mu_i}{\partial \beta_r} \right)^2 + \left(\frac{y_i}{\mu_i} + \frac{y_i - 1}{1 - \mu_i} \right) \frac{\partial^2 \mu_i}{\partial \beta_r^2} \right] - \frac{1}{\sigma^2_{(r)}} \\
-E \frac{\partial^2 l}{\partial \beta_r^2} &= -\sum_{i=1}^{n} \left[\left(-\frac{1}{\mu_i} + \frac{1}{\mu_i - 1} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_r} \right)^2 + \Big(1 + (-1) \Big) \frac{\partial^2 \mu_i}{\partial \beta_r^2} \right] + \frac{1}{\sigma^2_{(r)}} \\
&= -\sum_{i=1}^{n} \left[\frac{1}{\mu_i (1 - \mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 x_{ir}^2 \right] + \frac{1}{\sigma^2_{(r)}} \\
&= \sum_{i=1}^{n} \left[\frac{1}{\mu_i (1 - \mu_i)} \Big(\mu_i (1 - \mu_i) \Big)^2 x_{ir}^2 \right] + \frac{1}{\sigma^2_{(r)}} \\
&= \sum_{i=1}^{n} [x_{ir} \mu_i (1 - \mu_i) x_{ir}] + \frac{1}{\sigma^2_{(r)}}.
\end{aligned}
$$

In the above, we used the following:

$$
\begin{aligned}
\eta_i = \log(\frac{\mu_i}{1 - \mu_i}) = X_i \boldsymbol{\beta}^* \Rightarrow \frac{\partial \eta_i}{\partial \beta_r} =& X_{ir} \\
\mu_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \Rightarrow \frac{\partial \mu_i}{\partial \eta_i} =& \frac{e^{\eta_i} \cdot (1 + e^{\eta_i}) - e^{\eta_i} \cdot e^{\eta_i}}{(1 + e^{\eta_i})^2} \\
=& \frac{e^{\eta_i} \cdot [(1 + e^{\eta_i}) - e^{\eta_i}]}{(1 + e^{\eta_i})^2} \\
=& \frac{e^{\eta_i}}{1 + e^{\eta_i}} \cdot \frac{1}{1 + e^{\eta_i}} \\
=& \frac{e^{\eta_i}}{1 + e^{\eta_i}} \cdot (1 - \frac{e^{\eta_i}}{1 + e^{\eta_i}}) \\
=& \mu_i \cdot (1 - \mu_i).
\end{aligned}
$$

Using $\boldsymbol{X}^*$ introduced in the hierarchical prior setting, we find $\mathbf{W}^* = \text{diag}\{\mathbf{W}_y, W_0, \mathbf{W}_\beta, \mathbf{W}_\xi, \mathbf{W}_\theta\} = \text{diag}\{w_1, \ldots, w_n, \frac{1}{\sigma_0^2}, \frac{1}{\boldsymbol{\sigma}_\beta^2}, \frac{1}{\boldsymbol{\sigma}_\xi^2} \frac{1}{\boldsymbol{\sigma}_\theta^2}\}$.

2.3 EM Algorithm for Estimating $\boldsymbol{\phi}$

In contrast with classical GLM where the variance parameters are known, the variance parameters as well as other hyper-parameters in our case are unknown. EM algorithm is incorporated with classical GLM in order to get estimates of the unknown parameters (Gelman et al., 2008, 2014). It involves two steps, E-step and M-step, to obtain estimaters for $\boldsymbol{\phi}$ in our hierarchical model. The joint posterior of $\boldsymbol{\beta}^*$ and $\boldsymbol{\phi}$ can be written as:

$$\pi(\boldsymbol{\beta}^{*}, \boldsymbol{\phi}|\mathbf{y}^{*}) \propto \pi(\mathbf{y}^{*}|\boldsymbol{\beta})\pi(\beta_0|\sigma_0^2)\pi(\sigma_0^2|\nu_0, s_0^2)\pi(\boldsymbol{\beta}|\boldsymbol{\xi}, \boldsymbol{\sigma}_{\beta}^2)\pi(\boldsymbol{\xi}|\boldsymbol{\theta}, \boldsymbol{\sigma}_{\xi}^2)$$

$$\begin{aligned}
&\cdot \pi(\boldsymbol{\sigma}_{\beta}^2|\nu_{\beta}, \boldsymbol{s}_{\beta_p}^2)\pi(\boldsymbol{\theta}|\sigma_{\theta}^2)\pi(\sigma_{\theta}^2|\nu_{\theta}, s_{\theta}^2) \cdot \pi(\boldsymbol{\sigma}_{\xi}^2|\nu_{\xi}, s_{\xi}^2)\pi(\boldsymbol{s}_{\beta_p}^2|a, b) \\
&= \left[\prod_{i=1}^{n} \pi(\mathbf{y_i}|\boldsymbol{X}^{*}\boldsymbol{\beta}^{*})\right] \left[(2\pi)^{-\frac{1}{2}}(\sigma_0^2)^{-\frac{1}{2}} \exp\left(-\frac{\beta_0^2}{2\sigma_0^2}\right)\right] \left[\frac{(s_0^2\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} \frac{\exp\left(-\frac{\nu_0 s_0^2}{2\sigma_0^2}\right)}{(\sigma_0^2)^{1+\nu_0/2}}\right] \\
&\cdot \prod_p \prod_j \prod_s \left\{(2\pi)^{-\frac{1}{2}}(\sigma_{\beta_{jp}}^2)^{-\frac{1}{2}} \exp\left(-\frac{(\beta_{sjp}-\xi_{jp})^2}{2\sigma_{\beta_{jp}}^2}\right)\right\} \\
&\cdot \prod_p \prod_j \left\{\frac{(s_{\beta_p}^2\nu_{\beta}/2)^{\nu_{\beta}/2}}{\Gamma(\nu_{\beta}/2)} \frac{\exp\left(-\frac{\nu_{\beta}s_{\beta_p}^2}{2\sigma_{\beta_{jp}}^2}\right)}{(\sigma_{\beta_{jp}}^2)^{1+\nu_{\beta}/2}}\right\} \\
&\cdot \prod_p \prod_j \left\{(2\pi)^{-\frac{1}{2}}(\sigma_{\xi_p}^2)^{-\frac{1}{2}} \exp\left(-\frac{(\xi_{jp}-\theta_p)^2}{2\sigma_{\xi_p}^2}\right)\right\} \cdot \prod_p \left\{\frac{(s_{\xi}^2\nu_{\xi}/2)^{\nu_{\xi}/2}}{\Gamma(\nu_{\xi}/2)} \frac{\exp\left(-\frac{\nu_{\xi}s_{\xi}^2}{2\sigma_{\xi_p}^2}\right)}{(\sigma_{\xi_p}^2)^{1+\nu_{\xi}/2}}\right\} \\
&\cdot \prod_p \left\{(2\pi)^{-\frac{1}{2}}(\sigma_{\theta}^2)^{-\frac{1}{2}} \exp\left(-\frac{(\theta_p-0)^2}{2\sigma_{\theta}^2}\right)\right\} \cdot \left[\frac{(s_{\theta}^2\nu_{\theta}/2)^{\nu_{\theta}/2}}{\Gamma(\nu_{\theta}/2)} \frac{\exp\left(-\frac{\nu_{\theta}s_{\theta}^2}{2\sigma_{\theta}^2}\right)}{(\sigma_{\theta}^2)^{1+\nu_{\theta}/2}}\right] \\
&\cdot \prod_p \left\{\frac{b^a}{\Gamma(a)}(s_{\beta_p}^2)^{a-1}\exp(-bs_{\beta_p}^2)\right\} \cdot \text{constant.}
\end{aligned}$$

Thus, log of joint posterior density (L.P) is:

$$
\begin{aligned}
L.P = &-\frac{1}{2}\ln(\sigma_0^2) - \frac{\beta_0^2}{2\sigma_0^2} - \frac{\nu_0 s_0^2}{2\sigma_0^2} - \left(1+\frac{\nu_0}{2}\right)\ln(\sigma_0^2) \\
&+ \sum_p \sum_j \sum_s \left\{ -\frac{1}{2}\ln(\sigma_{\beta_{jp}}^2) - \frac{(\beta_{sjp}-\xi_{jp})^2}{2\sigma_{\beta_{jp}}^2} \right\} + \sum_p \sum_j \left\{ -\frac{1}{2}\ln(\sigma_{\xi_p}^2) - \frac{(\xi_{jp}-\theta_p)^2}{2\sigma_{\xi_p}^2} \right\} \\
&+ \sum_p \sum_j \left\{ \frac{\nu_\beta}{2}\ln(s_{\beta_p}^2) - \frac{\nu_\beta s_{\beta_p}^2}{2\sigma_{\beta_{jp}}^2} - (1+\frac{\nu_\beta}{2})\ln(\sigma_{\beta_{jp}}^2) \right\} \\
&+ \sum_p \left\{ -\frac{1}{2}\ln(\sigma_\theta^2) - \frac{\theta_p^2}{2\sigma_\theta^2} \right\} - \frac{\nu_\theta s_\theta^2}{2\sigma_\theta^2} - \left(1+\frac{\nu_\theta}{2}\right)\ln(\sigma_\theta^2) \\
&+ \sum_p \left\{ -\frac{\nu_\xi s_\xi^2}{2\sigma_{\xi_p}^2} - \left(1+\frac{\nu_\xi}{2}\right)\ln(\sigma_{\xi_p}^2) \right\} \\
&+ \sum_p \left\{ (a-1)\ln(s_{\beta_p}^2) - b s_{\beta_p}^2) \right\} \\
&+ \text{constant.}
\end{aligned}
$$

In E-step:

To obtain the EM estimates, we use normal approximation in order to get expected log of joint posterior density (E.L.P) (Gelman et al., 2014) as follows:

$$\beta_0 \sim N(\hat{\beta}_0, V_{\hat{\beta}_0}),\ \beta_{sjp} \sim N(\hat{\beta}_{sjp}, V_{\hat{\beta}_{sjp}}),\ \xi_{jp} \sim N(\hat{\xi}_{jp}, V_{\hat{\xi}_{jp}}),\ \text{ and }\ \theta_p \sim N(\hat{\theta}_p, V_{\hat{\theta}_p}),$$

where the means and variances of the normal distributions are the estimates obtained using IWLS as described in Section 2.2.

Thus,

$$
\begin{aligned}
E(\beta_0 - 0)^2 &= [E(\beta_0)]^2 + Var(\beta_0) \approx \hat{\beta}_0^2 + V_{\hat{\beta}_0}, \\
E(\beta_{sjp} - \xi_{jp})^2 &= [E(\beta_{sjp} - \xi_{jp})]^2 + Var(\beta_{sjp} - \xi_{jp}) \\
&\approx (\hat{\beta}_{sjp} - \hat{\xi}_{jp})^2 + V_{\hat{\beta}_{sjp}} + V_{\hat{\xi}_{jp}} - 2\text{Cov}(\hat{\beta}_{sjp}, \hat{\xi}_{jp}), \\
E(\xi_{jp} - \theta_p)^2 &= [E(\xi_{jp} - \theta_p)]^2 + Var(\xi_{jp} - \theta_p) \\
&\approx (\hat{\xi}_{jp} - \hat{\theta}_p)^2 + V_{\hat{\xi}_{jp}} + V_{\hat{\theta}_p} - 2\text{Cov}(\hat{\xi}_{jp}, \hat{\theta}_p), \\
\text{and} \quad E(\theta_p - 0)^2 &= [E(\theta_p)]^2 + Var(\theta_p) \approx \hat{\theta}_p^2 + V_{\hat{\theta}_p}.
\end{aligned}
$$

As a result, the expected log-joint posterior density that averages over $\boldsymbol{\beta}^*$ using the above approximations can be expressed as:

$$
\begin{aligned}
\text{E.L.P} = &-\frac{1}{2}\ln(\sigma_0^2) - \frac{\hat{\beta}_0^2 + V_{\hat{\beta}_0}}{2\sigma_0^2} - \frac{\nu_0 s_0^2}{2\sigma_0^2} - \left(1 + \frac{\nu_0}{2}\right)\ln(\sigma_0^2) \\
&+ \sum_p \sum_j \sum_s \left\{ -\frac{1}{2}\ln(\sigma_{\beta_{jp}}^2) - \frac{(\hat{\beta}_{sjp} - \hat{\xi}_{jp})^2 + V_{\hat{\beta}_{sjp}} + V_{\hat{\xi}_{jp}} - 2\text{Cov}(\hat{\beta}_{sjp}, \hat{\xi}_{jp})}{2\sigma_{\beta_{jp}}^2} \right\} \\
&+ \sum_p \sum_j \left\{ -\frac{1}{2}\ln(\sigma_{\xi_p}^2) - \frac{(\hat{\xi}_{jp} - \hat{\theta}_p)^2 + V_{\hat{\xi}_{jp}} + V_{\hat{\theta}_p} - 2\text{Cov}(\hat{\xi}_{jp}, \hat{\theta}_p)}{2\sigma_{\xi_p}^2} \right\} \\
&+ \sum_p \left\{ -\frac{\nu_\xi s_\xi^2}{2\sigma_{\xi_p}^2} - \left(1 + \frac{\nu_\xi}{2}\right)\ln(\sigma_{\xi_p}^2) \right\} \\
&+ \sum_p \sum_j \left\{ -\frac{\nu_\beta s_{\beta_p}^2}{2\sigma_{\beta_{jp}}^2} - \left(1 + \frac{\nu_\beta}{2}\right)\ln(\sigma_{\beta_{jp}}^2) \right\} \\
&+ \sum_p \left\{ -\frac{1}{2}\ln(\sigma_\theta^2) - \frac{\hat{\theta}_p^2 + V_{\hat{\theta}_p}}{2\sigma_\theta^2} \right\} - \frac{\nu_\theta s_\theta^2}{2\sigma_\theta^2} - \left(1 + \frac{\nu_\theta}{2}\right)\ln(\sigma_\theta^2) \\
&+ \sum_p \sum_j \left\{ \frac{\nu_\beta}{2}\ln(s_{\beta_p}^2) \right\} + \sum_p \left\{ (a-1)\ln(s_{\beta_p}^2) - b s_{\beta_p}^2 \right\} + \text{constant}.
\end{aligned}
$$

In M-step:

Taking partial derivatives w.r.t $\sigma_0^2, \sigma_{\beta_{jp}}^2, \sigma_{\xi_p}^2, \sigma_\theta^2$, and $s_{\beta_p}^2$, the maximum likelihood estimators are obtained by solving the following equations:

$$
\begin{aligned}
\frac{\partial \text{E.L.P}}{\partial \sigma_0^2} &= -\frac{1}{2}\frac{1}{\sigma_0^2} - \frac{\hat{\beta}_0{}^2 + V_{\hat{\beta}_0} + \nu_0 s_0^2}{2}\frac{(-1)}{(\sigma_0^2)^2} - (1 + \frac{\nu_0}{2})\frac{1}{\sigma_0^2} = 0, \\
\frac{\partial \text{E.L.P}}{\partial \sigma_{\beta_{jp}}^2} &= \left[-\left(\frac{\nu_\beta}{2} + 1 + \frac{N_{s\ in\ jp}}{2}\right)\frac{1}{\sigma_{\beta_{jp}}^2} - \frac{\nu_\beta \cdot \sigma_{\beta_p}^2(-1)}{2(\sigma_{\beta_{jp}}^2)^2}\right] \\
&\quad + \sum_{s\ \ in\ \ gene\ \ jp} \frac{(\hat{\beta}_{sjp} - \hat{\xi}_{jp})^2 + V_{\hat{\beta}_{sjp}} + V_{\hat{\xi}_{jp}} - 2\text{Cov}(\hat{\beta}_{sjp}, \hat{\xi}_{jp})}{2(\sigma_{\beta_{jp}}^2)^2} = 0, \\
\frac{\partial \text{E.L.P}}{\partial \sigma_{\xi_p}^2} &= \left[-\left(\frac{\nu_\xi}{2} + 1 + \frac{N_{j\ in\ p}}{2}\right)\frac{1}{\sigma_{\xi_p}^2} - \frac{\nu_\xi s_\xi^2}{2}\frac{(-1)}{(\sigma_{\xi_p}^2)^2}\right] \\
&\quad - \sum_{j\ in\ pathway\ p} \frac{(\hat{\xi}_{jp} - \hat{\theta}_p)^2 + V_{\hat{\xi}_{jp}} + V_{\hat{\theta}_p} - 2\text{Cov}(\hat{\xi}_{jp}, \hat{\theta}_p)}{2(\sigma_{\xi_p}^2)^2}(-1) = 0, \\
\frac{\partial \text{E.L.P}}{\partial \sigma_\theta^2} &= \left[-\left(\frac{\nu_\theta}{2} + 1 + \frac{P}{2}\right)\frac{1}{\sigma_\theta^2} - \frac{\nu_\theta s_\theta^2}{2}\frac{(-1)}{(\sigma_\theta^2)^2}\right] - \sum_p \frac{\hat{\theta}_p^2 + V_{\hat{\theta}_p^2}}{2(\sigma_\theta^2)^2}(-1) = 0, \text{ and} \\
\frac{\partial \text{E.L.P}}{\partial s_{\beta_p}^2} &= \left(\frac{\nu_\beta * N_{j\ in\ p}}{2} + a - 1\right)\frac{1}{s_{\beta_p}^2} - b - \sum_{j\ in\ pathway\ p} \frac{\nu_\beta}{2\sigma_{\beta_{jp}}^2} = 0,
\end{aligned}
$$

where $N_{s\ in\ jp}$ is the number of SNPs mapped to gene j in pathway p and $N_{j\ in\ p}$ is the number of genes in pathway p. The estimators obtained as solutions to the above equations are:

$$
\begin{aligned}
\hat{\sigma}_0^2 &= \frac{\hat{\beta}_0^2 + V_{\hat{\beta}_0} + \nu_0 s_0^2}{\nu_0 + 3}, \\
\hat{\sigma}_{\beta_{jp}}^2 &= \frac{\nu_\beta \cdot s_p^2 + \sum_{s \ \ in\ gene\ jp} [(\hat{\beta}_{sjp} - \hat{\xi}_{jp})^2 + V_{\hat{\beta}_{sjp}} + V_{\hat{\xi}_{jp}} - 2\mathrm{Cov}(\hat{\beta}_{sjp}, \hat{\xi}_{jp})]}{\nu_\beta + 2 + N_{s\ in\ jp}}, \\
\hat{\sigma}_{\xi_p}^2 &= \frac{\nu_\xi \cdot s_\xi^2 + \sum_{j \ \ in\ pathway\ p} [(\hat{\xi}_{jp} - \hat{\theta}_p)^2 + V_{\hat{\xi}_{jp}} + V_{\hat{\theta}_p} - 2\mathrm{Cov}(\hat{\xi}_{jp}, \hat{\theta}_p)]}{\nu_\xi + 2 + N_{j\ in\ p}}, \\
\hat{\sigma}_\theta^2 &= \frac{\nu_\theta \cdot \hat{s}_\theta^2 + \sum_p (\hat{\theta}_p^2 + V_{\hat{\theta}_p})}{\nu_\theta + 2 + P}, \quad \text{and} \\
\hat{s}_{\beta_p}^2 &= \frac{\nu_\beta \cdot N_{j\ in\ p} + 2a - 2}{2b + \sum_{j\ in\ pathway\ p} (\frac{\nu_\beta}{\hat{\sigma}_{\beta_{jp}}^2})} \quad .
\end{aligned}
$$

2.4 IWLS-EM Algorithm for Model Fitting

Putting the IWLS and EM algorithm steps together provides an algorithm to fit the model (2.1) as follows:

- Step 0: Assign initial values to $\boldsymbol{\beta}^*$ and $\boldsymbol{\phi}$, and use them to obtain $\mathbf{y}^*$ and $\mathbf{W}$.
- Step 1: Update $\boldsymbol{\beta}^*$ by performing an IWLS step to find the mode $\hat{\boldsymbol{\beta}}^*(\hat{\boldsymbol{\phi}})$ of $\boldsymbol{\beta}^*$ given current value of $\boldsymbol{\phi}$
- Step 2: Update $\boldsymbol{\phi}$ by performing an EM step to find the mode $\hat{\boldsymbol{\phi}}(\hat{\boldsymbol{\beta}}^*)$ of $\boldsymbol{\phi}$.

Steps 1 and 2 are iterated until convergence to obtain the final estimates of mode $\hat{\boldsymbol{\beta}}^*(\hat{\boldsymbol{\phi}})$ and its covariance matrix $\boldsymbol{\Sigma}$. Using the normal approximation of mode, we perform classical tests of association based on Z or t test statistics for the mean effects of pathways, genes, and SNPs in a hierarchical manner as follows.

2.5 Inference and Hierarchical FDR

First, all pathways under study are tested for association using $H_0 : \theta_p = 0$ for all p. The significant pathways are followed up by testing the genes contained therein using $H_0 : \xi_{jp} = 0$ for all j in each significant pathway p. Finally, the SNPs within the significant genes are tested using $H_0 : \beta_{sjp} = 0$ for all s in each significant gene j within each significant pathway p.

An important issue with such a large scale testing is multiplicity adjustment. Even though the issue is relatively less severe in a Bayesian hierarchical modeling compared to classical setting (Gelman et al., 2012), we address it using hierarchical FDR (Yekutieli, 2008). It involves applying a level-q standard Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to each family of hypotheses tested hierarchically. Here we have three levels of families — all pathways (level 1 family), all genes within each significant pathway (level 2 family), and all SNPs within each significant gene (level 3 family). Our interest lies in full tree FDR, i.e., the entire set of discoveries. For this, a theoretical upper bound exists and an approximation is given by the following (δ^* can be usually set as 1) (Yekutieli, 2008) :

$$q.\ \delta^*.\ \frac{\text{observed \# of discoveries } + \text{ observed \# of families tested}}{\text{observed \# of discoveries } + 1}. \tag{2.4}$$

So, by adjusting the q-value, we can achieve a desired FDR for the entire inference procedure.

2.6 Effective Number of Parameters

We also explore a variation of this hierarchical FDR procedure by adjusting the thresholds. The usual Benjamini-Hochberg thresholds are based on m_t, the number of hypothesis tests in a family Γ_t, i.e., the number of independent parameters being tested. In a Bayesian hierarchical model, the number of independent parameters for a family Γ_t may be less than m_t due to the dependence across the parameters both within and between the various levels of hierarchy. For such models, the *effective number of parameters* (p_D) is defined as the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest (Spiegelhalter et al., 2002). Specifically, $p_D = E_{\psi|y}[-2\log\{\pi(y|\psi)\}] + 2\log[\pi\{y|\bar{\psi}\}]$ for a family of hypotheses with the parameters of focus being denoted by ψ, where $\bar{\psi}$ is the posterior mean of ψ and $\pi(y|\psi)$ is the distribution of y at ψ. When the likelihood is approximated by a normal likelihood (as in GLM framework), p_D for the family Γ_t can be written approximately as $p_D \approx m_t - trace(-P''_{\bar{\psi}} V_{\bar{\psi}})$, where $P''_{\psi} = \partial^2 \log \pi(\psi)/\partial\psi^2$, $\pi(\psi)$ is the prior distribution of ψ, $V_{\psi}^{-1} = -\partial^2 \log \pi(\psi|y)/\partial\psi^2$, and $\pi(\psi|y)$ is the posterior distribution of ψ. A key point in the calculation of p_D is that it depends on the family of parameters (tests) ψ under focus, and varies as the focus of testing shifts from one family to another both within and between the different levels of hierarchy. The specific expression for each type of family can be derived by considering the general hierarchical normal model framework (Lindley and Smith, 1972) to be illustrated below. As $p_D \leq m_t$ for a family Γ_t, using p_D in the FDR thresholds instead of m_t can potentially increase power. P''_{ψ} is a diagonal matrix of prior variances since the prior distributions are Normal under our hierarchical structure.

2.6.1 Illustration of Calculation of V_{ψ}^{-1}

To illustrate how V_{ψ}^{-1} for any family of parameters under study is calculated for a three-level hierarchy, we consider a simple example consisting of only two pathways with each pathway consisting of two genes and each gene having two SNPs. The overall hierarchical structure can be described in the following way:

$$
\begin{aligned}
&\boldsymbol{Y} = A_1\boldsymbol{\beta} + \boldsymbol{e_1}, \quad \boldsymbol{e_1} \sim N(\mathbf{0}, C_1), \\
&\boldsymbol{\beta} = A_2\boldsymbol{\xi} + \boldsymbol{e_2}, \quad \boldsymbol{e_2} \sim N(\mathbf{0}, C_2), \\
&\boldsymbol{\xi} = A_3\boldsymbol{\theta} + \boldsymbol{e_3}, \quad \boldsymbol{e_3} \sim N(\mathbf{0}, C_3), \text{ and} \\
&\boldsymbol{\theta} \sim N(\mathbf{0}, C_4). \qquad (2.5)
\end{aligned}
$$

Here $\boldsymbol{\beta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\theta}$ are coefficient vectors previously introduced in the model. A_1 is the genotype matrix while A_2 and A_3 are matrices reflecting the membership of SNPs to genes and genes to pathways. A value of 1 in A_2 and A_3 indicates that the element for the corresponding row is mapped to the element for the corresponding column, otherwise, it is 0. $\boldsymbol{e_1}, \boldsymbol{e_2}$ and $\boldsymbol{e_3}$ are the error terms. C_1 is variance matrix of $\boldsymbol{e_1}$ in fitting response Y. C_2, C_3, and C_4 are prior variance matrices of $\boldsymbol{\beta}, \boldsymbol{\xi}$, and $\boldsymbol{\theta}$, respectively.

Let $\boldsymbol{\beta_{1,2}} = (\beta_1, \beta_2)$ be the vector of SNP effects in gene 1 and $\boldsymbol{\beta_{3,4}} = (\beta_3, \beta_4)$, $\boldsymbol{\beta_{5,6}} = (\beta_5, \beta_6)$, $\boldsymbol{\beta_{7,8}} = (\beta_7, \beta_8)$ be vectors of SNP effects in gene 2, gene 3, and gene 4, respectively. The gene effects are denoted by ξ_1, ξ_2, ξ_3, and ξ_4 with the first two for the genes in the first pathway and the other two for the genes in the second pathway.

Let A_{1i} denote a submatrix of the genotype matrix A_1 whose columns correspond to the SNPs in gene i for $i = 1, 2, 3, 4$. For example, A_{11} consists of the columns for SNP1 and SNP2 in gene 1 from A_1.

Let A_{2i} denote a submatrix of A_2 corresponding to the SNPs in gene i for $i = 1, 2, 3, 4$. For example, A_{21} contains the rows corresponding to SNP 1, SNP 2 and the column corresponding to gene 1 from A_2. C_{2i} is the corresponding covariance matrices for the group of SNPs under study. Similarly, A_{3i} denotes a submatrix of A_3 corresponding to the genes in pathway i for $i = 1, 2$. For example, A_{31} is obtained by taking the rows of A_3 that correspond to gene 1 and gene 2. C_{3i} is the corresponding covariance matrices for the group of genes under study. With these specifications, we can write:

$$\begin{aligned}
\boldsymbol{\beta_{1,2}} &= A_{21}\xi_1 + \boldsymbol{e_{21}},\ \boldsymbol{e_{21}} \sim N(\mathbf{0}, C_{21}), \\
\xi_1 &= A_{31}\boldsymbol{\theta} + e_{31},\ e_{31} \sim N(0, C_{31}) \\
\boldsymbol{\beta_{3,4}} &= A_{22}\xi_2 + \boldsymbol{e_{22}},\ \boldsymbol{e_{22}} \sim N(\mathbf{0}, C_{22}), \\
\xi_2 &= A_{32}\boldsymbol{\theta} + e_{32}, e_{32} \sim N(0, C_{32}), \\
\boldsymbol{\beta_{5,6}} &= A_{23}\xi_3 + \boldsymbol{e_{23}},\ \boldsymbol{e_{23}} \sim N(\mathbf{0}, C_{23}), \\
\xi_3 &= A_{33}\boldsymbol{\theta} + e_{33},\ e_{33} \sim N(0, C_{33}), \\
\boldsymbol{\beta_{7,8}} &= A_{24}\xi_4 + \boldsymbol{e_{24}},\ \boldsymbol{e_{24}} \sim N(\mathbf{0}, C_{24}),\ \text{and} \\
\xi_4 &= A_{34}\boldsymbol{\theta} + e_{34},\ e_{34} \sim N(0, C_{34}) \qquad (2.6)
\end{aligned}$$

When calculating the effective number of parameters at the pathway level, i.e., corresponding to θ, we consider $\boldsymbol{Y}$ from equation (2.5) being writen as:

$$\begin{aligned}
\boldsymbol{Y} &= \boldsymbol{e_1} + A_1[\boldsymbol{e_2} + A_2(\boldsymbol{e_3} + A_3\boldsymbol{\theta})] \\
&= \boldsymbol{e_1} + A_1\boldsymbol{e_2} + A_1A_2\boldsymbol{e_3} + A_1A_2A_3\boldsymbol{\theta}.
\end{aligned}$$

As a result,

$$E(\boldsymbol{Y}|\boldsymbol{\theta}) = A_1A_2A_3\boldsymbol{\theta},$$

$$Var(\boldsymbol{Y}|\boldsymbol{\theta}) = C_1 + A_1C_2A_1^T + A_1A_2C_3A_2^TA_1^T.$$

The posterior distribution of θ can be expressed as:

$$\begin{aligned}
&\pi(\boldsymbol{\theta}|\boldsymbol{Y}, A_1, A_2, A_3, C_1, C_2, C_3, C_4) \propto \pi(\boldsymbol{Y}|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}) \\
&\propto \exp\left\{-\frac{1}{2}\left[(\boldsymbol{Y} - A_1A_2A_3\boldsymbol{\theta})^T(C_1 + A_1C_2A_1^T + A_1A_2C_3A_2^TA_1^T)^{-1}(\boldsymbol{Y} - A_1A_2A_3\boldsymbol{\theta})\right]\right\} \\
&\cdot \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\theta} - \mathbf{0})^TC_4^{-1}(\boldsymbol{\theta} - \mathbf{0})\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\theta}^TA_3^TA_2^TA_1^T(C_1 + A_1C_2A_1^T + A_1A_2C_3A_2^TA_1^T)^{-1}A_1A_2A_3\boldsymbol{\theta} + \boldsymbol{\theta}^TC_4^{-1}\boldsymbol{\theta}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T\left[C_4^{-1} + A_3^TA_2^TA_1^T(C_1 + A_1C_2A_1^T + A_1A_2C_3A_2^TA_1^T)^{-1}A_1A_2A_3\right]\boldsymbol{\theta}\right\}.
\end{aligned}$$

This quadratic term of $\boldsymbol{\theta}$ implies that V_ψ^{-1} would be $V_\psi^{-1}(\theta) = [C_4^{-1} + A_3^TA_2^TA_1^T(C_1 + A_1C_2A_1^T + A_1A_2C_3A_2^TA_1^T)^{-1}A_1A_2A_3]$. Similar idea of finding the quadratic terms in posterior distributions of $\boldsymbol{\xi}$ and $\boldsymbol{\beta}$ can be applied in order to get effective number of parameters for a family of genes within a pathway and a family of SNPs within a gene. In particular, at SNP level with $\boldsymbol{\beta_{1,2}}$, and $\boldsymbol{\xi}$ given based on (2.6), we have:

$$\begin{aligned}
\boldsymbol{Y}|\boldsymbol{\beta_{1,2}}, \boldsymbol{\xi} &= \boldsymbol{e_1} + A_{11}\boldsymbol{\beta_{1,2}} + A_{12}\boldsymbol{\beta_{3,4}} + A_{13}\boldsymbol{\beta_{5,6}} + A_{14}\boldsymbol{\beta_{7,8}} \\
&= \boldsymbol{e_1} + A_{11}\boldsymbol{\beta_{1,2}} + A_{12}(\boldsymbol{e_{22}} + A_{22}\xi_2) + A_{13}(\boldsymbol{e_{23}} + A_{23}\xi_3) + A_{14}(\boldsymbol{e_{24}} + A_{24}\xi_4) \\
&= A_{11}\boldsymbol{\beta_{1,2}} + A_{12}A_{22}\xi_2 + A_{13}A_{23}\xi_3 + A_{14}A_{24}\xi_4 + \boldsymbol{e_1} + A_{12}\boldsymbol{e_{22}} + A_{13}\boldsymbol{e_{23}} + A_{14}\boldsymbol{e_{24}}, \\
E(\boldsymbol{Y}|\boldsymbol{\beta_{1,2}}, \boldsymbol{\xi}) &= A_{11}\boldsymbol{\beta_{1,2}} + A_{12}A_{22}\xi_2 + A_{13}A_{23}\xi_3 + A_{14}A_{24}\xi_4, \\
Var(\boldsymbol{Y}|\boldsymbol{\beta_{1,2}}, \boldsymbol{\xi}) &= C_1 + A_{12}C_{22}A_{12}^T + A_{13}C_{23}A_{13}^T + A_{14}C_{24}A_{14}^T.
\end{aligned}$$

The posterior distribution of $\boldsymbol{\beta}_{\mathbf{1,2}}$ is then written as:

$$\begin{aligned}
\pi(\boldsymbol{\beta}_{\mathbf{1,2}}|\boldsymbol{Y},\boldsymbol{\xi},\boldsymbol{\beta}) &\propto \pi(\boldsymbol{Y}|\boldsymbol{\beta}_{\mathbf{1,2}},\boldsymbol{\xi})\cdot\pi(\boldsymbol{\beta}_{\mathbf{1,2}}|\boldsymbol{\xi})\\
&\propto \exp\Bigg\{-\frac{1}{2}\Bigg[(\boldsymbol{Y}-A_{11}\boldsymbol{\beta}_{\mathbf{1,2}}-A_{12}A_{22}\xi_2-A_{13}A_{23}\xi_3-A_{14}A_{24}\xi_4)\Bigg]^T\\
&\cdot\Bigg(C_1+A_{12}C_{22}A_{12}^T+A_{13}C_{23}A_{13}^T+A_{14}C_{24}A_{14}^T\Bigg)^{-1}\\
&\cdot\Bigg[(\boldsymbol{Y}-A_{11}\boldsymbol{\beta}_{\mathbf{1,2}}-A_{12}A_{22}\xi_2-A_{13}A_{23}\xi_3-A_{14}A_{24}\xi_4)\Bigg]\Bigg\}\\
&\cdot\exp\Bigg\{-\frac{1}{2}(\boldsymbol{\beta}_{\mathbf{1,2}}-A_{21}\xi_1)^TC_{21}^{-1}(\boldsymbol{\beta}_{\mathbf{1,2}}-A_{21}\xi_1)\Bigg\}.
\end{aligned}$$

Collecting the quadratic terms for $\boldsymbol{\beta}_{\mathbf{1,2}}$, we get

$$\boldsymbol{\beta}_{\mathbf{1,2}}^TA_{11}^T(C_1+A_{12}C_{22}A_{12}^T+A_{13}C_{23}A_{13}^T+A_{14}C_{24}A_{14}^T)^{-1}A_{11}\boldsymbol{\beta}_{\mathbf{1,2}}+\boldsymbol{\beta}_{\mathbf{1,2}}^TC_{21}^{-1}\boldsymbol{\beta}_{\mathbf{1,2}}$$

and this implies that V_ψ^{-1} term for $\boldsymbol{\beta}_{\mathbf{1,2}}$ is

$$\begin{aligned}
&V_\psi^{-1}(\boldsymbol{\beta}_{\mathbf{1,2}})=C_{21}^{-1}+A_{11}^T(C_1+A_{12}C_{22}A_{12}^T+A_{13}C_{23}A_{13}^T+A_{14}C_{24}A_{14}^T)^{-1}A_{11}\\
&=C_{21}^{-1}+A_{11}^T(C_1+A_1C_2A_1^T-A_{11}C_{21}A_{11}^T)A_{11}.
\end{aligned}$$

Similarly the other V_ψ^{-1} elements are:

$$\begin{aligned}
V_\psi^{-1}(\boldsymbol{\beta}_{\mathbf{3,4}})&=C_{22}^{-1}+A_{12}^T(C_1+A_1C_2A_1^T-A_{12}C_{22}A_{12}^T)A_{12},\\
V_\psi^{-1}(\boldsymbol{\beta}_{\mathbf{5,6}})&=C_{23}^{-1}+A_{13}^T(C_1+A_1C_2A_1^T-A_{13}C_{23}A_{13}^T)A_{13},\\
V_\psi^{-1}(\boldsymbol{\beta}_{\mathbf{7,8}})&=C_{24}^{-1}+A_{14}^T(C_1+A_1C_2A_1^T-A_{14}C_{24}A_{14}^T)A_{14}.
\end{aligned}$$

Thus, we can generalize this formula for any family of SNPs of interest to get V_ψ^{-1} for this specific group as

$$
\begin{aligned}
V_\psi^{-1} &= [\text{prior variance matrix for this family of SNPs}]^{-1} \\
&\quad + (\text{subset of design matrix corresponding to this family})^T \\
&\quad \cdot \Big[C_1 + A_1 C_2 A_1^T - (\text{subset of design matrix corresponding to this family}) \\
&\quad \cdot (\text{prior variance matrix for this group}) \cdot (\text{subset})^T\Big]^{-1} \\
&\quad (\text{subset of design matrix corresponding to this group}).
\end{aligned}
$$

Finally, for finding the effective number of parameters at gene level, consider gene family $\boldsymbol{\xi_{1,2}} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$ as an example. With $\boldsymbol{\xi_{1,2}}$ and θ given,

$$\begin{aligned}
\boldsymbol{Y}|\boldsymbol{\xi_{1,2}}, \boldsymbol{\theta} &= \boldsymbol{e_1} + A_{11}(\boldsymbol{e_{21}} + A_{21}\xi_1) + A_{12}(\boldsymbol{e_{22}} + A_{22}\xi_2) + A_{13}(\boldsymbol{e_{23}} + A_{23}\xi_3) \\
&\quad + A_{14}(\boldsymbol{e_{24}} + A_{24}\xi_4) \\
&= \boldsymbol{e_1} + A_{11}\boldsymbol{e_{21}} + A_{11}A_{21}\xi_1 + A_{12}\boldsymbol{e_{22}} + A_{12}A_{22}\xi_2 + A_{13}\boldsymbol{e_{23}} \\
&\quad + A_{13}A_{23}(\boldsymbol{e_{33}} + A_{33}\boldsymbol{\theta}) + A_{14}\boldsymbol{e_{24}} + A_{14}A_{24}(\boldsymbol{e_{34}} + A_{34}\boldsymbol{\theta}), \\
E(\boldsymbol{Y}|\boldsymbol{\xi_{1,2}}, \boldsymbol{\theta}) &= A_{11}A_{21}\xi_1 + A_{12}A_{22}\xi_2 + A_{13}A_{23}A_{33}\boldsymbol{\theta} + A_{14}A_{24}A_{34}\boldsymbol{\theta} \\
&= \begin{pmatrix} A_{11} & A_{12} \end{pmatrix} \begin{pmatrix} A_{21} & \mathbf{0} \\ \mathbf{0} & A_{22} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} A_{13} & A_{14} \end{pmatrix} \begin{pmatrix} A_{23} & \mathbf{0} \\ \mathbf{0} & A_{24} \end{pmatrix} \begin{pmatrix} A_{33} & \mathbf{0} \\ \mathbf{0} & A_{34} \end{pmatrix} \boldsymbol{\theta}, \\
Var(\boldsymbol{Y}|\boldsymbol{\xi_{1,2}}, \boldsymbol{\theta}) &= C_1 + A_{11}C_{21}A_{11}^T + A_{12}C_{22}A_{12}^T + A_{13}C_{23}A_{13}^T + A_{13}A_{23}C_{33}A_{23}^TA_{13}^T \\
&\quad + A_{14}C_{24}A_{14}^T + A_{14}A_{24}C_{34}A_{24}^TA_{14}^T \\
&= C_1 + A_{11}C_{21}A_{11}^T + A_{12}C_{22}A_{12}^T + A_{13}C_{23}A_{13}^T + A_{14}C_{24}A_{14}^T \\
&\quad + A_{13}A_{23}C_{33}A_{23}^TA_{13}^T + A_{14}A_{24}C_{34}A_{24}^TA_{14}^T \\
&= C_1 + A_1C_2A_1^T + A_{13}A_{23}C_{33}A_{23}^TA_{13}^T + A_{14}A_{24}C_{34}A_{24}^TA_{14}^T \\
&= C_1 + A_1C_2A_1^T + \begin{pmatrix} A_{13} & A_{14} \end{pmatrix} \begin{pmatrix} A_{23} & \mathbf{0} \\ \mathbf{0} & A_{24} \end{pmatrix} \begin{pmatrix} C_{33} & \mathbf{0} \\ \mathbf{0} & C_{34} \end{pmatrix} \begin{pmatrix} A_{23} & \mathbf{0} \\ \mathbf{0} & A_{24} \end{pmatrix}^T \begin{pmatrix} A_{13} \\ A_{14} \end{pmatrix}.
\end{aligned}$$

As a result, the posterior distribution of $\boldsymbol{\xi_{1,2}}$ is then written as

$$\begin{aligned}
\pi(\boldsymbol{\xi_{1,2}}|\boldsymbol{Y},\boldsymbol{\theta}) &\propto \pi(\boldsymbol{Y}|\boldsymbol{\xi_{1,2}},\boldsymbol{\theta})\cdot\pi(\boldsymbol{\xi_{1,2}}|\boldsymbol{\theta})\\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{Y}-E(\boldsymbol{Y}|\boldsymbol{\xi_{1,2}},\boldsymbol{\theta})\right]^T\cdot Var^{-1}(\boldsymbol{Y}|\boldsymbol{\xi_{1,2}},\boldsymbol{\theta})\cdot\left[\boldsymbol{Y}-E(\boldsymbol{Y}|\boldsymbol{\xi_{1,2}},\boldsymbol{\theta})\right]\right\}\\
&\cdot\exp\left\{-\frac{1}{2}\left[\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix}-\begin{pmatrix}A_{31} & \mathbf{0}\\ \mathbf{0} & A_{32}\end{pmatrix}\boldsymbol{\theta}\right]^T\begin{pmatrix}C_{31} & \mathbf{0}\\ \mathbf{0} & C_{32}\end{pmatrix}^{-1}\left[\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix}-\begin{pmatrix}A_{31} & \mathbf{0}\\ \mathbf{0} & A_{32}\end{pmatrix}\boldsymbol{\theta}\right]\right\}.
\end{aligned}$$

Thus, the quadratic term for $\boldsymbol{\xi_{1,2}}$ is

$$\begin{aligned}
&-\frac{1}{2}\left\{\begin{pmatrix}\xi_1 & \xi_2\end{pmatrix}\begin{pmatrix}C_{31} & \mathbf{0}\\ \mathbf{0} & C_{32}\end{pmatrix}^{-1}\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix}\right\}\\
&+\left[\begin{pmatrix}A_{11} & A_{12}\end{pmatrix}\begin{pmatrix}A_{21} & \mathbf{0}\\ \mathbf{0} & A_{22}\end{pmatrix}\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix}\right]^T\cdot Var^{-1}(\boldsymbol{Y}|\boldsymbol{\xi_{1,2}},\boldsymbol{\theta})\cdot\left[\begin{pmatrix}A_{11} & A_{12}\end{pmatrix}\begin{pmatrix}A_{21} & \mathbf{0}\\ \mathbf{0} & A_{22}\end{pmatrix}\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix}\right].
\end{aligned}$$

This implies that $V_\psi^{-1}(\boldsymbol{\xi_{1,2}})$ for $\boldsymbol{\xi_{1,2}}$ is

$$\begin{pmatrix}C_{31} & \mathbf{0}\\ \mathbf{0} & C_{32}\end{pmatrix}^{-1}+\left[\begin{pmatrix}A_{11} & A_{12}\end{pmatrix}\begin{pmatrix}A_{21} & \mathbf{0}\\ \mathbf{0} & A_{22}\end{pmatrix}\right]^T\cdot Var^{-1}(\boldsymbol{Y}|\boldsymbol{\xi_{1,2}},\boldsymbol{\theta})\cdot\left[\begin{pmatrix}A_{11} & A_{12}\end{pmatrix}\begin{pmatrix}A_{21} & \mathbf{0}\\ \mathbf{0} & A_{22}\end{pmatrix}\right].$$

Similarly, $V_\psi^{-1}(\boldsymbol{\xi_{3,4}})$ for $\boldsymbol{\xi_{3,4}}$ is

$$\begin{pmatrix}C_{33} & \mathbf{0}\\ \mathbf{0} & C_{34}\end{pmatrix}^{-1}+\left[\begin{pmatrix}A_{13} & A_{14}\end{pmatrix}\begin{pmatrix}A_{23} & \mathbf{0}\\ \mathbf{0} & A_{24}\end{pmatrix}\right]^T\cdot Var^{-1}\cdot(\boldsymbol{Y}|\boldsymbol{\xi_{3,4}},\boldsymbol{\theta})\cdot\left[\begin{pmatrix}A_{13} & A_{14}\end{pmatrix}\begin{pmatrix}A_{23} & \mathbf{0}\\ \mathbf{0} & A_{24}\end{pmatrix}\right].$$

2.7 Software

The proposed method has been implemented in an R package BHPathway (Version 1.0) available at `http://www.utdallas.edu/~swati.biswas`.

CHAPTER 3

SIMULATION STUDY

We carry out simulations under different settings to investigate the proposed method BHPathway and compare with three commonly used pathway association methods, namely PLINK (Purcell et al., 2007), ALIGATOR (Holmans et al., 2009), and GRASS (Chen et al., 2010), which are available in an R package named SNPath (SNPath R package, 2018). The different settings vary in the number, size, and structure of pathways, number of null and non-null pathways, effect sizes, and minor allele frequencies (MAF) of SNPs, and these are chosen to allow us investigate various aspects that may affect the results. The settings will be described in detail in the following sub-sections.

In general, pathways with larger number of causal SNPs are assigned smaller effect sizes so that the powers of all methods are not close to 100% for small FDRs. In all settings, we use a logistic regression model to simulate case-control status with causal SNPs and their regression coefficients depending on the setting. We use various uniform distributions to generate effects of the causal SNPs. The effects of the null SNPs in non-null pathways as well as of all SNPs in null pathways are set to be 0. Thus, the SNP effects are generated from a mixture distribution of a point mass at 0 and uniform distributions. The data are deliberately simulated from a distribution different from the assumed model (2.1) to assess the impact of model mis-specification.

A sample consists of 500 cases and 500 controls. A total of 1000 replications are generated for each setting. The methods are compared for power of detecting associated pathways, and for this the inference is carried out using standard Benjamini-Hochberg FDR (Benjamini and

Hochberg, 1995) for all methods. Hierachical FDR is used for BHPathway in one setting to examine the power for detecting associated genes once a pathway is implicated. As other methods do not allow detection at the gene level, they are not considered in this gene-level power calculation.

3.1 Setting 1

We simulated six pathways with pathways 1 to 3 being non-null and the rest being null pathways. All pathways have the same structure, each consisting of 20 genes, and each gene consisting of 10 SNPs. A schematic diagram of the structures of the non-null pathways is shown in Figure 3.1. In each non-null pathway, 8 (out of 20) genes are associated. The number of non-null SNPs in each gene in pathways 1, 2, and 3 are varied, and are 8, 6, and 4, respectively. The corresponding effects of these non-null SNPs are generated from U(0, 0.1), U(0, 0.15), and U(0, 0.2) distributions. We set the MAF of each SNP to be 0.1. Figure 3.2 shows the powers for detecting the first three pathways plotted against average (empirical) FDR calculated over the three null pathways as an ROC curve. We see that BHPathway has higher powers for detecting all three associated pathways.

3.2 Setting 2

Here we consider realistic linkage disequilibrium (LD) patterns and generate one null and one non-null pathway based on structure of a KEGG pathway hsa04950. There are 13 genes and 192 SNPs in each pathway as shown in Figure 3.3. The MAF of SNPs range from $2.5 * 10^{-5}$ to 0.25. We use HAP-SAMPLE software (Wright et al., 2007) to simulate genotypes for the

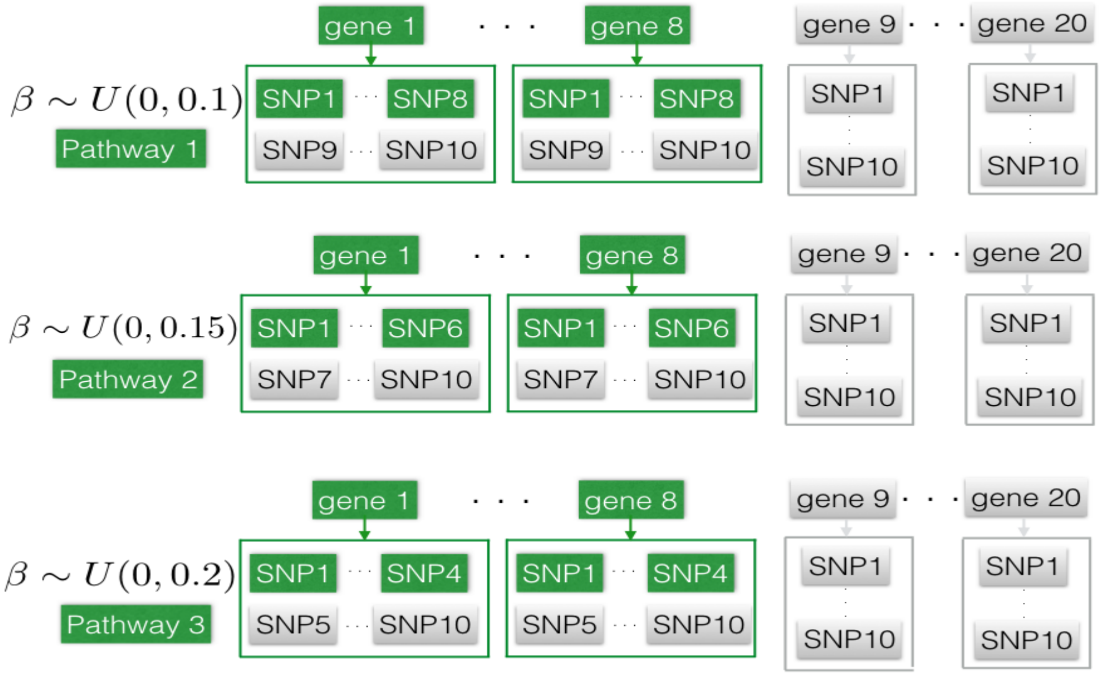


Figure 3.1. Setting 1: Structure of the three non-null pathways. There are three additional null pathways of the same structure with all SNPs being null (not shown). The components in green (dark shade) are non-null while the ones in gray (light shade) are null.

SNPs in this pathway that match with the CEU population of HapMap. In the associated pathway, about half of the genes are associated and in each associated gene, about half of the SNPs are associated. The effect sizes of the non-null SNPs are generated from U(0.15, 0.3) distribution. Under this setting, three sub-settings are considered for assigning the effect sizes to causal SNPs: (1) larger effect sizes are assigned to SNPs with larger MAFs (2) larger effect sizes are assigned to SNPs with smaller MAFs, and (3) causal SNPs are within a contiguous block with larger effect sizes assigned to SNPs with smaller MAFs. The rest of the features of the simulation are the same as in Setting 1. In Figures 3.4, 3.5, and 3.6, we plot the ROC curves for the three sub-settings. The powers are, in general, much higher in sub-setting 1

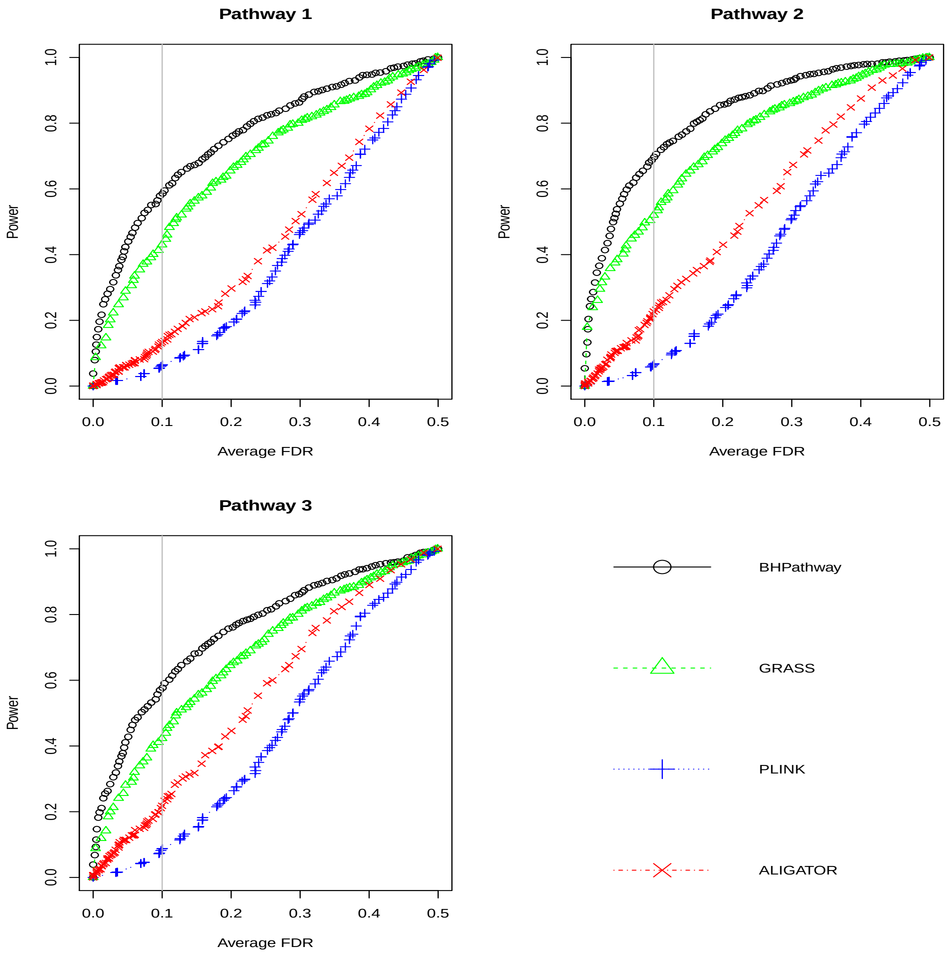
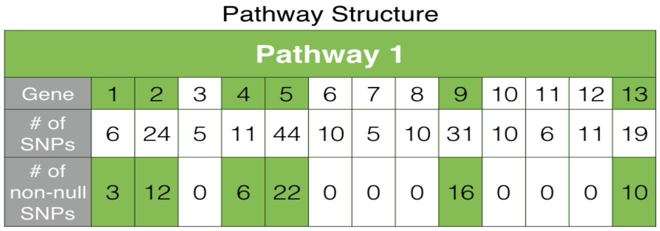


Figure 3.2. Setting 1 Results: Power for detecting the three non-null pathways.

compared to the other two, as expected. In all sub-settings, BHPathway has higher power than the other methods. Among the other methods, there is no method that performs best in all sub-settings. For example, GRASS performs better than ALIGATOR in sub-setting 3, however, the vice versa is true in sub-setting 2.

In addition to investigating the power at the pathway level, we also study the power of detection at the gene level when a pathway is implicated to be significant. For this, we use the hierarchical inference procedure described in Section 2.5 and choose a q-value such that



Pathway Structure

Pathway 1

Gene	1	2	3	4	5	6	7	8	9	10	11	12	13
# of SNPs	6	24	5	11	44	10	5	10	31	10	6	11	19
# of non-null SNPs	3	12	0	6	22	0	0	0	16	0	0	0	10

Figure 3.3. Setting 2: Structure of the non-null pathway. There is also a null pathway of the same structure with all SNPs being null (not shown). The components in green are non-null.

the hierarchical FDR in equation (2.4), averaged over 1000 replications, is approximately 0.1. We present two sets of results in each of Figures 3.7, 3.8, and 3.9 — (a) without and (b) with using the effective number of parameters. In these results, the number of times a gene is found to be significant is out of the total number of times the corresponding pathway is found to be significant, e.g., in Figure 3.7 (a), gene 1 was found to be significant 881 times out of 980 times that the pathway was found to be significant. The effective number of parameters at the pathway level is almost the same as the number of pathways tested in all settings (in this setting, it is 2), and thus at the pathway level, use of the effective number of parameters in inference does not make much difference in results. However, at the gene level, in Figures 3.7, 3.8, and 3.9, we see that using the effective number of parameters increases the power to detect the associated genes, and helps in distinguishing between the null and non-null genes once a pathway is found to be significant.

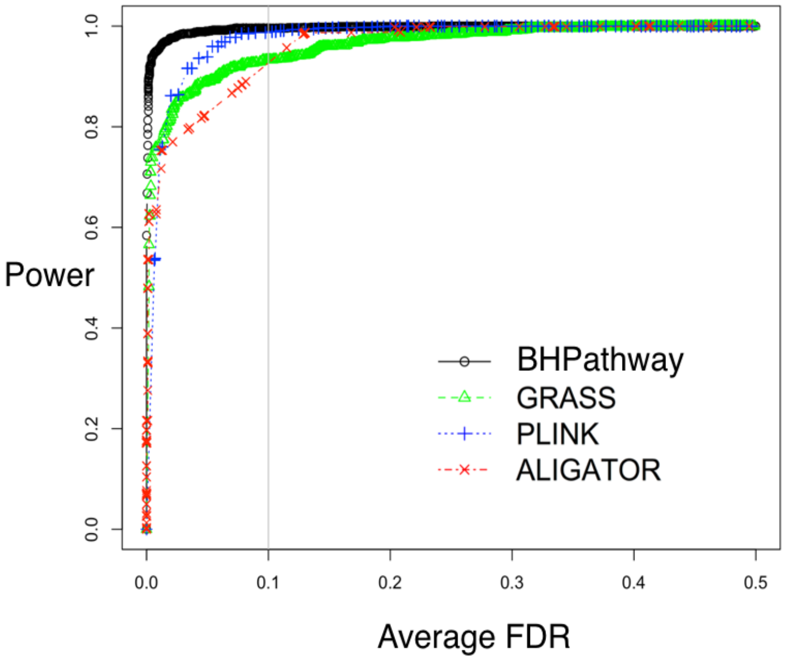


Figure 3.4. Setting 2, Sub-Setting 1 Results: Power for detecting the non-null pathway.

3.3 Setting 3

In this setting, we consider three KEGG pathway structures hsa00910, hsa00062, and hsa04130 consisting of 6, 14, and 19 genes, respectively, and simulate six pathways, three null and three non-null of these structures, as shown in Figure 3.10. The range of MAF of SNPs is from 8.33e-06 to 0.25. As in Setting 2, about half of the genes in each non-null pathway is associated and in each non-null gene, about half of the SNPs are associated. The effect sizes are generated from U(0.15, 0.2), U(0.1, 0.15), and U(0.05, 0.1) distributions for the non-null SNPs in hsa00910, hsa00062, and hsa04130, respectively. The causal SNPs are within a contiguous block with larger effect sizes assigned to SNPs with smaller MAFs. The

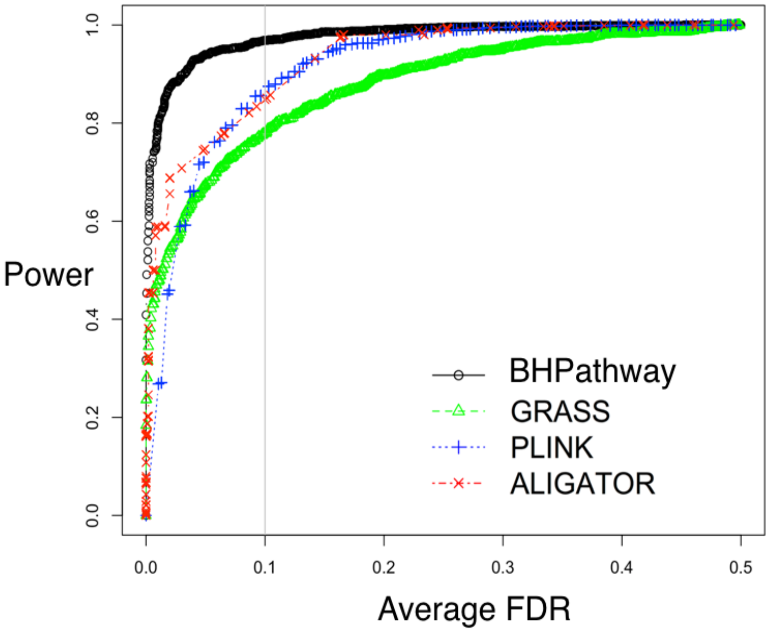


Figure 3.5. Setting 2, Sub-Setting 2 Results: Power for detecting the non-null pathway.

sample generation procedure for this and subsequent settings is the same as for Setting 2 using HAP-SAMPLE. Figure 3.11 shows the ROC curves for the three associated pathways. BHPathway has markedly higher powers than the other three methods especially for pathways 2 and 3.

3.4 Setting 4

This setting is based on the same pathway structures as in Setting 3. The difference is in the generation of associated SNPs. Here one-third (instead of one-half) of the genes are associated in each non-null pathway and within each associated gene, one-half of the SNPs

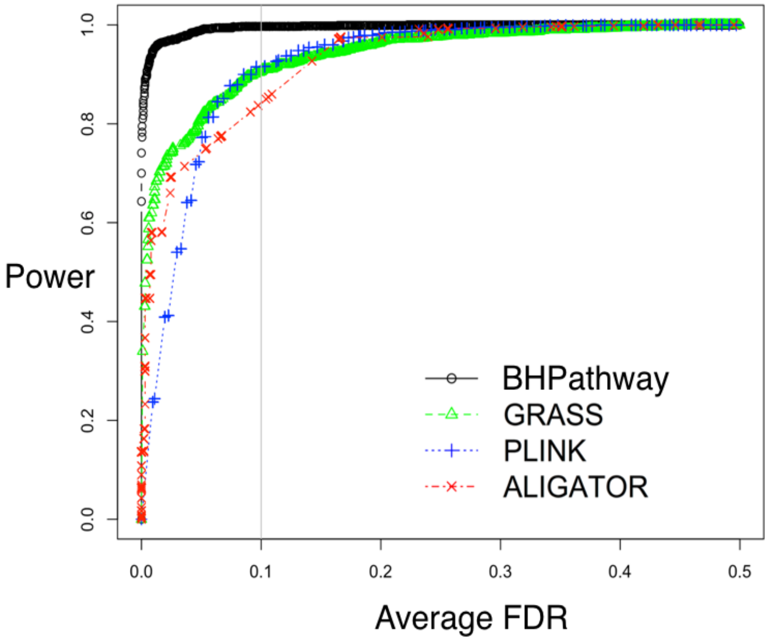
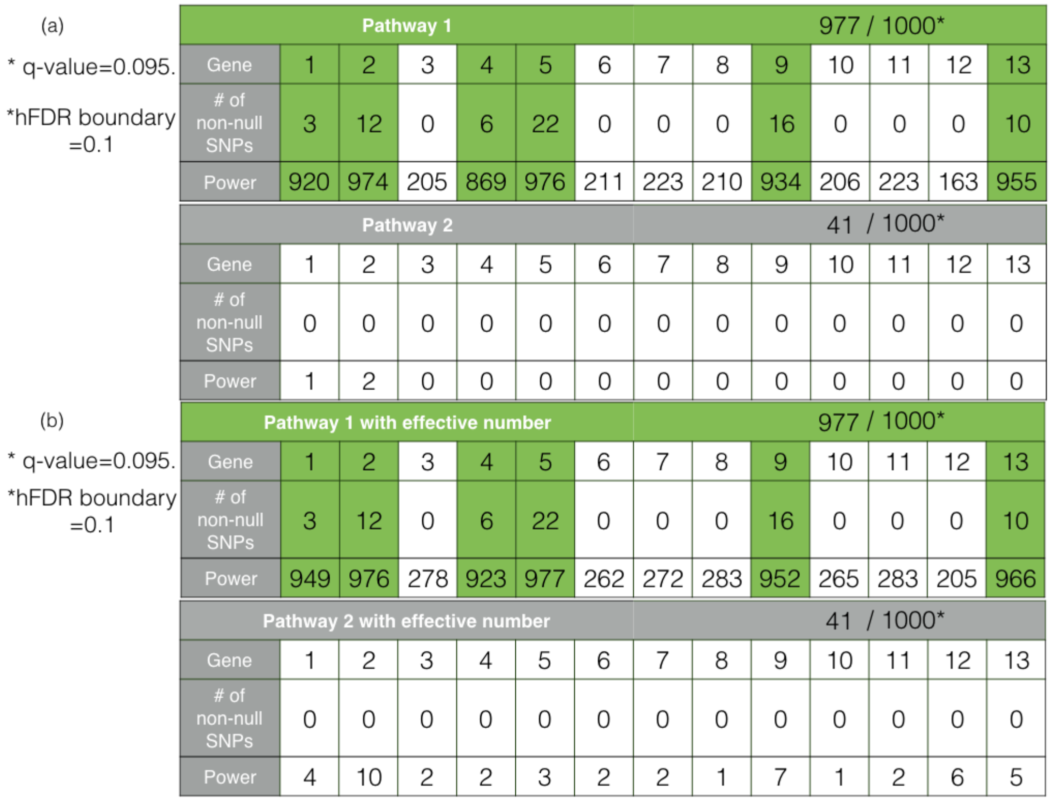


Figure 3.6. Setting 2, Sub-Setting 3 Results: Power for detecting the non-null pathway.

are associated as shown in Figure 3.12 . The uniform distributions used for generating the effect sizes are U(0.2, 0.4), U(0.1, 0.3), and U(0.05, 0.2) for hsa00910, hsa00062, and hsa04130, respectively, and SNPs with larger MAF are assigned larger effect sizes. The results are shown in Figure 3.13, which are somewhat different from the ones found in the previous settings. In particular, for pathway 1, GRASS has higher power than BHPathway at average FDR of 0.1 although they are similar at low FDR values. For the other two pathways, BHPathway maintains its power advantage over the other methods as seen in the preceding settings.



(a)

* q-value=0.095.

*hFDR boundary =0.1

Pathway 1							977 / 1000*						
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13
# of non-null SNPs	3	12	0	6	22	0	0	0	16	0	0	0	10
Power	920	974	205	869	976	211	223	210	934	206	223	163	955

Pathway 2							41 / 1000*						
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13
# of non-null SNPs	0	0	0	0	0	0	0	0	0	0	0	0	0
Power	1	2	0	0	0	0	0	0	0	0	0	0	0

(b)

* q-value=0.095.

*hFDR boundary =0.1

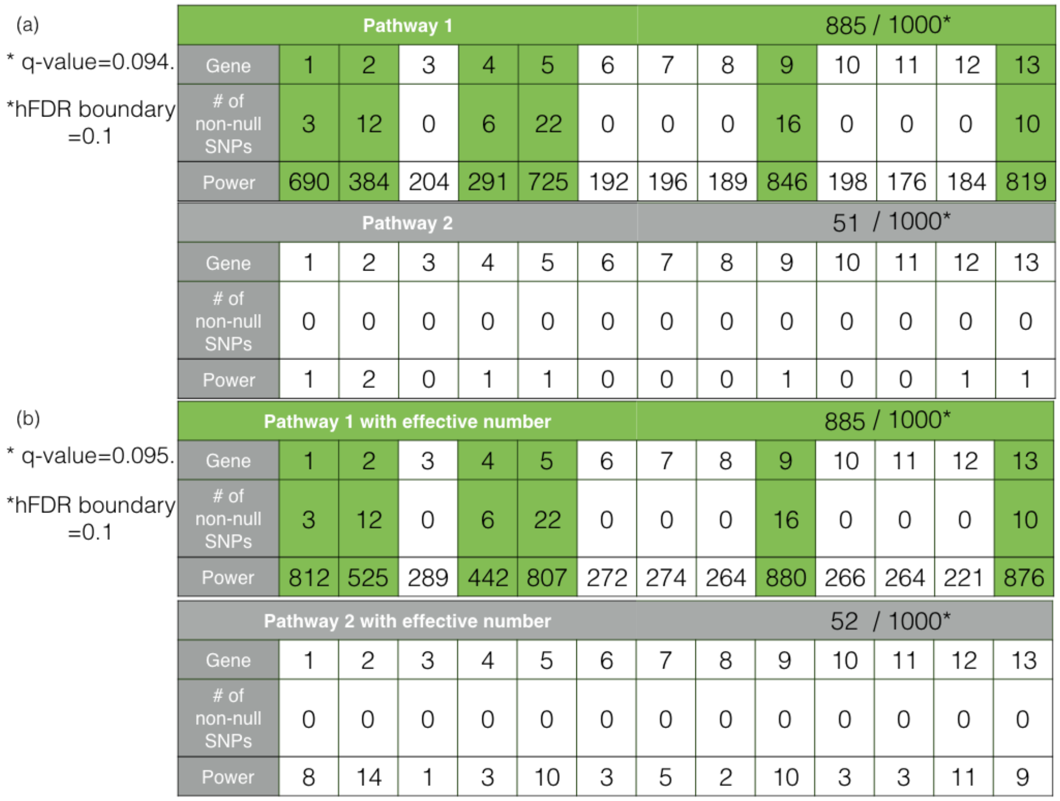
Pathway 1 with effective number							977 / 1000*						
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13
# of non-null SNPs	3	12	0	6	22	0	0	0	16	0	0	0	10
Power	949	976	278	923	977	262	272	283	952	265	283	205	966

Pathway 2 with effective number							41 / 1000*						
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13
# of non-null SNPs	0	0	0	0	0	0	0	0	0	0	0	0	0
Power	4	10	2	2	3	2	2	1	7	1	2	6	5

Figure 3.7. Setting 2, Sub-Setting 1 Results: The number of times each gene is detected once the pathway is detected (a) without and (b) with using the effective number of parameters. The numerator in the top row of each table is the number of times the pathway is found to be significant (out of 1000) and the numbers in the last row are the numbers of times the genes are found to be significant (out of the number of times the pathway is found to be significant). The hierarchical FDR of the entire inference procedure is about 0.1.

3.5 Setting 5

This setting is different from the previous ones in several respects. Here we consider a situation where only one pathway is to be tested. This is a relatively larger pathway following the structure of KEGG pathway hsa03022, which has 23 genes and 207 SNPs after matching with CEU population (as in Setting 2) and excluding SNPs with MAF < 0.01. One-half of the genes are set to be associated and within each associated gene, about one-half of the SNPs are set to be associated. The effect sizes for the non-null SNPs are generated from U(-0.1,



(a)

* q-value=0.094.

*hFDR boundary =0.1

Pathway 1						885 / 1000*							
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13
# of non-null SNPs	3	12	0	6	22	0	0	0	16	0	0	0	10
Power	690	384	204	291	725	192	196	189	846	198	176	184	819

Pathway 2						51 / 1000*							
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13
# of non-null SNPs	0	0	0	0	0	0	0	0	0	0	0	0	0
Power	1	2	0	1	1	0	0	0	1	0	0	1	1

(b)

* q-value=0.095.

*hFDR boundary =0.1

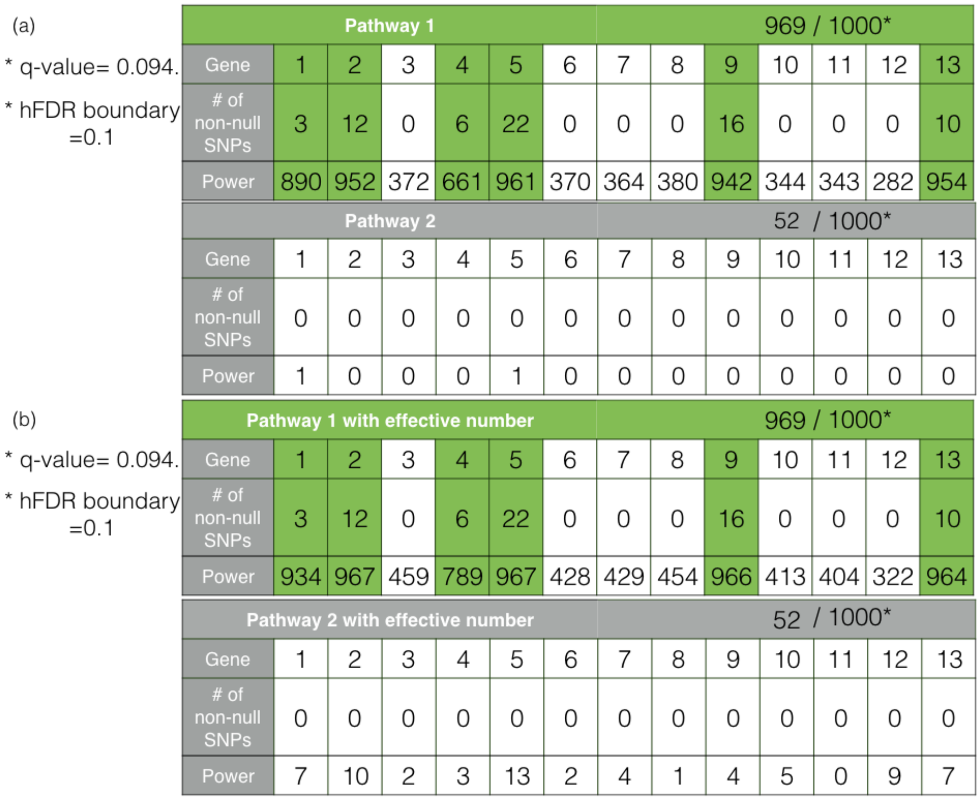
Pathway 1 with effective number						885 / 1000*							
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13
# of non-null SNPs	3	12	0	6	22	0	0	0	16	0	0	0	10
Power	812	525	289	442	807	272	274	264	880	266	264	221	876

Pathway 2 with effective number						52 / 1000*							
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13
# of non-null SNPs	0	0	0	0	0	0	0	0	0	0	0	0	0
Power	8	14	1	3	10	3	5	2	10	3	3	11	9

Figure 3.8. Setting 2, Sub-Setting 2 Results: The number of times each gene is detected once the pathway is detected (a) without and (b) with using the effective number of parameters. The numerator in the top row of each table is the number of times the pathway is found to be significant (out of 1000) and the numbers in the last row are the numbers of times the genes are found to be significant (out of the number of times the pathway is found to be significant). The hierarchical FDR of the entire inference procedure is about 0.1.

0.2) distribution with larger effect sizes assigned to SNPs were smaller MAF. Note that in this setting, we allow the effects to be negative as well as positive unlike in the previous settings, where all effects were positive.

As the inference procedure in BHPathway is mean-based, having SNP effects of opposite directions in a gene can cancel each other out nullifying the overall effect. To overcome this issue, we use a simple solution of checking the direction of effect of each SNP in a pre-processing step. For this, a simple logistic regression model is fitted for each SNP with



(a)

* q-value= 0.094.

* hFDR boundary =0.1

Pathway 1							969 / 1000*						
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13
# of non-null SNPs	3	12	0	6	22	0	0	0	16	0	0	0	10
Power	890	952	372	661	961	370	364	380	942	344	343	282	954

Pathway 2							52 / 1000*						
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13
# of non-null SNPs	0	0	0	0	0	0	0	0	0	0	0	0	0
Power	1	0	0	0	1	0	0	0	0	0	0	0	0

(b)

* q-value= 0.094.

* hFDR boundary =0.1

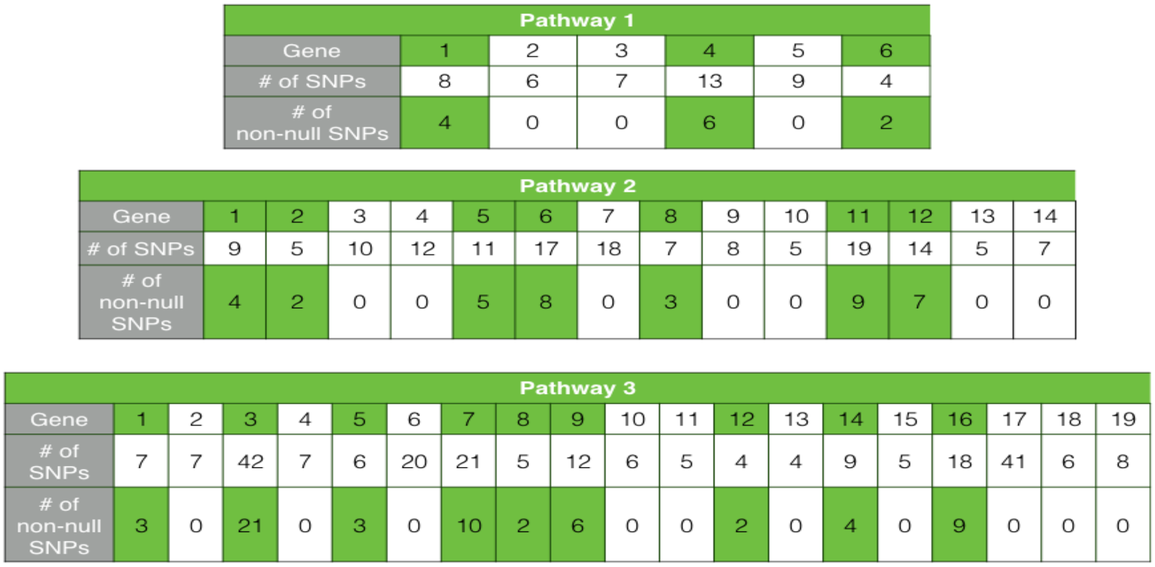
Pathway 1 with effective number							969 / 1000*						
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13
# of non-null SNPs	3	12	0	6	22	0	0	0	16	0	0	0	10
Power	934	967	459	789	967	428	429	454	966	413	404	322	964

Pathway 2 with effective number							52 / 1000*						
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13
# of non-null SNPs	0	0	0	0	0	0	0	0	0	0	0	0	0
Power	7	10	2	3	13	2	4	1	4	5	0	9	7

Figure 3.9. Setting 2, Sub-Setting 3 Results: The number of times each gene is detected once the pathway is detected (a) without and (b) with using the effective number of parameters. The numerator in the top row of each table is the number of times the pathway is found to be significant (out of 1000) and the numbers in the last row are the numbers of times the genes are found to be significant (out of the number of times the pathway is found to be significant). The hierarchical FDR of the entire inference procedure is about 0.1.

case/control status as the response, and the sign of the regression coefficient is checked. If the sign is negative, we interchange the minor and major alleles for that SNP so that the direction of the effect becomes positive. Note that this pre-processing step does not affect the strength of association and does not involve any statistical testing.

As there is only one pathway tested in this setting, FDR is not relevant. Instead, we estimate the type I error rate by simulating 1000 null pathways of the same structure, i.e., with all SNP effects set to be null. These are analyzed in exactly the same manner as the



Pathway 1						
Gene	1	2	3	4	5	6
# of SNPs	8	6	7	13	9	4
# of non-null SNPs	4	0	0	6	0	2

Pathway 2														
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13	14
# of SNPs	9	5	10	12	11	17	18	7	8	5	19	14	5	7
# of non-null SNPs	4	2	0	0	5	8	0	3	0	0	9	7	0	0

Pathway 3																			
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
# of SNPs	7	7	42	7	6	20	21	5	12	6	5	4	4	9	5	18	41	6	8
# of non-null SNPs	3	0	21	0	3	0	10	2	6	0	0	2	0	4	0	9	0	0	0

Figure 3.10. Setting 3: Structure of the three non-null pathways. There are also three null pathways of the same structure with all SNPs being null (not shown). The components in green are non-null.

non-null pathway. For both the null and non-null settings, we calculate the percentage of times (out of 1000) that the pathway is found to be significant, which gives us power and type I error rate at a given cutoff for p-value. The cutoff is then varied to get an ROC curve. In this setting, we do not include ALIGATOR as it requires at least two pathways to be tested together to give sensible results (because it tests for competitive null hypothesis). Figure 3.14 shows the results for the other three approaches. Here we see that BHPathway and PLINK have similar powers while GRASS performs worse.

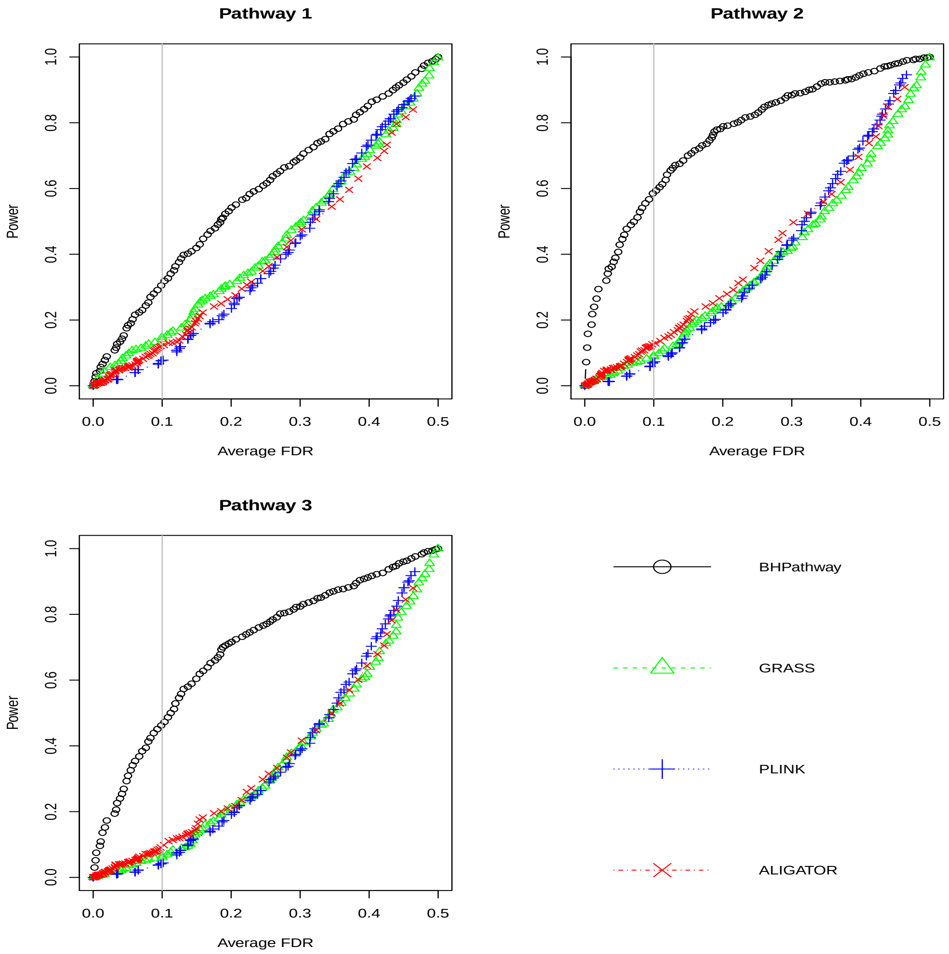
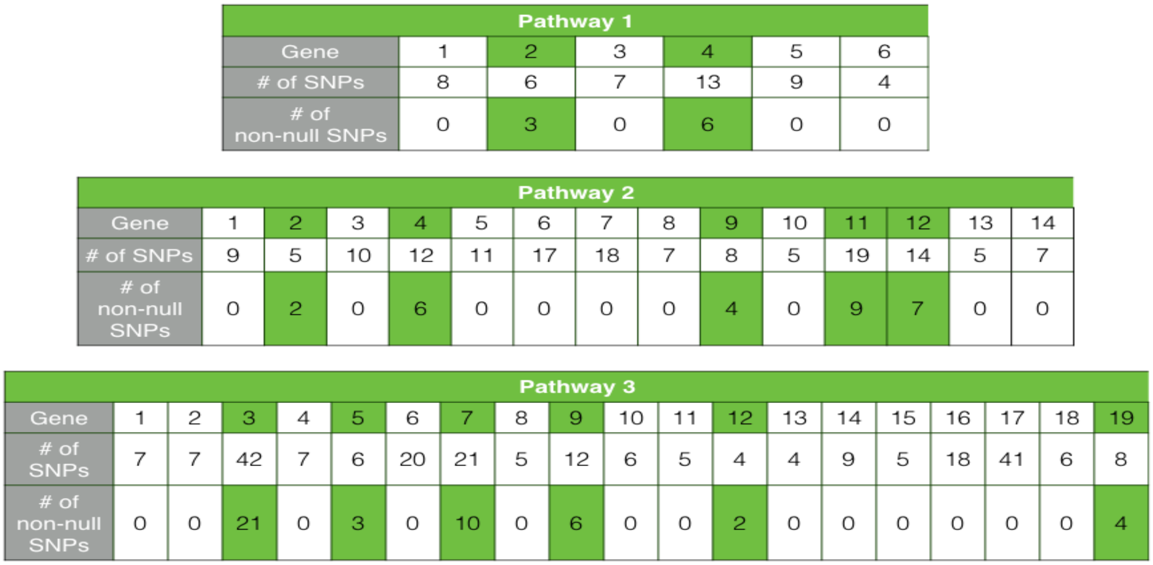


Figure 3.11. Setting 3 Results: Power for detecting the three non-null pathways.



Pathway 1						
Gene	1	2	3	4	5	6
# of SNPs	8	6	7	13	9	4
# of non-null SNPs	0	3	0	6	0	0

Pathway 2														
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13	14
# of SNPs	9	5	10	12	11	17	18	7	8	5	19	14	5	7
# of non-null SNPs	0	2	0	6	0	0	0	0	4	0	9	7	0	0

Pathway 3																			
Gene	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
# of SNPs	7	7	42	7	6	20	21	5	12	6	5	4	4	9	5	18	41	6	8
# of non-null SNPs	0	0	21	0	3	0	10	0	6	0	0	2	0	0	0	0	0	0	4

Figure 3.12. Setting 4: Structure of the three non-null pathways. There are also three null pathways of the same structure with all SNPs being null (not shown). The components in green are non-null.

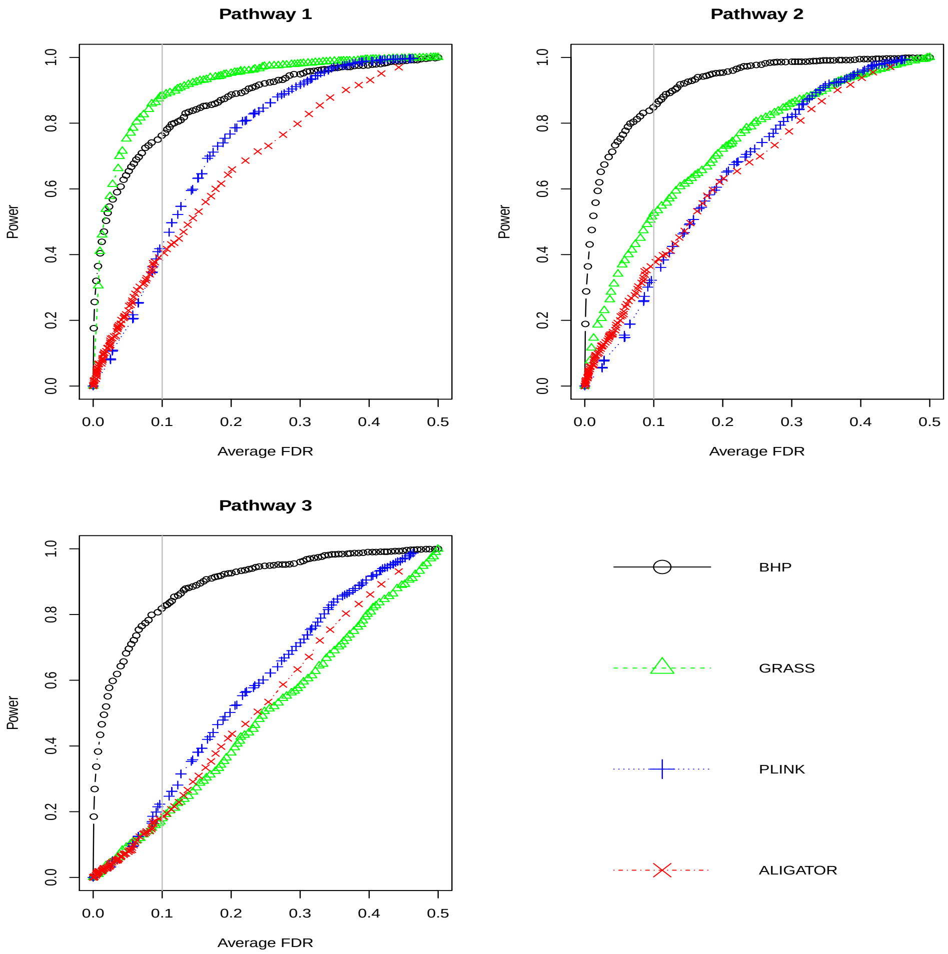


Figure 3.13. Setting 4 Results: Power for detecting the three non-null pathways.

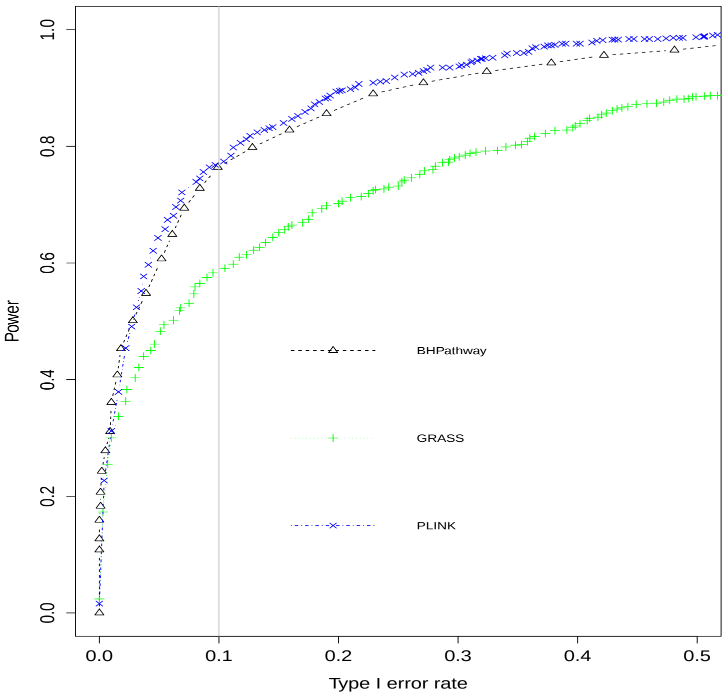


Figure 3.14. Setting 5 Results: Power for detecting the non-null pathway.

CHAPTER 4

REAL DATA APPLICATIONS

4.1 Application to Breast Cancer Data

We consider GWAS data CGEMS (Cancer Genetics Markers of Susceptibility) on breast cancer obtained from dbGaP (accession number phs00147.v3.p1). A total of 555,351 SNP genotypes were available. We apply commonly used quality control criteria to filter SNPs using PLINK (Purcell et al., 2007). In particular, SNPs with missing genotype rate higher than 10%, minor allele frequency lower than 1%, and failing the Hardy-Weinberg equilibruim test with p-values less than 0.001 were removed. Also, SNPs were pruned based on LD using R^2 threshold of 0.5 to obtain a final set of SNPs for analysis. We started with all pathways from the KEGG database and those with gene counts between 10 and 300 are selected giving 298 pathways for analysis (Menashe et al., 2010; Chen et al., 2010). The start and end points of all genes are obtained from the Ensembl genome browser (Ensembl Genome Browser, 2018). SNPs are mapped to genes if they locate within 10 kb upstream and downstream of a gene region. A total of 317,643 SNPs are finally used in the analysis.

There are 1,145 cases and 1,142 controls available. However, each SNP has some missing genotypes among the 2,287 individuals, and so we analyzed each pathway separately using only the subset of individuals who had no missing genotype on all the SNPs within the particular pathway. Thus, the total number of individuals with complete genotype used for analysis is different across the 298 pathways. Summary statistics about the numbers of genes and SNPs in each pathway and the number of individuals are provided in Table 4.1.

Table 4.1. Summary on gene counts, SNP counts, and subjects across 298 KEGG pathways for the Breast Cancer data.

Summary of gene count						
Summary Statistics	**Min**	**1st**	**Median**	**Mean**	**3rd**	**Max**
Overall for 298 pathways	7	33.25	60	71.48	92.75	243
Summary of SNP count						
Summary Statistics	**Min**	**1st**	**Median**	**Mean**	**3rd**	**Max**
Overall for 298 pathways	42	274.5	724	1066	1643	4190
Summary on number of individuals						
Summary Statistics	**Min**	**1st**	**Median**	**Mean**	**3rd**	**Max**
Overall for 298 pathways	218	630.2	1121	1118.8	1554.8	2149

In a preliminary single-SNP analysis using PLINK, we find that Z scores for SNPs could be both positive and negative. As the inference of BHPathway is mean-based, opposite effects may cancel out each other. So, for SNPs with negative effect, we interchanged the minor and major alleles. Note that this pre-processing step does not affect the strength of association of the SNP or its gene or pathway.

Table 4.2 shows the top two pathways that BHPathway found below 0.05 cutoff for p-value. Among them, the top pathway hsa04910 is Insulin Signaling pathway. This pathway and its components have been reported to be associated with breast cancer, and cancer, in general, in several molecular and gene expression studies (Rostoker et al., 2015; Poloz and Stambolic, 2015; Belfiore and Malaguarnera, 2011; Djiogue et al., 2013). A recent study conducted a gene-based analyses of genes in several insulin related pathways in relation to breast cancer and reported several significant genes (Ruiz-Narvaez et al., 2016). It has been reported that insulin signaling receptor is often overexpressed in tumor cells, particularly that of the breast (Poloz and Stambolic, 2015). Indeed therapies have been developed to target insulin-like growth factor-I receptor and/or insulin receptor pathways for breast cancer (Law et al., 2008). Our finding of significance of this pathway may be the first of its kind based on GWAS SNP data.

We also tested this pathway and its components using hierarchical FDR boundary values of 0.1 and 0.2 (as given in equation (2.4)), both with and without using the effective numbers of parameters. The pathway was significant at the FDR boundary of 0.1 (both with and without using effective number of parameters), however, none of its component genes were found to be significant even at 0.2 FDR value. The two top most significant genes were AKT3 and MAPK1 with BHPathway p-values of 0.043 and 0.082, respectively. These genes regulate processes such as cell proliferation, cell division, migration, and apoptosis (Rostoker et al., 2015). AKT3 has been implicated by several studies to be involved with growth of triple negative breast cancer, an aggressive subtype of breast tumor with a poor outcome, and therapeutic targeting of this gene has been suggested as a novel treatment option (Chin et al., 2014; O'Hurley et al., 2014; Grottke et al., 2016; Hu et al., 2018). MAPK1 has been shown to be associated with breast cancer and other cancers (Slattery et al., 2014; Reyes-Gibby et al., 2016; Li et al., 2015).

The second pathway found to be significant by BHPathway is hsa04977 named Vitamin Digestion and Absorption. This result is also consistent with literature reporting connection of vitamins with cancer (Deeb et al., 2007; Chou et al., 2011; Chen et al., 2014). This pathway was significant at hierarchial FDR of 0.1, however, its genes were not found to be significant.

The gene with the smallest p-value (of 0.06) is CUBN. An earlier study had found a SNP in this gene to be asssociated with breast cancer (Anderson et al., 2011). The gene with the next smallest p-value of 0.1 is SLC19A3, and this gene has been also implicated in some studies (Sweet et al., 2010; Cheuk et al., 2015; Ng et al., 2011). As Table 4.2 shows, GRASS and PLINK do not find these two pathways to be significant. However, GRASS and PLINK find several other pathways to be significant at 5% level as reported in Table 4.3. An interesting observation though is that none of the significant pathways detected by one method was found to be significant by another method. This suggests that perhaps the methods are complementary in the sense that they have different powers for detecting different types of pathways with different types of data. We did not apply ALIGATOR to these data as each pathway is tested individually while ALIGATOR, being a competitive test, requires at least two pathways to be tested together. Table 4.4 shows the pathways that BHPathway, GRASS, or PLINK found with p-values between 0.05 and 0.1.

Table 4.2. Pathways with p-values not exceeding 0.05 using BHPathway applied on the Breast Cancer data.

KEGG pathway ID	**BHPathway**	**GRASS**	**PLINK**
hsa04910	0.009	0.593	0.14
hsa04977	0.041	0.849	0.76

4.2 Application to Renal Cancer Data

We consider the National Cancer Institute's (NCI) GWAS data on renal cell carcinoma obtained from dbGaP (accession number phs000351.v1.p1). It consists of four studies with three being prospective cohort studies and one case-control study. The three cohort studies are

Table 4.3. Pathways with p-values not exceeding 0.05 using GRASS or PLINK applied on the Breast Cancer data.

KEGG pathway ID	**BHPathway**	**GRASS**	**PLINK**
hsa05146	0.764	0.008	0.62
hsa00900	0.325	0.01	0.4
hsa00770	0.611	0.014	0.3
hsa00030	0.702	0.022	0.63
hsa05323	0.634	0.029	0.1
hsa04380	0.735	0.031	0.77
hsa00600	0.404	0.039	0.19
hsa04621	0.622	0.041	0.24
hsa04062	0.767	0.044	0.21
hsa04012	0.875	0.044	0.58
hsa00561	0.321	0.681	0.002
hsa00340	0.13	0.418	0.003
hsa04930	0.646	0.841	0.007
hsa04914	0.228	0.266	0.011
hsa00601	0.356	0.542	0.014
hsa04916	0.75	0.124	0.01
hsa04015	0.85	0.281	0.01
hsa00330	0.423	0.851	0.02
hsa05206	0.73	0.22	0.03
hsa05211	0.349	0.437	0.03
hsa00515	0.591	0.759	0.03
hsa04340	0.703	0.98	0.03
hsa00052	0.396	0.097	0.04
hsa00590	0.213	0.362	0.04
hsa05100	0.759	0.83	0.04
hsa00140	0.904	0.84	0.04
hsa00565	0.797	0.563	0.05
hsa04973	0.945	0.737	0.05
hsa00630	0.703	0.919	0.05

Alpha-Tocopherol and Beta-Carotene Cancer Prevention Study (ATBC), American Cancer Society Cancer Prevention Study-II (CPS-II), and Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO) while the case-control study is NCI's US Kidney Cancer Study (USKC). We analyzed 610K subset of the renal cancer data genotyped on Illumina

610 BeadChips at the NCI Core Genotyping Facility. A total of 537,091 SNPs are mapped to the same 298 KEGG pathways that we previously analyzed for the breast cancer data. We applied the same filtering criteria as before to get 258,309 SNPs. There were a total of 1,312 cases and 3,424 controls with genotypes. However, as before, the total number of subjects used in the analysis of each pathway varied due to subjects having missing genotypes.

When we applied the methods to the combined dataset from all four studies, we got a highly unexpected result — all 298 pathways were significant using BHPathway, PLINK, and GRASS (most with extremely small p-values). As this result does not appear to be reasonable, we examined the four studies closely and found that USKC case-control data were actually collected using a complex sampling scheme (Colt et al., 2011; Zhang et al., 2017). Unless a statistical approach explicitly accounts for such a complex sampling design, the results can be biased and false positive rates can become uncontrollably high (Korn and Graubard, 1999; Zhang et al., 2017). As BHPathway as well as the other two methods do not account for complex sampling, we excluded this dataset and analyzed the three cohort studies, which reduced the sample size to 650 cases and 2,863 controls. Table 4.5 shows the summary statistics based on the combined dataset from three cohort studies only.

Table 4.6 shows the top two pathways found by BHPathway to be significant at 5% level. The first pathway hsa03050 is Proteasome pathway. This pathway has been implicated in the literature for its role in renal cancer and cancer, in general (Mani and Gelmann, 2005; Corn, 2007; Frezza et al., 2011). As in the breast cancer data analyses, we applied hierarchical FDR criteria to this pathway and found it to be significant at FDR boundary of 0.1 (both with and without using the effective number of parameters), however, none of its genes were significant. On increasing the hierarchical FDR boundary to 0.2 and using the effective number of parameters, we found 19 genes to be significant (none were significant without using the effective number of parameters). These genes are listed in Table 4.7. The top most significant gene in this pathway is PSMB7. A recent study reports that this gene along with some other constitutive proteasome genes are over-expressed in most cancer types (Rouette et al., 2016). In fact, three other genes reported in this study, namely, PSMB8, PSMB9, and PSMB10 are also listed in Table 4.7.

The second pathway in Table 4.6 is Inflammatory Bowel Disease and the most significant gene found in this pathway is IL4R with p-value of 0.15. The pathway was significant at hierarchical FDR boundary of 0.1, however, the gene was not significant even at FDR boundary value of 0.2. Some studies have shown association between renal cancer and this pathway and IL4R gene (Romano et al., 2016; Derikx et al., 2015; Obiri et al., 1993).

Table 4.8 shows the pathways that GRASS or PLINK found below 0.05 cutoff for p- value and Table 4.9 shows the pathways that BHPathway, GRASS, or PLINK found with p-values between 0.05 and 0.1 in the renal cancer data.

Table 4.4. Pathways with p-values between 0.05 and 0.1 using BHPathway, GRASS, or PLINK applied to the Breast Cancer data.

KEGG pathway ID	**BHPathway**	**GRASS**	**PLINK**
hsa04659	0.062	0.757	0.85
hsa00072	0.07	0.927	0.38
hsa04130	0.072	0.701	0.86
hsa04657	0.073	0.464	0.73
hsa00910	0.078	0.584	0.3
hsa00062	0.08	0.877	0.85
hsa04976	0.082	0.522	0.48
hsa04915	0.633	0.055	0.96
hsa00230	0.925	0.056	0.77
hsa04142	0.166	0.058	0.69
hsa05210	0.214	0.058	0.22
hsa01524	0.559	0.058	0.39
hsa03022	0.638	0.062	0.06
hsa00534	0.329	0.065	0.12
hsa03420	0.837	0.065	0.69
hsa04742	0.532	0.069	0.5
hsa04975	0.431	0.07	0.45
hsa05224	0.919	0.074	0.39
hsa05030	0.43	0.076	0.83
hsa03040	0.963	0.08	0.07
hsa04921	0.956	0.081	0.61
hsa04210	0.379	0.082	0.1
hsa04922	0.625	0.095	0.48
hsa00052	0.396	0.097	0.04
hsa00860	0.76	0.099	1
hsa03022	0.638	0.062	0.06
hsa00670	0.335	0.197	0.06
hsa05222	0.604	0.667	0.06
hsa05230	0.348	0.79	0.06
hsa00512	0.231	0.991	0.06
hsa03040	0.963	0.08	0.07
hsa04727	0.685	0.329	0.07
hsa05212	0.649	0.42	0.07
hsa05142	0.881	0.694	0.07
hsa04550	0.766	0.189	0.08
hsa03060	0.312	0.257	0.08
hsa04071	0.73	0.347	0.08
hsa01040	0.565	0.148	0.09
hsa03008	0.266	0.607	0.09
hsa05162	0.942	0.657	0.09
hsa00410	0.869	0.677	0.09

Table 4.5. Summary on gene and SNP counts across 298 KEGG pathways for Renal Cancer data.

Summary of gene count						
Summary Statistics	**Min**	**1st**	**Median**	**Mean**	**3rd**	**Max**
Overall for 298 pathways	9	34.5	63.5	74.7	99.75	256
Summary of SNP count						
Summary Statistics	**Min**	**1st**	**Median**	**Mean**	**3rd**	**Max**
Overall for 298 pathways	50	282	645.5	866.8	1268.2	3280
Summary on number of individuals						
Summary Statistics	**Min**	**1st**	**Median**	**Mean**	**3rd**	**Max**
Overall for 298 pathways	169	691.2	1003	1016	1369	1878

Table 4.6. Pathways with p-values not exceeding 0.05 using BHPathway applied to the Renal Cancer data.

KEGG pathway ID	**BHPathway**	**GRASS**	**PLINK**
hsa03050	0.027	0.421	0.61
hsa05321	0.051	0.208	0.39

Table 4.7. Genes in pathway hsa03050 found to be significant using the effective number of parameters at hierarchical FDR boundary of 0.2 in the Renal Cancer data.

Entrez Gene ID	**P-value**
5695	0.109
5719	0.117
5700	0.125
5689	0.128
5698	0.128
5684	0.15
9861	0.151
5682	0.152
5721	0.157
5696	0.158
5701	0.162
5699	0.163
11047	0.169
5709	0.170
5706	0.172
5688	0.174
5713	0.178
51371	0.178
5710	0.18

Table 4.8. Pathways with p-values not exceeding 0.05 using GRASS, or PLINK applied to the Renal Cancer data.

KEGG pathway ID	**BHPathway**	**GRASS**	**PLINK**
hsa04540	0.956	0	0
hsa04530	0.980	0	0
hsa05416	0.548	0.002	0.01
hsa04520	0.816	0.002	0.01
hsa04550	0.5	0.003	0.02
hsa04512	0.659	0.004	0.02
hsa04510	0.854	0.01	0.04
hsa04210	0.578	0.012	0.04
hsa00310	0.821	0.014	0.04
hsa04921	0.861	0.014	0.05
hsa04911	0.776	0.016	0.05
hsa00430	0.616	0.017	0.06
hsa05210	0.666	0.017	0.06
hsa03013	0.985	0.018	0.07
hsa05340	0.579	0.019	0.08
hsa04114	0.977	0.019	0.09
hsa04660	0.909	0.022	0.1
hsa04621	0.618	0.024	0.1
hsa04340	0.887	0.024	0.11
hsa00511	0.342	0.042	0.12
hsa00052	0.898	0.042	0.13
hsa04664	0.511	0.045	0.13
hsa00563	0.135	0.05	0.13

Table 4.9. Pathways with p-values between 0.05 and 0.1 using BHPathway, GRASS, or PLINK applied to the Renal Cancer data.

KEGG pathway ID	**BHPathway**	**GRASS**	**PLINK**
hsa04390	0.056	0.935	0.99
hsa04918	0.057	0.534	0.73
hsa00220	0.058	0.243	0.42
hsa00760	0.080	0.597	0.8
hsa04966	0.092	0.063	0.15
hsa05330	0.096	0.688	0.84
hsa04371	0.991	0.057	0.14
hsa04933	0.964	0.059	0.15
hsa05418	0.605	0.064	0.15
hsa05412	0.574	0.072	0.17
hsa01040	0.989	0.073	0.17
hsa00970	0.997	0.075	0.17
hsa04514	0.719	0.082	0.17
hsa00860	0.709	0.084	0.18
hsa04950	0.613	0.086	0.19
hsa00062	0.359	0.088	0.19
hsa05152	0.574	0.089	0.2

CHAPTER 5

DISCUSSION AND FUTURE WORK

5.1 Discussion

We have proposed a novel approach, BHPathway, for testing pathway association using GWAS data. Unlike most current approaches that rely on ad hoc ways of combining information across various levels of hierarchy, we propose a unified hierarchical model connecting all three levels of hierarchy naturally inherent in a pathway structure in a seamless manner. This is achieved by modeling the effects of each level conditional on the effects at the preceding level in a GLM framework. To handle the high dimensionality, the regression coefficients are regularized using hierarchical t priors. The computational intensity of fitting such a large and unified model is controlled by utilizing a combination of IWLS and EM algorithms to estimate the posterior modes and thereby forgoing the use of computationally intensive MCMC algorithms.

From our simulation studies, we find that BHPathway can have higher power than the other commonly used pathway analysis methods in many cases where there are multiple variants of modest size. On the other hand, it may have comparable or lower power in some cases. Thus, the methods may be complementary. In fact, this observation is also supported by our real data analyses wherein we found that the different approaches implicated different pathways. Thus, in a real data analysis, it may be worth applying several methods to utilize their strengths in uncovering different types of associations. Nonetheless, BHPathway has the additional advantage of being able to pinpoint the genes of interest when a pathway is found to be significant as we illustrated in both simulated and real data analyses. For this, the inference has to be carried out in a hierarchical manner using hierarchical FDR. Moreover, we found that the use of the effective number of parameters instead of the total number of parameters in the multiplicity adjustment using hierarchical FDR helps in increasing the power of detection at the gene level.

Missing data is a major challenge in the analysis of real data. In our datasets, we had to analyze each pathway individually even though, in principle, BHPathway can analyze several pathways jointly (as illustrated with simulated data). One possible way to handle missing genotype data is to make use of the imputed genotypes that come with the GWAS data from dbGaP. We attempted to do that, however, and found that the imputed set did not overlap with the observed genotype set so that two sets cannot be merged to find and fill-in the genotypes of SNPs/individuals missing in the observed data. In such cases, an alternative could be to abandon the observed genotype set and work with the imputed set only.

5.1.1 Sensitivity Analysis of Hyper-Parameters

The hyper-parameters $\nu_0 = 1,\ s_0 = 10,\ a = 6.25,\ b = 1,\ \nu_\beta = 1,\ \nu_\xi = 1,\ \nu_\theta = 1,\ s_\xi = 0.04$, and $s_\theta = 2.5$ is the set of values we currently choose after sensitivity check on how the change of these parameters affect the pathway p-values.

Increasing ν_β and ν_ξ causes very minor change on pathway p-values. We choose $\nu_\beta = 1,\ \nu_\xi = 1$ as they have been used earlier (Gelman et al., 2008; Yi and Ma, 2012).

Tuning $a, b, \nu_\theta, s_\theta$ for the top pathways that BHPathway found significant for the renal cancer data, there is barely any difference observed on the pathway p-values. We choose $\nu_\theta = 1,\ \nu_0 = 1,\ s_0 = 10$ as suggested by (Gelman et al., 2008).

Both $s^2_{\beta_p}$ and s^2_θ are the scale parameters from Inv-χ^2 distribution. For s_θ, it has been suggested to use a value of 2.5 by (Gelman et al., 2008). Unlike s^2_θ being a fixed value, $s^2_{\beta_p}$ is a random variable. As a result, $a = 6.25$ and $b = 1$ are chosen to yield a prior mean of $s^2_{\beta_p}$ from $Gamma(a, b)$ to be 2.5^2, which is 6.25.

Tuning s_ξ has a relatively stronger effect on the pathway p-value. For the majority of the pathways tested using the simulated data as well as real data, a smaller value of s_ξ leads to a smaller pathway p-value. As an example, we plot ROC curves in Figure 5.1 by varying s_ξ values in a simulated dataset. We see that $s_\xi = 0.04$ is a suitable choice. Similar pattern is present in the two real datasets we analyzed earlier. Table 5.1 shows the change in hsa00760 pathway p-value under different s_ξ values. We choose to use the s_ξ value of 0.04 as it exhibits a reasonable ability in distinguishing between the p-values of null and non-null genes in a pathway (results not shown).

5.2 Future Work

5.2.1 Overlapping SNPs/Genes in Multiple Genes/Pathways

It is a common situation that a SNP is mapped to multiple genes and/or a gene is mapped to multiple pathways. To handle this issue, a potential idea is described in the following (Zhang et al., 2014). In particular, we consider the case of a gene being mapped to three different pathways. The same ideas may be used for a SNP mapped to multiple genes.

Table 5.1. Pathway p-value for hsa00760 in the Renal Cancer data under varying s_ξ values.

s_ξ **value**	**Pathway p-value**
0.005	0.073
0.01	0.073
0.015	0.074
0.02	0.074
0.025	0.075
0.03	0.077
0.035	0.078
0.04	0.08
0.045	0.082
0.05	0.084
0.055	0.086
0.06	0.089
0.065	0.091
0.07	0.094
0.075	0.097
0.08	0.1
0.085	0.103
0.09	0.106
0.095	0.109
0.1	0.112

Let gene effect $\mu \sim N(\theta_*, \tau_*^2)$. Here θ_* and τ_*^2 are pathway-specific parameters. Their values are unclear because the gene maps to three pathways. We define μ_1, μ_2, and μ_3 as the proportions of μ mapped to pathways 1 to 3. Let $\mu_i = w_i \cdot \mu$ and $\mu_i \sim N(w_i\theta_*, w_i^2\tau_*^2)$ for $i = 1, 2, 3$ with $\sum w_i = 1$ and $\sum w_i^2 = 1$. Also, as μ_1, μ_2, and μ_3 represent effects of the gene in pathways 1 to 3, we have $\mu_i \sim N(\theta_i, \tau_i^2)$, $i = 1, 2, 3$. Thus, by comparing the prior means

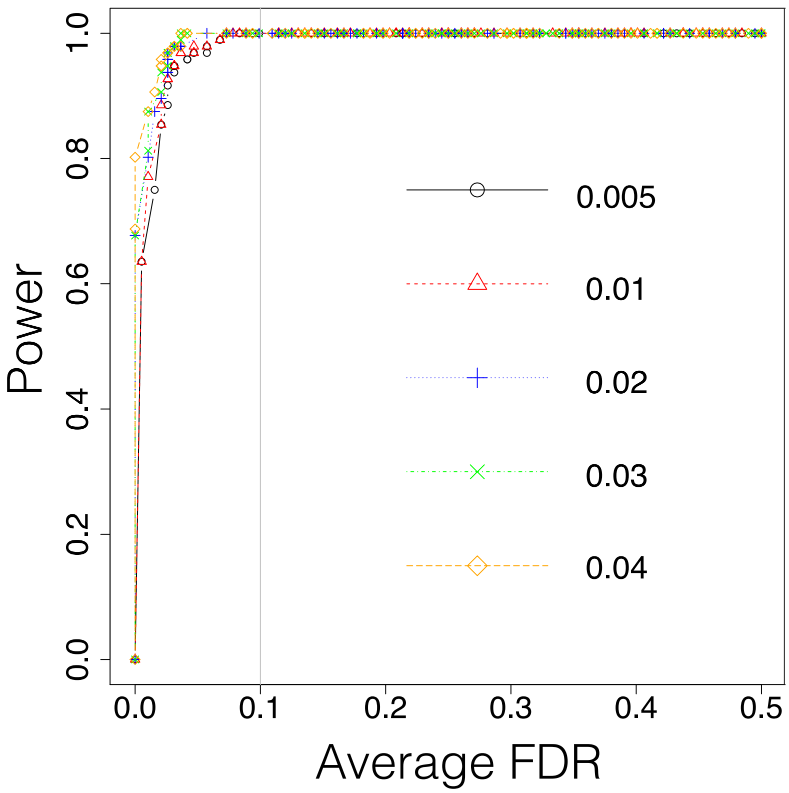


Figure 5.1. ROC curves with varying s_ξ values.

and prior variances of $\mu_1,\ \mu_2$, and μ_3, we have:

$$w_1 \cdot \theta_* = \theta_1, \quad w_1^2 \cdot \tau_*^2 = \tau_1^2,$$

$$w_2 \cdot \theta_* = \theta_2, \quad w_2^2 \cdot \tau_*^2 = \tau_2^2,$$

$$w_3 \cdot \theta_* = \theta_3, \quad w_3^2 \cdot \tau_*^2 = \tau_3^2,$$

$$w_1 \cdot \theta_* + w_2 \cdot \theta_* + w_3 \cdot \theta_* = \theta_*, \text{ and } \ w_1^2 \cdot \tau_*^2 + w_2^2 \cdot \tau_*^2 + w_3^2 \cdot \tau_*^2 = \tau_*^2.$$

The solutions to the left four equations are $w_i = \theta_i/\theta_*$ for $i = 1, 2, 3$ and $\theta_* = \theta_1 + \theta_2 + \theta_3$. The solutions to the right four equations are $w_i^2 = \tau_i^2/\tau_*^2$ for $i = 1, 2, 3$ and $\tau_*^2 = \tau_1^2 + \tau_2^2 + \tau_3^2$. However, even though we obtain solutions to θ_* and τ_*^2, the choice of w_i is not unique and is an open question.

5.2.2 Extension to Quantitative Trait

As our method is GLM based, it can easily incorporate phenotypes of continuous, categorical, or count types. Also, the effects of covariates can be incorporated.

In GLM, an additional *dispersion parameter* τ^2 can be present depending on the distribution of response. In that case, $\mathrm{var}(y_i) = \tau^2 V(\mu_i)$, where $V(\mu_i)$ is the *variance function*. Note that $\tau^2 = 1$ for binomial distribution (in the logistic regression model used earlier). When $\tau^2 \neq 1$, we can assign a uniform prior distribution for τ^2 over a finite, known interval. It can be estimated as in classical GLM (McCullagh and Nelder, 1989).

5.2.3 Extension to Multiple Dependent Phenotypes

To extend the method to deal with multiple responses, we consider a similar GLM-based hierarchical model for each response and add a random subject effect in each model to induce dependence between the responses. Suppose there are Q phenotypes under study, each having a distribution belonging to an exponential family. The observed responses are y_{iq}, $i = 1, \ldots, n$, $q = 1, \ldots, Q$. The model allows for some subjects to have missing values on some phenotypes. Under the GLM framework, for q^{th} phenotype, we define $\mu_{iq} = E(y_{iq})$, g_q as the link function, $\eta_{iq} = g_q(\mu_{iq})$ as the linear predictor, $V_q(\mu_{iq})$ as the variance function, and τ_q^2 the dispersion parameter. Let $\boldsymbol{\beta}_q$ be the vector of regression coefficients for effects of S SNPs on the q^{th} phenotype and β_{0q} be the corresponding intercept. Let u_i be the random effect of the ith subject. We model $\eta_{iq} = g_q(\mu_{iq}) = \beta_{0q} + \mathbf{x}_i' \boldsymbol{\beta}_q + u_i$, $i = 1, \ldots, n$, $q = 1, \ldots, Q$. Let $\mathbf{y}_q = (y_{1q}, \ldots, y_{nq})$, $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_Q)$, $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0Q})$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_Q)$, and $\boldsymbol{u} = (u_1, \ldots, u_n)$.

Assume that $\boldsymbol{\beta}_0 | \sigma_0^2 \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I})$, and $\boldsymbol{u} | \sigma_u^2 \sim N(0, \sigma_u^2 \boldsymbol{I})$ with $\sigma_u^2 \sim \text{Inv-}\chi^2(\nu_u, s_u^2)$, where ν_u and s_u^2 are fixed. The hierarchical prior structure for the Q sets of SNP effects in this model and also the effects of genes and pathways are assumed to be the same as in (2.1). This allows model parsimony as well as borrowing of information across the related phenotypes. Specifically, $\boldsymbol{\beta}_q$ has the same prior as that of the SNP effects in model (2.1). Thus, the gene effects ξ_{jp} and pathway effects θ_p are common to all phenotypes. However, note that as SNP effects vary by phenotype, by the virtue of the assumed hierarchical structure, it induces the gene and pathway effects to also vary by phenotype indirectly. The other hyper-parameters, including $\sigma_0^2, \sigma_{\beta jp}^2, \sigma_{\xi p}^2$, and σ_θ^2, and their priors are the same as in model (2.1). Let $\tilde{\boldsymbol{y}}_q$ denote the pseudodata vector for the qth phenotype, and $\tilde{\boldsymbol{y}} = (\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_Q)$.

Let $\boldsymbol{\beta}^* = (\beta_0, \boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\xi}, \boldsymbol{\theta})$, $\boldsymbol{\phi} = (\sigma_0^2, \boldsymbol{\sigma}_\beta^2, \sigma_u^2, \boldsymbol{\sigma}_\xi^2, \sigma_\theta^2, \mathrm{s}_\beta^2)$, and $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_Q^2)$. Now we proceed asbefore to find the conditional posterior mode of $\pi(\boldsymbol{\beta}^* | \boldsymbol{\phi}, \boldsymbol{\tau}^2, \mathbf{y})$. Consider a linear model $\boldsymbol{Y}^* = \boldsymbol{X}^* \boldsymbol{\beta}^* + \boldsymbol{\epsilon}^*$, where $\boldsymbol{Y}^* = (\tilde{\boldsymbol{y}},\ \mathbf{0})$ with $\mathbf{0}$ of order $(Q + QS + n + J + P)$, and $\boldsymbol{X}^*$ has a similar structure as in (2.2) but it additionally includes the usual design matrices $\boldsymbol{X}$ for the Q models appended together as a block diagonal matrix and also n rows and columns associated with $\boldsymbol{u}$. The matrix $\boldsymbol{X}$ corresponding to each phenotype will be exactly the same with dimension $n \times (S + 1)$. Further, $\boldsymbol{\epsilon}^*$ is distributed as $N(\mathbf{0}, \boldsymbol{W}^{-1})$, where $\boldsymbol{W} =$

diag $\{\mathbf{W}_{y1}, \dots, \mathbf{W}_{yQ}, \mathbf{W}_0, \mathbf{W}_\beta, \dots, \mathbf{W}_\beta, \mathbf{W}_u, \mathbf{W}_\xi, \mathbf{W}_\theta\}$ is of order $nQ + Q + SQ + n + J + P$. Each component of $\mathbf{W}$ is itself a diagonal matrix and $\mathbf{W}_{yq}$ has a similar formula as for one phenotype. The diagonal elements of $\mathbf{W}_\beta$ are reciprocals of the prior variances $\sigma^2_{\beta jp}$. Likewise, the diagonal elements of $\mathbf{W}_0$, $\mathbf{W}_u$, $\mathbf{W}_\xi$, and $\mathbf{W}_\theta$ are σ_0^{-2}, σ_u^{-2}, $\sigma_{\xi p}^{-2}$, and σ_θ^{-2}, respectively. As before, this linear model representation allows using an IWLS algorithm to obtain the approximate conditional posterior mode of $\boldsymbol{\beta}^*$ as $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\phi}, \boldsymbol{\tau}^2) = (\boldsymbol{X}^{*\prime}\boldsymbol{W}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{W}\boldsymbol{Y}^*$ and its approximate covariance matrix as $\boldsymbol{\Sigma}(\boldsymbol{\phi}, \boldsymbol{\tau}^2) = (\boldsymbol{X}^{*\prime}\boldsymbol{W}\boldsymbol{X}^*)^{-1}$.

The parameter $\boldsymbol{\phi}$ is estimated by the conditional posterior mode of $\pi(\boldsymbol{\phi}|\boldsymbol{\tau}^2, y)$. For this, we use an approximate EM algorithm as follows (Gelman et al., 2014). Taking $\boldsymbol{\beta}^*$ as "missing data" in the E-step, we find $H(\boldsymbol{\phi}|\boldsymbol{\phi}^{old}, \boldsymbol{\tau}^2, \boldsymbol{y})$, the expectation of $\log \pi(\boldsymbol{\phi}, \boldsymbol{\beta}^*|\boldsymbol{\tau}^2, \boldsymbol{y})$ with respect to the conditional posterior distribution of $\boldsymbol{\beta}^*$ given the current value of $\boldsymbol{\phi}$ (denoted by $\boldsymbol{\phi}^{old}$) and $\boldsymbol{\tau}^2$. Approximating this distribution by a normal distribution with mean $\hat{\boldsymbol{\beta}}^*(\boldsymbol{\phi}^{old}, \boldsymbol{\tau}^2)$ and variance matrix $\boldsymbol{\Sigma}(\boldsymbol{\phi}^{old}, \boldsymbol{\tau}^2)$, both obtained via IWLS described earlier, $H(\boldsymbol{\phi}|\boldsymbol{\phi}^{old}, \boldsymbol{\tau}^2, \boldsymbol{y})$ can be obtained in a closed-form (Gelman et al., 2014). This involves approximations such as $E(\beta_{sjpq} - \xi_{jp})^2 \approx (\hat{\beta}_{sjpq} - \hat{\xi}_{jp})^2 + \widehat{\text{var}}(\hat{\beta}_{sjpq} - \hat{\xi}_{jp})$. In the M-Step, we maximize H with respect to $\boldsymbol{\phi}$ to get the updated value $\hat{\boldsymbol{\phi}}(\hat{\boldsymbol{\beta}}^*, \boldsymbol{\tau}^2)$ whose elements are explicitly available as simple expressions, allowing easy scalability of the method. For example,

$$\hat{\sigma}^2_{\beta jp} = \frac{\sum_q \sum_s \left\{(\hat{\beta}_{sjpq} - \hat{\xi}_{jp})^2 + \widehat{\text{var}}(\hat{\beta}_{sjpq} - \hat{\xi}_{jp})\right\} + \nu_\beta s^2_{\beta p}}{QS_{jp} + \nu_\beta + 2}, \; \hat{\sigma}^2_\theta = \frac{\sum_p \left\{\hat{\theta}^2_p + \widehat{\text{var}}(\hat{\theta}_p)\right\} + \nu_\theta s^2_\theta}{P + \nu_\theta + 2}.$$

Thus, we now have an iterative algorithm for fitting the extended model following the same steps (IWLS and EM). A normal approximation for the mode $\hat{\boldsymbol{\beta}}^*(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\tau}}^2)$ with covariance matrix $\boldsymbol{\Sigma}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\tau}}^2)$ can be used to perform the needed hypotheses testing.

REFERENCES

Anderson, L. N., M. Cotterchio, D. E. Cole, and J. A. Knight (2011, Aug). Vitamin D-related genetic variants, interactions with vitamin D exposure, and breast cancer risk among Caucasian women in Ontario. *Cancer Epidemiol. Biomarkers Prev. 20*(8), 1708–1717.

Belfiore, A. and R. Malaguarnera (2011, Aug). Insulin receptor and cancer. *Endocr. Relat. Cancer 18*(4), R125–147.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Royal Stat Soc, Series B 57*, 289–300.

Cantor, R. M., K. Lange, and J. S. Sinsheimer (2010, Jan). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am. J. Hum. Genet. 86*(1), 6–22.

Casella, G. and R. L. Berger (2002). *Statistical Inference*, Volume 2. Duxbury Pacific Grove, CA.

Chen, L. S., C. M. Hutter, J. D. Potter, Y. Liu, R. L. Prentice, U. Peters, and L. Hsu (2010, Jun). Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am. J. Hum. Genet. 86*(6), 860–871.

Chen, P., C. Li, X. Li, J. Li, R. Chu, and H. Wang (2014, Apr). Higher dietary folate intake reduces the breast cancer risk: a systematic review and meta-analysis. *Br. J. Cancer 110*(9), 2327–2338.

Cheuk, I. W., V. Y. Shin, M. T. Siu, J. Y. Tsang, J. C. Ho, J. Chen, G. M. Tse, X. Wang, and A. Kwong (2015). Association of EP2 receptor and SLC19A3 in regulating breast cancer metastasis. *Am J Cancer Res 5*(11), 3389–3399.

Chin, Y. R., T. Yoshida, A. Marusyk, A. H. Beck, K. Polyak, and A. Toker (2014, Feb). Targeting Akt3 signaling in triple-negative breast cancer. *Cancer Res. 74*(3), 964–973.

Chou, Y. C., C. H. Chu, M. H. Wu, G. C. Hsu, T. Yang, W. Y. Chou, H. P. Huang, M. S. Lee, C. P. Yu, J. C. Yu, and C. A. Sun (2011). Dietary intake of vitamin B(6) and risk of breast cancer in Taiwanese women. *J Epidemiol 21*(5), 329–336.

Colt, J. S., K. Schwartz, B. I. Graubard, F. Davis, J. Ruterbusch, R. DiGaetano, M. Purdue, N. Rothman, S. Wacholder, and W. H. Chow (2011, Nov). Hypertension and risk of renal cell carcinoma among white and black Americans. *Epidemiology 22*(6), 797–804.

Corn, P. G. (2007, Nov). Role of the ubiquitin proteasome system in renal cell carcinoma. *BMC Biochem. 8 Suppl 1*, S4.

database of Genotypes and Phenotypes (last accessed July 01, 2018). `https://www.ncbi.nlm.nih.gov/gap`.

Davies, R. B. (1980). Algorithm as 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 29*(3), 323–333.

Deeb, K. K., D. L. Trump, and C. S. Johnson (2007, Sep). Vitamin D signalling pathways in cancer: potential for anticancer therapeutics. *Nat. Rev. Cancer 7*(9), 684–700.

Derikx, L. A., L. H. Nissen, J. P. Drenth, C. M. van Herpen, W. Kievit, R. H. Verhoeven, P. F. Mulders, C. A. Hulsbergen-van de Kaa, M. J. Boers-Sonderen, T. R. van den Heuvel, M. Pierik, I. D. Nagtegaal, F. Hoentjen, P. M. Kluin, M. Hogenes, A. F. Hamel, R. Natte, C. M. van Dijk, H. V. Kusters-Vandevelde, S. H. Sastrowijoto, A. P. Willig, G. Dijkstra, A. E. van der Meulen-de Jong, M. K. Vu, A. Cats, J. B. Haanen, C. J. van der Woude, M. G. Russel, B. Oldenburg, J. J. Meeuse, S. Corporaal, A. M. Zonneveld, P. J. Wahab, S. J. van den Hazel, W. G. Mares, R. J. Lieverse, M. A. Meijssen, K. Thuernau, D. Janik, H. van der Heide, I. J. Klompmaker, C. J. Bolwerk, R. Stuyt, A. A. van Bodegraven, C. Y. Ponsioen, R. L. West, R. N. Zeijen, T. J. Tang, P. J. Wismans, P. Dewint, J. Y. Lai, A. C. Tan, A. C. Depla, E. T. Keulen, L. E. Oostenbrug, M. E. Bartelink, G. W. Erkelens, J. Vecht, J. W. Tjhiep-Wensing, T. E. Romkens, W. A. de Boer, R. K. Linskens, M. L. Verhulst, A. H. Naber, G. C. Noomen, P. E. Dekkers, P. P. Viergever, J. R. Vermeijden, M. Willems, H. G. Lam, N. Mahmmod, B. C. Loffeld, J. M. Jansen, P. C. Stokkers, J. P. Kuijvenhoven, M. J. Wagtmans, C. H. Clemens, L. T. Vlasveld, P. J. Bus, R. P. Dahlmans, R. Beukers, P. C. Ter Borg, P. P. van der Veek, J. T. Sarneel, S. Vandebosch, E. Halet, M. C. Rijk, M. J. Grubben, P. J. Kil, L. P. Gilissen, F. L. Wolters, and M. Uiterwaal (2015, Nov). Better survival of renal cell carcinoma in patients with inflammatory bowel disease. *Oncotarget 6*(35), 38336–38347.

Djiogue, S., A. H. Nwabo Kamdje, L. Vecchio, M. J. Kipanyula, M. Farahna, Y. Aldebasi, and P. F. Seke Etet (2013, Feb). Insulin resistance and cancer: the role of insulin and IGFs. *Endocr. Relat. Cancer 20*(1), R1–R17.

Dudbridge, F. and B. P. Koeleman (2003). Rank truncated product of p-values, with application to genomewide association scans. *Genetic epidemiology 25*(4), 360–366.

Ensembl Genome Browser (last accessed July 01, 2018). `http://useast.ensembl.org/biomart/martview/6301ed9a9317400228e3d3c5a09bafda`.

Evangelou, M., F. Dudbridge, and L. Wernisch (2014, Mar). Two novel pathway analysis methods based on a hierarchical model. *Bioinformatics 30*(5), 690–697.

Frezza, M., S. Schmitt, and Q. P. Dou (2011, Dec). Targeting the ubiquitin-proteasome pathway: an emerging concept in cancer therapy. *Curr Top Med Chem 11*(23), 2888–2905.

Fridley, B. L. and J. M. Biernacka (2011, Aug). Gene set analysis of SNP data: Benefits, challenges, and future directions. *Eur. J. Hum. Genet. 19*(8), 837–843.

Gelman, A., J. B. Carlin, H. S. Stern., D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC.

Gelman, A., J. Hill, and M. Yajima (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness 5*, 189–211.

Gelman, A., A. Jakulin, M. Pittau, and Y.-S. Su (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat. 2*, 1360–1383.

Grottke, A., F. Ewald, T. Lange, D. Norz, C. Herzberger, J. Bach, N. Grabinski, L. Graser, F. Hoppner, B. Nashan, U. Schumacher, and M. Jucker (2016). Downregulation of AKT3 Increases Migration and Metastasis in Triple Negative Breast Cancer Cells by Upregulating S100A4. *PLoS ONE 11*(1), e0146370.

Holden, M., S. Deng, L. Wojnowski, and B. Kulle (2008, Dec). GSEA-SNP: Applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics 24*(23), 2784–2785.

Holmans, P., E. K. Green, J. S. Pahwa, M. A. Ferreira, S. M. Purcell, P. Sklar, M. J. Owen, M. C. O'Donovan, and N. Craddock (2009, Jul). Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet. 85*(1), 13–24.

Hu, X., J. Wang, W. He, P. Zhao, and C. Ye (2018, Mar). MicroRNA-433 targets AKT3 and inhibits cell proliferation and viability in breast cancer. *Oncol Lett 15*(3), 3998–4004.

KEGG database (last accessed July 01, 2018). `https://www.kegg.jp`.

Korn, E. and B. Graubard (1999). *Analysis of health surveys.* Wiley, New York.

Lange, K. (2003). *Mathematical and statistical methods for genetic analysis.* Springer Science & Business Media.

Law, J. H., G. Habibi, K. Hu, H. Masoudi, M. Y. Wang, A. L. Stratford, E. Park, J. M. Gee, P. Finlay, H. E. Jones, R. I. Nicholson, J. Carboni, M. Gottardis, M. Pollak, and S. E. Dunn (2008, Dec). Phosphorylated insulin-like growth factor-i/insulin receptor is present in all breast cancer subtypes and is related to poor survival. *Cancer Res. 68*(24), 10238–10246.

Li, X. W., M. Tuergan, and G. Abulizi (2015, Nov). Expression of MAPK1 in cervical cancer and effect of MAPK1 gene silencing on epithelial-mesenchymal transition, invasion and metastasis. *Asian Pac J Trop Med 8*(11), 937–943.

Lindley, D. V. and A. F. M. Smith (1972). Bayesian estimates for the linear model. *J Royal Stat Soc, Series B 34*, 1–41.

Mani, A. and E. P. Gelmann (2005, Jul). The ubiquitin-proteasome pathway and its role in cancer. *J. Clin. Oncol. 23*(21), 4776–4789.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher (2009, Oct). Finding the missing heritability of complex diseases. *Nature 461*(7265), 747–753.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Menashe, I., D. Maeder, M. Garcia-Closas, J. D. Figueroa, S. Bhattacharjee, M. Rotunno, P. Kraft, D. J. Hunter, S. J. Chanock, P. S. Rosenberg, and N. Chatterjee (2010, Jun). Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. *Cancer Res. 70*(11), 4453–4459.

Ng, E. K., C. P. Leung, V. Y. Shin, C. L. Wong, E. S. Ma, H. C. Jin, K. M. Chu, and A. Kwong (2011). Quantitative analysis and diagnostic significance of methylated SLC19A3 DNA in the plasma of breast and gastric cancer patients. *PLoS ONE 6*(7), e22233.

Obiri, N. I., G. G. Hillman, G. P. Haas, S. Sud, and R. K. Puri (1993, Jan). Expression of high affinity interleukin-4 receptors on human renal cell carcinoma cells and inhibition of tumor cell growth in vitro by interleukin-4. *J. Clin. Invest. 91*(1), 88–93.

O'Dushlaine, C., E. Kenny, E. A. Heron, R. Segurado, M. Gill, D. W. Morris, and A. Corvin (2009, Oct). The SNP ratio test: Pathway analysis of genome-wide association datasets. *Bioinformatics 25*(20), 2762–2763.

O'Hurley, G., E. Daly, A. O'Grady, R. Cummins, C. Quinn, L. Flanagan, A. Pierce, Y. Fan, M. A. Lynn, M. Rafferty, D. Fitzgerald, F. Ponten, M. J. Duffy, K. Jirstrom, E. W. Kay, and W. M. Gallagher (2014, Apr). Investigation of molecular alterations of AKT-3 in triple-negative breast cancer. *Histopathology 64*(5), 660–670.

Poloz, Y. and V. Stambolic (2015, Dec). Obesity and cancer, a case for insulin signaling. *Cell Death Dis 6*, e2037.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham (2007, Sep). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet. 81*(3), 559–575.

Ramanan, V. K., L. Shen, J. H. Moore, and A. J. Saykin (2012, Jul). Pathway analysis of genomic data: Concepts, methods, and prospects for future development. *Trends Genet. 28*(7), 323–332.

Reyes-Gibby, C. C., J. Wang, M. R. Silvas, R. Yu, S. C. Yeung, and S. Shete (2016, Feb). MAPK1/ERK2 as novel target genes for pain in head and neck cancer patients. *BMC Genet. 17*, 40.

Romano, M., F. DE Francesco, L. Zarantonello, C. Ruffolo, G. A. Ferraro, G. Zanus, A. Giordano, N. Bassi, and U. Cillo (2016, Apr). From Inflammation to Cancer in Inflammatory Bowel Disease: Molecular Perspectives. *Anticancer Res. 36*(4), 1447–1460.

Rostoker, R., S. Abelson, K. Bitton-Worms, I. Genkin, S. Ben-Shmuel, M. Dakwar, Z. S. Orr, A. Caspi, M. Tzukerman, and D. LeRoith (2015, Apr). Highly specific role of the insulin receptor in breast cancer progression. *Endocr. Relat. Cancer 22*(2), 145–157.

Rouette, A., A. Trofimov, D. Haberl, G. Boucher, V. P. Lavallee, G. D'Angelo, J. Hebert, G. Sauvageau, S. Lemieux, and C. Perreault (2016, Sep). Expression of immunoproteasome genes is regulated by cell-intrinsic and -extrinsic factors in human cancers. *Sci Rep 6*, 34019.

Ruiz-Narvaez, E. A., K. L. Lunetta, C. C. Hong, S. Haddad, S. Yao, T. D. Cheng, J. T. Bensen, E. V. Bandera, C. A. Haiman, M. A. Troester, C. B. Ambrosone, L. Rosenberg, and J. R. Palmer (2016). Genetic variation in the insulin, insulin-like growth factor, growth hormone, and leptin pathways in relation to breast cancer in African-American women: the AMBER consortium. *NPJ Breast Cancer 2*.

Schaid, D. J., J. P. Sinnwell, G. D. Jenkins, S. K. McDonnell, J. N. Ingle, M. Kubo, P. E. Goss, J. P. Costantino, D. L. Wickerham, and R. M. Weinshilboum (2012, Jan). Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. *Genet. Epidemiol. 36*(1), 3–16.

Shahbaba, B., C. M. Shachaf, and Z. Yu (2012, May). A pathway analysis method for genome-wide association studies. *Stat Med 31*(10), 988–1000.

Silver, M., G. Montana, and A. D. N. Initiative (2012, Jan). Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. *Stat Appl Genet Mol Biol 11*(1), Article 7.

Slattery, M. L., L. H. Hines, A. Lundgreen, K. B. Baumgartner, R. K. Wolff, M. C. Stern, and E. M. John (2014, Sep). Diet and lifestyle factors interact with MAPK genes to influence survival: the Breast Cancer Health Disparities Study. *Cancer Causes Control 25*(9), 1211–1225.

SNPath R package (last accessed July 01/2018). `https://research.fhcrc.org/hsu/en/software.html`.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. V. D. Linde (2002). Bayesian measures of model complexity and fit. *J Royal Stat Soc, Series B 64*, 583–639.

Sweet, R., A. Paul, and J. Zastre (2010, Dec). Hypoxia induced upregulation and function of the thiamine transporter, SLC19A3 in a breast cancer cell line. *Cancer Biol. Ther. 10*(11), 1101–1111.

Torkamani, A., E. J. Topol, and N. J. Schork (2008, Nov). Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics 92*(5), 265–272.

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

Wang, K., M. Li, and H. Hakonarson (2010, Dec). Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet. 11*(12), 843–854.

Wang, L., P. Jia, R. D. Wolfinger, X. Chen, B. L. Grayson, T. M. Aune, and Z. Zhao (2011, Mar). An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies. *Bioinformatics 27*(5), 686–692.

Weng, L., F. Macciardi, A. Subramanian, G. Guffanti, S. G. Potkin, Z. Yu, and X. Xie (2011). SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics 12*, 99.

Wright, F. A., H. Huang, X. Guan, K. Gamiel, C. Jeffries, W. T. Barry, F. P. de Villena, P. F. Sullivan, K. C. Wilhelmsen, and F. Zou (2007, Oct). Simulating association studies: A data-based resampling method for candidate regions or whole genome scans. *Bioinformatics 23*(19), 2581–2588.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics 89*(1), 82–93.

Yan, Q., H. K. Tiwari, N. Yi, W.-Y. Lin, G. Gao, X.-Y. Lou, X. Cui, and N. Liu (2014). Kernel-machine testing coupled with a rank-truncation method for genetic pathway analysis. *Genetic epidemiology 2014*(5), 447–456.

Yekutieli, D. (2008). Hierarchical false discovery rate-controlling methodology. *J Am Stat Assoc 103*, 309–316.

Yi, N. and S. Ma (2012). Hierarchical shrinkage priors and model fitting for high-dimensional generalized linear model. *Statistical applications in genetics and molecular biology 11*(6).

Yu, K., Q. Li, A. W. Bergen, R. M. Pfeiffer, P. S. Rosenberg, N. Caporaso, P. Kraft, and N. Chatterjee (2009, Dec). Pathway analysis by adaptive combination of p-values. *Genet. Epidemiol. 33*(8), 700–709.

Zhang, K., S. Cui, S. Chang, L. Zhang, and J. Wang (2010, Jul). i-GSEA4GWAS: A web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Res. 38*(Web Server issue), W90–95.

Zhang, L., J. S. Morris, J. Zhang, R. Z. Orlowski, and V. Baladandayuthapani (2014). Bayesian joint selection of genes and pathways: Applications in multiple myeloma genomics. *Cancer Inform 13*(Suppl 2), 113–123.

Zhang, Y., J. N. Hofmann, M. P. Purdue, S. Lin, and S. Biswas (2017, Sep). Logistic Bayesian LASSO for genetic association analysis of data from complex sampling designs. *J. Hum. Genet. 62*(9), 819–829.

BIOGRAPHICAL SKETCH

Lei Zhang was born in Quzhou, China. After completing her Bachelor's degree from Anhui University in 2010 and Master's degree from the University of Toledo in 2012, she joined The University of Texas at Dallas as a PhD student in August 2012. During her PhD study, she met her husband, Ruikai Cao, who was also a student at The University of Texas at Dallas and had a baby girl in September 2017. She will join GM Financial in the position of "Economist II - Modeler" after completion of her Ph.D.

CURRICULUM VITAE

Lei Zhang

July 05, 2018

Contact Information:

Email: `lxz096120@utdallas.edu`

Educational History:

- B.S., Statistics, Anhui University, 2010.
- M.S., Statistics, The University of Toledo, 2012.
- Ph.D., Statistics, The University of Texas at Dallas, 2018.

Employment History:

- Teaching and Research Assistant, The University of Texas at Dallas, Richardson, TX, August 2012 – May 2018.
- Intern Fellow(Advanced Analytics), BNSF Railway, Fort Worth, TX, May 2016 – August 2016.
- Teaching Assistant, The University of Toledo, Toledo, OH, August 2010 – May 2012.

Presentations:

- Poster presentation, Conference of Texas Statisticians, April, 2017.
- Presentation on "A Bayesian hierarchical model for pathway analysis with simultaneous inference on Pathway-Gene-SNP structure", ENAR Biometric Society Spring Meeting, March, 2017.
- Presentation on "A Bayesian hierarchical model for pathway analysis with simultaneous inference on Pathway-Gene-SNP structure", Joint Statistical Meetings, August, 2016.
- Poster presentation, Southern Regional Council on Statistics, June, 2016.

Honors and Awards:

- Small Grants travel award, The University of Texas at Dallas, 2016 and 2017.
- Member of The Honor Society of Phi Kappa Phi, 2014-2016.
- R.L. Anderson student poster award, Southern Regional Council on Statistics, 2016.
- Julia Williams Van Ness Merit Scholarship, 2015.

Professional Memberships:

- American Statistical Association
- American Mathematical Society