STATISTICAL METHODS FOR HIGH-DIMENSIONAL GENETIC AND GENOMIC DATA

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

CHONG WU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

ADVISED BY DRS. WEIHUA GUAN AND WEI PAN

June, 2018

Acknowledgements

There are several people that have earned my gratitude for their support and contribution to my time at the University of Minnesota.

First, I wish to thank my advisors Profs. Weihua Guan and Wei Pan. Prof. Guan taught me a lot about how to collaborate with researchers outside the department and gave me lots of freedom to determine my research topic. In my last three Ph.D. years, I mainly work with Prof. Pan. He gave me extensive personal and professional guidance and taught me a great deal about both scientific research and life in general. Importantly, he guided me and gave me the moral support during the most difficult times when finding an academic position.

Besides my advisors, I would like to thank the rest of my thesis committee: Profs. Baolin Wu, and Jim Pankow for their support and constructive comments on my research.

My sincere thanks also go to Profs. Haitao Chu, and Jim Hodges, who helped me a lot improve my presentation skills and get a Doctoral Dissertation Fellowship. My collaborators, Profs. Xiaotong Shen, Gongjun Xu, Ellen Demerath, Jun Chen, and my peers Jun Young Park, have all extended their support in a very special way, and I gained a lot from them.

Further, I gratefully acknowledge the funding sources and the supercomputing service at Minnesota Supercomputing Institute that made my Ph.D. work possible.

Lastly, I want to thank my parents and girlfriend for all their encouragement and love. Thank you.

Abstract

Modern genetics research constantly creates new types of high-dimensional genetic and genomic data and imposes new challenges in analyzing these data. This thesis deals with several important problems in analyzing high-dimensional genetic and genomic data, ranging from DNA methylation data to human microbiome data.

First, we introduce a site selection and multiple imputation method to impute missing data in covariates in epigenome-wide analysis of DNA methylation data, which can help us adjust potential confounders, such as cell type composition. Second, to overcome low power issue of human microbiome association studies, we propose a powerful data-driven approach by weighting the variables (taxa) in a manner determined by the data itself. The increased power of the new test not only decreases the sample size required for a human microbiome association study but also allows for new discoveries with existing datasets. Third, we propose an adaptive test on a high-dimensional parameter of a generalized linear model (in the presence of a low-dimensional nuisance parameter). Benefiting from its adaptivity, the proposed test maintains high statistical power under various high-dimensional scenarios. We further establish its asymptotic null distribution. Finally, we propose a novel pathway-based association test by integrating gene expression, gene functional annotations, and a main genome-wide association study dataset. We applied it to a schizophrenia GWAS summary association dataset and identified 15 novel pathways associated with schizophrenia, such as *GABA receptor complex* (GO:1902710), which could not be uncovered by the standard single SNP-based analysis or gene-based TWAS.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Modern genetics research constantly creates new types of high-dimensional genetic and genomic data and imposes new challenges in analyzing these data. For example, DNA methylation value can be highly influenced by tissue types, developmental stage, and disease status [3]. Failing to adjust for these confounding factors may result in questionable conclusions. However, incomplete data on confounding variables, such as tissue types, is a common and unsolved problem in applied research. This thesis deals with several important problems in analyzing high-dimensional genetic and genomic data: imputing missing covariate values in epigenome-wide analysis of DNA methylation data (Chapter 2), an adaptive association test for microbiome data (Chapter 3) and generalized linear models (GLMs, Chapter 4), and integrative pathway-based association analysis (Chapter 5).

DNA methylation is a widely studied epigenetic mechanism. It most often results in repression of gene expression and has been shown to change over time as part of the normal aging process and in response to environmental factors and also to be altered by diseases. Incorporating epigenetic variation into genetic studies of these diseases represents a powerful and effective way to account for environmental influences that gene-sequence-based studies of disease do not. Nowadays, epigenome-wide association studies (EWAS) is becoming increasingly popular to interrogate methylation change associated with phenotypes of interest. Facing many of the same challenges as epidemiology studies, missing data on confounding variables is a very common problem in applied research.

To adjust for confounders in EWAS, one idea is based on the principal component analysis (PCA) and infers confounders using surrogate variables [4, 5, 6, 7]. On the other hand, when the confounders are measured, they can be directly adjusted in a regression framework. However, incomplete data is a common problem in applied research. By using the correlation between DNA methylation values and the covariates, we can impute the missing covariate values based on a set of pre-selected CpG sites [8]. Chapter 2 proposes a new statistical method to impute missing covariate values in the analysis of DNA methylation data, which is based on the variable selection and multiple imputation (MI) approach [9]. Specifically, we apply variable selection to select a set of CpG sites from genome-wide methylation data and use the MI approach to impute missing covariates and perform association tests accordingly.

Similar to EWAS, human microbiome association studies aim to detect an association of human microbiome diversity with a phenotype of interest, such as disease status; this can improve our understanding of complex traits and diseases. A human body has more than 10 times as many microbes living in it as cells. These microorganisms play an important part in our overall health, such as protecting us from diseases and digesting food [10]. The genomes of human microbes, and the way they interact with the human host, are collectively termed the microbiome. Indeed, many human diseases have been linked to disorders of the human microbiome through microbiome association studies. Rauch et al. have demonstrated that the gut microbiome is important in the development of the human immune system, and abnormalities in microbial diversity are correlated with several inflammatory diseases as well as colon cancer, type 2 diabetes, and obesity [11]. An efficient statistical method statistical method in human microbiome association study may improve the statistical power and help better understand risk factors related to the diseases.

One popular method for testing the association between an overall microbiome composition and a phenotype of interest is to use a distance-based test, such as PERMANOVA [12]. Recently, a new method called microbiome regression based kernel association test (MiRKAT) was proposed [13]. Incorporating phylogenetic relationships among taxa, MiRKAT transforms a phylogenetic distance metric into a kernel to measure similarities among samples. Then a semi-parametric kernel machine regression framework is applied to evaluate the association. MiRKAT allows for an easy covariate

adjustment. While enticing, a severe limitation of kernel or distance-based methods is that they do not conduct variable selection or variable weighting, which is crucial for high-dimensional microbiome data so that noise accumulations can be alleviated. To overcome this limitation and improve statistical power, Chapter 3 proposes a powerful data-driven approach called MiSPU by weighting the variables (taxa) in a manner determined by the data itself. The proposed method is based on a newly defined measure combining microbial abundance with phylogenetic tree information, from which a powerful and highly adaptive test can be derived. The increased power of the new method not only decreases the sample size required for a microbiome association study but also allows new discoveries with existing datasets.

In the human microbiome association studies, we calculate the p-value for the MiSPU via a resampling method, which is highly computationally extensive. One intuitive follow-up research problem is to establish the asymptotic null distribution of MiSPU and calculate its p-value accordingly to save computational sources. This leads us to study the significance testing for regression coefficients in high-dimensional GLMs.

Many existing methods [1, 2] in this field are based on the sum-of-squares of the score vector for the parameters and are usually powerful against a moderately dense alternative hypothesis, where there are a relatively large proportion of signals. In contrast, if the non-zero signals are strong but sparse, the sum-of-squares of the score vector based tests lose substantial power while a test based on the supremum of the score vector is more powerful. Importantly, there are some intermediate situations in which neither type of test is powerful. Chapter 4 proposes an adaptive test on a high-dimensional parameter of a generalized linear model in the presence of a low-dimensional nuisance parameter. Benefiting from its adaptivity, the proposed test maintains high statistical power under various high dimensional situations. To apply the proposed test, we establish its asymptotic null distribution accordingly. In addition, we apply it and other existing tests to an Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, detecting possible associations between Alzheimer's disease and some gene pathways with a large number of single nucleotide polymorphisms (SNPs).

Besides proposing a new statistical association testing procedures such as methods discussed in Chapters 3 and 4 to improve statistical power, researchers propose integrative testing methods to improve power by incorporating external information. Recently,

by noting that an important class of genetic variants, termed expression quantitative trait loci (eQTLs) enrich among the GWAS trait-associated variants [14, 15], researchers propose transcriptome-wide association study (TWAS) and related methods [16, 17, 18] to integrate eQTL information with GWAS data to identify the genes associated with a complex trait. Though appealing, due to the limited sample sizes of eQTL and GWAS data, they may fail to identify some weakly associated genes with small effect sizes. More importantly, a group of functionally related genes as annotated in a biological pathway are often involved in the same disease susceptibility and progression [19]. To overcome these limitations, Chapter 5 extends integrative gene-based testing like TWAS to integrative pathway-based association analysis to identify pathways associated with complex traits and diseases. Specifically, we propose a new self-contained test that integrates eQTL data, GWAS individual-level or summary data, SNP LD information, and gene functional annotations as public pathway collections to identify pathways associated with a complex trait.

Chapter 6 summarizes the major work in this thesis and introduces some potential related future topics.

Chapter 2

Imputation of Missing Covariate Values in Epigenome-wide Analysis of DNA Methylation Data

DNA methylation is a widely studied epigenetic mechanism and alterations in methylation patterns may be involved in the development of common diseases. Unlike inherited changes in genetic sequence, variation in site-specific methylation varies by tissue, developmental stage, and disease status, and may be impacted by aging and exposure to environmental factors, such as diet or smoking. These non-genetic factors are typically included in epigenome-wide association studies (EWAS) because they may be confounding factors to the association between methylation and disease. However, missing values in these variables can lead to reduced sample size and decrease the statistical power of EWAS. This chapter proposes a site selection and multiple imputation (MI) method to impute missing covariate values and to perform association tests in EWAS. Then, we compare this method to an alternative projection-based method. Through simulations, we show that the MI-based method is slightly conservative, but provides consistent estimates for effect size. We also illustrate these methods with data from the Atherosclerosis Risk in Communities (ARIC) study to carry out an EWAS between methylation levels

and smoking status, in which missing cell type compositions and white blood cell counts are imputed. The main part of this chapter has been published in Wu et al. 2016 [20].

2.1 Introduction

DNA methylation of cytosine residues at CpG dinucleotides has particular of interest since it has a central role in normal human development and disease [21]. Epigenome-wide association studies (EWAS), analogous to genome-wide association studies (GWAS), are becoming increasingly common and popular to interrogate methylation change associated with disease status or related traits. Facing many of the same challenges as epidemiology studies, DNA methylation value can be highly related to the tissue, developmental stage, disease status and may be impacted by aging and exposure to environmental factors such as diet or smoking, which may lead to potential confounding in association tests [3]. For example, in some DNA methylation dataset, cases and controls typically differ in their cell-type composition, which can result in spurious associations—those that merely tag the cell type rather than reveal more fundamental biological signals of interest [22, 23].

To adjust for confounders in analysis of DNA methylation data, one approach is based on the principal component analysis (PCA) and infers confounders using surrogate variables [4, 5, 6, 7]. When the confounders are measured, they can be directly adjusted in a regression framework. However, incomplete data is a common problem in applied research. Restricting the analysis to complete cases will reduce the size of data and the power of association tests, which are critical for EWAS with stringent genome-wide significance level. Using the correlation between methylation values and the covariates, we can impute the missing covariate values based on a set of pre-selected CpG sites. For example, when a reference dataset is available, Houseman et al. [8] proposed a projection-based method to infer the underlying composition of cell populations with distinct DNA methylation profiles.

Here, we explore statistical methods to impute missing covariate values in analysis of DNA methylation data, which is based on the multiple imputation (MI) approach [9]. Specifically, we apply variable selection to select a set of CpG sites from genome-wide methylation data and use the MI approach to impute missing covariates and perform

association tests. We also extend the projection-based method [8] by treating the complete cases as the reference data. Through simulations, we show that the MI-based methods produce asymptotically unbiased estimates for the effect size, with slightly conservative power compared to the projection-based method. On the other hand, the projection-based method [8] has the correct type I error rate, but may produce biased estimates under the alternative hypothesis. We further apply these methods to the DNA methylation data in the Atherosclerosis Risk in Communities (ARIC) study. We compare the CpG sites identified for association with smoking status, using different imputation approaches for missing cell type compositions and whole blood cell counts (WBC) in the analysis.

2.2 Methods

2.2.1 Notation

We denote the DNA methylation measure Y. Here we assume all the measurements are known and on a "beta value" scale, which can be interpreted as the proportion of methylated molecules at a given locus. Note that the proposed methods can be directly applied to alternative methylation measures, e.g., the M-value [24]. Suppose we have p covariates (C), including the covariate of interest (X), such as smoking status in our example below, and potential confounders (Z), such as the cell-type composition and WBC. We assume for each covariate, say C_i , the C_i^{obs} is the observed part while C_i^{mis} is the missing part. We use C_c to denote the set of non-missing covariates. In this chapter, we consider two types of covariates with missing values, continuous and binary.

2.2.2 Variable Selection

Given the large number of CpG sites surveyed in an epigenome-wide study, it is infeasible to include all CpG sites in a standard multiple imputation model. Buuren and Groothuis suggested that 15 variables are generally sufficient for imputation [25]. Thus we use the following variable selection method, which includes a screen stage and a selection stage.

First, in the screen stage, for an incomplete covariate C_i, we compute test statistics for association between C_i^{obs} and DNA methylation level at each CpG site. We select

the most differentially methylated CpGs with p-value$< 10^{-7}$, up to 100 sites. The p-value cutoff is based on the Bonferroni correction for genome-wide significance using the HM450 chip, but other cutoff value can also be considered.

Second, in the selection stage, we use a forward stepwise with BIC criterion to select the final predictor set from the selected CpG sites. In the forward selection, the CpG sites are added to the model (linear for continuous covariate and logistic for binary) one at a time [26], with the site which produces the largest reduction in the BIC value. The selection will stop when 30 CpG sites are selected, if the BIC stopping rule is still not satisfied.

Note that the initial set of 100 CpG sites for variable selection and the 30 CpG sites for imputation are subjective. In practice, the actual numbers of sites can vary depending on the total number of CpG sites that are associated with the missing covariate, and the size of complete cases.

2.2.3 Multiple Imputation

For each incomplete covariate, the missing data are replaced by m independent imputation sets of values. Here, for continuous missing covariates, we consider two methods: predictive mean matching (MI-PMM), a general semi-parametric imputation method [27], and linear imputation method (MI-Norm) [9]. In MI-PMM, imputation is restricted to the observed values and can preserve non-linear relations even if the structural part of the imputation model is misspecified [25]. For categorical missing covariate variables, we consider using logistic regression (MI-Logreg) to impute the missing values.

In MI-Norm, for a specific covariate C, we regress C_i^{obs} on the set of selected K CpG sites $(Y_{i1}, Y_{i2}, \ldots, Y_{iK})$ and the covariates with complete information, say $Y^i = (Y_{i1}, Y_{i2}, \ldots, Y_{iK}, X, C_c)$. Here, we assume the missing covariate C_i^{mis} follows a normal distribution:

$$C_i^{mis} \sim N(\beta_i Y^i, \sigma_i^2),$$

where β_i and σ_i can be estimated from C_i^{obs} . Considering the uncertainty caused by missing data, we draw the imputation parameters β_i^*, σ_i^* from the joint posterior distribution [28]. Specifically, σ_i^* is drawn as $\sigma_i^* = \hat{\sigma}\sqrt{(n_{obs} - p)/g}$, where g is a random

draw from a Chi square distribution with $n_{obs} - p$ degrees of freedom, and p is the total number of predictors in the model. Then, β_i^* is drawn as $\beta_i^* = \hat{\beta} + (\sigma^*/\hat{\sigma})\, u_1 V^{1/2}$, where u_1 is a row vector of p independent random draws form a standard normal distribution, and V the estimated covariance matrix of $\hat{\beta}$. We impute the missing observations based on the normal distribution described above and the imputation parameters β_i^*, σ_i^*.

In MI-PMM, we match each missing C_i^{mis} to the respondent with the closet predicted mean and then impute using the respondents value directly [27].

In MI-Logreg, instead of fitting a linear model as in MI-Norm, we fit a logistic regression model. For each missing observation C_i^{mis}, let $p_i^* = \left(1 + \exp\left(-\beta_i^* Y^i\right)\right)^{-1}$, and an imputed value is drawn as 1 if $u_i < p_i^*$, otherwise as 0, where u_i is a random number from a standard uniform distribution [9].

We analyze each of the m imputed datasets separately to test for association between DNA methylation and trait of interest. We then use Rubins rule [28] to combine the m estimates into an overall estimate. The combined estimate $\hat{\beta}_i$ is the average of the individual estimates:

$$\hat{\beta}_i = \frac{1}{m} \sum_{k=1}^{m} \hat{\beta}_i^k.$$

The total variance of $\hat{\beta}_i$ is calculated as follows:

$$\mathrm{var}(\hat{\beta}_i) = W + \left(1 + \frac{1}{m}\right) B,$$

where W is the within-imputation variance ($W = 1/m \sum_{k=1}^{m} W_k$, uncertainty about the results from one imputed dataset) and B is the between-imputation variance ($B = 1/m \sum_{k=1}^{m} \left(\hat{\beta}_i^k - \hat{\beta}_i\right)^2$, uncertainty due to the missing information).

2.2.4 Projection-based Method

We tailor the method proposed by Houseman et al. [8] to a more general situation. Briefly, the Houseman algorithm identifies 100–300 CpG sites that discriminated cellular composition in sorted normal human cell populations (consisting of B cells, granulocytes, monocytes, NK, and CD4+ and CD8+ T cells). The method fits a linear model at each of these CpG sites using a reference dataset to estimate the coefficient

for each cellular component. It then uses a matrix projection approach to map these estimated coefficients to the relative proportions of each cellular component in the samples without cell type composition information. This method can be applied to other phenotypes besides cell type composition, if a reference dataset is available with both DNA methylation data and phenotype of interest. In the case of missing data, we use the set of samples with complete information (C_i^{obs}) as the reference dataset, and apply the projection approach to impute the missing values (C_i^{mis}).

2.3 Simulations

2.3.1 Simulation Settings

We conducted a simulation study to evaluate the performance of proposed methods for the type I error, statistical power, bias and coverage of the estimates. For each subject, the variable of interest, X, was generated from a uniform distribution between 0.1 and 0.9. A set of 3 covariate, Z_1, Z_2, and Z_3, were also simulated which may contain missing values.

For continuous confounders, we considered three simulation models (Figure 2.1), which generated methylation data for a total of 400 CpG sites on 2,000 subjects. Specifically, we simulated data for 50 CpG sites under model 1, 200 CpG sites under model 2, and 150 CpG sites under model 3. These CpG sites were analyzed together as from a single dataset. In our analyses, we used the first 250 CpG sites from model 1 and 2 to impute missing values in covariates Z for the projection-based method; for the two MI-based methods, up to 30 CpG sites among the 250 sites were selected to carry out multiple imputation. The association was then tested at each of the 400 CpG sites adjusting for imputed covariates. We further considered sample size of 500 and 1,000 subjects.

In simulation model 1 and 2, the covariates Z_1, Z_2, and Z_3 were correlated with X, but they were independent of X in model 3. We masked a proportion of samples as missing in these three covariates, with equal missing rate at 0.3 or 0.9. In simulation model 1, we simulated methylation levels at 50 CpG sites as follows:

$$Y^{(k)} = \beta^{(k)} Z + \gamma^{(k)} X + \epsilon, \qquad k = 1, 2, \ldots, 50,$$

where both β and γ are positive (Figure 2.1a). In model 2, we simulated methylation levels at 200 CpG sites using the same model above, with $\beta > 0$ but $\gamma = 0$ (Figure 2.1b). In these two models, the covariates Z confound the relationship between methylation level Y and variable X. We used the 50 CpG sites from model 1 to evaluate the statistical power of different imputation methods, and the 200 CpG sites from model 2 to evaluate the type I error in presence of confounding factors. In model 3, we simulated data at another 150 CpG sites where their methylation levels were correlated with covariates Z and X, but Z was independent of X (Figure 2.1c). Under this simulation model, we evaluated the statistical power of imputation methods when the missing covariates are independent risk factors.

To consider the effect of the categorical confounders, we further simulated binary covariates Z_4 (simulation model 4), which was correlated with X. In this model, we simulated methylation levels at 50 CpG sites as follows:

$$Y^{(k)} = \beta^{(k)} Z + \gamma^{(k)} X + \epsilon, \qquad k = 1, 2, \ldots, 50,$$

where $Z = (Z_1, Z_2, Z_3, Z_4)$ and both β and γ are positive (similar to simulation model 1). In model 5, we simulated methylation levels at 200 CpG sites using the same model above, with $\beta > 0$ but $\gamma = 0$ (similar to simulation model 2). In these two models, the covariate Z_4 confounds the relationship between methylation level Y and variable X, and is set as the missing covariate with 30% missing rate. Note that projection based method can still be applied in these two models by treating the binary covariate as continuous (a linear probability model).

We simulated 100 datasets and averaged the association test results of these 100 datasets and multiple CpG sites within each dataset to calculate bias and uncertainty of estimation. For each of imputation method, the coverage of estimation was calculated as the proportion of simulations that the 95% confidence interval contained the true value of covariate estimates. The power was calculated as the proportion of simulations that the association p-value is less than 0.05 in the simulation model 1, 3 and 4, and the type I error rate was calculated using data in model 2 and 5. We compared the results to those from the full data analysis, in which we assumed that we have all data without any missing value and adjust the covariates (Z) through a linear regression. We further

compared the results to those from the complete-case analysis, in which subjects with missing value were excluded.

2.3.2 Simulation Results

All the three imputation methods show non-inflated type I errors at different missing rates (Tables 2.1 and 2.2). At the significance level of 0.05, the two MI-based methods are slightly conservative (type I error $< 5\%$) and have wider coverage ($> 95\%$) than expected. The projection-based method has type I error rate close to the expected 5%. In contrast, the type I errors are inflated when the covariates Z are ignored in the association analysis and Z are correlated with both the variable of interest (X) and the methylation measure (Y) (not shown).

When the missing rate is relatively low (30%), the projection-based method has higher statistical power than the two MI-based methods (Figure 2.2a); when the missing rate is high (90%), the projection-based method and MI-PMM have similar power, which are higher than that of MI-Norm (Figure 2.2b). However, the projection-based method has down-wards bias in estimated effect size and narrower coverage less than 95% (Tables 2.1 and 2.2), especially when the missing rate is high (90%). When the effect size is large and missing rate is high, the coverage of estimated effect size from the projection-based method can be $< 80\%$, compared to the expected 95% (Table 2.2). In MI-based methods, the estimated effect size is very close to the true value and the coverage is slightly wider than those of the full data analysis (96% vs 95%). In addition, we compare the distribution of estimated effect size for different methods (Figure 2.3). The two MI-based methods yield consistent estimates for the true parameter values, but slightly large variation compared to full data analysis, especially when the missing rate is high (90%). The projection-based method consistently underestimates the effect size under the alternative hypothesis, but the variation of estimates is similar to that from full data analysis.

Note that simply excluding the subjects with missing covariates will lead to loss of power. Figure 2.2 also shows the statistical power of complete-case analysis, in which subjects with missing covariates are excluded. Comparing to the full data analysis or imputed data analysis (MI-Norm, MI-PMM, projection based method), complete-case analysis suffers from a loss of power, especially when the missing rate is high (Figure

2.2b).

We also evaluate the performance of various methods under different sample sizes ($n = 500$ or 1000). The results are consistent (not shown). When the missing covariate is binary, we applied the multiple-imputation based method (MI-Logreg), which shows similar performance as for continuous missing covariates (not shown).

When the covariates are only correlated with methylation levels but not the variable of interest, i.e., Z are independent predictors instead of confounders (simulation model 3), both the projection-based method and MI-based methods perform similarly to the full data analysis. All approaches have a controlled type I error, correct 95% coverage, and comparable statistical power (not shown).

2.4 ARIC Data Analysis: Association with Smoking Status

The Atherosclerosis Risk in Communities (ARIC) study is a prospective cohort study of cardiovascular disease risk in four U.S. communities. Between 1987 and 1989, 7,082 men and 8,710 women aged 45-64 years were recruited from Forsyth County, North Carolina; Jackson, Mississippi (African Americans only); suburban Minneapolis, Minnesota; and Washington County, Maryland. Before conducting the study, the institutional review boards of each participating university approved the protocol. After written informed consent including that for genetic studies was obtained, participants underwent a baseline clinical examination (Visit 1) and three subsequent follow-up clinical exams at intervals of roughly three years (Visits 2–4). An additional follow-up exam (Visit 5) was conducted in 2011–13.

Using the Illumina HM450 chip, bisulphite-converted DNA from 2,905 African American participants at Visit 2 (1990–92; $n = 2,504$) or Visit 3 (1993–95; $n = 441$) was measured for methylation status. Using Illumina GenomeStudio 2011.1, Methylation module 1.9.0 software, we determined the degree of methylation. Individuals ($n = 71$) were excluded from analysis if a pass rate for the DNA sample for the study participant was less than 99% (probes with a detection p-value > 0.01/all probes on the array), or the sample failed gender or genotype consistency checking. Probes on the HM450 chip for which the pass rate was less than 99% (sample with a detection p-value > 0.01 at

probe/all samples) were not analyzed ($n = 5,170$).

In an epigenome-wide analysis, we tested the association between DNA methylation (Y) and smoking status (never vs. current smoker) (X) at each CpG site using linear mixed effects model (LMM). The covariates (Z) included cell type composition, age, field center, gender and white blood cell count. Technical variables such as chip and chip position were included as random effects. Among the covariates, cell type composition and total white blood cell (WBC) counts contained considerable missing values, with $\sim 91\%$ and 14% missing respectively.

In the first analysis, we impute cell type composition using multiple imputation (30 imputations) and projection-based method, but only analyze samples with measured WBC. After imputation, we have 1,640 African American participants that are available for analysis. Using the projection-based method, we identify 2,552 CpG sites associated with smoking status with genome-wide significance (p-value $< 10^{-7}$). As a comparison, MI-Norm finds 1,362 significant CpG sites and MI-PMM finds 1,747 sites. Figure 2.4a shows a scatter plot of the $-\log_{10}$ transformed association p-values for the projection-based method and MI-PMM, suggesting smaller p-values for the projection-based method than MI-PMM in general. Figure 2.4b shows a comparison of the estimated coefficients at each CpG site. MI-PMM and the projection-based method show comparable estimates, especially at CpG sites with strong significance.

Our second analysis also imputes missing WBC values. The sample size after imputation increases to 1,932, which provides additional power for association tests. Figure 2.5 shows comparison of estimated coefficients and association p-values with and without imputed WBC. While the estimated coefficients are similar for the top signals, the association p-values are generally smaller in the second analysis.

2.5 Discussion

Missing data is a common problem in association studies, which can reduce the sample size and lead to decreased statistical power. Using genome-wide methylation data, we have proposed multiple imputation-based methods to impute missing covariates, which can be either continuous or categorical, using a set of selected CpG sites. We also extended an existing projection-based method for missing covariates imputation. Through

computer simulations, we evaluated performance of the 2 types of methods. Our results suggest that although the projection- based method has the correct type I error rate and greater power than the MI-based methods, it may yield biased estimate and coverage for the effect size. In contrast, the MI-based methods are slightly conservative, but can provide consistent estimates for the effect size.

One key assumption of the projection-based method is "biological orthogonality" [8], where the CpG sites used for missing phenotype imputation should not contain the sites that are associated with the covariate of interest. In practice, this assumption may not always be satisfied. In our simulations, 50 of the 250 CpG sites that are used to impute the missing covariates are correlated with the trait of interest, which leads to significant bias in the estimates of effect size. On the other hand, this assumption is not required for MI-based methods. In fact, the imputation model is recommended to include all variables that are in the analysis model to avoid estimation bias and maximize model certainty [29, 25].

The MI-based methods take into account of imputation uncertainty by adding random noise to the imputed values and averaging over multiple imputed datasets. In the simulation study, the MI-based methods show slightly conservative results compared to the full data analysis. This conservatism has been shown previously by some theoretical results [29]. The projection-based method can be viewed as an imputation approach with a single imputed dataset, and therefore ignores the uncertainty for imputed values. It can potentially underestimate the variance of estimated effect size and reduce the statistical power. However, our results show that the variance of estimates from the projection-based methods is similar to that from using the full data (Figure 2.4), possibly because the imputation that is based on many CpG sites has high accuracy and little imputation uncertainty compared to biological variation among samples. In our simulations, the correlation between imputed and true values is ~ 0.9 using the projection-based method.

For both the projection and MI-based methods, a pre-selected set of CpG sites is required which is associated with the covariates to be imputed. The projection-based method typically uses 100–500 CpG sites. In contrast, the MI-based method needs to limit the number of CpG sites to be much smaller than the sample size of full data, which can potentially reduce the imputation accuracy compared to the projection-based

method. In our analysis, we limited the number of CpG sites to be no larger than 30 in both MI-Norm and MI-PMM.

In addition to the imputation methods discussed in this chapter, another class of statistical methods can infer unobserved confounders based on singular value decomposition (SVD) of the residual (and coefficient) matrix, for example, the surrogate variable methods [4, 5] and the "reference-free" method [6]. These methods are robust to model specification in association analysis; however, they do ignore the partial information on the covariates provided by complete cases in the missing data scenario, which can lead to loss of statistical power.

In summary, we have proposed and evaluated imputation methods for missing values in covariates using high-dimen- sional DNA methylation data. The proposed methods will help to control for potential confounding and increase statistical power of epigenetic association studies. An R implementation of the methods is available at: `https://github.com/ChongWu-Biostat/MethyImpute`.

Table 2.1: Methods comparison for simulation model 1, missing rate is 30%, imputing time is 10 for MI method.

γ	Projection-based		MI-Norm		MI-PMM		Complete	
	Est	Cov	Est	Cov	Est	Cov	Est	Cov
0	0.0000	0.951	-0.0001	0.964	0.0001	0.963	0.0000	0.951
0.005	0.0048	0.951	0.0049	0.965	0.0050	0.964	0.0050	0.951
0.01	0.0095	0.946	0.0099	0.964	0.0100	0.965	0.0100	0.951
0.012	0.0114	0.943	0.0119	0.963	0.0120	0.965	0.0120	0.951
0.014	0.0134	0.941	0.0139	0.963	0.0140	0.966	0.0140	0.951
0.016	0.0153	0.939	0.0159	0.963	0.0160	0.966	0.0160	0.951
0.018	0.0172	0.937	0.0179	0.963	0.0180	0.966	0.0180	0.951
0.02	0.0191	0.935	0.0199	0.963	0.0199	0.964	0.0200	0.951
0.022	0.0210	0.932	0.0219	0.963	0.0220	0.964	0.0220	0.951

Est = Regression coefficient estimate while adjusting the confounders. Cov= 95% coverage rate. The effect size is the same for all CpG sites evaluated under the simulation model.

Table 2.2: Methods comparison for simulation model 1, missing rate is 90%, imputing time is 30 for MI method.

γ	Projection-based		MI-Norm		MI-PMM		Complete	
	Est	Cov	Est	Cov	Est	Cov	Est	Cov
0	0.0000	0.953	0.0002	0.963	0.0007	0.960	0.0000	0.951
0.005	0.0042	0.940	0.0051	0.966	0.0056	0.962	0.0050	0.951
0.01	0.0085	0.912	0.0100	0.965	0.0104	0.965	0.0100	0.951
0.012	0.0102	0.897	0.0119	0.962	0.0124	0.962	0.0120	0.951
0.014	0.0119	0.875	0.0139	0.960	0.0144	0.959	0.0140	0.951
0.016	0.0136	0.851	0.0158	0.959	0.0163	0.959	0.0160	0.951
0.018	0.0153	0.826	0.0177	0.961	0.0182	0.961	0.0180	0.951
0.02	0.0170	0.799	0.0197	0.960	0.0202	0.962	0.0200	0.951
0.022	0.0187	0.776	0.0217	0.960	0.0221	0.962	0.0220	0.951

Est = Regression coefficient estimate while adjusting the confounders. Cov= 95% coverage rate. The effect size is the same for all CpG sites evaluated under the simulation model.

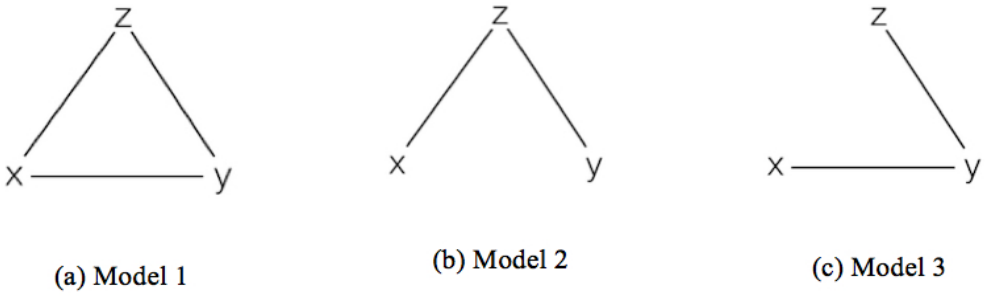

Figure 2.1: Simulation models 1–3. X: covariate of interest; Y: DNA methylation level; Z: covariate(s) that contain missing values.

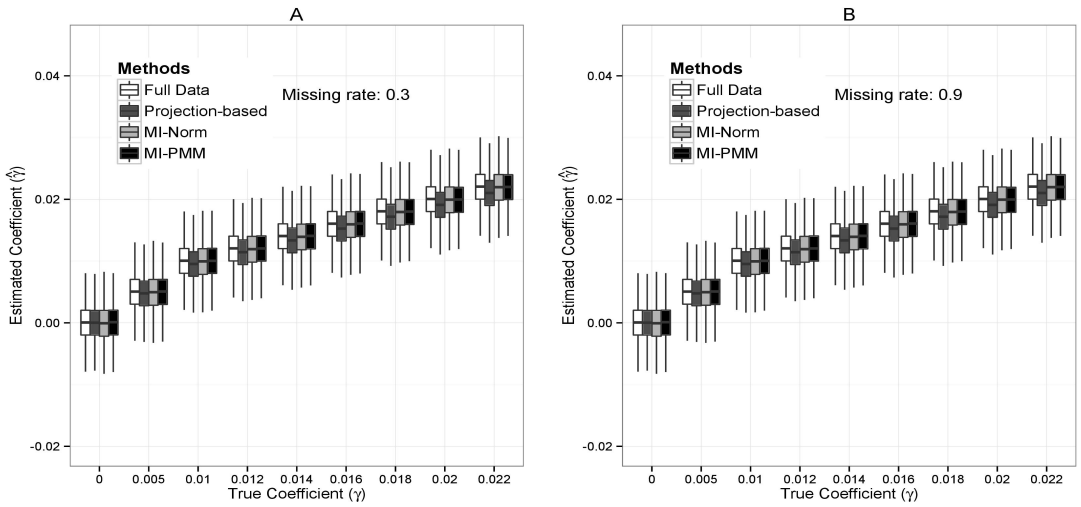

Figure 2.2: Power under different effect size for simulation model 1. Comparisons between resulting power using the projection-based method, MI-Norm, MI-PMM, full data (assuming that we have all data without any missing), and complete-case analysis (excluding subjects with missing values). (A) missing rate = 0.3; (B) missing rate = 0.9.

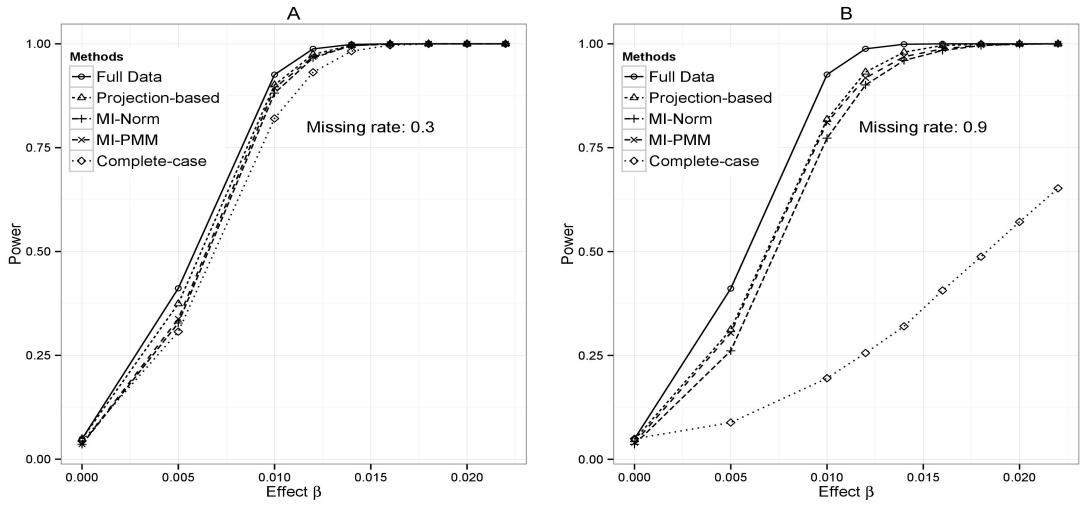

Figure 2.3: Boxplots for estimated effect size for simulation model 1. Comparisons between resulting estimated effect size using the projection-based method, MI-Norm, MI-PMM, full data (assuming that we have all data without any missing), and complete-case analysis (excluding subjects with missing values). (A) Missing rate = 0.3; (B) missing rate = 0.9.

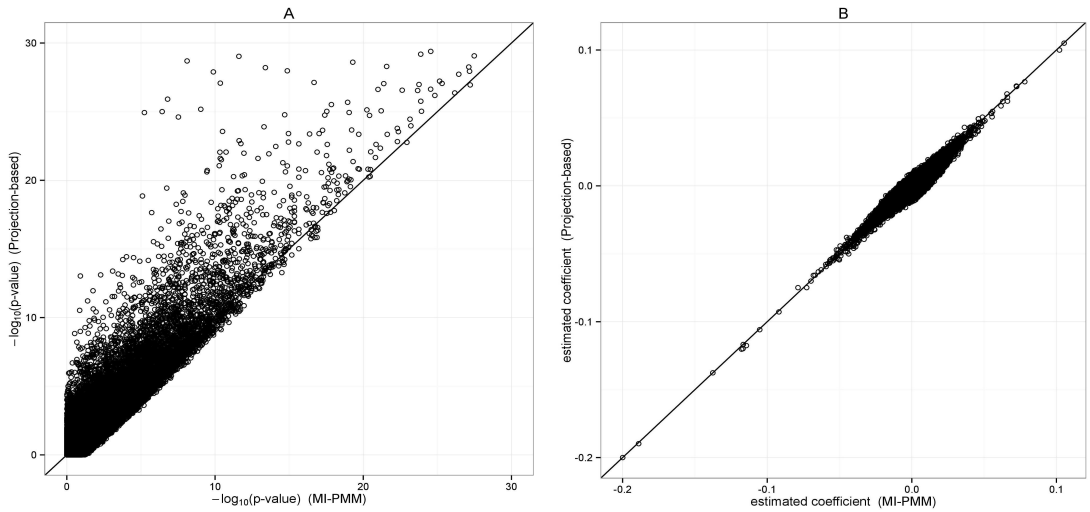

Figure 2.4: EWAS of smoking status: MI-PMM vs. the projection-based method. The figure is based on 1,640 participants after imputing the missing cell type composite using MI-PMM or projection-based method. (A) Comparison of $-\log_{10}$(p-value) from MI-PMM and from the projection-based method (p-values truncated at 10^{-30}); (B) Comparison of coefficient estimates from MI-PMM and from the projection-based method.

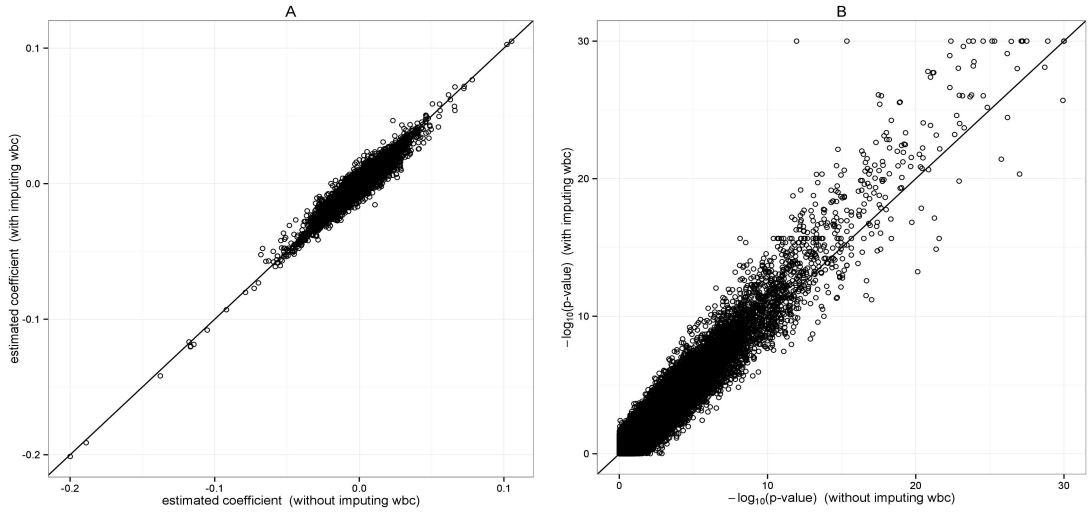

Figure 2.5: EWAS of smoking status with imputed WBC. Analysis is-based on MI-PMM. The data excluding observations with missing WBC contains 1,640 participants, while imputing missing WBC leads to 1,932 participants. (A) Comparison of coefficient estimate with and without imputing missing WBC values; (B) Comparison of $\log_{10}$(p-value) with and without imputing missing WBC values (p-values truncated at 10^{-30}).

Chapter 3

An Adaptive Association Test for Microbiome Data

There is an increasing interest in investigating how the compositions of microbial communities are associated with human health and disease. Although existing methods have identified many associations, a proper choice of a phylogenetic distance is critical for the power of these methods. To assess an overall association between the composition of a microbial community and an outcome of interest, this chapter presents a novel multivariate testing method called aMiSPU, that is joint and highly adaptive over all observed taxa and thus high powered across various scenarios, alleviating the issue with the choice of a phylogenetic distance. The main part of this chapter has been published in Wu et al. 2016 [30].

3.1 Introduction

A variety of microbial communities (i.e. microbiotas) and their genomes (i.e. microbiome) exist throughout the human body [31] and play an important role in one's overall health, such as food digestion and nutrition, development and regulation of the immune system, and prevention of invasion and growth of pathogens [32]. On the other hand, disruptions of the human microbial communities are associated with a wide range of human diseases, such as liver cancer [33], obesity [34], colorectal cancer [35], inflammatory bowel disease [36], type 2 diabetes [37], and antibiotics-associated diarrhea [38].

Understanding the association between human microbiotas and diseases might help diagnose disease and develop personalized medicine [39] to restore a disturbed microbial ecosystem to a healthy state, for instance, using a personalized synthetic community and complementary set of nutrients [32].

Recent advances in sequencing technologies have made it feasible to profile microbiotas in a large number of samples via targeted sequencing of the 16S rRNA gene [40], and extend the study of human genome to the human microbiome, which consists of the collection of the microbial genomes at the various sites of the human body, seen as an "extended" human genome [10]. Many human microbiome studies aim to detect a possible association of the human microbiome with a phenotype, such as a disease status, called outcome (of interest) here, after adjusting for potential confounders. These association studies not only can improve our understanding of the non-genetic components of complex traits and diseases, but also might open up an entirely new way for drug development. Although univariate tests (on a single taxon one by one) are widely used in analysis of differential abundance, multivariate tests (on multiple taxa jointly and simultaneously) have become increasingly popular due to their higher statistical power in aggregating multiple weak associations and reducing the burden of multiple testing. Furthermore, many univariate tests critically depend on some strong parametric assumptions on the distributions or mean-variance functional forms for microbiome data, leading to inflated type I errors when the assumptions are violated [41]. In contrast, no such assumption is imposed in our proposed multivariate test, which, coupled with a proposed permutation procedure for p value calculation, is essentially semi-parametric and applicable to even small samples. In this chapter, we mainly focus on multivariate tests.

One popular method for testing the association between an overall microbiome composition and an outcome of interest is to use a distance- or dissimilarity-based test, such as PERMANOVA [12]. Via the standard pipelines such as QIIME and mothur [42, 43], the 16S sequence tags are usually clustered into operational taxonomic units (OTUs), which can be considered surrogates for biological taxa within a specified amount of sequence divergence allowed for each OTU. At 97% similarity, these OTUs represent common 'species'. A specific distance measure is chosen to measure the dissimilarity between each pair of samples, taking into account the phylogeny among taxa. Then the

pairwise distance is compared to the distribution of the outcome of interest for evaluating the association between the overall microbiome composition and the outcome. Recently, a new method called microbiome regression based kernel association test (MiRKAT) was proposed [13]. Incorporating phylogenetic relationships among taxa, MiRKAT transforms a phylogenetic distance metric into a kernel to measure similarities among samples. Then a semi-parametric kernel machine regression framework is applied to evaluate the association. MiRKAT allows for an easy covariate adjustment and extensions to other types of the outcome. By the correspondence between the distance-based association testing and kernel machine regression [44, 13], MiRKAT is closely related to distance-based methods, such as PERMANOVA. In addition, MiRKAT provides an omnibus test that combines several relevant kernels to be more robust across different scenarios. However, the choice of kernels has to be decided by the end user, and more importantly, no automatic taxon selection or weighting implemented in the framework.

Up till now, numerous distance measures have been developed to depict community differences between two samples. Among many possible distance metrics, the UniFrac-type distance metrics are most popular, which account for phylogenetic relationships among microbial taxa [45, 46, 47]. There are several different versions of UniFrac. The unweighted UniFrac distance [45], which is defined as the fraction of the branch length of the tree that leads to descendants from either sample, but not both, is a qualitative diversity measure and is very efficient in detecting abundance changes in rare taxa given that more prevalent species are likely to be present in all individuals. In contrast, the weighted UniFrac distance [46], which weights the branches of a phylogenetic tree based on the abundance differences, is more sensitive to changes in abundant taxa. The generalized UniFrac distance [47] was introduced to unify the weighted and unweighted versions by striking a balance in weighting between relative differences and absolute differences. Many other distances ignoring phylogenetic information are also available. Bray-Curtis distance [48], for example, quantifies the taxonomic dissimilarity between two samples on the basis of the OTU counts only.

Noise accumulation is a vital problem for high dimensional data. For example, due to noise accumulation in estimating the population centroids in a high dimensional feature space, classification using all features can be as bad as a random guess [49]. A severe limitation of kernel or distance-based methods is that they do not conduct

variable selection or variable weighting, which is crucial for high-dimensional microbiome data so that noise accumulations can be alleviated. In particular, with the dimension much larger than the sample size, some and even most of microbial taxa may not be associated with the outcome; without variable selection or weighting, using all the taxa for distance or kernel calculations simply contributes noises, leading to power loss as to be shown. Therefore, differential weighting of the microbial taxa according to their importance can potentially improve the power of microbiome association test. We thus propose a data-driven approach to achieve adaptive weighting of the taxa based on the data. The proposed method is based on a generalized taxon proportion combining microbial abundance information with phylogenetic tree information and an adaptive test called aSPU test, which is based on a family of sum of powered score (SPU) tests [50], incorporating variable weighting. Each SPU test is indexed as SPU(γ) by an integer $\gamma > 0$ that controls the extent of weighting on the variables; we call the corresponding tests MiSPU(γ) (microbiome based sum of powered score) and aMiSPU (adaptive MiSPU) when applied to microbiome data. We will demonstrate through numerical simulations and analysis of real data that aMiSPU can be easily applied and much more powerful than existing tests in most scenarios with well-controlled type I error rates. Although aMiSPU was inspired by the aSPU test, the two differ in whether and how to accommodate unique features of microbial data. In particular, we propose the generalized taxon proportion to combine microbial abundance information and phylogenetic tree information simultaneously. As shown in numerical simulations, directly applying the aSPU test with OTU abundances generally failed to achieve high power. Finally, an R package *MiSPU* that implements MiSPU with a C++ version of UniFrac distance calculation has been developed.

3.2 Methods

3.2.1 Data and Notation

Suppose n samples have been collected, each with a microbial community profile. For sample i, let Y_i denote an outcome of interest, which can be binary (e.g., disease status) or continuous. Let $X_i = (X_{i1}, \ldots, X_{ip})$ be the p covariates, such as age, gender and other clinical and environmental variables that we want to adjust for, and $Z_i = (Z_{i1}, \ldots, Z_{im})$

be the abundances of m taxa derived from observed q OTUs for the ith sample. Note that an OTU represents a common species while a taxon is a group of one or more species. Here, we assume that each of $Z_{i1}, \ldots, Z_{iq}$ is the count of the OTU in sample i and Z_{ik}, $q+1 \leq k \leq m$, is the sum of the counts of the OTUs belonging to taxon k in sample i. The evolutionary relationships among these OTUs and taxa are given by a rooted phylogenetic tree, which contains all q OTUs (as leaf nodes) and $m-q$ taxa (as internal nodes). Suppose b_k is the distance from the root of the phylogenetic tree to taxon k, and $p_{ik} = Z_{ik}/\sum_{j=1}^{q} Z_{ij}$ is the proportion of taxon k in sample i. The goal is to test for a possible association between the overall microbial community composition and the outcome of interest after adjusting for the covariates.

3.2.2 A New Class of Tests—MiSPU

The MiSPU and aMiSPU tests are introduced in this subsection. Figure 3.1 illustrates the overall structure of the tests, detailing the input (a rooted phylogenetic tree, a sample for OTU counts, an outcome of interest and possibly some covariates) and the three key steps: calculating a generalized taxon proportion for each taxon, calculating the test statistics, and applying a residual permutation scheme to obtain the p values.

One major characteristic of microbial composition data is that taxa are related as described by a phylogenetic tree. Phylogenetic distance measures that account for phylogenetic relationships among taxa can be much more powerful than those ignoring evolutionary information [47]. Among these, UniFrac distances are most popular. Consider two samples i and j, the unweighted UniFrac distance, which considers only species presence or absence, is a qualitative measure and defined as [45]:

$$d_{ij}^{U} = \frac{\sum_{k=1}^{m}\{b_k|I(p_{ik}>0) - I(p_{jk}>0)|\}}{\sum_{k=1}^{m} b_k},$$

where $I(\cdot)$ is the indicator function. In contrast, the weighted UniFrac, which uses OTU abundance information, is a quantitative measure [46]:

$$d_{ij}^{W} = \frac{\sum_{k=1}^{m} b_k|p_{ik} - p_{jk}|}{\sum_{k=1}^{m} b_k|p_{ik} + p_{jk}|}.$$

Our basic observation is that, phylogenetic distance metrics, which account for the

relationship among taxa via a phylogenetic tree, measures the distance among samples using all the variables (i.e. taxa) without variable selection or variable weighting. Since the dimension of microbial data is usually high, much larger than the number of samples, many taxa may provide only weak or no signals. Using a phylogenetic distance without variable weighting or variable selection may or may not be powerful. Instead, corresponding to the unweighted and weighted UniFrac distances, for each sample i and taxon k, we define the corresponding generalized taxon proportions as

$$Q_{ik}^u = b_k I(p_{ik} > 0), \qquad Q_{ik}^w = b_k p_{ik},$$

respectively. Note that the raw weighted UniFrac distance [46] between two samples is exactly the same as the L_1 distance of the weighted generalized taxon proportion between the two samples.

Inspired by a multivariate test for association analysis of rare variants [50], we construct a class of versatile score-based tests such that for a given scenario, at least one of the tests is powerful. Then we combine these tests to maintain high power across a wide range of scenarios. Specifically, for a binary outcome, we use a logistic regression model:

$$\text{Logit}[Pr(Y_i = 1)] = \beta_0 + \beta' X_i + \sum_{k=1}^{m} Q_{ik}\varphi_k,$$

where Q_{ik} is either Q_{ik}^u or Q_{ik}^w.

For a continuous outcome, we use a linear model:

$$Y_i = \beta_0 + \beta' X_i + \sum_{k=1}^{m} Q_{ik}\varphi_k + \epsilon_i,$$

where ϵ_i is an error term with mean 0 and variance σ^2.

We are interested in testing the null hypothesis $H_0 : \varphi = (\varphi_1, \ldots, \varphi_m)' = 0$; that is, there is no association between any taxa and the outcome of interest under H_0. The

score vector $U = (U_1, \ldots, U_m)'$ for φ is [44, 51, 52, 50]:

$$U = \sum_{i=1}^{n} (Y_i - \hat{\mu}_{i,0}) Q_{i\cdot},$$

where $Q_{i\cdot} = (Q_{i1}, Q_{i2}, \ldots, Q_{im})$ and $\hat{\mu}_{i,0}$ is the predicted mean of the outcome of interest (Y_i) under H_0. Note that a general weighted score-based test can be written as

$$T_G = w'U = \sum_{k=1}^{m} w_k U_k,$$

where $w = (w_1, \ldots, w_m)'$ is a vector of weights for the m generalized taxon proportions. Most existing association tests use the score vector U to construct a test statistic, because of the closed-form of the score vector U and that most information in data is contained in U. Therefore, we use U to construct the weights for the score vector U. Under H_0, we have $U \sim N(0, Cov(U|H_0))$ asymptotically, suggesting that a larger $|U_k|$ offers stronger evidence to reject $H_{0,k} : \varphi_k = 0$. Specifically, we choose $w = (U_1^{\gamma-1}, \ldots, U_m^{\gamma-1})'$ to weight the score vector for the generalized taxon proportions, leading to a MiSPU test:

$$T_{\text{MiSPU}(\gamma)} = w'U = \sum_{k=1}^{m} U_k^{\gamma}.$$

Since $\gamma = 1$ essentially treats all the variables equally important while association directions of the generalized taxon proportions may vary, $\gamma = 1$ often yields low power and thus is excluded here. Importantly, as γ increases, the MiSPU(γ) test puts more weights on the larger components of U while gradually ignoring the remaining components. As γ goes to infinity, we have

$$T_{\text{MiSPU}(\infty)} \propto ||U||_{\infty} = \max_{k=1}^{m} |U_k|.$$

We simply define $T_{\text{MiSPU}(\infty)} = \max_{k=1}^{m} |U_k|$. Note that the two versions of Q_{ik}, i.e. Q_{ik}^w and Q_{ik}^u, yield the weighted MiSPU$_w$ and unweighted MiSPU$_u$ respectively.

We use a permutation scheme [50] to calculate the p value as the following.

1. Fit the null linear or logistic regression model by regressing Y on the covariates

X under H_0 to obtain $\hat{\mu}_{i,0} = E(Y_i|H_0)$ and residuals $r_i = Y_i - \hat{\mu}_{i,0}$.

2. Permute the residuals $r = \{r_i|i = 1, \ldots, n\}$ to obtain a permuted set $r^{(b)}$.

3. Regress Q on the covariates X to obtain the residuals $\hat{Q}$.

4. Calculate the new score based on the permuted residuals as $U^{(b)} = \sum_{i=1}^{n} \hat{Q}_i . r_i^{(b)}$ and the corresponding null statistic $T_{\text{MiSPU}}^{(b)} = T_{\text{MiSPU}}(U^{(b)})$.

5. Calculate the p value as $\left[\sum_{b=1}^{B} I\big(|T_{\text{MiSPU}}^{(b)}| \geq |T_{\text{MiSPU}}|\big) + 1\right]/(B+1)$ after B permutations.

It would be desirable to data-adaptively choose the value of γ and the version of the generalized taxon proportion since the optimal choice of them depends on the unknown true association patterns. Similar to the adaptive SPU (aSPU) test [50], we propose an adaptive MiSPU (aMiSPU) test, which combines the p values of multiple MiSPU tests with various values of γ and two versions of Q_{ik}. Suppose that we have some candidate values of γ in Γ, e.g., $\Gamma = \{2, 3, \ldots, 8, \infty\}$ as used in our later simulations and real data analysis; then our combining procedure is to take the minimum p value:

$$
\begin{aligned}
T_{\text{aMiSPU}_u} &= \min_{\gamma \in \Gamma} P_{\text{MiSPU}_u(\gamma)}; \\
T_{\text{aMiSPU}_w} &= \min_{\gamma \in \Gamma} P_{\text{MiSPU}_w(\gamma)}; \\
T_{\text{aMiSPU}} &= \min\{P_{\text{aMiSPU}_u}, P_{\text{aMiSPU}_w}\}.
\end{aligned}
$$

Note that we take the minimum p value of aMiSPU$_u$ and aMiSPU$_w$ to form the final aMiSPU test. T_{aMiSPU_u}, T_{aMiSPU_w} and T_{aMiSPU} are no longer a genuine p value, but we can use the permutation to estimate its p value, using the same set of the null statistics used to calculate the p values for the MiSPU tests [50].

We comment on the choice of Γ and the version of the generalized taxon proportion. Depending on how many taxa are truly associated with the outcome of interest, one may use a smaller or larger γ. For example, if more of the taxa are not associated, a larger γ would be desirable. In our numerical simulations and real data analysis, we have found that $\Gamma = \{2, 3, \ldots, 8, \infty\}$ often suffices. MiSPU(8) often gives almost the same results as those of MiSPU(∞), suggesting no need to use other larger γ's. In practice, we

suggest using the aMiSPU test that combines the strengths (and possibly weaknesses) of various MiSPU tests; the aMiSPU test can be regarded as a rigorous means for multiple testing adjustment with the use of several MiSPU tests, while the results of MiSPU tests may shed light on the underlying association patterns. For example, if a MiSPU with the unweighted generalized taxon proportion gives the most significant p value, it may indicate the outcome of interest is more likely to be associated with the abundance changes in rare taxa; if some odd γ's yield more significant results than even γ's, then most or all of the large associations are in the same direction.

Although we focus on rRNA sequencing data, the proposed method can be applied to metagenomic whole genome shotgun sequencing data as well. Via MEGAN [53], DNA reads (or contigs) can be summarized as OTUs and their counts. Using a standard algorithm, species-specific sequences are assigned to OTUs or taxa near the leaves of a phylogenetic tree, whereas widely conserved sequences are assigned to taxa closer to the root [53]. Once we have OTU abundance data and a phylogenetic tree, aMiSPU can be applied as before.

3.2.3 Taxon Selection

A limitation of most multivariate tests is their inability for variable selection: even if the null hypothesis is rejected, they may not give any information on which taxa are (or are not) likely to be associated with the outcome of interest. We note that the aMiSPU test can be used to rank the importance of the taxa. First, if $P_{\text{aMiSPU}_u} < P_{\text{aMiSPU}_w}$, we use unweighted generalized taxon proportion in the subsequent analysis; otherwise, we use the weighted one. For ease of exposition, suppose we choose the weighted one. Second, we estimate the optimal value of $\hat{\gamma} = \text{argmin}_{\gamma\in\Gamma} P_{\text{MiSPU}_w(\gamma)}$ chosen by the aMiSPU$_w$ test. If $\hat{\gamma} = \infty$, we can easily find the most significant taxon. Third, suppose $\hat{\gamma} < \infty$, we asses the relative contribution of each taxon r to the aMiSPU$_w$ test as $\mathcal{C}_r = |U_r|^{\hat{\gamma}} / \sum_{j=1}^{m} |U_r|^{\hat{\gamma}}$. Fourth, we rank the taxon based on their $\mathcal{C}_r$ values, and we can select a few top k_1 taxa, such as $k_1 = 1$, or such that the sum of their relative contributions $\sum_{r=1}^{k_1} \mathcal{C}_r \geq \alpha_1$ with $\alpha_1 = 0.7$, say; the choice of k_1 or α_1 determines the trade-off between increasing true positives and increasing false positives.

3.2.4 The `MiSPU` Package and Implementation

We implemented the MiSPU and aMiSPU tests in an R statistical software package called `MiSPU`, in which a C++ version of UniFrac distances faster than the `GUniFrac` R package is also provided. The package is available on GitHub (`https://github.com/ChongWu-Biostat/MiSPU`) and CRAN. We applied MiRKAT available in the `MIRKAT` R package developed by Ni Zhao and Michael Wu at website `http://research.fhcrc.org/wu/en/software.html`. The SPU and aSPU tests are available in the R package `aSPU` on CRAN.

3.2.5 Simulation Settings

We used a phylogenetic tree of OTUs from a real throat microbiome dataset [54], which consists of 856 OTUs after discarding singleton OTUs. The simulation settings was similar to that used in [13]. Specifically, we generated the OTU counts for each individual via the following steps.

1. Based a real throat microbiome dataset [54], the estimated OTU proportions $(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_{856})$ as well as the estimated overdispersion parameter $\hat{\theta}$ were obtained via maximum likelihood.

2. For sample i, the observed OTU proportions were randomly generated from a Dirichlet distribution: $(p_{1i}, p_{2i}, \ldots, p_{856i}) \sim \text{Dirichlet}(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_{856}, \hat{\theta})$.

3. The total count of OTUs for sample i, say n_i, was randomly drawn from a negative binomial distribution with mean 1000 and size 25. This step mimicked varying total reads per sample.

4. For sample i, the observed OTU counts were randomly generated from a multinomial distribution: $(Z_{i1}, Z_{i2}, \ldots, Z_{i856}) \sim \text{Multinomial}(n_i; p_{1i}, p_{2i}, \ldots, p_{856i})$.

The procedure of generating simulated data is available as a function in R package `MiSPU`. We considered several simulation scenarios that differed in how some OTUs were related to the outcome of interest.

Under simulation scenario 1, we partitioned the 856 OTUs into 20 clusters (lineages) by performing partition around medoids (PAM) based on the cophenetic distance matrix. The abundance of these 20 OTU clusters varied tremendously, such that each OTU

cluster corresponded to some possible bacterial taxa. We assumed that the outcome of interest depended on the abundance cluster that constituted 6.7% of the total OTU reads. Then we simulated dichotomous outcomes as follows:

$$\text{Logit}\left(E(Y_i|X_i, Z_i)\right) = 0.5 f_{\text{scale}}(X_{1i} + X_{2i}) + \beta f_{\text{scale}}\Big(\sum_{j\in A} Z_{ij}\Big),$$

where β was the effect size and $f_{\text{scale}}(Z_{i1})$ standardized the sample mean of $Z_{i\cdot}$'s to 0 and standard deviation to 1. For continuous outcomes, we simulated under the model

$$Y_i = 0.5 f_{\text{scale}}(X_{1i} + X_{2i}) + \beta f_{\text{scale}}\left(\sum_{j\in A} Z_{ij}\right) + \epsilon_i,$$

where $\epsilon_i \sim N(0, 1)$. X_{1i} and X_{2i} were the covariates to be adjusted for, and A was the index set of the selected OTU cluster. X_{1i} was generated from a Bernoulli distribution $\text{Bin}(1, 0.5)$, while X_{2i} was from a standard normal distribution $N(0, 1)$. To consider the effect of potential confounders, we studied the case where X_{2i} and Z_i were correlated; specifically, $X_{2i} = f_{\text{scale}}\left(\sum_{j\in A} Z_{ij}\right) + N(0, 1)$. And we varied the effect size β to mimic different magnitudes of association.

Under simulation scenario 2, we partitioned all the OTUs into 40 clusters and assumed the outcome was associated with the abundance cluster with only 3 OTUs. Under simulation scenarios 3, 4 and 5, we assumed that the outcome of interest was associated with the abundance cluster with 24.8%, 16.6% and 1.5% of the total OTU reads, respectively. Under simulation scenario 6, we assumed the outcome was associated with randomly selected 50 OTUs.

For all the simulation scenarios, we considered using MiSPU_u and MiSPU_w with $\gamma = 2, 3, \ldots, 8$. We combined the MiSPU tests to get aMiSPU_u, aMiSPU_w and aMiSPU. We compared aMiSPU with MiRKAT with the weighted and unweighted UniFrac kernels (K_w and K_u, respectively), the Bray-Curtis Kernel (K_{BC}), and a generalized UniFrac Kernel with $\alpha = 0.5$ (K_5), respectively. Additionally, we also applied the optimal MiRKAT, which combines the above four kernels.

Throughout the simulations, the sample size and test significance level were fixed at 100 and $\alpha = 0.05$, respectively. The results were based on 1,000 independent replicates

for $\beta \neq 0$ and 10,000 independent simulations for $\beta = 0$.

3.3 Results

In this section, we present the simulation results from MiSPU and MiRKAT, as well as the results for three real datasets.

3.3.1 Numerical Simulation Results for Type I Error and Power

To save space, we focus on a few simulation set-ups with a binary outcome. First, the type I error rates of MiSPU and aMiSPU across different simulation set-ups were satisfactorily controlled when the confounders were suitably adjusted (Table 3.1). When the covariates were independent of the microbiome composition, MiSPU and aMiSPU controlled type I error rate well no matter whether we adjusted for X or not. In comparison, when X and Z were correlated, adjusting for X was necessary: failing to adjust for X led to an inflated type I error. We only showed some results here and the type I errors at the nominal $\alpha = 0.01$ level were also investigated with the same conclusion (not shown)

Figure 3.2 shows statistical power with a binary outcome in simulation scenario 1, in which a phylogenetic cluster with 6.7% OTUs was associated with the outcome. For all the tests considered, the power increased when the effect size increased. Due to the upweighting of the more likely to be informative microbial taxa, a MiSPU$_w$ was much more powerful than a MiRKAT test, regardless of whether X and Z were correlated or not. Because only a few taxa were related to the outcome of interest, a MiSPU(γ) test with a larger γ performed slightly better than that with a smaller γ. Nevertheless, MiSPU$_w$(2) still performed much better than any MiRKAT. Compared to the MiSPU$_w(\infty)$, the aMiSPU$_w$ combining different weights with various γ values lost some power but still maintained power considerably higher than that of many other tests. As expected, by ignoring phylogenetic information of the microbiome data, the SPU and the aSPU tests [50] failed to achieve high power (not shown). Since there were some abundant OTUs in the informative cluster A, the unweighted UniFrac suffered from a loss of power and led to the failure of aMiSPU$_u$ to improve power. However, aMiSPU combining aMiSPU$_u$ and aMiSPU$_w$ lost only little power as compared

to aMiSPU$_w$.

Figure 3.3 shows the statistical power with a binary outcome in simulation scenario 2, where a small phylogenetic cluster with only contains 3 OTUs was associated with the outcome. We again show the empirical power curves when X and Z were independent (Figure 3.3a) and when X and Z were correlated (Figure 3.3b). Results were similar to those of simulation scenario 1, except that the aMiSPU$_u$ performed better than aMiSPU$_w$. aMiSPU, which combines aMiSPU$_u$ and aMiSPU$_w$, lost only little power as compared to the best choice of MiSPU, but remained much more powerful than any of MiRKAT. As expected, the weighted UniFrac kernel was the least powerful.

Other simulations (scenario 3,4, and 5) showed consistently that aMiSPU generally outperformed MiRKAT and aSPU when a phylogenetic cluster was associated with the outcome). However, when some randomly selected OTUs were associated with the outcome (scenario 6), the aSPU test was the winner; however, we comment that this scenario may not be realistic.

In practice, the true state of nature can vary from case to case. The simulation results showed that the power of MiRKAT essentially depends on the chosen kernel; a poor choice of the kernel leads to a tremendous loss of power. In contrast, MiSPU uses the generalized taxon proportion Q_{ik} and puts higher weight on more likely to be informative taxa, achieving much higher power than MiRKAT in most situations. The performance of MiSPU is also dependent on the choice of γ and the version of generalized taxon proportion: a better choice leads to higher power. However, aMiSPU alleviates this problem by combining MiSPUs with different γ and the two versions of the generalized taxon proportion, and is the overall winner over a wide range of different scenarios.

Univariate testing on each OTU or taxon one by one incurs a heavy burden for a correction for multiple testing; often the easy-to-use but conservative Bonferroni method is applied, leading to reduced power. Compared to multivariate testing methods, such as MiSPU and MiRKAT, the power of the nonparametric Kruskal-Wallis test [55, 56] was very low (Figure 3.2A). Even worse, many parametric univariate tests, due to their strong parametric assumptions on the distributions or parametric specifications on the mean-variance forms for the OTU counts, may have inflated false positive rates, as pointed out by others [57, 41]. For example, in our simulations under scenario 1, the

empirical type I error rates for DESeq2 [58] and metagenomeSeq-fitZig [59] were inflated. Accordingly, we did not further investigate their power properties. Relevantly and importantly, univariate tests encounter a so-called curse of compositionality problem: since the increased (or decreased) relative abundance of some OTUs necessarily lead to other (null or unmodifieded) OTUs to have opposite changes in their relative abundance, false positives result for some null OTUs. In contrast, multivariate joint testing methods, such as PERMANOVA, MiRKAT and aMiSPU, do not suffer from this curse of compositionality problem.

3.3.2 Numerical Simulation Results for Taxon Selection

Beyond an overall assessment of association, several methods [55, 56, 58, 59] have been developed for identifying specific OTUs driving a detected association. For example, since the compositions of potentially pathogenic bacteria across healthy and disease populations might be different, identifying such bacteria is of interest. One byproduct of the aMiSPU test is to rank the importance of the taxa. We evaluated this approach using simulated data under scenario 1 with an effect size equal to 2, and compared the results to those of the other metagenomic tools, metagenomeSeq-fitZig [59], a Kruskal-Wallis (KW) test as used in LEFSe (linear discriminant analysis effect size) [55] and STAMP [56], and DESeq2 [58], a representative for RNA-seq analysis.

Simulation results under scenario 1 were summarized in Table 3.2. The informative OTU set contained 57 OTUs. On average, the taxon set selected by aMiSPU contained 58.5 OTUs, 27.2 of which were truly informative. In contrast, fitZig [59] selected 157 OTUs and only 12.3 OTUs were truly informative. Perhaps due to the failure to consider the fact that most OTUs in a microbiome association study were rare, DESeq2 and KW test performed poorly with a too small mean number of true positives. Under scenario 1, we chose a relatively abundant OTU cluster that contained 57 OTUs to be related to the outcome. As expected, incorporating phylogenetic tree information helped us select truly informative abundant OTUs, thus aMiSPU performed better. In contrast, with only a moderate effect size for each informative OTU, a univariate association test was much less powerful in identifying informative OTUs.

3.3.3 Analysis of a Gut Microbiome Dataset for Gender and Diet Effects

Diet strongly affects human health, partly by modulating gut microbiome composition. Wu et al. [60] investigated the association of dietary and environmental variables with the gut microbiota, where the diet information was converted into a vector of micro-nutrient intakes. In this cross-sectional study, 98 healthy volunteers were enrolled and habitual long-term diet information was collected using food frequency questionnaire. The questionnaires were converted to intake amounts of 214 micro-nutrients, which was further normalized via residual method to standardize for caloric intake. Stool samples were collected; DNA samples were analyzed and denoised prior to taxonomic assignment. The denoised sequences were then analyzed by the QIIME pipeline [43] with the default parameter settings in the QIIME pipeline, yielding 3071 OTUs after discarding the singleton OTUs.

Increasing evidence suggests that there is sex difference in the human gut microbiome, which in turns modulates many pathological and physiological processes [61, 62]. However, no significant sex effect was detected using PERMANOVA based on this dataset [60]. We thus re-analyzed the dataset for gender effect by applying MiRKAT and MiSPU with $100,000$ permutations. Using MiRKAT, we found the p values from weighted UniFrac, unweighted UniFrac and Bray-Curtis kernel to be 0.035, 0.039, and 0.087 respectively. The optimal MiRKAT generated a p value of 0.080, failing to reject the null hypothesis even at the $\alpha = 0.05$ significance level. In comparison, MiSPU$_w(2)$, MiSPU$_w(3)$, MiSPU$_w(8)$, MiSPU$_w(\infty)$ provided the p values of 0.011, 0.0018,0.0022, 0.0022, respectively. MiSPU$_w(3)$, provided the most significant p value, suggesting that there is a sparse association pattern between gut microbiome composition and the gender status and the large associations between gender and and one or few microbial taxa were in the same direction. The aMiSPU, combining the weighted, unweighted generalized taxon proportions and $\gamma = \{2, 3, \ldots, 8, \infty\}$, yielded a p value of 0.0058, rejecting the null hypothesis at the $\alpha = 0.01$ significance level, suggesting an association between gender status and microbiome composition. Note that perhaps due to the relatively high signal sparsity, previous studies [60, 63] using distance based methods [12] failed to find any association. Unlike MiRKAT and distance based analyses, the aMiSPU test can be used for taxon selection. Since MiSPU$_w(3)$ provided the most significant p value,

we used the weighted generalized taxon proportion and $\hat{\gamma} = 3$. We found that a taxon in *Bacteroides* explained more than 90% relative contributions. The top 4 taxa all came from the *Bacteroides*, suggesting that gender was likely associated with *Bacteroides*, but independent with other enterotypes (Figure 3.4).

One goal of the study is to identify nutrients that are associated with the gut microbiome composition. We re-analyzed the data from the gut samples by using MiRKAT [13] and aMiSPU. Specifically, we applied the optimal MiKRAT test to analyze the association between each nutrient and microbial community composition by combining the weighted and unweighted UniFrac distances, generalized UniFrace distance with $\alpha = 0.5$ and the Bray-Curtis distance (after being transformed to the corresponding similarity matrices). We further applied aMiSPU$_u$ and aMiSPU$_w$ with $\gamma = 2, 3, \ldots, 8, \infty$. Then we combined aMiSPU$_u$ and aMiSPU$_w$ for aMiSPU. Figure 3.5 shows that there was no uniformly most powerful test; depending on the unknown truth, including specific association directions and effect sizes, a given test may or may not be most powerful. Perhaps due to the sparse association between some of the nutrients and microbial community composition, aMiSPU$_u$ detected some signals undiscovered by others.

3.3.4 Analysis of a Gut Microbiome Dataset for Association with IBD

The disruption of the gut microbiota is thought to have an important effect on the etiology of inflammatory bowel diseases (IBDs) such as Crohn's disease (CD) and ulcerative colitis (UC). Willing et al. [36] explored the composition of the IBD gut microbiome and identified some IBD-associated bacterial signatures. In this cohort study, 40 twin pairs who were concordant or discordant for CD or UC were collected and the compositions of microbial communities in feces samples were determined via 454 pyrotag sequencing. Sequences were checked for quality and those that were less than 200 base pairs in length, contained incorrect primer sequences, or contained more than 1 ambiguous base were discarded [36].

We tested the association between the disease status and the overall microbiome composition via MiRKAT and MiSPU using 10,000 permutations. MiRKAT yielded the p values from weighted UniFrac, unweighted UniFrac and Bray-Curtis kernels to be 0.223, 0.059, and 0.475 respectively. The optimal MiRKAT generated a p value of 0.144, failing to reject the null hypothesis even at the $\alpha = 0.10$ significance level.

In comparison, MiSPU$_u(2)$, MiSPU$_u(3)$, MiSPU$_u(\infty)$ provided the p values of 0.036, 0.053,0.084, respectively. The aMiSPU test, combining the weighted and unweighted generalized taxon proportions and $\gamma \in \{2, 3, \ldots, 8, \infty\}$, yielded a p value of 0.097, slightly smaller than 0.10, rejecting the null hypothesis at 0.10 significance level. None of these tests could reject the null hypothesis at the $\alpha = 0.05$ significance level, perhaps due to the small sample size. Note that, perhaps the disease status was more likely to be associated with abundance changes in rare taxa, MiSPU$_u$ provided a more significant p value than that of MiSPU$_w$.

3.3.5 Analysis of a Throat Microbiome Dataset for Smoking Effects

Cigarette smokers have an increased risk of infection involving the respiratory tract. Recently, a microbiome-profiling study was conducted to investigate the smoking effect on the oropharyngeal and nasopharyngeal bacterial communities [54]. In brief, they analyzed the upper airway bacterial colonization in 29 healthy cigarette smokers compared with 33 non-smokers. For each DNA sample, 102 of the bacterial rRNA gene were PCR-amplified using individually barcoded primer sets. Then pyrosequences were denoised prior to taxonomic assignment [64]. Using QIIME pipeline [43], sequences were clustered at 97% similarity level into OTUs. They excluded the samples with fewer than 500 reads and OTUs with only one read, leading to 60 remaining samples and 856 OTUs. Gender (p value$<$ 0.05) and antibiotic use within the last 3 months were collected.

In a previous analysis [13], MiKRAT was applied to test association between smoking and microbial community composition while adjusting for gender and antibiotic status effect. Using MiRKAT, we found the p values from weighted UniFrac, unweighted UniFrac and Bray-sCurtis kernels to be 0.0048, 0.014, and 0.002 respectively. The optimal MiRKAT generated a p value of 0.0031 [13]. In comparison, MiSPU$_w(2)$, MiSPU$_w(7)$, MiSPU$_w(8)$, MiSPU$_w(\infty)$ yielded the p values of 0.0147, 0.0011,0.0013, 0.0012 respectively. MiSPU(8) and MiSPU(∞) provided almost the same p values, further confirming that no need to use other larger γ's. MiSPU$_w(7)$ provided the most significant p value, suggesting that there was a sparse association pattern and the large associations between smoking status and one or few microbial taxa were in the same direction. The aMiSPU$_w$, combining all the MiSPU$_w$ tests with $\gamma = 2, 3, \ldots, 8, \infty$,

yielded a p value of 0.0029. The aMiSPU$_u$, combining all the MiSPU$_u$ tests with $\gamma = 2, 3, \ldots, 8, \infty$, yielded a p value of 0.0431, less significant than that from aMiSPU$_w$ and suggesting that some abundant taxa may be correlated with the smoking status. The aMiSPU test, combining aMiSPU$_w$ and aMiSPU$_u$, yielded a p value of 0.0050, confirming the results of the previous analysis, though slightly larger than that of the optimal MiRKAT.

3.4 Discussion

We have proposed and studied a class of MiSPU tests and an adaptive version (aMiSPU) for an overall association between a microbial community and an outcome of interest. The aMiSPU test is based on the score vector for a new variable called generalized taxon proportion, which combines taxon abundance information with phylogenetic tree information, rendering it both computationally efficient and general to cover a wide range of applications with binary or quantitative outcomes and possible covariates. Our major contribution is that, by recognizing the limitation of the existing methods without variable selection or variable weighting, we propose the use of the two versions of the generalized taxon proportion to simultaneously account for the effects of relative abundances of microbial taxa and that of branch lengths in a phylogenetic tree, and apply many possible weights indexed by a single parameter $\gamma \geq 2$ to weight the taxa differentially. This approach can maintain high power in a wide range of scenarios.

Besides assessing the overall association with a microbial community, one may be interested in finding possible taxa driving a detected association. Unlike MiRKAT [13] and distance-based methods [12, 47, 65], which are unable for taxon selection, the proposed aMiSPU test can be used to rank the importance of taxa and thus provide some insights on which taxa are likely to be associated with the outcome of interest.

A few modifications or extensions are possible. First, in our current implementation of MiSPU, we propose the use of a generalize taxon proportion and weight it based on its corresponding score component; we may explicitly consider some interactions among the taxa. Second, we take the minimum p value to combine the results of multiple MiSPU tests. Instead, we may apply other methods that may perform better in some scenarios [66]. Finally, though we focused on a binary and continuous outcome of interest, it might

be of interest and possible to extend MiSPU to cases with a multivariate, longitudinal or survival outcome in a general framework of regression.

In summary, we have evaluated the MiSPU and aMiSPU tests extensively using both simulated and real data, revealing its excellent performance across many situations. As noted, aMiSPU maintains high power across a wide range of scenarios, though the identity of the most powerful MiSPU test is expected to change with the varying scenario. In comparison with other multivariate joint tests, we found that aMiSPU was often much more powerful, and thus we recommend its use in practice. An R package `MiSPU` implementing the aMiSPU test and a C++ version of UniFrac distance calculation is available on GitHub (`https://github.com/ChongWu-Biostat/MiSPU`) and CRAN.

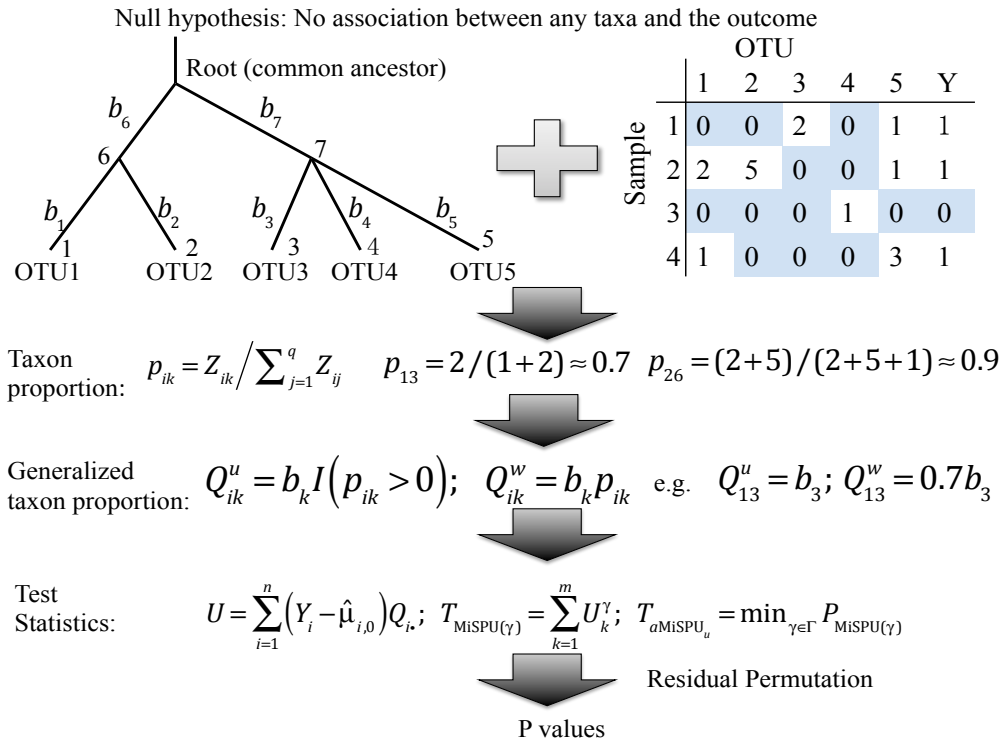

Figure 3.1: Schematic description of the use and steps in aMiSPU. Input data consist of a rooted phylogenetic tree, and a sample of OTU counts, an outcome of interest and possibly some covariates.

Table 3.1: Empirical type I error rates for MiSPU and aMiSPU for scenario 1 with a binary outcome. Type I error rate was evaluated for situations in which the covariates were independent of the OTUs ($X \perp\!\!\!\perp Z$) or correlated with the OTUs ($X \sim Z$) based on 10,000 simulated datasets at $\alpha = 0.05$. *Inflated type I error rates.

	$\text{aMiSPU}_w(2)$	$\text{aMiSPU}_w(\infty)$	aMiSPU_w	aMiSPU_u	aMiSPU
$X \perp\!\!\!\perp Z$, adjust X	0.052	0.050	0.052	0.049	0.048
$X \perp\!\!\!\perp Z$, no adjust X	0.051	0.051	0.052	0.049	0.049
$X \sim Z$, adjust X	0.043	0.038	0.043	0.049	0.040
$X \sim Z$, no adjust X	0.091*	0.119*	0.112*	0.053	0.088*

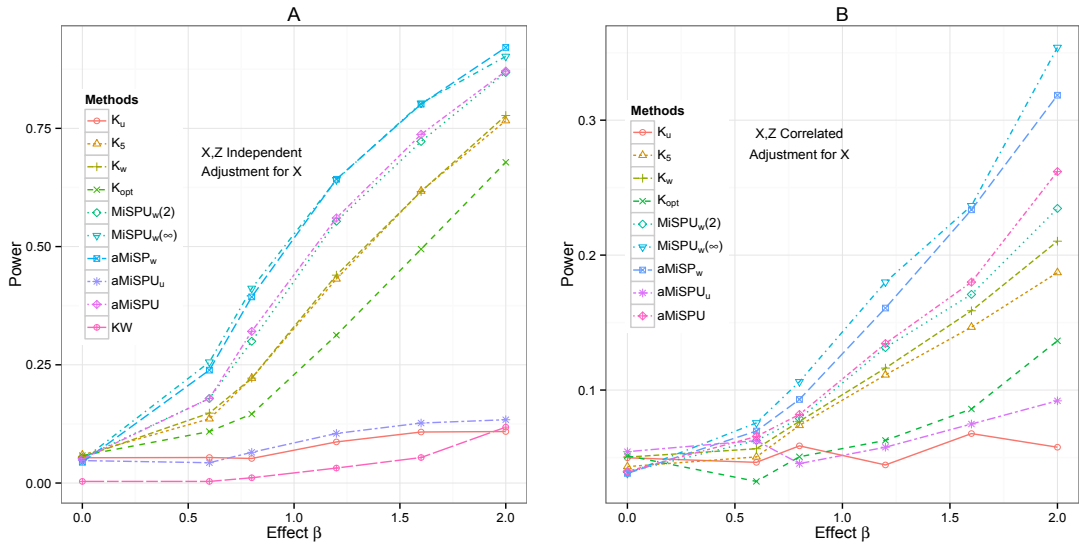

Figure 3.2: Type I error and power comparison for scenario 1 with a binary outcome. A selected phylogenetic cluster (6.7%) of the OTUs were associated with the outcome. Results were shown for (A) X and Z were independent and (B) X and Z were correlated. K_u, K_5 and K_w represent MiRKAT results from the unweighted UniFrac kernel, unweighted UniFrac kernel and generalized UniFrac kernels with $\alpha = 0.5$, respectively. K_{opt} represents the simulation results for optimal MiRKAT considering Bray-Curtis kernel, unweighted UniFrac kernel, weighted UniFrac kernel and generalized UniFrac kernel. $\text{MiSPU}_w(2)$, $\text{MiSPU}_w(\infty)$ and aMiSPU_w represent MiSPU_w test with $\gamma = 2, \infty$ and aMiSPU_w summarizing $\gamma = 2, 3, \ldots, 8, \infty$, respectively. aMiSPU_u and aMiSPU represent the test summarizing $\gamma = 2, 3, \ldots, 8, \infty$ and combining aMiSPU_u and aMiSPU_w, respectively. KW represents Kruskal-Wallis test. Results were presented at $n = 100$.

Table 3.2: Sample means (SDs in parentheses) of the total number of selected OTUs (Total), and of the numbers of true positives (TP) and false positives (FP) based on 1,000 simulation replications under scenario 1, by fitZig [59], DESeq2 [58], KW test, aMiSPU with $\alpha_1 = .7$, or aMiSPU with $k_1 = 1$. For fitZig, DESeq2 and KW test, cutoff 5×10^{-5}, 0.05, 0.05 were chosen, respectively.

	Method	Total	TP	FP
$X \perp\!\!\!\perp Z$	fitZig	157.0 (49.4)	12.3 (4.7)	144.6 (45.7)
	DESeq2	20.4 (4.8)	3.4 (1.5)	17.0 (4.4)
	KW	25.8 (6.1)	3.5 (1.6)	22.3 (5.7)
	aMiSPU with $\alpha_1 = .7$	234.1 (241.5)	42.6 (18.5)	191.6 (234.8)
	aMiSPU with $k_1 = 1$	58.5 (76.9)	27.2 (20.6)	31.3 (78.1)

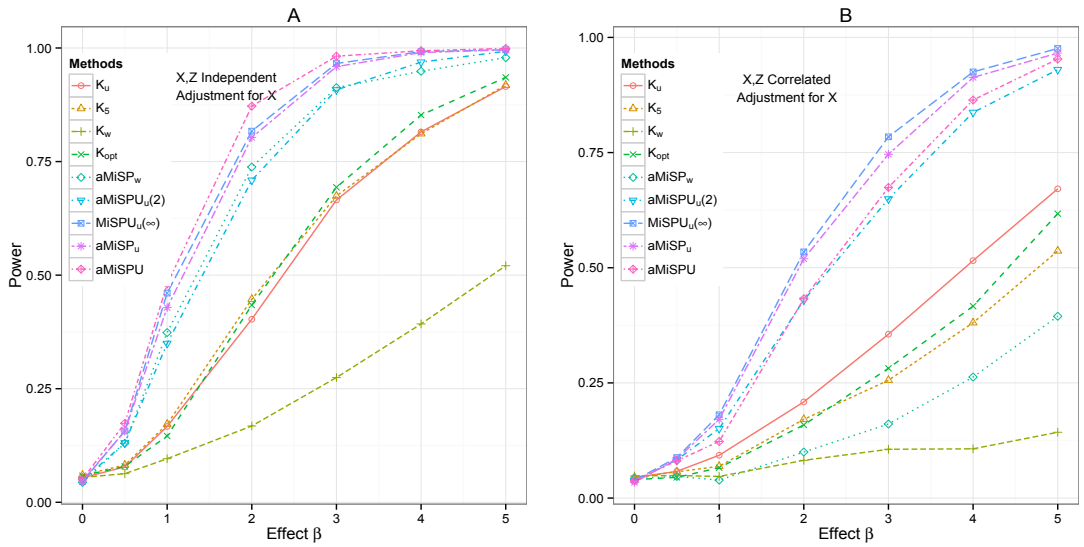

Figure 3.3: Type I error and power comparison for scenario 2 with a binary outcome. A selected phylogenetic cluster (0.35%) of the OTUs were associated with the outcome. Results were shown for X and Z were independent (A) or correlated (B). K_u, K_5 and K_w represent MiRKAT results from the unweighted UniFrac kernel, unweighted UniFrac kernel and generalized UniFrac kernels with $\alpha = 0.5$, respectively. K_{opt} represents the simulation results for optimal MiRKAT considering Bray-Curtis kernel, unweighted UniFrac kernel, weighted UniFrac kernel and generalized UniFrac kernel. $\text{MiSPU}_u(2)$, $\text{MiSPU}_u(\infty)$ and aMiSPU_u represent MiSPU_u test with $\gamma = 2, \infty$ and aMiSPU_u summarizing $\gamma = 2, 3, \ldots, 8, \infty$, respectively. aMiSPU_w and aMiSPU represent the test summarizing $\gamma = 2, 3, \ldots, 8, \infty$ and combining aMiSPU_u and aMiSPU_w, respectively. Results were presented at $n = 100$.

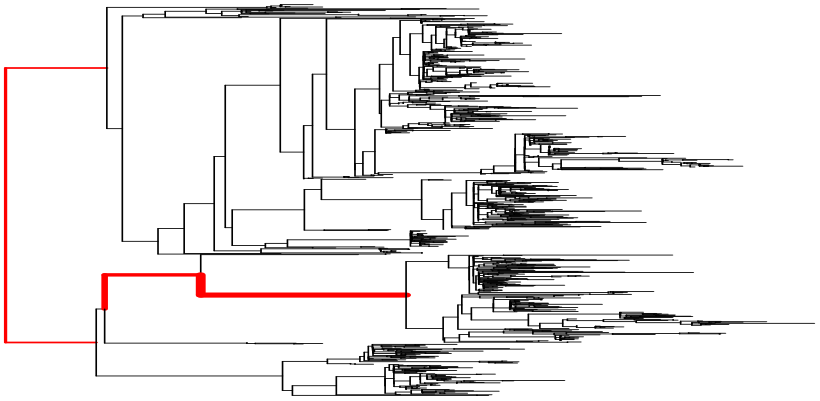

Figure 3.4: Phylogenetic tree of Bacteroides enterotypes for a gut microbiome dataset. Black edges stands for non-associated signals, while red edges stands for the associated signals. The width of the edges stands for the magnitude of the association.

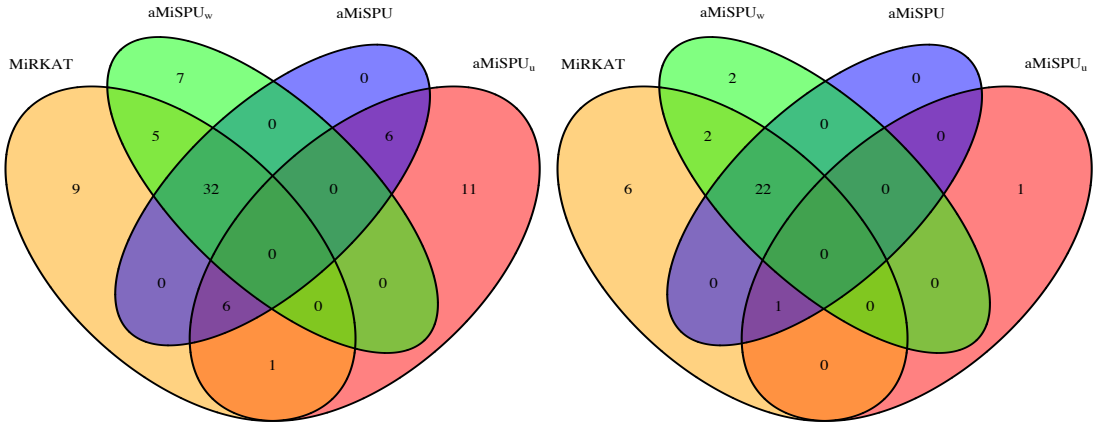

Figure 3.5: Venn diagram of detected associations for the gut microbiome dataset. 214 nutrients are included in the testing. Results were shown for p value cutoff 0.05 (A) and 0.01 (B). MiRKAT represents the results for optimal MiRKAT considering Bray-Curtis kernel, unweighted UniFrac kernel, weighted UniFrac kernel and generalized UniFrac kernel. aMiSPU$_w$ represents a test combining MiSPU$_w$ with $\gamma = 2, \infty$. aMiSPU$_u$ and aMiSPU represent the test summarizing $\gamma = 2, 3, \ldots, 8, \infty$ and combining aMiSPU$_u$ and aMiSPU$_w$, respectively.

Chapter 4

An Adaptive Test on High-dimensional Parameters

Significance testing for high-dimensional generalized linear models (GLMs) has been increasingly needed in various applications, however, existing methods are mainly based on a sum of squares of the score vector and only powerful under certain alternative hypotheses. In practice, depending on whether the true association pattern under an alternative hypothesis is sparse or dense or between, the existing tests may or may not be powerful. In this chapter, we propose an adaptive test on a high-dimensional parameter of a GLM (in the presence of a low-dimensional nuisance parameter), which can maintain high power across a wide range of scenarios. To evaluate its p-value, its asymptotic null distribution is derived. We conduct simulations to demonstrate the superior performance of the proposed test. In addition, we apply it and other existing tests to an Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, detecting possible associations between Alzheimer's disease and some gene pathways with a large number of single nucleotide polymorphisms (SNPs). The main part of this chapter has been accepted by Statistica Sinica.

4.1 Introduction

Generalized linear models (GLMs) [67] have been increasingly used in high-dimensional settings due to the surge of high-dimensional data in many fields, ranging from business to genetics. One topic of intensive interest is significance testing on regression coefficients in high-dimensional GLMs. For example, genome-wide association studies (GWASs) have led to the discovery of many genetic variants, mostly single nucleotide polymorphisms (SNPs), associated with common and complex diseases. Given the number of SNPs tested in GWASs, a univariate test must meet a stringent threshold for statistical significance (with p-value $< 5 \times 10^{-8}$) and thus is often underpowered. When failing to identify any or a sufficient number of associated SNPs based on the univariate test, one may be interested in directly testing a genetic marker set with possibly a large number of SNPs to both gain statistical power and enhance biological interpretation.

In these applications, the dimension of the parameters to be tested, p, is often close to or higher than the sample size, n. For low dimensional situations with $p \ll n$, traditional multivariate tests, such as the likelihood ratio test and the Wald test, have been widely used [67]; however, the power of both the Wald test and the likelihood ratio test tend to diminish quite rapidly as p increases [68]. These tests even break down completely when $p > n$ since the maximum likelihood estimates (MLEs) of the parameters are not uniquely determined. To deal with these difficulties, several tests for high-dimensional data have been proposed accordingly [1, 2, 68, 69, 70]. In particular, Zhong and Chen [69] proposed a modified F-test in high-dimensional linear regression models, allowing $p \to \infty$ as $n \to \infty$; Lan et al. [70] extended the test to GLMs with a general random design matrix. Meanwhile, Goeman et al. [68] proposed a test statistic for high-dimensional linear models and Goeman et al. [1] derived its asymptotic distribution for a fixed p in GLMs. Guo and Chen [2] further modified Goeman's test statistic [1] to a simpler form and allowed both n and $p \to \infty$. In a penalized regression framework, several inference methods for a low-dimensional sub-vector of a high-dimensional regression coefficient vector have been developed [71, 72, 73], which however differs from the goal of testing on a high-dimensional parameter here and thus will not be further discussed.

The existing methods are mainly based on the sum-of-squares of the score vector for

the parameters of interest and are usually powerful against alternative hypotheses with moderately dense signals/association patterns, where there is a relatively large proportion of associated (i.e. non-null) parameters. In contrast, if the nonzero associations are strong but sparse, the sum-of-squares-type tests lose substantial power while a test based on the supremum of the score vector is more powerful. Importantly, as to be shown in the simulation section, there are some intermediate situations in which neither type of the above tests is powerful. In practice, it is often unclear which type of tests should be applied since the underlying truth is unknown.

In this chapter, we develop an adaptive test that would yield high statistical power under various high-dimensional scenarios, ranging from highly dense to highly sparse signal situations. The main idea is that, since we do not know which and how many parameters being tested are associated with the response, we first construct a class of sum of *powered* score tests such that hopefully at least one of them would be powerful for a given situation. The proposed adaptive test then selects the one with the most significant testing result with a proper adjustment for multiple testing. To apply the proposed test, we establish its asymptotic null distribution. In particular, we derive the joint null distribution of the individual powered score test statistics, which converge to either a multivariate normal distribution or an extreme value distribution. The joint asymptotic null distribution for the proposed tests is used to calculate asymptotics-based p-values, a more convenient and faster alternative to other computing-intensive resampling methods such as the bootstrap.

4.2 Some Existing Tests

Suppose n identical and independently distributed (i.i.d.) samples $\{(Y_i, Z_i, X_i) : i = 1, 2, \ldots, n\}$ have been collected, for which we have an n-vector response (outcome of interest) Y, an $n \times q$ matrix $\mathbb{Z}$ for q covariates, and an $n \times p$ matrix $\mathbb{X}$ for p variables of interest. For subject i, let $Z_i = (Z_{i1}, \ldots, Z_{iq})$ be the q covariates, such as age, gender, and other clinical variables that we want to adjust for, and $X_i = (X_{i1}, \ldots, X_{ip})$ be the p-dimensional variables of interest. Without loss of generality, we assume that $E(\mathbb{X}) = 0$ as otherwise $\mathbb{X}$ can be re-centered by its mean. Assuming a generalized linear model,

we have

$$E(Y|\mathbb{X},\mathbb{Z}) = g^{-1}(\mathbb{X}\beta + \mathbb{Z}\alpha), \tag{4.1}$$

where p-vector β and q-vector α are unknown parameters, and g is the canonical link function. We are interested in testing

$$H_0 : \beta = \beta_0 \qquad \text{versus} \qquad H_1 : \beta \neq \beta_0, \tag{4.2}$$

while treating α as the nuisance parameter. We target the situation with "small q, large p and large n".

The best-known tests for low-dimensional data are the Wald test and the likelihood ratio test; however, the power of both the Wald test and the likelihood ratio test diminishes quite rapidly as the dimension p increases [68]. More importantly, in a high-dimensional situation with $p > n$, these tests break down completely since the MLEs for the parameters no longer exist uniquely. [68] derived the following test statistic for testing hypothesis (4.2) based on the score vector

$$T_{\text{Goe}} = U^\intercal U - \text{trace}(\mathcal{I}),$$

where U and $\mathcal{I}$ are the score vector and observed information matrix for β under the null hypothesis, respectively. Ignoring some constant, T_{Goe} equals to

$$T_{\text{Goe2}} = n^{-1}(Y - \mu_0)^\intercal \mathbb{X}\mathbb{X}^\intercal (Y - \mu_0),$$

where μ_0 is the expectation of Y under the null hypothesis. Goeman et al. [68] calculated the p-value of this test statistic via permutations or moment matching. Goeman et al. [1] modified T_{Goe} with the following statistic

$$T_{\text{GT}} = \frac{(Y - \hat{\mu}_0)^\intercal \mathbb{X}\mathbb{X}^\intercal (Y - \hat{\mu}_0)}{(Y - \hat{\mu}_0)^\intercal \mathbb{D} (Y - \hat{\mu}_0)},$$

where $\hat{\mu}_0$ and $\mathbb{D}$ are the maximum likelihood estimate of μ_0 under the null hypothesis and a diagonal $n \times n$ matrix equal to the diagonal of $\mathbb{X}\mathbb{X}^\intercal$, respectively. Goman et al. [1] derived its asymptotic null distribution for fixed p. Since the denominator of T_{GT}

increases the variance and thus adversely affects the power, Guo and Chen[2] proposed the following test statistic

$$T_{\text{HDGLM}} = n^{-1}(Y - \hat{\mu}_0)^\intercal(\mathbb{X}\mathbb{X}^\intercal - \mathbb{D})(Y - \hat{\mu}_0),$$

and further derived the asymptotic normal distribution of T_{HDGLM} for diverging $p \to \infty$ as $n \to \infty$ under some assumptions.

Remark 1. To our knowledge, most high-dimensional tests are based on a sum-of-squared score vector, which have also been used in GWASs with large n and small p. For instance, Pan [51] proposed a sum-of-squared-score test (similar to T_{Goe2}) for testing the association between multiple SNPs and the outcome of interest in GLMs. Another similar test is SKAT [74].

4.3 New Method

For the purpose of presentation, we first consider the case without nuisance parameters, then the case with nuisance parameters.

4.3.1 Testing Without Nuisance Parameters

In this subsection, we assume the GLM (4.1) with $\alpha = 0$. Many existing tests are based on the score vector $U = (U_1, \ldots, U_p)^\intercal$ for β, which, up to some constant, has elements

$$U_j = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_{0i})X_{ij}, \qquad 1 \le j \le p,$$

with $\mu_{0i} = g^{-1}(X_i\beta_0)$.

For notation simplicity, we write $S_{ij} = (Y_i - \mu_{0i})X_{ij}$ for $1 \le i \le n$ and $1 \le j \le p$. As to be demonstrated later, depending on the unknown association effects β to be tested, different tests may be more powerful. Inspired by Pan et al. [50], we would use U to construct some weights to upweight more informative components of the score vector,

proposing a sum of powered score (SPU) test statistic with power index $0 < \gamma < \infty$ as

$$L(\gamma, \mu_0) = \sum_{j=1}^{p} w_j U_j = \sum_{j=1}^{p} U_j^{\gamma-1} U_j = \sum_{j=1}^{p} U_j^{\gamma} = \sum_{j=1}^{p} \Big(\frac{1}{n} \sum_{i=1}^{n} S_{ij} \Big)^{\gamma},$$

where $w_j = U_j^{\gamma-1}$ can be considered as a data-dependent weight.

Note that $\gamma = 2$ yields a sum-of-squares-type test statistic, which is similar to the existing tests reviewed in the previous section. As an even integer $\gamma \to \infty$, we have $L(\gamma, \mu_0) \propto L(\gamma, \mu_0)^{1/\gamma} \to \max_{1 \le j \le p} \big| \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu_{0i}) X_{ij} \big|$, thus we define $L(\infty, \mu_0)$ as

$$L(\infty, \mu_0) = \max_{1 \le j \le p} \frac{n \left(\frac{1}{n} \sum_{i=1}^{n} S_{ij} \right)^2}{\sigma_{jj}},$$

where $\mathbf{\Sigma} = (\sigma_{kj})_{p \times p}$, and $\sigma_{kj} = \mathrm{Cov}[S_{ik}, S_{ij}]$ for $1 \le k, j \le p$. Note that here the covariance matrix $\mathbf{\Sigma}$ is defined *unconditionally* on the covariates and consequently it does not depend on the subject index i. See Remark 7 for more discussion.

The class of the SPU tests cover several tests used in GWASs as special cases. For example, for large n and small p, $L(2, \mu_0)$ is like SKAT with a linear kernel [74]; $L(1, \mu_0)$ is a burden test in genetic rare variant assoication analysis [75]. As to be shown in simulations, if most variables of $\mathbb{X}$ are associated with the response Y with similar effect sizes and the same association direction, then a burden test like $L(1, \mu_0)$ would yield high statistical power. In contrast, in a situation with only moderately dense signals or with different association directions, $L(\gamma, \mu_0)$ with an even integer $\gamma \ge 2$ would be more powerful. In particular, the supremum based test statistic, $L(\infty, \mu_0)$ yields high statistical power if only few variables are strongly associated with Y (i.e. a highly sparse non-zero components of β). In short, the power of $L(\gamma, \mu_0)$ depends on the unknown true association pattern (i.e. value of β), such as signal sparsity and magnitudes. To choose the most powerful test automatically, we propose the following adaptive test to combine the multiple tests accordingly:

$$T_{\mathrm{aSPU}} = \min_{\gamma \in \Gamma} P_{\mathrm{SPU}(\gamma, \mu_0)},$$

where $P_{\mathrm{SPU}(\gamma, \mu_0)}$ is the p-value of $L(\gamma, \mu_0)$ test. For simplicity, we write $L(\gamma, \mu_0)$, $\mathrm{SPU}(\gamma, \mu_0)$ and $\mathrm{SPU}(\gamma)$ exchangeably. Taking the minimum p-value is a simple and

effective way to approximate the most powerful test [50]. Note that T_{aSPU} is no longer a genuine p-value and we need to derive its asymptotic null distribution to facilitate calculating its p-value.

Remark 2. The optimal value of γ for the test statistic $L(\gamma)$ to achieve the highest power depends on the specific alternative. We aim to choose a Γ set to maintain high power of the aSPU test under a wide range of scenarios. The supremum based test statistic for high-dimensional two-sample testing has been studied in [76]; from their Theorem 2, the power of the supremum based test converges to 1 if the signal is strong with a high sparsity level; see also related discussions in [77] and [78]. When the signal is dense with a constant effect size, $L(1)$ is most powerful [79]. $L(2)$ is a sum-of-squares-type test that has been widely used and studied. By default, we recommend include $\gamma = 1, 2, \infty$ and a small subset of moderate values of γ in Γ. More generally, as recommended in [79], we use $\Gamma = \{1, 2, \ldots, \gamma_u, \infty\}$ with a γ_u such that $L(\gamma_u)$ gives similar results to that of $L(\infty)$; we find in the simulation studies that often $\gamma_u = 6$ or 8 suffices and the performance of the aSPU test is robust to such a choice of γ_u.

Remark 3. Our proposed test is an extension of the original aSPU test [50] to high-dimensional GLMs; the original aSPU test was proposed for analysis of rare variants with large n and small p. For simplicity, we use the same name "aSPU" for our proposed test here. Since the asymptotic properties of the adaptive aSPU test for GLMs have not been studied, we derive its asymptotic null distribution in a high-dimensional setting, based on which the asymptotic p-values of $L(\gamma, \mu_0)$ and T_{aSPU} can be calculated.

Next we derive the asymptotic properties under the null hypothesis. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists some constant C such that $|a_n| \leq C|b_n|$ holds for all $n \geq N$, and write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$. Under $H_0 : \beta = \beta_0$, we first derive some asymptotic approximations to the mean and the variance of $L(\gamma, \mu_0)$ for $\gamma < \infty$, and then establish the asymptotic distribution of $L(\gamma, \mu_0)$. The following assumptions are needed.

C1. The eigenvalues of $\mathbf{\Sigma}$ are bounded, that is, $B^{-1} \leq \lambda_{\min}(\mathbf{\Sigma}), \lambda_{\max}(\mathbf{\Sigma}) \leq B$ for some finite constant B, where $\lambda_{\min}(\mathbf{\Sigma})$ and $\lambda_{\max}(\mathbf{\Sigma})$ denote the minimum and maximum eigenvalues of matrix $\mathbf{\Sigma}$, respectively. Moreover, the absolute value of any corresponding

correlation element is strictly smaller than 1; in other words, $\max_{1\le i\neq j\le p}|\sigma_{ij}|/\sqrt{\sigma_{ii}\sigma_{jj}} < 1-\xi$ for some constant $\xi > 0$.

C2. Given a set of multivariate random vectors $W = \{W^{(j)} : j \ge 1\}$, for integers $a < b$, let χ_a^b be the σ-algebra generated by $\{W^{(m)} : m \in [a,b]\}$. The α-mixing coefficient $\alpha_W(s)$ is defined as $\sup\{|Pr(A\cap B) - Pr(A)Pr(B)| : 1 \le t < p, A \in \chi_1^t, B \in \chi_{t+s}^{\infty}\}$. We assume $W = \{W^{(j)} = (S_{ij}, i = 1,\dots,n) : j \ge 1\}$ is α-mixing such that $\alpha_W(s) \le M\delta^s$, where $\delta \in (0,1)$ and M is some constant.

C3. Under $H_0 : \beta = \beta_0$, $E\left[(S_{ij})^3\right] = 0$ for $1 \le j \le p$.

C4. $(\log p)/n^{1/4} = o(1)$.

C5. There exist some constants η and $K > 0$ such that $E\left[\exp\left\{\eta(S_{ij})^2/\sigma_{jj}\right\}\right] \le K$ for $1 \le j \le p$.

Remark 4. Assumptions C1, C4, and C5 are mild conditions, and are used to establish the weak convergence of $L(\infty, \mu_0)$. Cai et al. [76] used exactly the same assumptions (C1, C4, and C5) when deriving the limiting distribution of a supremum-type test statistic for high-dimensional two-sample mean testing. Assumption C2 assumes an α-mixing-type weak dependence structure of the data, which has been widely used in spatial statistics and time series. For high-dimensional two-sample mean testing, a similar mixing condition has been used in [79] and [80]. Alternatively, we may consider the weak dependence structure adopted in [2], where a factor-type model for $S_i = (S_{i1},\dots,S_{ip})^\intercal$ is assumed. Intuitively, many random vectors, e.g., any ergodic and aperiodic Markov chain, meet the α-mixing weak dependence condition. Another example is for random vectors $X = (X_1, X_2, \dots)^\intercal$, where X_i and X_j are independent with $|i-j| > C$ for some constant C; then $\alpha_X(s) = 0$ if $s > C$, satisfying the α-mixing assumption as well. This type of structure has also been used for estimating a high-dimensional covariance matrix [81]. In addition, because the correlations among variables (i.e. SNPs) in our motivating genome-wide association study data decay to zero as their physical distances on the same chromosome increase (while the SNPs from different chromosomes are usually independent), the α-mixing assumption fits the application well and thus will be used in this chapter.

We write $L(\gamma,\mu_0) = \sum_{j=1}^p L^{(j)}(\gamma,\mu_0)$ with $L^{(j)}(\gamma,\mu_0) = \left(\frac{1}{n}\sum_{i=1}^n S_{ij}\right)^\gamma$, then denote $\mu(\gamma) = \sum_{j=1}^p \mu^{(j)}(\gamma)$ with $\mu^{(j)}(\gamma) = E\left(L^{(j)}(\gamma,\mu_0)\right)$, and $\sigma^2(\gamma) = \mathrm{var}(L(\gamma,\mu_0))$.

PROPOSITION 1. *Under assumptions C1, C3, and* $H_0 : \beta = \beta_0$, $\mu(1) = 0$ *and*

$$\mu(\gamma) = \begin{cases} \frac{\gamma!}{d!2^d} n^{-d} \sum_{j=1}^{p} \sigma_{jj}^d + o(pn^{-d}), & \text{if } \gamma = 2d, \\ o(pn^{-(d+1)}), & \text{if } \gamma = 2d+1, \end{cases}$$

where $\sigma_{jj} = E[(S_{ij})^2]$.

PROPOSITION 2. *Under assumptions C1–C3 and* H_0, $\sigma^2(1) = \frac{1}{n} \sum_{1 \le i,j \le p} \sigma_{ij} + o(pn^{-1})$ *and for* $\gamma \ge 2$,

$$\sigma^2(\gamma) = \mu(2\gamma) - \sum_{j=1}^{p} \{\mu^{(j)}(\gamma)\}^2 + \frac{1}{n^\gamma} \sum_{i \ne j} \sum_{\substack{2c_1+c_3=\gamma \\ 2c_2+c_3=\gamma \\ c_3>0}} \frac{(\gamma!)^2}{c_3! c_1! c_2! 2^{c_1+c_2}} \sigma_{ii}^{c_1} \sigma_{jj}^{c_2} \sigma_{ij}^{c_3} + o(pn^{-\gamma})$$

where $\sigma_{ij} = E[S_{ki} S_{kj}]$.

Note that the order of $\sigma^2(\gamma)$ is $pn^{-\gamma}$. Then we derive the following result to approximate the correlations among the $L(\gamma, \mu_0)$.

PROPOSITION 3. *Under assumptions C1–C3 and* $H_0 : \beta = \beta_0$, *for any finite and positive integers* $s, t \in \Gamma$, *we have*
(i) if $s + t$ *is even,*

$$Cov\{L(t, \mu_0), L(s, \mu_0)\}$$
$$= \mu(t+s) - \sum_{i=1}^{p} \mu^{(i)}(t) \mu^{(i)}(s) + \frac{1}{n^c} \sum_{i \ne j} \sum_{\substack{2c_1+c_3=t \\ 2c_2+c_3=s \\ c_3>0}} \frac{t! s!}{c_3! c_1! c_2! 2^{c_1+c_2}} \sigma_{ii}^{c_1} \sigma_{jj}^{c_2} \sigma_{ij}^{c_3} + o(pn^{-(t+s)/2}).$$

(ii) if $s + t$ *is odd,* $Cov\{L(t, \mu_0), L(s, \mu_0)\} = o(pn^{-(t+s)/2})$.

Let Γ be a candidate set of γ with $\infty \in \Gamma$. We further define $R = (\rho_{st})$, where $\rho_{ss} = 1$ for $s \in \Gamma \setminus \{\infty\}$ and $\rho_{st} = \text{Cov}\{L(s, \mu_0), L(t, \mu_0)\}/\{\sigma(s)\sigma(t)\}$ for $s \ne t \in \Gamma \setminus \{\infty\}$. In particular, $\rho_{st} = o(1)$ when $s + t$ is odd. Then we introduce Theorem 1, which describes the asymptotic distribution of $L(\gamma, \mu_0)$.

THEOREM 1. *Under assumptions C1–C5 and the null hypothesis H_0, we have:*
(i) For set $\Gamma' = \Gamma \setminus \{\infty\}$, the vector of the standardized test statistics $[\{L(\gamma, \mu_0) - \mu(\gamma)\}/\sigma(\gamma)]^{\intercal}_{\gamma\in\Gamma'}$ converges weakly to a normal distribution $N(0, R)$ as n, $p \to \infty$.
(ii) When $\gamma = \infty$, let $a_p = 2\log p - \log\log p$, for any $x \in \mathbb{R}$, $Pr\{L(\infty, \mu_0) - a_p \le x\} \to \exp\{-\pi^{-1/2}\exp(-x/2)\}$.
(iii) $[\{L(\gamma, \mu_0) - \mu(\gamma)\}/\sigma(\gamma)]^{\intercal}_{\gamma\in\Gamma'}$ is asymptotically independent with $L(\infty, \mu_0)$. That is, the joint distribution of $[\{L(\gamma, \mu_0) - \mu(\gamma)\}/\sigma(\gamma)]^{\intercal}_{\gamma\in\Gamma'}$ and $L(\infty) - a_p$ converges weakly to the product of the limiting distributions given in (i) and (ii).

Remark 5. Testing without nuisance parameters can be treated as a special case of testing with nuisance parameters. The methods described in the following subsection can be used for calculating the p-values for testing without nuisance parameters by replacing $\hat{\mu}_0$ with μ_0.

4.3.2 Testing With Nuisance Parameters

In this subsection, we consider testing on a high-dimensional regression coefficient vector in the presence of a low-dimensional nuisance parameter, which is a common task in practice. For example, in a study of complex disease, we usually have both SNP data and other demographic variables, which may confound the association between the SNPs and the outcome of interest. One may be interested only in genetic effects while adjusting for demographic variables, hence the coefficients for demographic variables are treated as low-dimensional nuisance parameters, which have to be estimated. Here, we are interested in testing hypothesis (4.2) under GLM (4.1).

Let $\mu_0(\alpha) = \mu_0 = g^{-1}(\mathbb{Z}\alpha + \mathbb{X}\beta_0)$ and $\hat{\mu}_0 = g^{-1}(\mathbb{Z}\hat{\alpha} + \mathbb{X}\beta_0)$, where the MLE $\hat{\alpha}$ is obtained under the null hypothesis. Since μ_0 is unknown, we use $\hat{\mu}_0$ and the test statistic $L(\gamma, \hat{\mu}_0)$ accordingly. To derive its asymptotic distribution, the following additional assumptions are needed.

C6. The dimension of nuisance parameters α, q, is fixed, and each covariate in $\mathbb{Z}$ is bounded almost surely. We assume $E(X_{ij}|\mathbb{Z}) \neq 0$ only holds for $j \in P_0$ with the size of P_0, p_0, satisfying $p_0 = O(p^{\eta})$ for a small positive η. We further assume the consistent and asymptotic normal MLE $\hat{\alpha}$ under the null hypothesis [82].

C7. There exist some positive constants K_1 and K_2 such that $K_1 < E[\epsilon_{0i}^2|\mathbb{Z} = z] < K_2$ almost everywhere for z in the support of the probability density of Z, where $\epsilon_{0i} = Y_i - \mu_{0i}$, $1 \leq i \leq n$. We further assume $E[\epsilon_{0i}|\mathbb{X}, \mathbb{Z}] = 0$.

C8. We assume $p/n^2 = o(1)$.

C9. The conditionally α-mixing coefficient $\alpha_{W|\mathcal{F}}(s)$ is defined as $\sup\{|Pr(A \cap B|\mathcal{F}) - Pr(A|\mathcal{F})Pr(B|\mathcal{F})| : 1 \leq t < p, A \in \chi_1^t, B \in \chi_{t+s}^{\infty}\}$, where $\mathcal{F}$ is a sub-σ-algebra of W. We assume $W = \{W^{(j)} = (X_{ij}, i = 1, \ldots, n) : j \geq 1\}$ is conditionally α-mixing given $\mathbb{Z}$ such that $\alpha_{W|\sigma(\mathbb{Z})}(s) \leq M\delta^s$, where $\delta \in (0, 1)$ and M is some constant.

Remark 6. Assumption C6 states that the dimension of nuisance parameters, q, is fixed as $n \to \infty$, which is appropriate in many applications, including GWASs of interest here. However, this assumption may not be appropriate in some applications. For example, in testing gene-environmental interactions, the main effects are treated as nuisance parameters, which may be high-dimensional [83]. Note that, we assume that each X_j is already centered and has sample mean 0, partially making it reasonable to assume $E[X_{ij}|\mathbb{Z}] \neq 0$ only for $j \in P_0$ with the size of P_0 in a small order of p (i.e. $p_0 = O(p^{\eta})$). This assumption is technically needed to prove Theorem 2. For finite γ, we can relax the assumption to $p_0 = O(p^{1/2-\delta})$, where δ is a small constant. If we are concerned about the validity of this assumption, we can regress each X_j on $\mathbb{Z}$ and use its residuals as the new X_j to approximately satisfy $E[X_{ij}|\mathbb{Z}] = 0$ for any $j = \{1, 2, \ldots, p\}$. Assumption C7 is common in GLMs, for instance, as assumption G in [84] and assumption 3.3 in [2]. Assumption C8 is an updated version of C4 and somewhat restrictive, which however is technically needed to prove Theorem 2. Note that, instead of considering only the sum-of-squares-type statistic (with $\gamma = 2$) similar to the HDGLM [2], here we derive the asymptotic distributions for any finite γ and $\gamma = \infty$, for which a stronger assumption is therefore used. However, this assumption may be relaxed: as to be shown in simulations, the asymptotic distribution still performed well for more general high dimensional situations, and we leave this interesting problem to future work. Conditionally α-mixing is introduced by [85] and assumption C9 is an updated version of C2 to adjust the case of nuisance parameters.

Although the estimated parameter $\hat{\alpha}$ does complicate the derivations, we still have the following theorem similar to Theorem 1.

THEOREM 2. *Under assumptions C1–C9 and the null hypothesis H_0, we have:*
(i) For set $\Gamma' = \Gamma \setminus \{\infty\}$, $[\{L(\gamma, \hat{\mu}_0) - \mu(\gamma)\}/\sigma(\gamma)]^{\intercal}_{\gamma\in\Gamma'}$ converges weakly to the normal distribution $N(0, R)$ specified in Theorem 1 as $n, p \to \infty$.
(ii) When $\gamma = \infty$, let $a_p = 2\log p - \log\log p$, for any $x \in \mathbb{R}$, $Pr\{L(\infty, \hat{\mu}_0) - a_p \le x\} \to \exp\{-\pi^{-1/2}\exp(-x/2)\}$.
(iii) $[\{L(\gamma, \hat{\mu}_0) - \mu(\gamma)\}/\sigma(\gamma)]^{\intercal}_{\gamma\in\Gamma'}$ is asymptotically independent with $L(\infty, \hat{\mu}_0)$.

Remark 7. In a GLM, conditional on $\mathbb{Z}$ and $\mathbb{X}$, we usually have $\mathrm{Cov}[S_{ik}, S_{ij}|\mathbb{Z}, \mathbb{X}] \neq \mathrm{Cov}[S_{i'k}, S_{i'j}|\mathbb{Z}, \mathbb{X}]$ for $i \neq i'$. In our derivations, we treat $\mathbb{Z}$ and $\mathbb{X}$ as random and assume the data are independently and identically distributed, which makes σ_{kj} well defined (unconditionally); and we derive the unconditional version of the asymptotic null distribution.

Since $\mu(\gamma)$, $\sigma(\gamma)$, and R can be approximated according to Propositions 1–3, respectively, the p-values for individual $L(\gamma, \hat{\mu}_0)$ can be calculated via either a normal or an extreme value distribution. We illustrate how to calculate the p-value for aSPU. Define $L_O = [\{L(\gamma, \hat{\mu}_0) - \mu(\gamma)\}/\sigma(\gamma) : \text{odd } \gamma \in \Gamma']$ and $L_E = [\{L(\gamma, \hat{\mu}_0) - \mu(\gamma)\}/\sigma(\gamma) : \text{even } \gamma \in \Gamma']$. By Proposition 3, $\mathrm{Cov}(L(t), L(s))$ is a small order term if $t + s$ is odd, implying L_O and L_E are asymptotically uncorrelated. By Theorem 2, L_O and L_E converge jointly and weakly to a multivariate normal distribution as n, $p \to \infty$, implying L_O and L_E are asymptotically independent. Further, by Theorem 2, $L(\infty, \hat{\mu}_0)$ is asymptotically independent of both L_O and L_E. Then we can calculate the p-value for aSPU via the following procedure.

Step 1 Define $t_O = \max_{\text{odd } \gamma\in\Gamma'} |\{L(\gamma, \hat{\mu}_0) - \mu(\gamma)\}/\sigma(\gamma)|$ and $t_E = \max_{\text{even } \gamma\in\Gamma'}\{L(\gamma, \hat{\mu}_0) - \mu(\gamma)\}/\sigma(\gamma)$ as the observed test statistic from the data and calculate the p-value for t_O and t_E as $p_O = Pr[\max_{\text{odd } \gamma\in\Gamma'} |\{L(\gamma, \hat{\mu}_0) - \mu(\gamma)\}/\sigma(\gamma)| > t_O]$ and $p_E = Pr[\max_{\text{even } \gamma\in\Gamma'}\{L(\gamma, \hat{\mu}_0) - \mu(\gamma)\}/\sigma(\gamma) > t_E]$. Use function *pmvnorm()* in R package *mvtnorm* to calculate the multivariate normal tail probabilities p_O and p_E.

Step 2 Calculate the p-value p_∞ of $L(\infty, \hat{\mu}_0)$ based on its asymptotic extreme value distribution.

Step 3 By the asymptotic independence, the asymptotic p-value for the aSPU test is $p_{\text{aSPU}} = 1 - (1 - p_{\min})^3$, where we have $p_{\min} = \min\{p_O, p_E, p_\infty\}$.

The above discussion assumes that the covariance matrix $\mathbf{\Sigma}$ is known. In practice, $\mathbf{\Sigma}$ has to be estimated. We may apply an existing method, such as the banding and thresholding technique, to estimate a high-dimensional sparse covariance matrix [81, 86]; see [87] for an excellent review. Under the α mixing assumption C2, σ_{ij} is close to zero when $|i - j|$ is large and thus we may apply the banding approach of [81] to estimate covariance matrix $\mathbf{\Sigma}$. Specifically, we first calculate the sample covariance matrix $\mathbb{S} = (s_{ij})$, where $s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(Y_k - \hat{\mu}_{0k})^2 X_{ki}X_{kj}$, then we further calculate the bandable covariance matrix with bandwidth k_n as $\hat{\mathbf{\Sigma}}_{k_n} = (s_{ij}I(|i - j| \leq k_n))$. Theoretically optimal bandwidth k_n and minimax risk rates of $\hat{\mathbf{\Sigma}}_{k_n}$ have been studied in [81]. Since a theoretically optimal k_n is determined by some unknown hyper-parameters, we use five-fold cross-validation to select an optimal bandwidth k_n [81, 86]. Following [79], under the assumptions in Theorem 2, we can show that $\hat{\mu}(\gamma)$ and $\hat{\sigma}^2(\gamma)$ estimated based on the bandable covariance matrix $\hat{\mathbf{\Sigma}}_{k_n}$ satisfy $\hat{\mu}(\gamma) = \{1+o(1)\}\mu(\gamma)$ and $\hat{\sigma}^2(\gamma) = \{1 + o(1)\}\sigma^2(\gamma)$ for properly chosen $k_n = o(n^{1/2})$.

With a relatively small sample size, five-fold cross-validation may select a smaller bandwidth than the optimal one and thus yield an underestimated $\hat{\sigma}^2(\gamma)$ and a smaller p-value. Alternatively, we propose to use the parametric bootstrap to estimate $\hat{\mu}(\gamma)$, $\hat{\sigma}^2(\gamma)$, and R. We first fit a null model under H_0 to obtain $\hat{\mu}_{0i} = \hat{E}(Y_i|Z_i, H_0)$, then simulate a new set of responses $Y_i^{(b)}$ from the corresponding model for $b = 1, 2, \ldots, B$. For example, for a binary outcome of interest, generate $Y_i^{(b)} \sim \text{Bin}(1, \hat{\mu}_{0i})$. We refit the model with $\{Y_i^{(b)} : i = 1, 2, \ldots, n\}$ and calculate the corresponding test statistic $L(\gamma, \hat{\mu}_0)^{(b)}$. Then $\hat{\mu}(\gamma) = \sum_{b=1}^{B} L(\gamma, \hat{\mu}_0)^{(b)}/B$, $\hat{\sigma}^2(\gamma) = \sum_{b=1}^{B}(L(\gamma, \hat{\mu}_0)^{(b)} - \hat{\mu}(\gamma))^2/(B-1)$ and $\hat{R} = \text{cor}(L(\Gamma, \hat{\mu}_0))$, where cor is the sample correlation. Unlike that the accuracy of a usual resampling method is bounded by the number of resampling B and thus a large B is needed for calculating a very small p-value, we can use a relatively small B to calculate $\hat{\mu}(\gamma)$, $\hat{\sigma}^2(\gamma)$, R and then an asymptotic p-value. Although estimating the mean and covariance matrix differently, the above two methods are still based on the asymptotics to calculate the p-values, hence are called asymptotics-based methods in the following. In contrast, we can also simply use the parametric bootstrap to calculate

the p-values (without direct use of the asymptotic results), which will be more time-consuming (requiring a large B for a highly significant p-value) but may perform better for finite samples; in the sequel, by default, the parametric bootstrap refers to this way of calculating the p-values.

Remark 8. The optimal value of γ for the test $L(\gamma, \hat{\mu}_0)$ to achieve the highest power depends on the true alternative. As to be shown in the numerical results, when the signal β is highly dense with the same sign, $L(1, \hat{\mu}_0)$ is more powerful than the competing tests. $L(2, \hat{\mu}_0)$ performs similarly to the tests of [2] since they have similar test statistics. There are some other situations, under which $L(2, \hat{\mu}_0)$ is not as powerful as other $L(\gamma, \hat{\mu}_0)$ tests, and therefore in these cases, the proposed test is more powerful than the competing tests. When the signal is strong and highly sparse, $L(\infty, \hat{\mu}_0)$ is more powerful. Due to the nature of its adaptiveness, the power of the aSPU test is often either the highest or close to the highest.

4.4 Numerical Results

4.4.1 Simulations

We conducted extensive simulations to compare the performance of the proposed adaptive test with two existing methods, the HDGLM [2] and the GT [1], due to their popularity and the availability of their computer code.

We set the sample size $n = 200$ and the dimension of β $p = 2000$, though other values were also considered. We generated a data matrix $\mathbb{X}_{n\times p}$ from a multivariate normal distribution; that is, we had independent $X_i \sim N(0, \mathbf{\Xi})$ for $i = 1, 2, \ldots, n$. We show the results with unit variances and a blocked first-order autoregressive correlation matrix $\mathbf{\Xi} = (\mathbf{\Xi}_{ij})$ with $\mathbf{\Xi}_{ij} = 0.4^{|i-j|}$ if $|i - j| \leq 3$ and 0 otherwise.

We further generated a data matrix with two covariates $\mathbb{Z}$ from a normal distribution $N(0, 0.5)$. The outcome Y was generated from a logistic regression model as in GLM (4.1) with a logit link function, $\alpha = (1, 1)^\intercal$, and $\beta = 0$ or $\neq 0$, corresponding to the null hypothesis H_0 or an alternative hypothesis H_1 respectively. Here, we mainly focused on the results for a binary outcome since in our real data application the response is binary and it is generally more challenging than that for a continuous outcome. Under

H_1, $\lfloor ps \rfloor$ elements of β were set to be non-zero, where $s \in [0, 1]$ controlled the degree of signal sparsity. We varied s to mimic varying sparsity levels, covering from highly sparse signals at $s = 0.001$ to less sparse and then to moderate dense at $s = 0.1$, finally to dense and highly dense signals at $s = 0.7$, respectively. The indices of non-zero elements in β were assumed to be uniformly distributed in $\{1, 2, \ldots, p\}$, and their values were constant at c. We varied s, c, n and p to evaluate the performance of the new method under various situations. We used the parametric bootstrap [50] to obtain a 'bronze-standard' (slightly inferior to a 'gold standard', where the true p-value is known) analysis, to which we compared the asymptotic results based on Theorem 2. In all simulations, we treated $\mathbf{\Sigma}$ as unknown and thus estimated $\mathbf{\Sigma}$, then calculated the means and covariances of the SPU test statistics according to Propositions 1–3. For each set-up, we simulated 1,000 datasets and averaged the testing results of these 1,000 datasets. The nominal significance level was set to $\alpha = 0.05$. For the aSPU test, the candidate set of γ was by default set to be $\Gamma = \{1, 2, \ldots, 6, \infty\}$.

Table 4.1 shows the type I error rates and power for $s = 0.1$. The results outside and inside parentheses in Table 1 were calculated from asymptotics- and parametric bootstrap-based methods, respectively; the results based on the two methods were very close to each other, confirming the results in Theorem 2. We further studied the performance of the asymptotics-based method under different sparsity levels ($s = 0.001, 0.05, 0.7$) and dimension $p = 4000$.

Figure 4.1 shows the empirical power for different methods under high-dimensional scenarios. When the signals were extremely sparse at $s = 0.001$, as expected, the supremum-type test SPU(∞) and aSPU performed much better than the competing tests, the GT and the HDGLM, in terms of power. When the signal non-sparsity increased from 0.001 to 0.05, the aSPU test performed similarly to the sum-of-squares-type tests, such as the GT and the HDGLM, and it was much more powerful than the supremum-type test SPU(∞). As the signals became more dense at $s = 0.1$, the aSPU test was the most powerful, closely followed by the SPU(1) and SPU(2) tests. At $s = 0.7$, the aSPU test remained to be the winner, and the SPU(1) test was more powerful than the sum-of-squares-type and supremum-type tests. Under all the situations considered, the aSPU consistently maintained high power, being either the winner or close to the winner.

Next, we analyzed the sensitivity of the aSPU test to the choice of Γ. Figure 4.2 shows the results for aSPU with $\Gamma_1 = \{1, 2, \ldots, 4, \infty\}$, $\Gamma_2 = \{1, 2, \ldots, 6, \infty\}$, $\Gamma_3 = \{1, 2, \ldots, 8, \infty\}$, and $\Gamma_4 = \{1, 2, \ldots, 10, \infty\}$ under different scenarios. As shown in Figure 2, the aSPU test was relatively robust to the choice of Γ.

In summary, due to the nature of its adaptiveness, the aSPU test either achieved the highest power or was close to the winner under various scenarios, validating its consistently good performance across a wide range of scenarios.

4.4.2 Real Data Analysis

Alzheimer's disease (AD) is the most common form of dementia, affecting many millions around the world. The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal multisite observational study of healthy elders, mild cognitive impairment (MCI), and AD [88]. It is jointly funded by the National Institutes of Health (NIH) and industry via the Foundation for the NIH and the Principal Investigator of this initiative is Michael W. Weiner, VA Medical Center and University of California. The major goal of ADNI is to test whether serial MRI, positron emission tomography (PET), and other biological markers can be combined to measure the progression of MCI and early AD. ADNI has recruited more than $1,500$ subjects, ages range from 55 to 90, to participate in the research. For latest information, see `www.adni-info.org`.

One objective of ADNI is to elucidate genetic susceptibility to AD. Due to a relatively small sample size and usually small genetic effect sizes, applying a univariate test to the ADNI data failed to identify any SNP passing the genome-wide significance level at 5×10^{-8} [89], and even a much larger meta-analysis of 74,046 individuals only identified very few genome-wide significant SNPs [90]. Hence, it is natural to consider possible associations at the pathway or even chromosome level, which may be more powerful through effect aggregation and a reduced burden of multiple testing, and shed light on the underlying genetic architecture.

We ran quality control steps first. To be specific, we filtered out all SNPs with a minor allele frequency < 0.05, those with a genotyping rate < 0.95, and those with a Hardy-Weinberg equilibrium test p-value $< 10^{-5}$. For testing polygenic effects (on chromosome level), we pruned SNPs with a criterion of linkage disequilibrium $r^2 > 0.1$ using a sliding window of size 200 SNPs and a moving step of 20. For pathway-level

analysis, we pruned SNPs with a criterion of linkage disequilibrium $r^2 > 0.8$ using a sliding window of size 50 SNPs and a moving step of 5. We imputed the missing SNPs via a Michigan Imputation Server [91] with the 1000 Genomes Project European ancestry samples as the reference panel. For covariates, we included gender, years of education, handedness, age, and intracranial volume measured at baseline. To better demonstrate the possible power differences among the different tests, we applied the tests at either the chromosome or pathway level.

First, we conducted polygenic testing at the chromosome level. The family-wise nominal significance level was set at 0.05, yielding a $0.05/22 \simeq 0.0023$ significance cutoff for each chromosome after the Bonferroni adjustment. Table 4.2 shows some representative results for both asymptotics and parametric bootstrap-based p-values for each test. Most asymptotic p-values of the proposed SPU and aSPU tests were close to their parametric bootstrap-based ones, indicating good approximations by asymptotics. The aSPU test gave significant p-values (< 0.0023) for 5 chromosomes. In contrast, The HDGLM [2] yielded significant p-values for only two chromosomes. As expected, the p-values of HDGLM were close to that of SPU(2) since the two test statistics are similar. Perhaps due to dense and weak signals on these chromosomes, the supremum type test SPU(∞) was not significant in any chromosome while the burden test SPU(1) was often more significant. However, in some situations, SPU(γ) with a larger γ might perform better. For example, for chromosome 5, perhaps due to moderately sparse and weak signals, SPU(3) gave the most significant p-value. Another example was for chromosome 14, SPU(3) yielded a significant result, while HDGLM gave a non-significant one. A meta-analysis of 74,046 individuals identified 2 SNPs at the genome-wide significance level on chromosome 14 [90], validating that chromosome 14 was not a false positive finding by SPU(3). Due to its adaptiveness, the aSPU test often yielded more significant results than the HDGLM across the chromosomes.

Next we conducted a pathway-based analysis. We retrieved a total of 214 pathways from the KEGG database [92]. As in practice [93], we restricted our analysis to pathways of at most 200 genes and at least 10 genes, and excluded the pathways with less than 1000 SNPs, leading to 141 pathways for the following analysis. We set a $0.05/141 \simeq 3\times10^{-4}$ significance cutoff for each pathway after the Bonferroni adjustment. Figure 4.3 compares the p-values of the asymptotics- and parametric bootstrap-based

methods, showing that the p-values of the former method were close to those of the latter, validating the good performance of the asymptotic results in Theorem 2 for real data analyses. The Pearson correlations of the p-values between the two methods ranged from 0.965 to 0.998. Table 4.3 shows 10 KEGG pathways with p-values less than 3×10^{-4} by either aSPU or GT or HDGLM. The three tests identified 10, 0, 1 significant pathways, respectively. The KEGG Alzheimer's disease pathway (hsa05010) can be treated as a true positive since the common variant in the APOE gene (one gene in the KEGG Alzheimer's disease pathway) alone explains 6% of total AD phenotypic variance [94]. For HSA05010 pathway, only the aSPU test gave a signficant p-value $< 3 \times 10^{-4}$, however, not by either GT (p-value= 0.0038) or HDGLM (p-value= 0.0014). Sporadic amyotrophic lateral sclerosis (ALS) is an age-associated disease and there are some evidence showing that ALS and AD are triggered by some common factors [95], while acute myeloid leukemia has been discovered to be associated with AD by other studies [96], lending some support for other two identified pathways (HSA05014 and HSA05221). Perhaps due to very strong but sparse signals in these three pathways, aSPU could identified these three pathways while GT and HDGLM failed.

In summary, the two real data applications here demonstrate that our proposed aSPU test was competitive and can be potentially useful in practice due to its adaptiveness.

4.5 Discussion

We have proposed a highly adaptive association test on a high-dimensional parameter in a GLM in the presence of a low-dimensional nuisance parameter. Its asymptotic null distribution is established, facilitating its asymptotic p-value calculations. At the first glance, the technical details of proving Theorems 1 and 2 are similar to those in a previous paper [79], however, the problem is more challenging here due to the presence of nuisance parameters.

As shown in both simulations and real data analyses, the proposed aSPU test is powerful across a wide range of scenarios considered. In comparison, two other existing tests, HDGLM [2] and GT [1], based on the sum of squares of the score vector, performed similarly to SPU(2), all of which were powerful only in situations with moderately dense

signals, but less powerful than some other SPU tests when the signals were either highly dense or highly sparse. In contrast, by combining multiple SPU tests, the aSPU test maintained high power across various scenarios. In addition to polygenic testing, we also applied the proposed aSPU test to pathway or gene set analysis, demonstrating its potential usefulness in practice. An R package *GLMaSPU* implementing the proposed test is publicly available on GitHub and CRAN; to facilitate its use, we have also created an online website at `http://wuchong.org/GLMaSPU.html`.

Table 4.1: Empirical type I error rates and power (%) of various tests in simulations with $n = 200$ and $p = 2000$. The sparsity parameter was $s = 0.1$, leading to 200 non-zero elements in β with a constant value c. The results outside and inside parentheses were calculated from asymptotics- and parametric bootstrap-based methods, respectively.

c	0	0.03	0.05	0.07	0.1	0.15
SPU(1)	5 (5)	33 (32)	59 (59)	73 (74)	84 (86)	92 (92)
SPU(2)	6 (5)	18 (15)	44 (39)	65 (61)	81 (78)	91 (89)
SPU(3)	4 (5)	28 (30)	58 (59)	76 (76)	89 (90)	96 (96)
SPU(4)	4 (6)	11 (14)	33 (36)	55 (58)	74 (75)	87 (87)
SPU(5)	4 (5)	15 (18)	37 (41)	59 (62)	78 (81)	88 (89)
SPU(6)	3 (6)	7 (11)	18 (24)	36 (43)	53 (59)	70 (72)
SPU(∞)	5 (5)	7 (7)	8 (9)	13 (16)	19 (22)	21 (25)
aSPU	5 (5)	22 (25)	53 (57)	75 (77)	90 (90)	96 (96)

Table 4.2: The p-values of various tests for ADNI data. The results outside and inside parentheses were calculated from the asymptotics- and parametric bootstrap-based methods, respectively.

Test	Chromosome (number of SNPs)			
	5 (3445)	13 (2071)	14 (1878)	21 (840)
SPU(1)	0.01 (0.01)	2×10^{-4} (6×10^{-4})	0.002 (0.002)	1×10^{-4} (2×10^{-4})
SPU(2)	0.03 (0.04)	0.11 (0.10)	0.25 (0.22)	0.15 (0.14)
SPU(3)	0.004 (0.003)	7×10^{-5} (7×10^{-4})	5×10^{-4} (2×10^{-3})	5×10^{-4} (2×10^{-3})
SPU(4)	0.11 (0.09)	0.14 (0.13)	0.30 (0.28)	0.33 (0.02)
SPU(5)	0.01 (0.02)	5×10^{-4} (3×10^{-3})	0.001 (0.005)	6×10^{-3} (0.01)
SPU(6)	0.32 (0.29)	0.22 (0.20)	0.28 (0.25)	0.38 (0.32)
SPU(∞)	0.95 (0.87)	0.66 (0.57)	0.07 (0.12)	0.27 (0.23)
aSPU	0.02 (0.03)	3×10^{-4} (9×10^{-4})	0.003 (0.006)	7×10^{-4} (5×10^{-4})
HDGLM	0.04 (0.04)	0.14 (0.12)	0.29 (0.25)	0.20 (0.17)

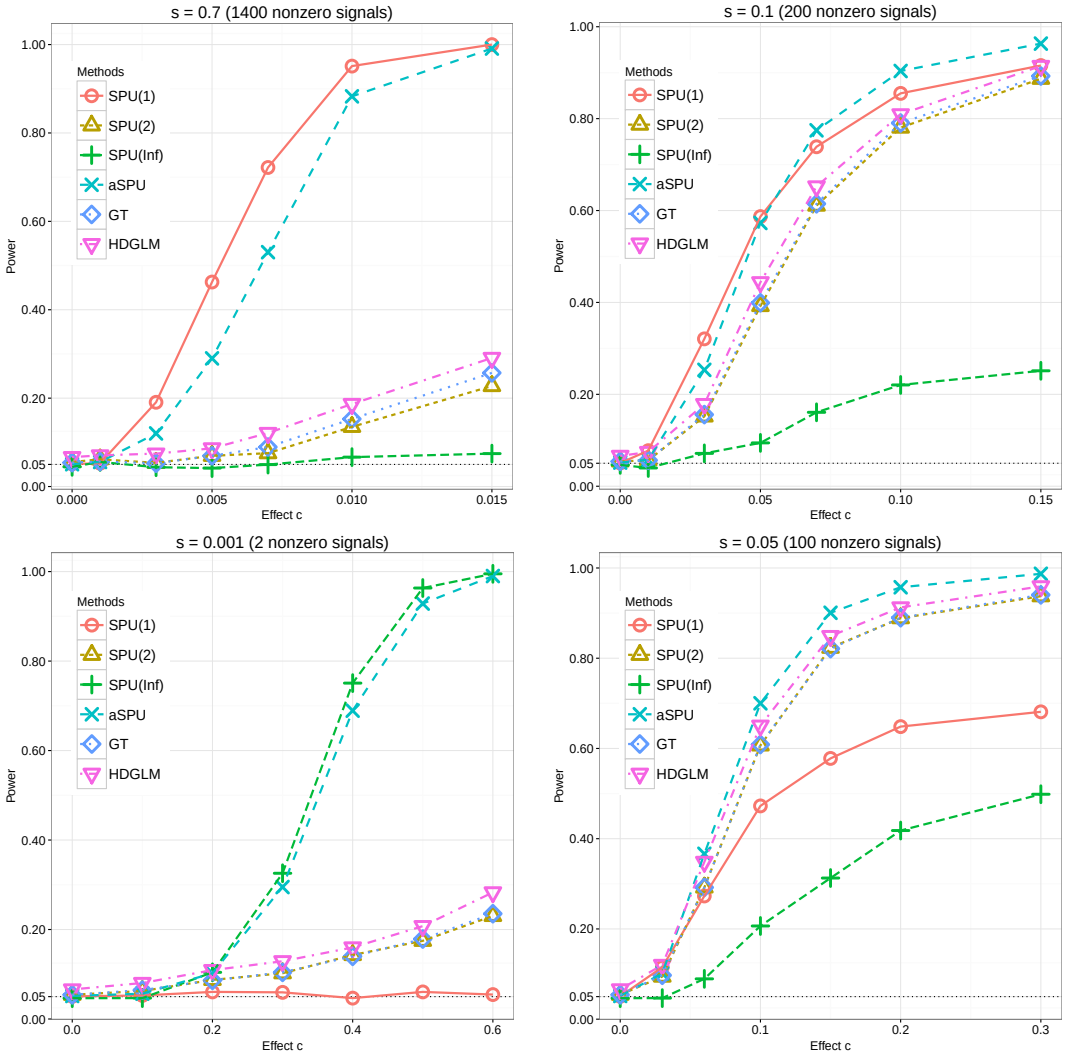

Figure 4.1: Empirical powers of SPU(1), SPU(2), SPU(∞), aSPU, GT [1], and HDGLM [2]. The signal sparsity parameter s varies from 0.001 to 0.7. We set $n = 200$ and $p = 2000$.

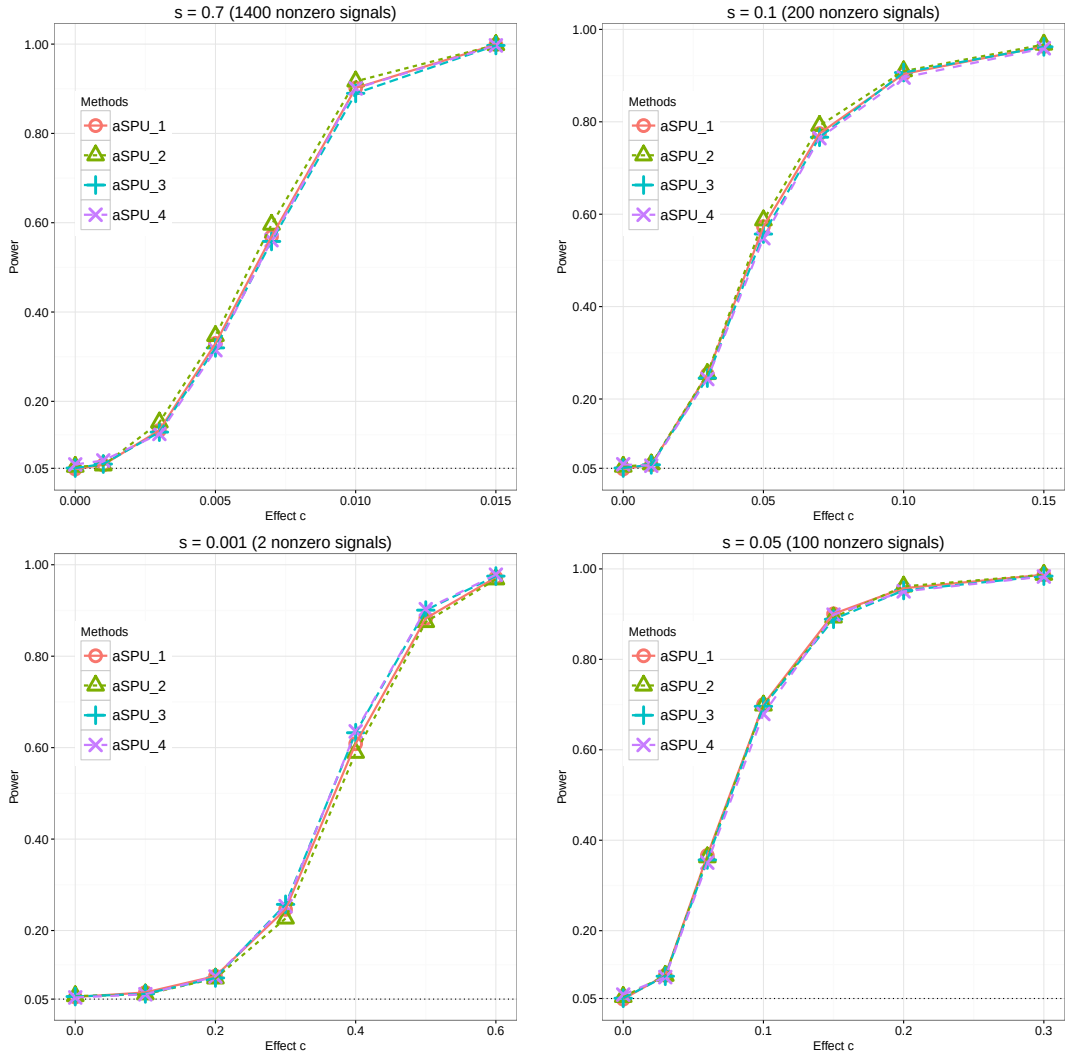

Figure 4.2: Empirical powers of aSPU with different Γ set. aSPU_1, aSPU_2, aSPU_3, aSPU_4 represent aSPU with $\Gamma_1 = \{1, 2, \ldots, 4, \infty\}$, $\Gamma_2 = \{1, 2, \ldots, 6, \infty\}$, $\Gamma_3 = \{1, 2, \ldots, 8, \infty\}$, and $\Gamma_4 = \{1, 2, \ldots, 10, \infty\}$, respectively. The signal sparsity parameter s varies from 0.001 to 0.7. We set $n = 200$ and $p = 2000$.

Table 4.3: Results of the ADNI data analysis: the significant KEGG pathways with p-values $< 3 \times 10^{-4}$ by any of aSPU, GT and HDGLM.

				p values		
KEGG ID	Pathway Name	# Genes	# SNPs	aSPU	GT	HDGLM
hsa05010	Alzheimer's disease	151	7251	0.0E+00	3.8E-03	1.4E-03
hsa05014	Amyotrophic lateral sclerosis	52	2503	0.0E+00	2.3E-03	3.2E-04
hsa05221	Acute myeloid leukemia	55	2024	0.0E+00	2.6E-03	7.6E-04
hsa04520	Adherens junction	72	6179	9.0E-09	4.4E-01	4.7E-01
hsa00071	Fatty acid degradation	40	1110	5.3E-08	1.6E-02	8.0E-03
hsa00830	Retinol metabolism	61	1256	2.1E-07	4.1E-03	7.9E-04
hsa00350	Tyrosine metabolism	38	1194	4.0E-07	7.7E-03	2.4E-03
hsa00982	Drug metabolism	70	1472	2.2E-05	3.6E-02	2.6E-02
hsa00534	Heparin	26	1630	6.4E-05	6.2E-04	1.1E-05
hsa00980	Metabolism of xenobiotics	68	1576	1.6E-04	9.5E-02	9.1E-02

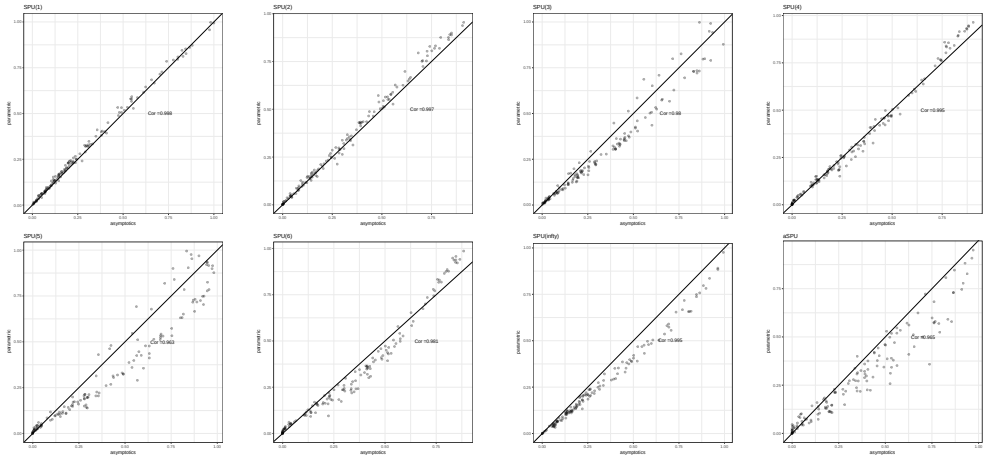

Figure 4.3: Comparison between the asymptotics- and the parametric bootstrap-based p-values of SPU(γ) and aSPU.

Chapter 5

Integrating eQTL Data with GWAS Summary Statistics in Pathway-based Analysis

Many genetic variants affect complex traits through gene expression, which can be exploited to boost statistical power and enhance interpretation in genome-wide association studies (GWASs) as demonstrated by the transcriptome-wide association study (TWAS) approach. Furthermore, due to polygenic inheritance, a complex trait is often affected by multiple genes with similar functions as annotated in gene pathways. Here we extend TWAS from gene-based analysis to pathway-based analysis: we integrate public pathway collections, expression quantitative trait locus (eQTL) data and GWAS summary association statistics (or GWAS individual-level data) to identify gene pathways associated with complex traits. The basic idea is to weight the SNPs of the genes in a pathway based on their estimated cis-effects on gene expression, then adaptively test for association of the pathway with a GWAS trait by effectively aggregating possibly weak association signals across the genes in the pathway. The p-values can be calculated analytically and thus fast. We applied our proposed test with the KEGG and GO pathways to two schizophrenia (SCZ) GWAS summary association datasets, denoted SCZ1 and SCZ2 with about 20,000 and 150,000 subjects respectively. Most of the significant pathways identified by analyzing the SCZ1 data were reproduced by the

SCZ2 data. Importantly, we identified 15 novel pathways associated with SCZ, such as *GABA receptor complex* (GO:1902710), which could not be uncovered by the standard single SNP-based analysis or gene-based TWAS. The newly identified pathways may help us gain insights into the biological mechanism underlying SCZ. Our results showcase the power of incorporating gene expression information and gene functional annotations into pathway-based association testing for GWAS. The main part of this chapter has been published in Wu and Pan, 2018 [97].

5.1 Introduction

Although genome-wide association studies (GWASs) have been remarkably successful in identifying genetic variants associated with complex traits and diseases, only a small to modest proportion of heritability for most complex traits and diseases can be explained by the identified genetic variants [98]. Furthermore, since the majority of identified variants are found in non-coding regions that are not in linkage disequilibrium (LD) with coding exons, a mechanistic understanding of how these variants influence traits is generally lacking [99, 15]. However, it is now known that an important class of variants, termed expression quantitative trait loci (eQTLs), affect complex traits by regulating gene expression levels; there is an enrichment of eQTLs among the GWAS trait-associated variants [14, 15]. Accordingly, transcriptome-wide association study (TWAS) and related methods [16, 17, 18] were proposed to integrate eQTL data with GWAS data to identify the genes associated with a complex trait. These methods may iimprove statistical power to detect associations relative to traditional SNP-based GWAS and gene-based tests that ignore information on gene expression regulation. Nevertheless, due to the limited sample sizes of eQTL data and GWAS data, they may fail to identify some more weakly associated genes with smaller effect sizes. On the other hand, genes do not work in isolation; instead, a group of functionally related genes as annotated in a biological pathway are often involved in the same disease susceptibility and progression [19]. Gene-based analysis testing each gene one-by-one may miss an important pathway if each gene in the pathway has only a small effect size, but in aggregation they contribute substantially. Hence, association analysis of a group of

functionally related genes, called *pathway-based analysis*, has been proposed and applied in practice to boost statistical power and improve interpretability over gene-based analysis for GWAS [100, 101, 102, 103, 104].

Here, we extend integrative gene-based testing like TWAS to integrative pathway-based association analysis to identify pathways associated with complex traits and diseases. Specifically, we propose a new self-contained test that integrates eQTL-derived weights, GWAS individual-level or summary data, SNP LD information, and gene functional annotations as public pathway collections to identify pathways associated with a complex trait (Figure 5.1). As in TWAS, we first estimate the cis-effects of the SNPs in each gene on its expression level, then adaptively test for association between a pathway and a trait by effectively aggregating possibly weak association signals across the genes in the pathway.

We note that our methodology differs from existing approaches. In principle, existing pathway-based analysis methods can be applied in a two-step approach. After obtaining the p-value for each gene by applying TWAS or a related method, an existing pathway analysis method, such as gene set enrichment analysis (GSEA) [105] or DAVID [106], can be applied to identify significant pathways. As to be shown later, a two-step approach, critically depending on the output of a gene-based test, may lose power as compared to our integrated single-step method. Furthermore, many existing pathway methods, including GSEA and DAVID, belong to the category of competitive tests, which compare the p-values of the genes in a given pathway with the p-values of other background genes to determine the significance level, while our method is a self-contained test with a null hypothesis that none of any genes in the pathway is associated with the disease; it is known that a self-contained test is often more powerful [107]. In addition, all the existing pathway analysis methods are only for GWAS data alone while failing to take advantage of eQTL information, leading to power loss and difficulties in interpreting the findings.

Our study was motivated by analyses of schizophrenia (SCZ) GWAS summary data. SCZ is a major chronic and severe mental disorder that is associated with considerable morbidity and mortality [108] and affects about 1% of the population. Although the high heritability of SCZ has been demonstrated by previous studies [109], to date, one of the largest GWAS meta-analyses, conducted by the Schizophrenia Working Group

of the Psychiatric Genomics Consortium (PGC), has only identified 128 independent associations spanning 108 conservatively defined loci [110]. To improve the statistical power and interpretability of the results, [16] applied TWAS to the PGC GWAS summary data and identified 157 significant genes, of which 35 did not overlap with a genome-wide significant locus within 500 kb. However, the pathophysiology of SCZ remains largely unknown and thus it is hard to develop new drugs with high efficacy and low side effects. Identifying SCZ-associated pathways is a crucial step for mechanistic understanding of SCZ and thus developing new drugs. Here, we performed gene- and pathway-based analyses to identify SCZ-associated genes and pathways, providing insights into the underlying mechanism of SCZ.

We reanalyzed two SCZ GWAS summary datasets, which were downloaded from the PGC website (see URLs): a meta-analyzed SCZ GWAS dataset with 8,832 cases and 12,067 controls, denoted as SCZ1 [111], and a more recent and larger one with 36,989 cases and 113,075 controls, denoted as SCZ2 [110]. First, we focused on gene-based analysis. By noting that TWAS is the same as the weighted Sum test with gene expression derived weights [18], we applied some more powerful tests, such as the weighted sum of squared score (SSU) test and the weighted adaptive sum of powered score (aSPU) test [50]. We analyzed the SCZ1 data and identified 51, 108, and 87 significant genes by applying TWAS, (weighted) SSU, and (weighted) aSPU, respectively. Among these identified genes, about 90% genes contained genome-wide significant SNPs within 500 kb in the SCZ2 data, constituting a highly significant and intuitive support for the identified loci. We then applied these tests to the SCZ2 data and identified 75 novel SCZ genes, of which 50 have not been reported in the literature yet. These results further confirm that both weighted SSU and weighted aSPU can improve statistical power to identify more associated genes over that of TWAS. Second, we conducted pathway-based analysis by applying our proposed approach with the Kyoto Encyclopedia of Genes and Genomes (KEGG) [112] and Gene Ontology (GO) [113] candidate pathways to the SCZ1 and SCZ2 data. Most of the significant pathways identified by analyzing the SCZ1 data were confirmed by the SCZ2 data. When analyzing the SCZ2 data, a two-step approach combining TWAS and an existing pathway method, DAVID, identified only one significant pathway, *sequence-specific DNA binding* (GO:0003700), which was also identified by our proposed method. Importantly, by analyzing the SCZ2 data we identified 15

novel significant SCZ-associated pathways, such as pathway *GABA receptor complex* (GO:1902710), which were missed by the gene-based TWAS or aSPU analysis. Hence, pathway-based analysis, as a complementary tool to gene-based analysis, may identify some pathways in which individual genes may have only too weak effects to be detected but their aggregated effects are strong. Overall, our results showcase the increased power of integrating GWAS summary data, eQTL data, reference LD information, and gene functional annotations to gain insights into the genetic basis of complex traits.

5.2 Material and Methods

5.2.1 Datasets

We downloaded two publicly available SCZ GWAS summary datasets from the PGC website (see URLs): the SCZ1 data, which contains the meta-analyzed summary statistics based on 20,899 individuals [111], and the SCZ2 data based on 150,064 individuals [110]. The sets of gene expression-derived weights and the 1000 Genomes Project reference panel were downloaded from the TWAS website (see URLs). Following the TWAS set-up, we removed the SNPs with the strand-ambiguous alleles (A/T, G/C) from the GWAS summary data. Two pathway collections, GO and KEGG, were downloaded from the Molecular Signatures Database (see URLs).

5.2.2 Review of TWAS and Related Methods

We review TWAS and its related methods, which take GWAS summary statistics, a set of gene expression-derived weights, and SNP LD information as input. Since all the methods are gene-based by testing the genes one by one, for the purpose of presentation we only need to consider a single gene.

For a given gene, we only consider a region around it (i.e. its coding region extended by a certain distance, say $\pm$500 kb, upstream and downstream from its TSS and TES respectively) for its *cis*-effects. Let $Z = (Z_1, \ldots, Z_p)'$ be a vector of z-scores of the SNPs for the gene based on the GWAS summary data, or constructed from the GWAS individual-level data. The null hypothesis H_0 to be tested is that the SNPs in a given SNP set (of a gene or a pathway) are not associated with a GWAS trait. With $W =$

$(\hat{w}_1, \ldots, \hat{w}_p)'$, a vector of the estimated *cis*-effects of the SNPs on gene expression based on a reference eQTL dataset, TWAS tests on H_0 using the weighted z-scores. Note that, with GWAS individual-level data, TWAS can be interpreted as testing for association between imputed gene expression and the GWAS trait; however, with GWAS summary data, $W'Z$ may be regarded as an imputed $z - score$ for the gene, but not imputed expression level. It turns out that TWAS is equivalent to the weighted Sum test [51, 18]. Because the Sum test implicitly assumes that all variants have an equal effect size and the same effect direction, the Sum test and thus TWAS, as discussed in the previous studies [51, 74, 50], may lose statistical power if the true association effects are sparse (i.e. with many 0s) or the effect directions are different. Note that, due to the usually small sample size of the eQTL dataset, there are always estimation errors with the estimated cis-effects W. More generally, any more powerful tests, such as the weighted SSU test or the weighted aSPU test, can be applied [18]. In particular, the SPU(γ) tests are possible candidates to use, covering some existing ones as special cases [50]. For example, SPU(1) equals to the Sum test, while SPU(2) equals to SSU and a kernel machine regression-based test (also known as SKAT [74] in rare variant analysis) with a linear kernel. As to be confirmed later, the SPU(2) test may yield higher statistical power than TWAS (or SPU(1)). Generally, the SPU(γ) tests with $\gamma \in \Gamma = \{1, 2, \ldots, 6, \infty\}$ can be applied, and their results can be combined by the adaptive aSPU test [50].

Since not all SNPs with non-zero weights (derived from the reference eQTL dataset) were presented in the GWAS summary data, we used the ImpG-Summary software [114] to impute missing z-scores to the 1000 Genomes Project reference panel accordingly. Because the correlations among Z can be approximated by LD among the SNPs [115, 116], we used the 1000 Genomes Project reference panel (European ancestry) (or other panels for other ethnic/racial groups) to estimate the LD and thus the correlation matrix for Z. In this study, we used five sets of gene expression reference weights that were based on the following four eQTL datasets: microarray gene expression data measured in peripheral blood from 1,245 unrelated subjects from the Netherlands Twin Registry (NTR), microarray expression array data measured in blood from 1,264 individuals from the Young Finns Study (YFS), RNA-seq measured in adipose tissue from 563 individuals from the Metabolic Syndrome in Men study (METSIM), and RNA-seq measured in the dorsolateral prefrontal cortex from 621 individuals from CommonMind Consortium

(CMC) [116]. The weights for differentially spliced introns were further constructed by analyzing CMC data (CMC-introns) [116]. All these weights were downloaded from the TWAS website (see URLs). To account for multiple testing, we applied the Bonferroni correction for each set of weights to maximize the consistency with the previously published results [116] and not to over-penalize the use of additional (and often highly correlated) gene expression-derived weights. Specifically, we reported the number of significant genes after correcting for the number of genes tested within the use of each of the five gene expression sets (YFS, NTR, METSIM, CMC, and CMC-introns; 5004 genes on average with none-zero weights and being tested).

5.2.3 A New Pathway-based Test

Given a pathway, we would like to test the null hypothesis H_0 that none of the genetic variants in the pathway is associated with a trait. We introduce a new pathway-based test to integrate gene functional annotations and a reference eQTL dataset with GWAS data. Figure 5.1 illustrates the workflow of our new pathway-based analysis. As a comparison, we also describe a two-step approach combining an existing integrative gene-based test (like TWAS) and an existing pathway analysis method (like DAVID), in which a gene-based p-value is calculated for each gene before they are combined in pathway analysis.

Given a pathway S^*, we first remove the genes whose gene expression-derived SNP weights are all 0, resulting in a subset S containing n genes. We partition its z-score vector $Z = (Z_{1\cdot}', \ldots, Z_{n\cdot}')'$ into the z-score sub-vectors for the genes, say for gene g (with k_g SNPs) as $Z_{g\cdot} = (Z_{g1}, \ldots, Z_{gk_g})'$. For each gene g, we standardize the gene expression derived weights $W_{g\cdot}$ by $W^s_{gi} = W_{gi} / \sum_{i=1}^{k_g} |W_{gi}|$ such that the weights of the genes are in a similar scale to avoid one or few genes (e.g. with large expression levels) dominate. The standardized weights for the gene set S are $W^s = (W^s_{1\cdot}, \ldots, W^s_{n\cdot})'$ with $W^s_{g\cdot} = (W^s_{g1}, \ldots, W^s_{gk_g})$. We propose the following test statistics:

$$\text{PathSPU}(\gamma) = \sum_{g=1}^{n} \sum_{k=1}^{k_g} (W^s_{gk} Z_{gk})^\gamma,$$

$$\text{aSPUpath2} = \min_{\gamma \in \{1,2\}} P_{\text{PathSPU}(\gamma)},$$

where $P_{\text{PathSPU}(\gamma)}$ is the $p-$value of the PathSPU(γ) test. Because PathSPU(1) and PathSPU(2) are independent [117], we can obtain the p-value of aSPUpath2 via the following steps:

1. Calculate the p-values, $p_1 = P_{\text{PathSPU}(1)}$ and $p_2 = P_{\text{PathSPU}(2)}$, based on the theory that PathSPU(1) and PathSPU(2) asymptotically follow a normal distribution and a mixture of χ^2 distribution under H_0, respectively [51].

2. Take the minimum p-value of PathSPU(1) and PathSPU(2), that is $p_{\min} = \min(p_1, p_2)$.

3. By the asymptotic independence of PathSPU(1) and PathSPU(2), the p-value for the aSPUpath2 is $p_{\text{aSPUpath2}} = 1 - (1 - p_{\min})^2$.

The aSPUpath2 test is new in two aspects: first, unlike many other pathway-based methods aggregating information from only SNP data [118, 102], aSPUpath2 incorporates information in a reference eQTL dataset, thus increasing the power and providing mechanistic insights; second, unlike many other methods, for example fastBAT [102], which are non-adaptive and thus only powerful under some specific alternatives, aSPUpath2 adaptively combines information and thus can maintain relatively high power across a wider range of situations. Finally, we note that aSPUpath2 is a special case of a more general and adaptive pathway-based test called aSPUpath [101, 118], motivated by the following two considerations. First, unlike aSPUpath, the p-value of aSPUpath2 can be calculated analytically and thus fast, though a simulation-based method can be equally applied; as to be demonstrated in the results section, the analogical method provides a good approximation to the simulation-based method. Second, aSPUpath2 is tailored to identifying pathways containing many associated genes or SNPs with only weak effects that cannot be detected by single SNP- or single gene-based analysis, for which it is more powerful.

Hence, aSPUpath2 can be used either alone or as a fast screening procedure for the more time-consuming and more general aSPUpath test.

We extracted candidate pathways from two gene functional annotation sources, KEGG and GO, which were downloaded from the MSigDB database ([105]; see URLs). Because a small pathway gives results not much different from a gene-based analysis, whereas the biological function of a large pathway is not specific, we restricted our analyses to the pathways containing between 10 and 200 genes, which is widely adopted in

pathway-based analysis [93, 101]. On average, we analyzed 4,220 gene sets for each set of weights. To account for multiple testing, we applied the Bonferroni correction within each set of weights and used a slightly conservative cutoff $0.05/5000 = 1 \times 10^{-5}$. Owing to the non-independence nature of many pathways, the Bonferroni correction might be over-conservative here.

5.2.4 Other Existing Pathway-based Tests

In principle, an existing pathway analysis method, in couple with a gene-based test, can be applied in a two-step approach. We compared our new method with this two-step approach using two popular pathway analysis methods, i-GSEA4GWAS [119] and DAVID [106], to further illustrate the power of our proposed test. Specifically, for i-GSEA4GWAS, we uploading the p-values for the genes (calculated by TWAS or SSU or aSPU) for a given pathway to the i-GSEA4GWAS server (see URLs). For DAVID, we uploaded to the DAVID server (see URLs) the significant genes identified by TWAS or SSU or aSPU as the gene list and used the genes we analyzed as the background.

5.2.5 Web Resources

The URLs for data presented herein are as follows:

- DAVID server: `https://david.ncifcrf.gov`;
- iGSEA4GWAS server: `http://gsea4gwas.psych.ac.cn`;
- MSigDB: `http://software.broadinstitute.org/gsea/msigdb/collections.jsp#C3`;
- NHGRI-EBI GWAS Catalog: `http://www.ebi.ac.uk/gwas/home`;
- PGC summary data: `https://www.med.unc.edu/pgc/downloads`;
- TWAS website: `http://gusevlab.org/projects/fusion`.

5.3 Results

5.3.1 TWAS and Related Methods Identify Known and Novel SCZ-associated Genes

First we applied TWAS (i.e. the weighted Sum test), the (weighted) SSU and (weighted) aSPU tests (that integrate gene expression-derived weights) to the SCZ1 data [111] of 20,899 individuals to identify SCZ-associated genes. Then we looked for genome-wide significant SNPs around these genes in the larger SCZ2 data [110] of 150,064 individuals for partial validation. Table 5.1 summarizes the numbers of the significant genes identified by the methods with the SCZ1 data. TWAS, SSU, and aSPU identified 51, 108, 87 significant genes (after taking the union of the results using the five sets of weights), respectively. Among these 87 significant genes identified by aSPU, 64 (around 70%) and 79 (around 90%) contained the genome-wide significant SNPs (p-value $< 5 \times 10^{-8}$) within 500 kb in the SCZ1 data and the SCZ2 data respectively, offering a highly significant validation of the identified loci. For TWAS and SSU, we have the similar proportions of the genes containing the genome-wide significant SNPs in both the SCZ1 and SCZ2 data. Clearly, SSU and aSPU identified more associated genes than TWAS. Compared to TWAS, SSU and aSPU can still maintain high power if many of the weighted SNPs in a gene are not associated with a trait or their associations are in different directions. Since we do not know the sparsity level and association directions of the underlying association patterns, we used the adaptive aSPU test. Here, perhaps due to the denser association patterns (i.e. with many associated SNPs), SSU identified a larger number of SCZ-associated genes than aSPU.

Then, we applied TWAS, SSU, and aSPU to the SCZ2 data, listing the number of significant genes identified by each method in Table 5.2. Here, we analyzed the whole SCZ2 data, which were based on 36,989 cases and 113,075 controls, while Gusev et al. (2016) [116] analyzed the non-overlapping case-control samples with 34,241 cases and 45,604 controls. This data difference led to our findings slightly different from their published ones [116]: applying TWAS to the SCZ2 data, we identified 202 significant genes, while Gusev et al. (2016) [116] identified 157 significant genes. Because the sample size of the SCZ2 is much larger than that of the SCZ1, applying to the SCZ2 data identified a much larger number of significant genes by each method. Again, SSU

and aSPU appeared to be more powerful than TWAS in terms of the number of the identified significant associations. However, because under different scenarios different tests may be more powerful, each test identified some unique genes missed by the other tests.

Overall, we identified 410 significant (and unique) genes by the three methods based on analyzing the SCZ2 data, of which 142 did not overlap with any genome-wide significant SNPs within ±500 kb in the SCZ2 data. Next, to consider the effects of different sets of weights (25,018 tests in total), we used a more stringent cutoff ($0.05/25,018 = 2 \times 10^{-6}$) to report the highly significant genes. We report the new associations that are more than 500 kb away from any genome-wide significant SNPs in the SCZ2 data. Note that TWAS, SSU, and aSPU identified 23, 68, and 32 highly significant genes, respectively, showcasing the increased discovery power of applying other tests over TWAS. Table 5.3 reports 32 highly significant genes identified by aSPU. We searched the NHGRI-EBI GWAS Catalog ([120]; see URLs) to determine if these significant genes have been reported by other studies. Among these 32 genes, 10 have been reported by other studies. On the other hand, among the 75 significant genes identified by any method, 20 genes, such as *FOXN2* (MIM: 143089; [121]), *MSRA* (MIM: 601250; [122], and *PAX5* (MIM: 167414; [123]), have been reported by other studies. Overall, these 75 newly identified genes represent a class of discoveries that would have been missed by the standard single SNP-based test, due to not only their power differences, but also the distal locations of the genome-wide significant SNPs.

5.3.2 New Pathway Method Identifies Known and Novel SCZ-associated Pathways

We applied the new pathway test aSPUpath2 to both the SCZ1 and SCZ2 data. Figure 5.2 compares its p-values from the asymptotics- and Monte Carlo simulation-based methods, showing that the asymptotics gave a good approximation to the gold standard but time-consuming simulation-based method. The correlation of $-\log_{10}$ p-values between these two methods for PathSPU(1), PathSPU(2), and aSPUpath were 0.9989, 0.9981, and 0.9972, respectively. Because the simulation-based method is computationally demanding while the asymptotics-based method is accurate and much faster, we used the asymptotics-based method to calculate the p-values of aSPUpath2 for the

subsequent analysis.

We gave the significant gene sets identified by anlalyzing SCZ1 and SCZ2 data as the SCZ1- and SCZ2-based significant gene sets. For simplicity, we denote them as the *SCZ1* and *SCZ2* gene sets, respectively. Our new method aSPUpath2 with the CMC-based weights identified 33 significant pathways, of which 24 (around 80%) contained the significant genes in the *SCZ1* gene set while 31 (around 94%) contained the significant ones in the *SCZ2* gene set. In particular, aSPUpath2 with the CMC-based weights identified six significant pathways that contained at least one significant gene in the *SCZ2* gene set but no significant genes in the *SCZ1* gene set, such as pathways *synapse organization* (GO:0050808, p-value $= 1.14 \times 10^{-6}$), *response to transforming growth factor beta* (GO:0071559, p-value $= 1.83 \times 10^{-6}$), *transforming growth factor beta receptor signaling pathway* (GO:0007179, p-value $= 4.28 \times 10^{-6}$), and *positive regulation of transforming growth factor beta production* (GO:0071636, p-value $= 5.65 \times 10^{-6}$). There exist some biological findings partially supporting these identified pathways that would be otherwise missed by gene-based analysis. Multiple members of transforming growth factor (TGF) beta superfamily play some roles in the developing nervous system [124]. Alteration in TGF-β1 expression has been observed in SCZ patients [125]. Synapse is an important component in the nervous system and SCZ patients were found to have enriched mutations in the genes belonging to the postsynaptic density at glutamatergic synapses [126]. In contrast, aSPUpath2 with the YFS-based weights identified 19 significant pathways, all of which contained at least one significant gene in both the *SCZ1* and *SCZ2* gene sets. Perhaps due to that the CMC-based gene expression was measured from the brain tissue and were more closely related to SCZ, while the YFS-based ones from the blood, the CMC-based weights were more informative. Overall, it was confirmed that pathway-based analysis is useful as a complementary tool to gene-based analysis, offering insights into the genetic basis of complex traits.

As an adaptive test, aSPUpath2 can maintain high power under various scenarios. For example, based on the SCZ1 data, for pathway *nuclear speck* (GO:0016607) with the CMC-based weights, there were 300 marginally significant and negatively associated SNPs (z-score < -1.96) and 309 marginally and positively associated SNPs (z-score > 1.96) among 5741 SNPs with non-zero weights. The varying association directions

among marginally significant SNPs led to a non-significant p-value$= 3.0 \times 10^{-3}$ of PathSPU(1). In contrast, because PathSPU(2) was robust to varying association directions, it yielded a significant p-value$= 2.1 \times 10^{-8}$. By combining the results of PathSPU(1) and PathSPU(2), aSPUpath2 yielded a significant p-value$= 4.1 \times 10^{-8}$. Furthermore, this pathway contained at least two significant genes in both the *SCZ1* and *SCZ2* gene sets, supporting the significance of the pathway. For pathway *regulation of cellular senescence* (GO:2000772) with the CMC-based weights, there were 86 marginally and negatively associated SNPs (z-score < -1.96) and 45 marginally but positively associated SNPs (z-score > 1.96) among 1516 SNPs with non-zero weights. The associations in different directions were not completely canceled out since the number of the negatively associated SNPs was almost twice as that of the positively associated SNPs. PathSPU(1) yielded a significant p-value ($= 1.9 \times 10^{-7}$), while PathSPU(2) yielded a non-significant p-value ($= 2.4 \times 10^{-3}$). Again by combining information from the two tests, aSPUpath2 yielded a significant p-value ($= 3.8 \times 10^{-7}$). This pathway also contained at least one significant gene in both the *SCZ1* and *SCZ2* gene sets. Generally, as any non-adaptive test, PathSPU(1) or PathSPU(2) may lose statistical power under different situations; however, by contrast, aSPUpath2 that data-adaptively aggregates information can maintain relatively high power across a wide range of situations.

Then we analyzed the SCZ2 data. The new test aSPUpath2 with the CMC- and YFS-based weights identified 235 and 242 significant pathways, respectively. Table 5.4 shows the 6 significant KEGG pathways identified by aSPUpath2 with the CMC-based weights. All of these significant pathways covered at least one significant gene in the *SCZ2* gene set while three pathways, *Alzheimer's disease* (hsa05010, p-value $= 2.4 \times 10^{-8}$), *systemic lupus erythematosus* (hsa05322, p-value $= 0.0$), and *hypertrophic cardiomyopathy* (hsa05410, p-value $= 2.3 \times 10^{-9}$), have been reported by other studies to be associated with SCZ [127, 128].

Table 5.5 shows the significant and novel pathways containing no significant genes in the SCZ2 gene set but detected by aSPUpath2 with either the CMC- or the YFS-based weights. Perhaps due to that the CMC-based weights were derived from the brain tissue and thus more relevant to SCZ than the YFS-based weights, using the CMC-based weights identified 12 significant and novel pathways, while using the YFS-based identified only three. Some existing studies partially supported the newly identified pathways.

For example, GABA system plays an important role in orchestrating the synchronicity of local networks and affects cognitive and emotional behavior [129]. Further, cognitive symptoms in SCZ are attributed to a cortical GABAergic deficit [129], partially supporting that pathway *GABA receptor complex* (GO:1902710) is possibly related to SCZ. Overall, these 15 newly identified pathways represent a class of discoveries that would have been missed by gene-based analysis.

5.3.3 Comparisons Between aSPUpath2 and Other Methods

With the application to the SCZ2 data with the CMC-based weights, we compared our proposed method with the two-step approach combining a gene-based test and an existing pathway analysis method, including the popular DAVID [106] or i-GSEA4GWAS [119]. We also compared it with the more general and standard aSPUpath [101].

We applied DAVID [106] with the CMC-based weights and identified one significant pathway: *transcription factor activity, sequence-specific DNA binding* (GO:0003700, Benjamini-corrected p-value $= 4.2 \times 10^{-3}$). This pathway was excluded in our earlier analysis because it contained more than 200 genes; when applied, aSPUpath2 could identify this pathway as well (p-value $= 4.5 \times 10^{-7}$). We also applied i-GSEA4GWAS [119] but failed to identify any significant pathways. In addition to the two-step nature of the above two pathway methods (thus depending on the output or performance of the gene-based testing in the first step), in contrast to the one-step approach of aSPUpath2, they also differ with respect to their null hypotheses being tested: both DAVID and i-GSEA4GWAS belong to the category of "competitive tests" testing for the enrichment of the associated genes in the pathway being tested as compared to other pathways, while our aSPUpath2 method is a "self-contained test" as a global test for identifying whether there is (are) any significant gene(s) in the pathway; due to the difference between the null hypotheses being tested, a self-contained test is in general more powerful than a corresponding competitive test.

Figure 5.3 shows the running times for aSPUpath2 and aSPUpath. Due to the computational constraint, we ran at most $B = 10^6$ simulations to calculate the p-values for aSPUpath. For the simulation-based method, the running time increased rapidly with the number of simulations, for which a larger value is required for a more significant

p-value. In contrast, since the p-values of aSPUpath2 was calculated by the asymptotics-based method, the running time was invariant to the p-values. aSPUpath with the CMC-based weights identified 179 significant pathways, of which 139 (around 80%) were also identified by applying aSPUpath2 with the CMC-based weights, constituting a highly significant overlap between their results. Furthermore, aSPUpath2 identified a total of 235 significant pathways, showcasing possibly higher statistical power over aSPUpath for the SCZ2 data. In summary, aSPUpath2 is several orders faster than aSPUpath, more so for large and highly significant pathways, and can be more powerful for densely associated pathways (i.e. those containing many associated SNPs/genes), thus we recommend using aSPUpath2 either alone or as a fast screening procedure for the more time-consuming and more general aSPUpath test.

5.3.4 Simulations

We conducted simulation studies to evaluate and compare the performance of our proposed new aSPUpath2 test with the aSPUpath test. We generated simulated data to mimic real data: we used the GO Biological Process pathways and CMC-derived SNP weights, and simulated z-scores as GWAS summary statistics for SNPs. Specifically, for a given pathway S^* in the GO Biological Process pathway database, we first removed the genes whose CMC-derived SNP weights were all 0, resulting in a subset S containing n genes and p SNPs with none-zero weights. We generated a z-score vector from a multivariate normal distribution, $Z \sim N(\mu, \Sigma)$, where $\mu = (\mu_1, \ldots, \mu_p)'$ was the mean and Σ was the LD matrix based on the 1000 Genomes Project reference panel (European ancestry), respectively. Note that z-scores are expected to have a multivariate normal distribution asymptotically. To save computing time, we assumed that the SNPs from different chromosomes were independent and only considered the pathways with less than 2000 SNPs. In total, we considered 1905 pathways. Further, we defined SNP j was associated or informative with the corresponding $\mu_j = \text{sign}(W_j)c$, where W_j was the CMC-derived weight for SNP j, $c \neq 0$ was some positive constant, and $\text{sign}(a)$ gave the sign of a; in contrast, SNP j was non-informative with $\mu_j = 0$. Note that we also considered non-constant $|\mu_j|$ for associated SNPs. To evaluate type I error rates, we considered the null case (set-up A) with no informative SNP ($\mu = 0$). To evaluate power, we further considered the following four set-ups under different situations:

set-up B, 50% SNPs in each gene were informative; set-up C, 10% SNPs in each gene were informative; set-up D, only one SNP in each gene was informative; and set-up E, only one SNP in 20% of the genes in the pathway was informative. Other SNPs were set to non-informative and we varied the true association strength c to generate power curves for set-up B to E. After generating a z-score vector for each pathway, we applied both the aSPUpath2 and aSPUpath tests. The entire procedure was repeated about 38,000 times (i.e. 20 per pathway) for set-up A. For other set-ups, with different c, we repeated the entire procedure about 1,900 times (1 per pathway) and fixed the nominal significance level at $\alpha = 0.05$.

Table 5.6 shows the empirical type I error rates, indicating that the PathSPU(1), PathSPU(2), and aSPUpath2 could control their type I rates satisfactorily under various nominal significance levels.

Figure 5.4 shows statistical power under set-ups B to E. In set-up B, because 50% of the SNPs in the pathway were informative with dense association signals, PathSPU(1) was expected to be most powerful as confirmed in Figure 5.4; since aSPUpath2 combined the information from both the PathSPU(1) and PathSPU(2), aSPUpath2 also achieved high power close to PathSPU(1). When the association signals were less dense with only 10% of the SNPs as informative (set-up C), all the tests performed similarly, though aSPUpath2 and PathSPU(1) had a slight edge over aSPUpath and PathSPU(2) respectively. When most SNPs (set-up D) or most genes were not associated with the trait (set-up E), aSPUpath was expected to be more powerful than aSPUpath2 because aSPUpath2 is tailored to identifying dense associations of pathways containing many associated SNPs/genes with only weak effects. In other simulation set-ups with varying $|\mu_j|$ for associated SNPs and/or different proportions of associated SNPs/genes, we obtained similar results (not shown). Note that, by theory, there is no uniformly most powerful test for pathway analysis; aSPUpath is more general and thus expected to be high powered across a wider range of scenarios than aSPUpath2, which is tailored for and more powerful for detecting dense association signals like in set-up A. However, aSPUpath2 is much faster than aSPUpath. Hence, as mentioned earlier, we recommend using aSPUpath2 either alone to detect densely associated pathways, or as a fast screening procedure for aSPUpath if one is interested in both densely and sparsely associated pathways.

5.4 Discussion

In this work, we have presented a powerful and adaptive method that integrates genetic and transcriptional variations to identify pathways associated with a complex trait. Using gene expression to construct weights and then adaptive weighting to identify significant pathways has some potential advantages. First, a pathway may be a more interpretable biological unit than a single SNP or gene, and may shed light into biological mechanisms underlying a trait or disease. Second, pathway-based analysis, complementary to gene-based analysis, and as demonstrated here, can identify important pathways that may be missed by gene-based analysis. Since different tests will be powerful under different underlying true association patterns, in particular, our proposed test may maintain relatively high statistical power across a wider range of situations due to its adaptive nature of aggregating association information across the genes in a pathway. Third, our proposed method is similar to other integrative gene-based methods, such as TWAS [16], PrediXcan [17] and aSPU [18], that incorporate eQTL information into GWAS analysis. However, differing from that the above integrative methods are gene-based, our method aggregates information across the genes to identify significant pathways. Importantly, unlike TWAS and PrediXcan, which use a simple weighted linear combination of genetic variants (or their z-scores) to construct test statistics, our approach adaptively (and non-linearly) weights the genetic variants and thus aggregates information based on the underlying association patterns to increase discovery power. As shown in our applications, our method could identify some important pathways that were missed by the above integrative gene-based tests, even followed with a standard pathway analysis. Finally, we note that our proposed approach is in the category of "self-contained tests", in which we are interested in identifying any pathway containing one or more genes or SNPs associated with a trait. This is different from the "competitive tests", such as DAVID and GSEA, that would detect pathways enriched with associated genes or SNPs as compared to background pathways.

Application of our proposed and other integrative gene-based methods to two SCZ summary data not only recapitulated many known genes or pathways but also identified many new ones. Specifically, we identified 75 significant genes without any known associated SNPs within 500 kb, of which 50 have not been reported in any studies yet.

It is possible that some of these significant genes represent new findings that have been missed due to the lower statistical power in other standard single SNP- or gene-based test without incorporating gene expression data. Furthermore, some pathways may contain only genes with small effect sizes, which may not be detected even by integrative gene-based tests like TWAS, but may be by our proposed pathway test. Here, we identified 15 novel significant pathways associated with SCZ, such as pathway *GABA receptor complex* (GO:1902710), which could be missed by gene-based TWAS or aSPU. Taken together, our results showcase the power of incorporating reference gene expression data into gene-based or pathway-based association testing for GWAS. The newly identified genes and pathways may help us gain insights into the biological mechanism underlying SCZ.

Although in this study we have mainly focused on SCZ and applied the various methods to two GWAS summary datasets, it is natural to apply our method to other complex traits with either individual-level or summary data. We expect that applying our proposed and other integrative methods like TWAS to other existing GWAS data may identify more novel associations and shed more light on the underlying biological mechanisms. We note that our proposed methodology can be applied with other endophenotype-derived weights [130] or even without weights (i.e. all SNPs with an equal weight).

Finally we comment on our view that TWAS is a weighted Sum test and its related issues, which are also discussed by [131] and in `http://hakyimlab.org/post/vulnerabilities/`. Although TWAS was originally proposed to identify GWAS associations through gene expression, any such discovery based on a single eQTL/GWAS dataset is at most only suggestive to mediating effects of gene expression. As discussed in [18], in spite of the connections of TWAS with two-stage least squares and Mendelian randomization (MR), due to the adopted strong assumptions that are likely to be violated in practice, cautions should be taken to avoid extrapolating any discovered GWAS associations to causal effects mediated through gene expression. Hence, we simply regard TWAS as a special case of weighted association testing. In this view, we yield a few benefits while avoiding possible over-interpretation of an association as a causal effect. First, due to some well-known limitations of the Sum test and inherent errors in estimating the cis-effects (i.e. weights) of genetic variants with usually small eQTL

datasets, modifications to TWAS may lead to more powerful analysis methods, such as based on the SSU/SPU(2) and aSPU tests (Xu et al 2017a). Other tests, like aSPU, with a more flexible weighting scheme, may also identify associations through other non-gene expression-mediated mechanisms. Second, in addition to gene expression, other molecular or clinical intermediate phenotypes can be used to construct weights for weighted GWAS association analysis [130].

The proposed statistical tests are implemented in R package `aSPU2` that is currently publicly available on GitHub (and will be put on CRAN); the online manual and example computer code are publicly available at `wuchong.org/aspupath2.html`.

Table 5.1: The numbers of the significant genes identified by analyzing the SCZ1 data for each single set of the weights and their union across these weights. The numbers a/b/c in each cell indicate the numbers of (a) the significant genes; (b) the significant genes covering at least one genome-wide significant SNP within ±500 kb in the SCZ1 data; (c) the significant genes covering at least one genome-wide significant SNP within ±500 kb in the SCZ2 data.

	YFS	NTR	METSIM	CMC-introns	CMC	Combined
TWAS	14/11/14	13/8/13	8/5/7	18/10/13	16/10/13	51/31/43
SSU	31/25/26	27/19/26	24/14/23	27/17/23	39/25/34	108/67/95
aSPU	29/26/26	23/16/22	21/16/21	26/18/21	28/22/25	87/64/79

Table 5.2: The numbers of the significant genes identified by analyzing the SCZ2 data for each single set of the weights and their union across these weights. The numbers a/b/c in each cell indicate the numbers of (a) the significant genes; (b) the significant genes covering at least one genome-wide significant SNP within ±500 kb in the SCZ1 data; (c) the significant genes covering at least one genome-wide significant SNP within ±500 kb in the SCZ2 data.

	YFS	NTR	METSIM	CMC-introns	CMC	Combined
TWAS	63/19/46	49/22/39	43/11/32	56/17/37	69/21/50	202/63/142
SSU	127/40/94	78/32/59	108/32/76	100/22/61	124/32/85	381/108/255
aSPU	105/40/83	69/34/60	87/33/72	85/24/55	110/34/82	314/110/234

Table 5.3: The significant and novel genes overlapping with no known GWAS risk variants within ±500 kb as identified by aSPU applied to the SCZ2 data. The validated gene-trait associations appeared in the following references: [1] [132]; [2] [110].

Weight	Gene	CHR	P0	P1	aSPU	TWAS	SSU	Most sig. SNP	Validation
YFS/NTR	*MAP7D1*	1	36621801	36646448	5.0E-07	5.2E-07	6.7E-07	3.3E-07	
YFS	*CNN3*	1	95362507	95392834	9.0E-07	7.1E-02	7.4E-08	9.4E-07	
CMC-introns	*GABPB2*	1	151043079	151091007	7.0E-07	4.5E-07	4.5E-07	5.6E-08	
CMC-introns	*TBC1D5*	3	17198653	17784240	1.7E-06	3.1E-06	1.7E-06	5.5E-08	[1]
YFS	*IK*	5	140026643	140042064	1.4E-06	6.4E-07	4.3E-07	3.6E-07	
CMC-introns	*CXXC5*	5	139028300	139062680	1.0E-07	2.1E-09	1.9E-08	1.5E-06	
YFS	*TMCO6*	5	140019012	140024993	1.8E-06	1.1E-04	4.2E-07	3.6E-07	
METSIM	*DND1*	5	140050379	140053171	8.0E-07	3.6E-06	1.1E-06	3.6E-07	
CMC	*ZMAT2*	5	140080031	140086239	1.5E-06	1.6E-05	1.9E-07	3.6E-07	
YFS	*ABCB1*	7	87133175	87342611	1.8E-06	5.5E-04	2.5E-07	1.4E-07	[1]
METSIM/CMC-introns	*ZDHHC2*	8	17013538	17082308	2.0E-07	5.0E-06	7.3E-08	1.1E-07	[1]
CMC-introns	*FGFR1*	8	38268655	38326352	3.0E-07	8.5E-07	7.6E-07	2.3E-07	
YFS	*ENDOG*	9	131580753	131584956	1.2E-06	6.4E-07	9.2E-07	1.9E-06	
METSIM	*PKN3*	9	131464802	131483197	6.0E-07	8.2E-01	2.2E-08	1.9E-06	
CMC	*TEK*	9	27109146	27230172	4.0E-07	1.1E-07	6.4E-08	4.7E-07	[1]
YFS	*ZDHHC5*	11	57435219	57468659	2.0E-07	2.2E-07	1.2E-07	6.7E-08	[2]
NTR/METSIM	*CLIP1*	12	122755979	122907179	1.7E-06	1.9E-04	9.2E-08	3.8E-06	[1]
CMC	*CCDC92*	12	124420954	124457163	1.0E-06	6.8E-03	6.5E-06	4.1E-07	
YFS/NTR	*PPP2R3C*	14	35554678	35591519	6.0E-07	2.7E-07	3.2E-07	1.5E-07	
METSIM	*KIAA0391*	14	35591052	35743271	5.0E-07	6.6E-07	3.0E-07	1.5E-07	[1]
METSIM	*PCNX*	14	71374122	71582099	1.0E-06	3.8E-06	1.8E-07	1.6E-07	[1]
CMC-introns	*AP3B2*	15	83328032	83378635	1.0E-07	1.7E-06	8.5E-08	5.5E-08	
METSIM/CMC-introns	*NMRAL1*	16	4511694	4524896	5.0E-07	1.6E-07	9.3E-06	2.8E-07	
CMC	*CORO7*	16	4404542	4466962	6.0E-07	5.2E-07	4.8E-07	2.8E-07	
CMC	*CPNE7*	16	89642175	89663654	1.0E-07	5.0E-08	4.4E-07	1.1E-07	
YFS/CMC	*CHMP1A*	16	89710838	89724193	1.4E-06	1.5E-02	9.1E-08	1.1E-07	
CMC-introns	*TCF25*	16	89939993	89977792	5.0E-07	1.4E-02	6.8E-08	1.1E-07	
CMC-introns	*CDK10*	16	89753075	89762772	1.7E-06	2.1E-01	9.4E-08	1.1E-07	
CMC	*RPL13*	16	89627064	89633237	1.5E-06	7.8E-04	1.6E-07	1.1E-07	
YFS	*PRPSAP2*	17	18743398	18834581	1.0E-07	5.6E-08	5.3E-07	7.8E-07	
METSIM	*KCNG2*	18	77623668	77660184	1.9E-06	4.5E-03	5.1E-08	2.2E-07	[1]
CMC	*SNRNP70*	19	49588464	49611870	1.0E-07	2.1E-03	1.7E-08	2.2E-07	

Table 5.4: The significant KEGG pathways identified by aSPUpath2 with the CMC-based weights for the SCZ2 data.

ID	Pathway name	PathSPU(1)	PathSPU(1)	aSPUpath2	# sig. genes
hsa05322	Systemic lupus erythematosus	2.6E-04	5.5E-10	1.1E-09	16
hsa05410	Hypertrophic cardiomyopathy	8.3E-02	1.5E-09	2.9E-09	2
hsa05414	Dilated cardiomyopathy	4.0E-01	3.1E-08	6.3E-08	2
hsa04120	Ubiquitin mediated proteolysis	6.1E-02	2.9E-07	5.8E-07	5
hsa05010	Alzheimer's disease	6.7E-01	9.1E-07	1.8E-06	5
hsa05016	Huntington's disease	4.7E-01	2.3E-06	4.5E-06	5

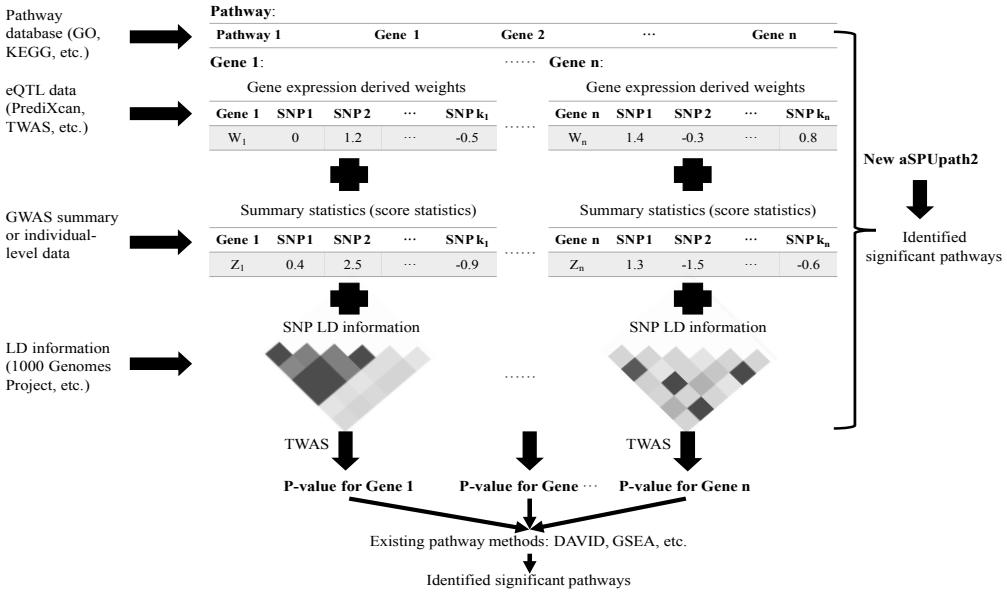

Figure 5.1: Workflow of pathway-based analysis.

Table 5.5: The significant and novel gene sets containing no significant genes as identified by aSPUpath2 with the CMC- or YFS-based weights.

ID	Description	# genes	PathSPU(1)	PathSPU(2)	aSPUpath2	Weights
GO:1902710	GABA receptor complex	18	9.6E-03	0.0E+00	0.0E+00	CMC
GO:1901661	quinone metabolic process	29	7.4E-01	1.0E-08	2.0E-08	YFS
GO:0043162	ubiquitin-dependent protein catabolic process	18	5.8E-01	4.4E-08	8.8E-08	CMC
GO:0016339	calcium-dependent cell-cell adhesion	27	5.7E-01	1.1E-07	2.2E-07	CMC
GO:0030315	T-tubule	45	7.3E-02	1.1E-07	2.3E-07	CMC
GO:0007528	neuromuscular junction development	36	4.7E-01	2.9E-07	5.7E-07	CMC
GO:0003143	embryonic heart tube morphogenesis	62	5.5E-03	4.7E-07	9.5E-07	CMC
GO:0007569	cell aging	67	2.0E-04	8.1E-07	1.6E-06	CMC
GO:0035050	embryonic heart tube development	73	2.3E-02	8.8E-07	1.8E-06	CMC
GO:0004181	metallo-carboxypeptidase activity	27	7.3E-01	9.7E-07	1.9E-06	CMC
hsa00590	Arachidonic acid metabolism	56	3.0E-01	1.2E-06	2.5E-06	YFS
GO:0051279	regulation of release of sequestered calcium ion into cytosol	75	2.7E-06	2.2E-05	5.3E-06	CMC
GO:0072665	protein localisation to vacuole	46	3.5E-01	3.0E-06	6.1E-06	CMC
GO:0010880	regulation of release of sequestered calcium ion into cytosol by sarcoplasmic reticulum	25	5.6E-06	3.0E-06	6.1E-06	CMC
GO:1901800	positive regulation of proteasomal protein catabolic process	98	1.5E-03	4.4E-06	8.7E-06	YFS

Table 5.6: Empirical type I error rates of our proposed pathway-based tests with some varying nominal significance levels α under simulation set-up A.

α	0.05	0.01	0.001
PathSPU(1)	4.9×10^{-2}	9.8×10^{-3}	1.2×10^{-3}
PathSPU(2)	5.3×10^{-2}	1.1×10^{-2}	1.3×10^{-3}
aSPUpath2	4.4×10^{-2}	1.0×10^{-2}	1.2×10^{-3}

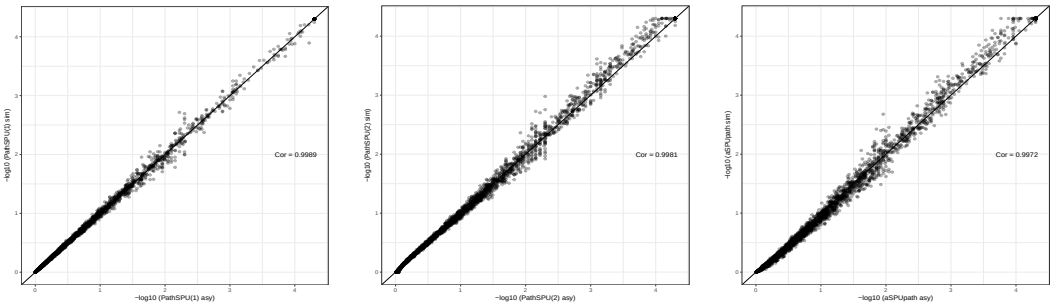

Figure 5.2: Comparison between the asymptotics- and simulation-based p-values of PathSPU(1) (left), PathSPU(2) (middle), and aSPUpath (right) based on the SCZ2 data with the GO Biological Process pathways.

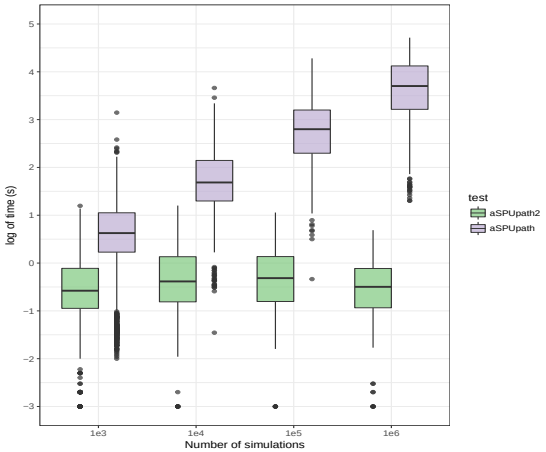

Figure 5.3: Comparison between running times of aSPUpath2 and aSPUpath for the SCZ2 data with the pathways in the GO Biological Process.

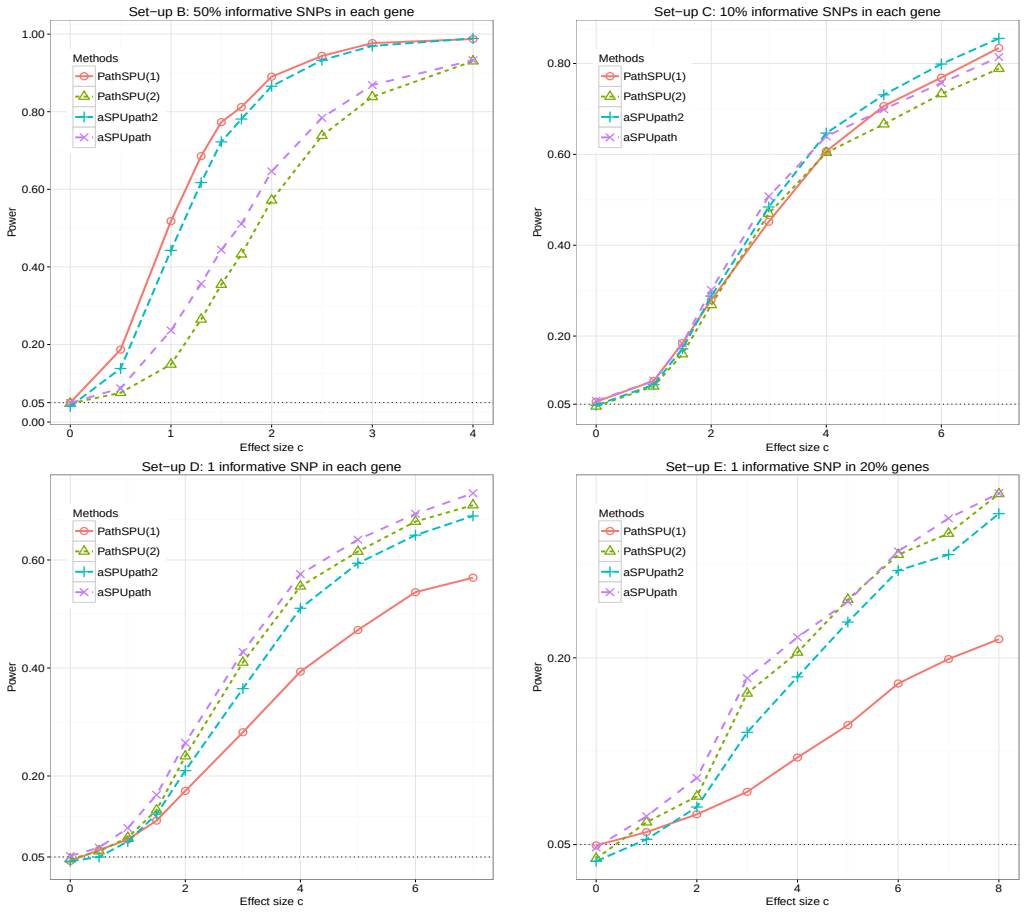

Figure 5.4: Empirical power at $\alpha = 0.05$ under different simulation set-ups (B–E).

Chapter 6

Conclusion and Future Work

6.1 Summary of Major Findings

This thesis introduced several new statistical methods to solve challenges imposed by new types of high-dimensional genetic and genomic data.

Chapter 2 proposed a two-stage (site selection stage plus multiple imputation stage) method to impute missing data in covariates in the epigenome-wide association studies (EWAS), which can help us adjust potential confounders, such as cell type composition. We applied our new method with data from the Atherosclerosis Risk in Communities (ARIC) study to carry out an EWAS between methylation levels and smoking status, in which missing cell type compositions and white blood cell counts are imputed.

Chapter 3 proposed a powerful data-driven approach called aMiSPU by weighting the variables (taxa) in a manner determined by the data itself to overcome low power issue of human microbiome association studies. Our simulations and real-data analyses demonstrated that the aMiSPU test was often more powerful than several competing methods while correctly controlling type I error rates. To help other researchers use our new method, we put the R package *MiSPU* on CRAN.

Chapter 4 proposed an adaptive test on a high-dimensional parameter of a GLM (in the presence of a low-dimensional nuisance parameter), which can maintain high power across a wide range of high-dimensional situations. We further established its asymptotic null distribution. In addition, we applied it and other competing tests to an Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, detecting possible

associations between Alzheimer's disease and some gene pathways.

In Chapter 5, we proposed a novel pathway-based association test by integrating gene expression, gene functional annotations, and a main genome-wide association study dataset. The basic idea was to weight the SNPs of the genes in a pathway based on their estimated *cis*-effects on gene expression, then adaptively test for association of the pathway with a GWAS trait by effectively aggregating possibly weak association signals across the genes in the pathway. We applied it and other competing methods to a schizophrenia (SCZ) GWAS summary association dataset and identified 15 novel pathways associated with SCZ, such as *GABA receptor complex* (GO:1902710), which could not be uncovered by the standard single SNP-based analysis or gene-based TWAS. The newly identified pathways may help us gain insights into the biological mechanism underlying SCZ.

6.2 Future Research

Modern genetics research constantly creates new types of high-dimensional data and imposes new problems on statistical genetics. In this connection, we have several aims for future research on statistical genetics.

- Hypothesis tests on high-dimensional parameters. Chapter 4 introduced an adaptive test for testing a high-dimensional parameters under GLMs with low-dimensional nuisance parameters. However, testing high-dimensional parameter under GLMs with high high-dimensional nuisance parameters is largely untouched. We plan to develop a new adaptive test to address this problem.

- Integrative analysis of GWAS and other data. It remains challenging to boost statistical power of GWAS to identify more risk variants or loci that can account for "missing heritability". Chapter 5 introduced a new way to integrate gene expression, gene functional annotations, and a main genome-wide association study dataset. In the future, we plan to integrate other dateset to boost statistical power. For example, recent biotechnological advances have made it feasible to experimentally measure the three-dimensional organization of the genome, including enhancer-promoter interactions in high resolutions. Due to the well known critical

roles of enhancer-promoter interactions in regulating gene expression programs, such data have been applied to link GWAS risk variants to their putative target genes, gaining insights into underlying biological mechanisms. However, their direct use in GWAS association testing is yet to be exploited. We plan to integrate enhancer-promoter interactions into GWAS association analysis to both boost statistical power and enhance biological interpretability.

- Precision medicine. Every day, millions of people take medications that do not help them. This fact motivates precision medicine, which tailors disease prevention and treatment to a patient based on his/her genetic, genomic, microbiomic, environmental, and lifestyle information. Stratifying patients based on their personal characteristics instead of a few clinical manifestations is an important step in precision medicine. My plan is to develop new classification methods based on multiple sources of personal characteristics and high-dimensional data to improve its effectiveness in predicting the treatment outcome.

In the future, I will continue developing innovative statistical methods for high-dimensional data. Additionally, I will continue to provide open-source and user-friendly software such that researchers can use our new methods easily.

References

[1] Jelle J Goeman, Hans C Van Houwelingen, and Livio Finos. Testing against a high-dimensional alternative in the generalized linear model: asymptotic type 1 error control. *Biometrika*, 98(2):381–390, 2011.

[2] Bin Guo and Song Xi Chen. Tests for high dimensional generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1079–1102, 2016.

[3] Christoph Bock. Analysing and interpreting dna methylation data. *Nature Reviews Genetics*, 13(10):705–719, 2012.

[4] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

[5] Andrew E Teschendorff, Joanna Zhuang, and Martin Widschwendter. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, 27(11):1496–1505, 2011.

[6] Eugene Andres Houseman, John Molitor, and Carmen J Marsit. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30(10):1431–1439, 2014.

[7] James Zou, Christoph Lippert, David Heckerman, Martin Aryee, and Jennifer Listgarten. Epigenome-wide association studies without the need for cell-type composition. *Nature methods*, 11(3):309–311, 2014.

[8] Eugene Andres Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T

Kelsey. Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):1, 2012.

[9] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011.

[10] Ilseung Cho and Martin J Blaser. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260–270, 2012.

[11] M Rauch and SV Lynch. The potential for probiotic manipulation of the gastrointestinal microbiome. *Current Opinion in Biotechnology*, 23(2):192–201, 2012.

[12] Brian H McArdle and Marti J Anderson. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82(1):290–297, 2001.

[13] Ni Zhao, Jun Chen, Ian M Carroll, Tamar Ringel-Kulka, Michael P Epstein, Hua Zhou, Jin J Zhou, Yehuda Ringel, Hongzhe Li, and Michael C Wu. Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. *The American Journal of Human Genetics*, 96(5):797–807, 2015.

[14] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter ACt Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

[15] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

[16] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, 2016.

[17] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L

Nicolae, Nancy J Cox, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 2015.

[18] Zhiyuan Xu, Chong Wu, Peng Wei, and Wei Pan. A powerful framework for integrating eqtl and GWAS summary data. *Genetics*, 207(3):893–902, 2017.

[19] Matthias Heinig, Enrico Petretto, Chris Wallace, Leonardo Bottolo, Maxime Rotival, Han Lu, Yoyo Li, Rizwan Sarwar, Sarah R Langley, Anja Bauerfeind, et al. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature*, 467(7314):460–464, 2010.

[20] Chong Wu, Ellen W Demerath, James S Pankow, Jan Bressler, Myriam Fornage, Megan L Grove, Wei Chen, and Weihua Guan. Imputation of missing covariate values in epigenome-wide analysis of dna methylation data. *Epigenetics*, 11(2):132–139, 2016.

[21] Stephen B Baylin and Peter A Jones. A decade of exploring the cancer epigenome—biological and translational implications. *Nature Reviews Cancer*, 11(10):726–734, 2011.

[22] Devin C Koestler, Brock C Christensen, Margaret R Karagas, Carmen J Marsit, Scott M Langevin, Karl T Kelsey, John K Wiencke, and E Andres Houseman. Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics*, 8(8):816–826, 2013.

[23] Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31(2):142–147, 2013.

[24] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(1):587, 2010.

[25] Stef Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.

[26] Shelley Derksen and HJ Keselman. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2):265–282, 1992.

[27] Roderick JA Little. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296, 1988.

[28] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[29] Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558, 1994.

[30] Chong Wu, Jun Chen, Junghi Kim, and Wei Pan. An adaptive association test for microbiome data. *Genome Medicine*, 8(1):56, 2016.

[31] Human Microbiome Project Consortium et al. A framework for human microbiome research. *Nature*, 486(7402):215–221, 2012.

[32] David A Relman. The human microbiome and the future practice of medicine. *JAMA*, 314(11):1127–1128, 2015.

[33] Eran Segal, Claude B Sirlin, Clara Ooi, Adam S Adler, Jeremy Gollub, Xin Chen, Bryan K Chan, George R Matcuk, Christopher T Barry, Howard Y Chang, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nature Biotechnology*, 25(6):675–680, 2007.

[34] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, 2009.

[35] Jiyoung Ahn, Rashmi Sinha, Zhiheng Pei, Christine Dominianni, Jing Wu, Jianxin Shi, James J Goedert, Richard B Hayes, and Liying Yang. Human gut microbiome and risk of colorectal cancer. *Journal of the National Cancer Institute*, 105(24):1907–1911, 2013.

[36] Ben P Willing, Johan Dicksved, Jonas Halfvarson, Anders F Andersson, Marianna Lucio, Zongli Zheng, Gunnar Järnerot, Curt Tysk, Janet K Jansson, and Lars Engstrand. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology*, 139(6):1844–1854, 2010.

[37] Fredrik H Karlsson, Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed. Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99–103, 2013.

[38] Benjamin P Willing, Shannon L Russell, and B Brett Finlay. Shifting the balance: antibiotic effects on host–microbiota mutualism. *Nature Reviews Microbiology*, 9(4):233–243, 2011.

[39] Justin L Sonnenburg and Michael A Fischbach. Community health care: therapeutic opportunities in the human microbiome. *Science Translational Medicine*, 3(78):78ps12, 2011.

[40] Roger S Lasken. Genomic sequencing of uncultured microorganisms from single cells. *Nature Reviews Microbiology*, 10(9):631–640, 2012.

[41] Siddhartha Mandal, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, 26:27663, 2015.

[42] Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H

Parks, Courtney J Robinson, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, 2009.

[43] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, 2010.

[44] Wei Pan. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genetic Epidemiology*, 35(4):211–216, 2011.

[45] Catherine Lozupone and Rob Knight. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, 2005.

[46] Catherine A Lozupone, Micah Hamady, Scott T Kelley, and Rob Knight. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 73(5):1576–1585, 2007.

[47] Jun Chen, Kyle Bittinger, Emily S Charlson, Christian Hoffmann, James Lewis, Gary D Wu, Ronald G Collman, Frederic D Bushman, and Hongzhe Li. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28(16):2106–2113, 2012.

[48] Edward W Beals. Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. *Advances in Ecological Research*, 14(1):1–55, 1984.

[49] Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36(6):2605–2637, 2008.

[50] Wei Pan, Junghi Kim, Yiwei Zhang, Xiaotong Shen, and Peng Wei. A powerful and adaptive association test for rare variants. *Genetics*, 197(4):1081–1095, 2014.

[51] Wei Pan. Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genetic Epidemiology*, 33(6):497–507, 2009.

[52] Michael C Wu, Peter Kraft, Michael P Epstein, Deanne M Taylor, Stephen J Chanock, David J Hunter, and Xihong Lin. Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, 2010.

[53] Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. MEGAN analysis of metagenomic data. *Genome Research*, 17(3):377–386, 2007.

[54] Emily S Charlson, Jun Chen, Rebecca Custers-Allen, Kyle Bittinger, Hongzhe Li, Rohini Sinha, Jennifer Hwang, Frederic D Bushman, and Ronald G Collman. Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PloS One*, 5(12):e15216, 2010.

[55] Nicola Segata, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S Garrett, and Curtis Huttenhower. Metagenomic biomarker discovery and explanation. *Genome Biology*, 12(6):R60, 2011.

[56] Donovan H Parks, Gene W Tyson, Philip Hugenholtz, and Robert G Beiko. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics*, 30(21):3123–3124, 2014.

[57] Paul J McMurdie and Susan Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, 10(4):e1003531, 2014.

[58] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):1–21, 2014.

[59] Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12):1200–1202, 2013.

[60] Gary D Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A Keilbaugh, Meenakshi Bewtra, Dan Knights, William A Walters, Rob Knight,

et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.

[61] Janet GM Markle, Daniel N Frank, Steven Mortin-Toth, Charles E Robertson, Leah M Feazel, Ulrike Rolle-Kampczyk, Martin von Bergen, Kathy D McCoy, Andrew J Macpherson, and Jayne S Danska. Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science*, 339(6123):1084–1088, 2013.

[62] Daniel I Bolnick, Lisa K Snowberg, Philipp E Hirsch, Christian L Lauber, Brian Parks, Aldons J Lusis, Rob Knight, J Gregory Caporaso, and Richard Svanbäck. Individual diet has sex-dependent effects on vertebrate gut microbiota. *Nature Communications*, 5:4500, 2014.

[63] Andrew H Moeller, Patrick H Degnan, Anne E Pusey, Michael L Wilson, Beatrice H Hahn, and Howard Ochman. Chimpanzees and humans harbour compositionally similar gut enterotypes. *Nature Communications*, 3:1179, 2012.

[64] Christopher Quince, Anders Lanzén, Thomas P Curtis, Russell J Davenport, Neil Hall, Ian M Head, L Fiona Read, and William T Sloan. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, 6(9):639–641, 2009.

[65] Brendan J Kelly, Robert Gross, Kyle Bittinger, Scott Sherrill-Mix, James D Lewis, Ronald G Collman, Frederic D Bushman, and Hongzhe Li. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. *Bioinformatics*, 31(15):2461–2468, 2015.

[66] Wei Pan, Fang Han, and Xiaotong Shen. Test selection with application to detecting disease association with multiple snps. *Human Heredity*, 69(2):120–130, 2010.

[67] Peter McCullagh and John A Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

[68] Jelle J Goeman, Sara A Van De Geer, and Hans C Van Houwelingen. Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):477–493, 2006.

[69] Ping-Shou Zhong and Song Xi Chen. Tests for high-dimensional regression coefficients with factorial designs. *Journal of the American Statistical Association*, 106(493):260–274, 2011.

[70] Wei Lan, Hansheng Wang, and Chih-Ling Tsai. Testing covariates in high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 66(2):279–301, 2014.

[71] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[72] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[73] Arend Voorman, Ali Shojaie, and Daniela Witten. Inference in high dimensions with the penalized score test. *arXiv preprint arXiv:1401.2678*, 2014.

[74] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[75] Stephan Morgenthaler and William G Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1):28–56, 2007.

[76] T Tony Cai, Weidong Liu, and Yin Xia. Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):349–372, 2014.

[77] David Donoho and Jiashun Jin. Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science*, 30(1):1–25, 2015.

[78] Jiashun Jin and Tracy Ke. Rare and weak effects in large-scale inference: methods and phase diagrams. *Statistical Sinnica*, 26(1):1–34, 2016.

[79] Gongjun Xu, Lifeng Lin, Peng Wei, and Wei Pan. An adaptive two-sample test for high-dimensional means. *Biometrika*, 103(3):609–624, 2016.

[80] Song Xi Chen, Jun Li, and Ping-Shou Zhong. Two-sample tests for high dimensional means with thresholding and data transformation. *arXiv preprint arXiv:1410.2848*, 2014.

[81] Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

[82] Ludwig Fahrmeir and Heinz Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368, 1985.

[83] Xinyi Lin, Seunggeun Lee, David C Christiani, and Xihong Lin. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*, 14(4):667–681, 2013.

[84] Jianqing Fan, Rui Song, et al. Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

[85] BLS Prakasa Rao. Conditional independence, conditional mixing and conditional association. *Annals of the Institute of Statistical Mathematics*, 61(2):441–460, 2009.

[86] Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

[87] T Tony Cai, Zhao Ren, Harrison H Zhou, et al. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59, 2016.

[88] Clifford R Jack, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L Whitwell, Chadwick Ward, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

[89] Junghi Kim, Yiwei Zhang, and Wei Pan. Powerful and adaptive testing for multi-trait and multi-snp associations with GWAS and sequencing data. *Genetics*, 203(2):715–731, 2016.

[90] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45(12):1452–1458, 2013.

[91] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, et al. Next-generation genotype imputation service and methods. *Nature Genetics*, 48(10):1284–1287, 2016.

[92] Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(Database Issue):D355–D360, 2009.

[93] Colm O'Dushlaine, Lizzy Rossin, Phil H Lee, Laramie Duncan, Neelroop N Parikshak, Stephen Newhouse, Stephan Ripke, Benjamin M Neale, Shaun M Purcell, Danielle Posthuma, et al. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature Neuroscience*, 18(2):199–209, 2015.

[94] Perry G Ridge, Shubhabrata Mukherjee, Paul K Crane, John SK Kauwe, et al. Alzheimer's disease: analyzing the missing heritability. *PloS One*, 8(11):e79771, 2013.

[95] Xiaochuan Wang, Julie Blanchard, Inge Grundke-Iqbal, Jerzy Wegiel, Han-Xiang Deng, Teepu Siddique, and Khalid Iqbal. Alzheimer disease and amyotrophic lateral sclerosis: an etiopathogenic connection. *Acta Neuropathologica*, 127(2):243–256, 2014.

[96] Jun-ichi Satoh. Molecular network of microRNA targets in Alzheimer's disease brains. *Experimental Neurology*, 235(2):436–446, 2012.

[97] Chong Wu and Wei Pan. Integrating eqtl data with gwas summary statistics in pathway-based analysis with application to schizophrenia. *Genetic Epidemiology*, 42(3):303–316, 2018.

[98] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[99] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

[100] Kai Wang, Mingyao Li, and Maja Bucan. Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, 81(6):1278–1283, 2007.

[101] Wei Pan, Il-Youp Kwak, and Peng Wei. A powerful pathway-based adaptive test for genetic association with common or rare variants. *The American Journal of Human Genetics*, 97(1):86–98, 2015.

[102] Andrew Bakshi, Zhihong Zhu, Anna AE Vinkhuyzen, W David Hill, Allan F McRae, Peter M Visscher, and Jian Yang. Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Scientific Reports*, 6:32894, 2016.

[103] Jin Li, Zhi Wei, Xiao Chang, Christopher J Cardinale, Cecilia E Kim, Robert N Baldassano, Hakon Hakonarson, International IBD Genetics Consortium, et al.

Pathway-based genome-wide association studies reveal the association between growth factor activity and inflammatory bowel disease. *Inflammatory Bowel Diseases*, 22(7):1540–1551, 2016.

[104] Lie Li, Xinlei Wang, Guanghua Xiao, and Adi Gazdar. Integrative gene set enrichment analysis utilizing isoform-specific expression. *Genetic Epidemiology*, 41:498–510, 2017.

[105] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[106] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009.

[107] Jelle J Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

[108] Jari Tiihonen, Jouko Lönnqvist, Kristian Wahlbeck, Timo Klaukka, Leo Niskanen, Antti Tanskanen, and Jari Haukka. 11-year follow-up of mortality in patients with schizophrenia: a population-based cohort study (FIN11 study). *The Lancet*, 374(9690):620–627, 2009.

[109] Patrick F Sullivan, Mark J Daly, and Michael O'donovan. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics*, 13(8):537–551, 2012.

[110] Stephan Ripke, Benjamin M Neale, Aiden Corvin, James TR Walters, Kai-How Farh, Peter A Holmans, Phil Lee, Brendan Bulik-Sullivan, David A Collier, Hailiang Huang, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014.

[111] Stephan Ripke, Colm O'Dushlaine, Kimberly Chambert, Jennifer L Moran, Anna K Kähler, Susanne Akterin, Sarah E Bergen, Ann L Collins, James J Crowley, Menachem Fromer, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*, 45(10):1150–1159, 2013.

[112] Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[113] Gene Ontology Consortium et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(Database Issue):D258–D261, 2004.

[114] Bogdan Pasaniuc, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev, Joseph Pickrell, Joel Hirschhorn, David P Strachan, Nick Patterson, and Alkes L Price. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20):2906–2914, 2014.

[115] Il-Youp Kwak and Wei Pan. Adaptive gene-and pathway-trait association testing with GWAS summary statistics. *Bioinformatics*, 32(8):1178–1184, 2016.

[116] Alexander Gusev, Nick Mancuso, Hilary K Finucane, Yakir Reshef, Lingyun Song, Alexias Safi, Edwin Oh, Steven McCaroll, Benjamin Neale, Roel Ophoff, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature genetics*, 50(4):538–548, 2018.

[117] Andriy Derkach, Jerry F Lawless, Lei Sun, et al. Pooled association tests for rare genetic variants: a review and some new results. *Statistical Science*, 29(2):302–321, 2014.

[118] Il-Youp Kwak and Wei Pan. Adaptive gene-and pathway-trait association testing with GWAS summary statistics. *Bioinformatics*, 32(8):1178–1184, 2015.

[119] Kunlin Zhang, Sijia Cui, Suhua Chang, Liuyan Zhang, and Jing Wang. 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 Research*, 38(Web Server Issue):W90–W95, 2010.

[120] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(Database Issue):D896–D901, 2017.

[121] Psychiatric Genomics Consortium Cross-Disorder Group. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*, 381(9875):1371–1379, 2013.

[122] X Ma, W Deng, X Liu, M Li, Z Chen, Z He, Y Wang, Q Wang, X Hu, DA Collier, et al. A genome-wide association study for quantitative traits in schizophrenia in China. *Genes, Brain and Behavior*, 10(7):734–739, 2011.

[123] Sandra K Loo, Corina Shtir, Alysa E Doyle, Eric Mick, James J McGough, James McCracken, Joseph Biederman, Susan L Smalley, Rita M Cantor, Stephen V Faraone, et al. Genome-wide association study of intelligence: additive effects of novel brain expressed genes. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(4):432–440, 2012.

[124] Paweł Kapelski, Maria Skibińska, Małgorzata Maciukiewicz, Dorota Zaremba, Maria Jasiak, and Joanna Hauser. Family association study of transforming growth factor beta1 gene polymorphisms in schizophrenia. *Psychiatria Polska*, 50(4):761–770, 2016.

[125] Yong-Ku Kim, Aye-Mu Myint, Bun-Hee Lee, Chang-Su Han, Heon-Jeong Lee, Dae-Jin Kim, and Brian E Leonard. Th1, th2 and th3 cytokine alteration in schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 28(7):1129–1134, 2004.

[126] Jeremy Hall, Simon Trent, Kerrie L Thomas, Michael C ODonovan, and Michael J Owen. Genetic risk for schizophrenia: convergence on synaptic pathways involved in plasticity. *Biological Psychiatry*, 77(1):52–58, 2015.

[127] Jing Qin Wu, Melissa J Green, Erin J Gardiner, Paul A Tooney, Rodney J Scott, Vaughan J Carr, and Murray J Cairns. Altered neural signaling and immune

pathways in peripheral blood mononuclear cells of schizophrenia patients with cognitive impairment: A transcriptome analysis. *Brain, Behavior, and Immunity*, 53:194–206, 2016.

[128] Danielle M Santarelli, Natalie J Beveridge, Paul A Tooney, and Murray J Cairns. Upregulation of dicer and microrna expression in the dorsolateral prefrontal cortex brodmann area 46 in schizophrenia. *Biological Psychiatry*, 69(2):180–187, 2011.

[129] Uwe Rudolph and Hanns Möhler. Gabaa receptor subtypes: Therapeutic potential in down syndrome, affective disorders, schizophrenia, and autism. *Annual Review of Pharmacology and Toxicology*, 54:483–507, 2014.

[130] Zhiyuan Xu, Chong Wu, Wei Pan, Alzheimer's Disease Neuroimaging Initiative, et al. Imaging-wide association study: Integrating imaging endophenotypes in GWAS. *Neuroimage*, 159:159–169, 2017.

[131] Michael Wainberg, Nasa Sinnott-Armstrong, David Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, Johan LM Bjorkegren, Manuel A Rivas, et al. Vulnerabilities of transcriptome-wide association studies. *bioRxiv*, 206961, 2017.

[132] Fernando S Goes, John McGrath, Dimitrios Avramopoulos, Paula Wolyniec, Mehdi Pirooznia, Ingo Ruczinski, Gerald Nestadt, Eimear E Kenny, Vladimir Vacic, Inga Peters, et al. Genome-wide association study of schizophrenia in Ashkenazi Jews. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 168(8):649–659, 2015.