Statistical tools for the multi-omics analysis of microbiome data

Angela Zhang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Michael Wu, Chair

Wei Sun

Timothy Thornton

Yue Cui

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

Statistical tools for the multi-omics analysis of
microbiome data

Angela Zhang

Chair of the Supervisory Committee:
Michael Wu
Department of Biostatistics

The human microbiome consists of trillions of bacteria, archaea, and viruses that exist on virtually every organ in the body. The microbiome plays a fundamental role in human health and has been implicated in several different diseases and conditions such as cardiovascular disease and certain cancers. Understanding the functional role of the microbiome can lead to increased understanding of these complex diseases and result in the development of more effective treatments. Although advances in technology have allowed for the inexpensive processing and analysis of high-throughput data, several statistical challenges exist in the analysis of microbiome data.

In my dissertation, I will present three projects that address the statistical challenges of high-dimensionality, multi-omics data integration, batch effects/other covariate adjustment, and the visualization of microbiome data. In Project 1, We address the issues of high-dimensionality and data integration by proposing a new procedure for testing the cumulative metabolic effect of the microbiome using a weighted variance component test framework. In this setup, we focus on metabolic pathways and recognize that metabolism can be represented by metagenomics (metabolic potential) and metabolomics (metabolic output). In Project 2, we address the issue of batch effects and high-dimensionality by outlining a two-step adjustment of the principal coordinates (PCs) of the microbial taxa data. In the first

step, we project the mean effect of the unwanted covariates out of the PCs. In the second step, we adjust out the second moment of the same covariates from the PCs by assuming a linear relationship between the covariates and the variance of the PCs. Finally, in Project 3, we propose an effect modification testing procedure for evaluating interactions between microbial taxa and environmental factors on an outcome of interest. We address concerns of data integration and high-dimensionality by using a variance component test framework with LASSO-selected variables to assess the effect modification of the microbiome on environmental variables.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Mike Wu, for mentoring me during my dissertation. In addition to strengthening my skills as a biostatistician, our weekly meetings provided an opportunity for Mike and I to check in with each other and discuss topics ranging from our experiences growing up as second generation Chinese Americans to strategies for employing modern lingo in his course curriculum. I will always cherish the conversations we had during these meetings; it was daunting to be a student in a top-ranked program and I often questioned my place among so many talented and motivated colleagues. Mike, in his typical tough love and relaxed attitude, bolstered my confidence in this program. His continuous guidance, support and enthusiasm carried me through all aspects of my PhD studies.

I also want to thank Julia Yue Cui. I'm incredibly grateful to have worked with her since the beginning of my graduate studies. I feel like I started with her when she was just starting out in academia and I am honored to have witnessed the tremendous and awe-inspiring progress she has made in her academic career. She is an inspiration to me. It was both joyous and insightful working with her. Her passion for her academic field encouraged me to seek out projects that stimulated my own interests and led me to my niche in the field of microbiome research. My time in her lab gave me a firsthand perspective on developing methods that directly addressed the statistical needs of microbiome researchers.

I am grateful for Wei Sun, Tim Thornton, and Neelendu Dey for agreeing to be on my thesis committee. Your questions and suggestions during our meetings allowed me to reach my full potential during my dissertation.

I would also like to thank the students. In addition to all the fun I had with all of them, it was gratifying that we were always open with our struggles in graduate school.

We never hesitated to help each other with coursework and roadblocks in our research. My involvement in department committees such as the Admissions Committee, the Equity, Diversity and Inclusion Committee and the Biostatistics Activities and Events Squad allowed me to realize my passion in connecting people and improving my community both in the field of statistics and beyond. I formed meaningful connections with you all and I know I have made lifelong friends (and potential collaborators).

Lastly, I want to thank my friends and family for their neverending support during my time in Seattle. Their confidence in me was always ironclad even when I had my own doubts. I'm especially grateful to Asher, who cheered me on and was always by side these past five years. Despite his unfamiliarity with statistics, he was always enthusiastic to hear about my projects and my day-to-day life as a graduate student. More importantly, his encyclopedic knowledge of synonyms and artistic finessing of words definitely added sophistication to my manuscripts and other papers.

Graduate school can feel awfully isolating, especially during difficult times, but it's always helpful for me to take a step back and acknowledge the wonderful support system I had throughout my five years in the program. I know very deeply that the completion of my dissertation depended on the love and encouragement I received from all of you.

DEDICATION

To all the loved ones who help me reach my full potential

Chapter 1

INTRODUCTION

1.1 Importance of the microbiome in human health

The microbiome consists of trillions of bacteria, fungi, and viruses and inhabits virtually every part of the human body [18]. In addition to its immense size, the microbiome is also incredibly diverse; the Human Microbiome Project estimates that 500 to 1000 species reside in our microbiome at any given time [76]. Previous literature, based on "back of the envelope calculations", estimated a tenfold ratio between microbial cells and host cells in the human body. However, current estimates, generated from more updated information, has projected a 1:1 ratio instead [66]. Despite this decrease, the biological importance of the microbiome remains. Each microbial species contains thousands of genes, allowing the total metagenome, i.e. the combined genomes of all microbial species, to have increased flexibility compared to the human genome. A comprehensive meta-analysis, covering more than 3,500 samples from 13 studies, reported almost 46 million non-redundant microbial genes in the gut and oral microbiome alone [73].

Many factors influence the composition of the microbiome. For example, urban and rural individuals have significant differences in microbial compositions despite similarities in species richness between these two groups [70]. Furthermore, environmental degradation caused by increased rapid urbanization has been hypothesized to increase the prevalence of non-communicable diseases, such as allergic diseases and skin conditions, due to dysbiosis of the skin microbiome [59]. Diet also plays a pivotal role in the composition of the gut microbiome. A diet high in fat and lower in fiber can result in depleted levels of short chain fatty acids (SCFAs) in the gut microbiome. Decreased levels of SCFAs can subsequently lead to chronic inflammation that increases the risk for certain cancers [61]. The composition of

the microbiome can change dramatically over time. Infants, whose diet is primarily composed of breast milk or formula, have microbiomes dominated by species like *Bifidobacterium spp.* that possess the ability to metabolize oligosaccharides. As the infant's diet becomes more complex, the diversity of the microbiome increases and more closely represents an adult microbiome [38].

The previous finding exemplifies that although certain factors, such as host genome, can determine its initial composition, the microbiome is a dynamic organ. Environmental and lifestyle changes can lead to compositional changes to the microbiome, some of which are clinically significant. Inflammatory bowel disease (IBD), consisting of Crohn's disease and ulcerative colitus, is perhaps the most well-studied disease connected to the microbiome. In particular, patients with IBD have fewer species in their microbiome and have a decrease of SCFAs [34]. In addition to IBD, certain cancers have also been linked to the microbiome. *Helicobacter pylori*, a well-documented bacteria directly connected to the formation of stomach ulcers, has been associated with gastric cancers. Additionally, bacteria such as *Salmonella typhi* and *Helicobacter spp.* have been linked to biliary cancer. The aforementioned bacterial species are thought to create a chronic inflammatory state that leads to carcinogenesis [25].

As demonstrated by its role in human diseases and conditions, dysbiosis of the microbiome can lead to far-reaching consequences in the human body. Current research has provided strong evidence that the gut microbiome affects not only proximal organs such as the liver but also distal organs like the brain [10, 75]. With regards to the gut-liver axis, microbially derived metabolites from the large intestine, notably SCFAs and secondary bile acids (BAs), enter the liver through the portal vein via enterohepatic circulation. High-fat diets can lead to an increase of pro-inflammatory secondary BAs that are metabolized from conjugated primary BAs and are linked to chronic inflammation of the colon. High-fiber diets, on the other hand, can result in an increase of anti-inflammatory SCFAs that are hypothesized to have anti-cancer properties as well [91]. The microbiome plays a profound role in the host nervous system as well. The healthy development and maturation of both the enteric and central nervous system hinge upon the bacterial colonization of gut through the maintained

protection of the intestinal barrier [6]. Lastly, in addition to the microbiome's role in each individual organ, the gut-liver axis can interact with the gut-brain axis as well. With respect to the gut-brain-liver axis, compositional changes in the microbiome as well as increased expression of inflammation genes in the liver transcriptome were linked to an increased susceptibility of Alzheimer's disease in mouse disease models [94].

1.2 Data complexities in the analysis of microbiome data

1.2.1 Motivation

So far we have demonstrated the complexity of the human microbiome, as well as its overarching importance in human health. Increased understanding of this system has great potential in the development of improved treatments for many human diseases and conditions. As one of the first of its kind, the Human Microbiome Project characterized the microbiomes of the skin, nose, mouth, gastrointestinal tract, and urogenital regions of 300 healthy individuals [76]. Since then, innovations in sequencing technologies have allowed for a deeper and more inexpensive analysis of the microbiome. Metagenomic data can be characterized by both taxonomic diversity and functional diversity. Taxonomic diversity refers to the operational taxonomic units (OTUs) found in the samples; OTUs are genetically similar clusters that are the standard unit of analysis in 16S rRNA sequencing data [1]. Functional diversity, on the other hand, refers to the gene and pathway content of the microbiome. Regarded as an improvement from 16S rRNA sequencing, shotgun sequencing can provide information regarding protein and pathway activity in the microbiome. In addition to metagenomic data, recent studies have also begun to collect metabolomics and proteomics data in order to better understand the dynamic chemical processes that occur within the microbiome [28].

Despite the popularity and importance of microbiome studies, analyzing microbiome data poses several unique challenges. There are three main motivations for a multi-omics analysis of microbiome data. The first is to understand the relationships between different

[1] In the context of this dissertation, we will use species, taxa, and OTUs interchangeably.

-omics data sets. For example, it may be of interest to determine which bacterial species are associated with an increase in the production of SCFAs in order to develop probiotics with anti-inflammatory properties. The second objective is to determine if the cumulative effects of the -omics data are associated with a specific outcome of interest, such as disease status. Metabolism, which is the process of chemical reactions in the body, can be represented by both metagenomics (metabolic potential) and metabolomics (metabolic output). We may have increased performance if we consider both types of data in our analysis. Lastly, the final rationale for developing integrative analysis techniques is to better understand the effect modification of the microbiome on host and environmental characteristics. The microbiome does not act as an isolated ecosystem; interactions with its host and environment can lead to fundamental changes to both systems [56]. Diseases are multi-faceted and it is often of interest to determine its underlying and interconnecting mechanisms from both the host and the microbiome.

Analyzing microbiome data can be difficult due its high-dimensional nature; each person may have up to thousands of unique OTUs and millions of unique genes within their microbiota. Typical statistical methods are either too computationally intensive or not even applicable in cases when the number of species or genes exceed the number of individuals in the study. Adding to this difficulty, microbiome data is also sparse and compositional. Concerning sparsity, the makeup of the microbiome is highly variable between samples and can essentially can serve as an additional fingerprint for an individual [16]. Despite its sheer number, many species are rare and only found in a handful of individuals, resulting in a large proportion of zeros in the data. With regards to compositionality, the total number of microbial reads in a sample are determined arbitrarily by the capacity of the high-throughput sequencing machine. Consequently, sequencing counts do not represent an absolute abundance of any one species but rather its proportion. Methods developed for host gene expression data, such as edgeR and DESeq2 are often inappropriate for microbiome data and can lead to inflated false positive rates [19, 24].

Although statistical inference, driven primarily by hypothesis testing, is vital in the

analysis of microbiome data, many scientists are also motivated by data-driven approaches. Data visualization is one such application for exploratory approaches that can be used to generate hypotheses for future studies. The high-dimensional nature of microbiome data, as mentioned previously, can make it difficult for data visualization. Bivariate plots, such as grouped box plots, can be used to explore the association between a clinical variable of interest and the abundance of a single species, although this quickly becomes infeasible for hundreds of species. Instead of a species-by-species approach, scientists instead rely on summary statistics, namely alpha diversity and beta diversity, to visualize changes that occur in the microbiome. Species richness, a metric of alpha diversity, refers to the number of unique species in an unit. Beta diversity, on the other hand, is focused on the pairwise differences between microbiomes. Plots from Principal Coordinates Analysis (PCoA) utilize beta diversity measures to distinguish compositional differences between groups.

1.2.2 Previous methods and specific challenges

Taken together, careful considerations need to be made in analyzing and visualizing microbiome data. In the last section, we outlined several motivations for the development of statistical tools for the multi-omics analysis of microbiome data. In this section, I will summarize and identify weaknesses in current methods for analyzing metagenomic and metabolomic data.

Pathway-based approaches can serve as a potential solution for addressing high-dimensionality in metabolomic data. In short, we can reduce the dimensionality of the problem by combining metabolites into pathways. Methods for pathway analysis can be divided into three generations and are mostly adapted from gene-based approaches. Over-representation analysis (ORA) is perhaps the most widely-used method from the first generation. ORA determines which pathways are over-represented in a group of metabolites that have been determined by a pre-defined criteria, i.e. metabolites with p-values < 0.05 from a specific hypothesis test. A significant p-value, obtained through the Fisher's exact test, suggests that this pathway may be dysfunctionally regulated [51]. Functional class scoring (FCS) represents the second

generation of pathway-based methods and can be seen as an extension of ORA that uses data from all metabolites instead of the pre-determined subset. Gene set enrichment analysis (GSEA) is the most popular method in this class [72] and the most implemented method among all three generations. Finally, topology-based approaches comprise the third generation of pathway-based approaches and utilize correlations between metabolites to determine networks of metabolites perturbed by an outcome of interest.

It is important to note that the previously described methods all utilize the competitive null hypothesis and may not fully address our main objective of detecting metabolic differences between phenotypes. Instead, we consider popular methods, like variance component and kernel-based approaches, that utilize the self-contained null hypothesis. Similar to competitive null methods, variance component (VC) methods for pathway analysis were originally developed for gene variant association testing. Sequence Kernel Association Test (SKAT), is by far the most popular method and combines both the vC and kernel framework in its approach. Consider the following model:

$$y_i = \alpha_0 + \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{G}_i\boldsymbol{\beta} + \epsilon_i \tag{1.1}$$

y_i is our outcome of interest and α_0 and $\boldsymbol{\alpha}$ are coefficients for the intercept and variables ($\boldsymbol{X}_i$) we want to adjust out. $\boldsymbol{G}_i$ is set of metabolites in a specific pathway. ϵ_i is the error term and follows a standard normal distribution. Within the context of metabolites, SKAT assumes that each metabolite within a pathway follows an arbitrary distribution with mean 0 and variance τ. Hence, the null hypothesis of testing $\tau = 0$ is equivalent to testing $\boldsymbol{\beta} = 0$. The following test statistic, Q, is then:

$$Q = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{G}\boldsymbol{G}^\top (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \tag{1.2}$$

Here, $\hat{\boldsymbol{\mu}}$ is estimated under the null model. Q asymptotically follows a mixture of χ^2_1 distributions. A significant p-value signifies that the pathway, as represented by its metabolites, is associated with the outcome of interest. We can extend this VC framework by consider-

ing kernel functions that model non-linear effects between the metabolites and the outcome [85]. Kernel functions, $\boldsymbol{K}(i, i')$, measure the similarity between subjects i and i' based on their metabolites. We can replace $\boldsymbol{G}\boldsymbol{G}^\top$, known as the linear kernel, by any general kernel function in the test statistic:

$$Q = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{K} (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \tag{1.3}$$

Up to this point, we have outlined methods for pathway analysis that only utilize one data set. The popularity and availability of multi-omics data highly motivates the development of data integration techniques. While there are a plethora of methods for microbiome data integration, such as marginal correlation analysis and regression-based approaches, they do not test if multiple data sets are associated with an outcome. As stated previously, metabolic pathways can be represented by both the metabolome and the metagenome. We may have greater statistical power if we can test the cumultative effects of metabolism, as represented by the metabolome and the metagenome, than if we test each data set individually.

Data integration can also be interpreted as the effect modification of the microbiome on a host or environmental trait. For example, it was found that a HIV pre-exposure prophylaxis (PrEP) treatment was not effective in women who had a specific vaginal microbiome composition. Further investigations revealed that a specific organism within the vagina had the ability to metabolize the drug, thus neutralizing its effects [79]. Understanding the role of the microbiome on host mechanisms can lead to the development of more effective treatments. We can model the interaction effects in the following manner:

$$y_i = \boldsymbol{X}_i \boldsymbol{\alpha} + \boldsymbol{G}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{\gamma} + \epsilon_i \tag{1.4}$$

Specifically, $\boldsymbol{\alpha}$ are the coefficients for the host or environmental variables, i.e. PrEP treatment, $\boldsymbol{\beta}$ are the coefficients for microbial taxa, $\boldsymbol{\gamma}$ are the coefficients for the interaction between the microbiome and host. While VC approaches can be applied to this model, careful considerations need to be made when modeling the main effects. Since $\boldsymbol{G}$ is often

high-dimensional, ordinary least squares (OLS) based methods are not applicable and ridge regression based methods are subject to highly inflated type 1 error when the effect size of individual taxa are large.

Lastly, we will discuss the effects of potentially confounding factors, such as batch effects, on obfuscating the true relationship between the microbiome and an outcome. Batch effects are common across all types of -omics study and are derived from technical variation (e.g personnel, time of day) not related to the study. Instead of obtaining results that show contrasts between phenotypes, we may instead see that differences are driven primarily by batch effects. Subsequently, batch effects can lead to inflated type 2 error when testing for an association between the microbiome and an outcome of interest. Several methods proposed for batch correction in microbiome data are extensions of microarray-based approaches. One such method, ComBat, utilizes both parametric and empirical Bayes methods to correct for batch effects in microarray studies [30]. Although ComBat can be applied to microbiome data, it is has been found to only partially reduce batch effects, especially if the batch effects are not independent from the biological signal [17]. Furthermore, ComBat assumes Gaussian linear models in its framework, which may be too simplistic to model the complex distributional traits of microbiome data. Additionally, many methods for batch effect correction are primarily developed for association testing and there is a dearth of methods that adjust for batch effects and other covariates in the visualization of microbiome data.

1.3 Organization of thesis

In this dissertation, we will present statistical methods aimed for the multi-omics analysis of microbiome data. These tools will primarily address four key challenges: 1) data integration of -omics data sets; 2) high-dimensionality; 3) batch effects and other potential confounding factors and 4) visualization of microbiome data.

We first address the challenges of high-dimensionality and data integration in Chapter 2. Innovations in high-throughput technology now allow for the rapid profiling of the metabolome and metagenome – the gene content of the bacteria – for characterizing micro-

bial metabolism. Due to small sample sizes and high-dimensionality of the data, pathway analysis of metabolomic data represents the standard for metagenomic analysis. We propose a weighted variance component framework to test if the joint effect of genes and metabolites in a metabolic pathway is associated with an outcome of interest. This setup allows us to represent multiple aspects of metabolism via the metabolic potential (metagenomics) and the metabolic output (metabolomics). This approach allows for analytic p-value calculation, correlation between data types, and optimal weighting.

In Chapter 3, we address the problem of adjusting for unwanted covariates in the visualization of high-dimensional data. Non-Euclidean distance measures, like the Bray-Curtis dissimilarity measure, can be used to quantify and visualize the relationship between samples based on high-dimensional data, such as microbiome data. Unwanted covariates, such as batch effects, can obscure the true signal of the experiment, e.g. overall compositional differences between control and treatment groups. Here, we propose a two-step process for adjusting the mean and variance of both continuous and categorical confounders from the principal coordinates of microbiome data. Using multiple real data sets, we apply this approach to various methods such as principal coordinates analysis (PCoA), hierarchical clustering, and permutational multivariate analysis of variance (PERMANOVA). Furthermore, we extend our two-step process to existing methods for evaluating multi-omics data (e.g. metabolomics and metagenomics).

We propose an effect modification testing procedure to address the challenge of evaluating interactions between microbial taxa and environmental factors on an outcome of interest in Chapter 4. Variance component test methods have typically been proposed for assessing interaction effects involving high-dimensional models. Many existing methods, however, tend to give inflated type I errors due to challenges in estimating the null model with high-dimensional main effects. Least squares based approaches fails especially when the number of microbial species exceed the number of samples, and ridge regression based approaches give biased estimates due to the bias-variance tradeoff. Instead, we can apply the typical variance component framework by considering only a subset of the variables obtained through LASSO

variable selection. We will also employ more complex kernels, such several microbiome-specific kernels, in evaluating this procedure.

Lastly, we will provide a dissertation summary and future directions in Chapter 5. The field of microbiome research is growing exponentially, thus increasing the demand for the development of statistical methods for more complex studies. We will briefly outline motivations for method development in microbiome genome-wide association studies (mbGWAS), longitudinal studies and prediction.

Chapter 2

TESTING PROCEDURE FOR THE JOINT ANALYSIS OF MULTIOMICS DATA

2.1 Introduction

Metabolism represents a central biological component that underlies a wide range of processes in cells and broader organisms. Altered metabolism and metabolic pathways represent a hallmark of many diseases and health conditions include diabetes, inflammatory bowel diseases, and certain cancers, among countless others [31, 3, 13]. Consequently, many studies across a range of high-throughput technologies are all focused on understanding the relationship between metabolic processes and a specific outcome of interest. Such studies include everything from genetic associations studies and gene expression profiling studies to metagenomic studies and metabolomic profiling studies. However, each of these different data types reflects just one aspect of metabolism and we may be able to achieve greater power if we instead consider the cumulative effect of metabolism.

Dysbiosis of the microbiome has been linked to a variety of diseases and conditions in humans. The composition of metagenome - the gene content for the community of microbes - can reveal host-microbe interactions that play important roles in host metabolism. The relationship between the host and the microbiome are mainly governed by the interactions of microbially-produced metabolites. In the context of metabolism, the metagenome can be defined as the metabolic potential while the metabolome represents the actual metabolic output. Advances in technology have allowed for the inexpensive processing and analysis of high-throughput data in microbiome studies. As a result, many studies have started collecting both metagenome and metabolomic data with the objective of relating metabolic pathways to outcomes.

Due to small sample sizes and high-dimensionality of metabolomics and metagenomics data, pathway analysis (wherein the effect of multiple genes or metabolites on an outcome is cumulatively assessed) of metabolomics data is commonly conducted and also represents the standard for metagenomics analysis. In addition to reducing the dimensionality of the data, there are several other advantages to taking a pathway approach as opposed to a gene-by-gene approach. In short, pathways are not generally driven by a single gene. In terms of biological relevance, an individual gene with a high level of differential expression may not be as revealing than a group of genes with moderate differential expression.

Several methods have been proposed for pathway-based analyses in metagenomics and metabolomics studies. Simple linear regression models can be used to represent the relationship between the metagenomic pathway and an outcome of interest. Methods for pathway analysis in metabolomic data have largely been adapted from gene-based approaches. One widely used example is over-representation analysis (ORA); in ORA, a list of over-represented pathways is generated from a subset of metabolites (e.g. determined to be statistically significant by some metric) [20]. As an extension of ORA, second and third generations of methods have also been developed for metabolomic pathway analysis. Originally designed for microarray analysis, functional class scoring (FCS) represents the second generation of tools and takes an *a priori* set of metabolites (e.g. all metabolites in a specific pathway) and compares it to an ordered list of metabolites (e.g. all statistically significant metabolites corrected for multiple testing and ranked by log-fold change) to determine whether the *a priori* group of genes are randomly distributed within the list [**?**]. Network-based analyses make up the third generation and have been shown to be superior to the previous methods since they consider metabolite interactions; topology-based approaches compare known cellular processes and pathways to a network of metabolite to determine if such pathways are altered in a disease condition.

We note that the aforementioned methods all utilize the competitive null hypothesis. There are some major caveats to this setup since it treats metabolites as a sampling unit instead of individuals. Typically, our main objective to detect changes across phenotypes,

which the competitive null hypothesis may not sufficiently address [51]. Several methods have been developed under the self-contained null hypothesis, which assumes that no metabolites are associated with a phenotype. Averaging-based approaches such as principal components analysis (PCA)-based methods have been used to select metabolites that capture the most variation in the data set. [89, 14]. Although these methods help overcome high-dimensionaltiy in microbiome data, PCA approaches suffer from lack of interpretability and lack of consensus on the necessary number of loadings to use in the analysis. Variance component- and kernel-based approaches have also been proposed for metabolomic pathway analysis. Instead of focusing on individual metabolites, these methods test if a group of metabolites, such as those found in a specific pathway, are associated with an outcome [92]. All these methods suffer an important weakness and are not suitable for multi-omics analysis since they only use one data set. In the current context of microbiome research, researchers are interested in relating outcomes to multiple data sets that collectively represent different aspects of metabolism.

Several methods have been proposed for an integrative approach to microbiome data analysis [29]. The most popular strategy is to perform a marginal correlation analysis between two biological features. The relationship between these two data sets can be quantified by some pre-specified correlation measure (e.g. Pearson's correlation, Spearman Rank Correlation) [23]. Although not as popular as marginal correlation analyses, regression-based methods have also been suggested as a potential strategy for elucidating the relationships between two features. This integrative strategy can be framed as a regression model where one feature type, such as the pathway abundance of the metagenomic data, is defined as the response variable and the other feature type, such as the metabolite levels associated with that pathway, can act as the predictor variables. These methods are not limited to only one predictor variable, although sparsity and rank constraints must be employed for multiple response variables [32]. One major limitation to regression-based approach is that they require that one feature type be defined as the predictor and the other as the response variable. This can pose a problem, especially when the underlying biology is not well understood.

Furthermore, both marginal correlation-based and regression-based approaches only assess the relationship between feature types. They do not test if an outcome is associated with multiple data sets. For example, these methods are not applicable if we are interested in the relationship between metabolic pathways, represented both metagenomics and metabolomics data, and disease type.

To address this question, we propose a weighted variance component (VC) framework to test if the joint effect of genes and metabolites in a metabolic pathway is associated with an outcome of interest. This setup allows us to represent multiple aspects of metabolism via the metabolic potential (metagenomics) and the metabolic output (metabolomics). Since the contribution of each data set to the outcome is unknown, we treat our test statistic as a weighted average of the metagenomics score test statistic and metabolomics score test statistic. We perform a grid search on this weighted sum and introduce two potential methods for combination testing for the vector of p-values generated from the grid search. This approach allows for analytic p-value calculation, correlation between data types, and optimal weighting.

The remainder of this chapter is organized as follows. In the next section, we introduce notation and present our weighted testing procedure by describing the test statistic and the asymptotic null distribution. Then in Section 3, we illustrate the advantages of our proposed approach through simulations and demonstrate the increase in power when considering both representations of metabolism. In Section 4, we show the utility of the weighted VC framework with data from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) by discerning the subset of metabolic pathways significantly associated with Type 2 diabetes (T2D). We conclude this chapter with a brief discussion in Section 5.

2.2 Weighted variance component test

2.2.1 Model

Suppose there are n subjects that have both metabolomic and metagenomic data for a specific pathway. Let y_i be a continuous or dichotomous trait of subject i. We consider the following regression model:

$$g[(Y_i)] = \alpha_0 + X_i\boldsymbol{\alpha} + W_i\boldsymbol{\beta} + \gamma z_i \tag{2.1}$$

$g(\cdot)$ is a link function that is the identity function for continuous outcomes and the logistic function for dichotomous outcomes. $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)$ are regression coefficients for covariates, $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, that we want to adjust out, such as age and sex. $W_i = (w_{i1}, w_{i2}, \ldots, w_{im})$ is vector of metabolite levels and z_i is the metagenome abundance for a specific pathway. $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)$ and γ are the regression coefficients for the m metabolites and metagenome abundances respectively.

It is possible for Z_i and W_i to be correlated. For example, if we have a large value for z_i, we could expect high levels of metabolites for that particular pathway. We can linearly transform Equation (2.2) as:

$$g[(Y_i)] = \alpha_0 + X_i\boldsymbol{\alpha} + W_i'\boldsymbol{\beta} + \gamma z_i \tag{2.2}$$

$\boldsymbol{W'} = (W_1', W_2', \ldots, W_n') = (\mathbf{I} - \boldsymbol{M})\boldsymbol{W}$ and $\boldsymbol{M} = \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$ where $\boldsymbol{M}$ is a projection matrix onto the column space of $\boldsymbol{Z}$.

2.2.2 Weighted variance component test

Evaluating if the joint effect of the metabolome and the metagenome has an effect on the outcome corresponds to testing the null hypothesis: $H_0 : \beta_j = \gamma = 0; j = 1, \ldots, m$. However, this framework assumes independence between the metagenomic and metabolomic datasets and is underpowered for pathways with many metabolites. Instead, suppose that $\boldsymbol{\beta}$ is a

random variable where each β_j is independent and has mean 0 and variance $\rho\tau$. Additionally, we suppose that γ is a random variable with mean 0 and variance $(1-\rho)\tau$. We can see that testing $H_0 : \tau = 0$ is equivalent to $H_0 : \beta_j = \gamma = 0$. We can use the weighted variance component test for this hypothesis.

Let $\boldsymbol{W'}$ be a $n \times k$ matrix of the uncorrelated metabolite values and let $\boldsymbol{Z}$ be a n-length vector of the pathway counts. Furthermore, let $0 \leq \rho \leq 1$. For each ρ, we can write the score test statistic as a weighted average of the metabolic potential (metagenome) and the metabolic output (metabolome).

$$Q_\rho = \rho(Y - \hat{\boldsymbol{\mu}})^\top \mathbf{z}\mathbf{z}^\top (Y - \hat{\boldsymbol{\mu}}) + (1-\rho)(Y - \hat{\boldsymbol{\mu}})^\top \boldsymbol{W'}\boldsymbol{W'}^\top (Y - \hat{\boldsymbol{\mu}}) \tag{2.3}$$

$$Q_\rho = \rho Q_{\text{genom}} + (1-\rho) Q_{\text{metab}} \qquad 0 \leq \rho \leq 1 \tag{2.4}$$

In addition to being computationally advantageous, the score test is preferable over the likelihood ratio (LR) test because it only requires fitting the null model; the LR distribution can be unstable, especially when the estimate is near the boundary. Furthermore, p-degrees-of-freedom LR tests can be less powerful when the number of taxa increases and are not even applicable when the number of species exceeds the sample size.

2.2.3 Minimum p-value method

We can see that when $\rho = 1$, the test statistic only involves the metabolomic data and when $\rho = 0$, the test statistic only involves the metagenomic data. In practice, ρ is unknown and needs to estimated for maximal power. To address this, we can perform a grid search; we set a grid of weights, $0 < \rho_1 \ldots < \rho_b < 1$, to obtain a p-value for each ρ_b. In order to select ρ that maximizes power, Lee et. al. proposed an optimal test procedure for the generalized SKAT method [39]. Based on this step, we define our test statistic in the following manner:

$$T = \min_{0 \leq \rho \leq 1} p_\rho \tag{2.5}$$

Q_{metab} and Q_{genom} can be approximated as a mixture of chi-square distributions. For a given ρ, Q_ρ is a mixture of two quadratic forms and is asymptotically equivalent to a mixture of two independent χ^2 random variables:

$$a(\rho)\eta_0 + (1-\rho)\kappa \tag{2.6}$$

$\eta_0 \sim \chi^2_1$ and κ approximately follows a mixture of χ^2. Let $q_\rho(T)$ be the quantile function for Q_ρ. Since κ can be approximated with moment matching, we can calculate the p-value of T:

$$\begin{aligned}
p\text{-value} &= 1 - \Pr\{Q_{\rho_1} < q_{\rho_1}(T), \ldots, Q_{\rho_b} < q_{\rho_b}(T)\} \\
&= 1 - \Pr\{a(\rho_1)\eta_0 + (1-\rho_1)\kappa < q_{\rho_1}(T), \ldots\} \\
&= 1 - \Pr\left[\kappa < \min_{0\le w\le 1} \frac{q_{\rho_v}(T) - a(\rho_v)\eta_0}{1-\rho_v}\right] \\
&= 1 - \left[\Pr\left(\kappa < \min_{0\le w\le 1} \frac{q_{\rho_v}(T) - a(\rho_v)\eta_0}{1-\rho_v}\right)\middle|\eta_0\right] \qquad \text{Iterated expectations}
\end{aligned}$$

2.2.4 Cauchy combination test

There have been multiple methods prescribed for combination testing but they often ignore correlation between p-values and can be computationally intensive [35, 5, 15]. The Cauchy combination test is a computationally efficient method used for combining individual p-values. There are several advantages to this method and it is well-suited to dealing with challenges arising from correlation and high-dimensionality. Under the null hypothesis, p-values are uniformly distributed; we can use this fact to transform them into Cauchy-distributed variables. Let $(p_0, p_{0.01}, p_{0.02}, \ldots, p_1)$ be a vector of p-values from the grid search in the weighted VC test and let w_ρ be the weights associated with each p_ρ. The test statistic for the Cauchy combination test is then a weighed sum of Cauchy-transformed p-values.

$$T = \sum_{\rho\in\{0,0.01,\ldots,1\}} w_\rho \tan\{(0.5 - p_\rho)\pi\} \tag{2.7}$$

Regardless of the correlation structure of p_ρ, T has a standard Cauchy distribution under the null. Interestingly, it has been shown that correlation between p-values has minimal impact on the tail of the distribution, making this method very powerful for highly-correlated p-values. With $w = \sum_\rho w_\rho$, the p-value of T can be approximated as:

$$p \text{ value} \approx 1/2 - \arctan T/w/\pi \tag{2.8}$$

2.2.5 Extension to kernel-based approaches

So far, we have assumed a linear association between the outcome of interest and the metagenome and metabolome for each pathway. Kernel functions are a particularly powerful approach for modeling non-linear and other complex relationships. We can reframe our proposed weighted VC approach in terms of kernel functions:

$$g[(Y_i)] = \alpha_0 + X_i\boldsymbol{\alpha} + f(W_i) + g(z_i) \tag{2.9}$$

Let $f, h \in \mathcal{F}$ a functional space generated by positive semidefinite kernel functions, $K(\cdot,\cdot)$. Furthermore, let $f(W_i) \sim F_0(0, \tau_1 K_1)$ and $h(z_i) \sim F_2(0, \tau_2 K_2)$ where F_k is an arbituary distribution with mean 0 and variance $\tau_k K$. The use of kernel functions allows more complex models to represent the relationship between an outcome and a specific metabolic pathway. At its basis, the kernel function measures the similarity between subjects i and i' based on their data. Let $(\phi_1, ...\phi_b)$ be a basis for $\mathcal{F}$. Taking the metabolomics data as an example, we can represent $f(W_i)$ as a linear combination of the basis functions.

$$f(W_i) = \sum_{j=1}^{J} \beta_j \phi_j(W_i) = \boldsymbol{\beta}^\top \boldsymbol{\phi}(W_i) \tag{2.10}$$

$$K(W_i, W_{i'}) = < \boldsymbol{\phi}(W_i), \boldsymbol{\phi}(W_{i'}) > \tag{2.11}$$

By using the kernel trick, we can implicitly define the basis space by using its inner

product, $K(W_i, W_{i'})$, instead. As a result, more complex models can be specified as long as there is a corresponding kernel function. The test statistic Q_ρ can be rewritten as

$$Q_\rho = \rho(Y - \hat{\boldsymbol{\mu}})^\top \boldsymbol{W}'\boldsymbol{W}'^\top (Y - \hat{\boldsymbol{\mu}}) + (1-\rho)(Y - \hat{\boldsymbol{\mu}})^\top \mathbf{z}\mathbf{z}^\top (Y - \hat{\boldsymbol{\mu}}) \tag{2.12}$$

$$Q_\rho = \rho(Y - \hat{\boldsymbol{\mu}})^\top \boldsymbol{K}_1 (Y - \hat{\boldsymbol{\mu}}) + (1-\rho)(Y - \hat{\boldsymbol{\mu}})^\top \boldsymbol{K}_2 (Y - \hat{\boldsymbol{\mu}}) \tag{2.13}$$

$\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are kernel functions for the metabolomic and metagenomic data respectively. In the original model, $\boldsymbol{W}'\boldsymbol{W}'^\top$ and $\mathbf{z}\mathbf{z}^\top$ are examples of linear kernels. Although we evaluate the performance of our method with the linear kernel, using a kernel that better captures relationship between the metabolic pathway and the outcome of interest can increase the power of our testing procedure.

2.3 Simulations

2.3.1 Simulation settings

We performed several simulations to assess the performance of the weighted VC method. We simulated metagenomic and metabolomic datasets from the empirical cumulative distribution functions of fecal whole-genome shotgun sequencing and capillary electrophoresis time-of-flight mass spectrometry (CE-TOFMS)-based metabolomics collected from 616 participants at the National Cancer Center Hospital in Tokyo, Japan. Metabolites were grouped based on the corresponding Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway given by the metagenomics data. In order to gauge the performance of the proposed method over pathways of varying size, we simulated data from two pathways: glycerophospholipid metabolism (map00564) which consisted of 6 metabolites and ABC transporters (map02010) which contained 37 metabolites. We tested the performance of our method on both continuous and dichotomous cases and used a grid of $\boldsymbol{\rho} = (0, 0.01, 0.02, \ldots, 0.99, 1.00)$ for our weights.

2.3.2 Type I Error

To evaluate type I error for continuous and dichotomous outcomes, we simulated 5000 datasets each for $n = (50, 100, 200, 300, 400, 500)$. In the following models below, we set $\boldsymbol{\alpha} = (1, 0.5, 0.5)$ and $\mathbf{X}$ as the matrix of covariates that we wanted to adjust out. We evaluated the performance of our proposed method on both continuous and dichotomous outcomes (Y_i). In order to test the performance of adjusting for both continuous and dichotomous covariates, X_1 was generated from a standard normal distribution and X_2 was generated from Bernoulli(0.5). The error terms, ϵ_i, were generated from independent standard normal distributions. We defined type 1 error as the proportion of p-values less than $\boldsymbol{\alpha} = (0.01, 0.05)$. Under the null model, we assumed the following models for continuous and dichotomous outcomes respectively:

$$Y_i = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \epsilon_i \quad \text{Continuous outcome} \tag{2.14}$$

$$\text{logit}[P(Y_i = 1)] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 \quad \text{Dichotomous outcome} \tag{2.15}$$

Type I performance of the weighted VC method for these simulations are shown in Table 2.1. Although we only show type 1 error results from the Cauchy combination test (Cauchy), our results from the minimum p-value method (Min-p) are very similar. Overall, the weighted VC method agrees well with $\alpha = (0.01, 0.05)$ for both pathways and types of outcomes. Our proposed method tends to be slightly conservative for small sample sizes in the continuous case, although this deviation is small. We note that the weighted VC test has slightly inflated type 1 error in the dichotomous outcome scenario, but this modest deviation decreases as sample size increases.

	Continuous				Dichotomous			
	map00564		**map02010**		**map00564**		**map02010**	
	0.01	**0.05**	**0.01**	**0.05**	**0.01**	**0.05**	**0.01**	**0.05**
n = 50	0.0076	0.0474	0.0086	0.0438	0.0084	0.0562	0.0104	0.0594
n = 100	0.0090	0.0500	0.0080	0.0478	0.0078	0.0512	0.0118	0.0542
n = 200	0.0086	0.0432	0.0094	0.0436	0.0110	0.0502	0.0108	0.0528
n = 300	0.0114	0.0520	0.0118	0.0490	0.0100	0.0512	0.0084	0.0532
n = 400	0.0100	0.0486	0.0106	0.0526	0.0100	0.0504	0.0120	0.0496
n = 500	0.0096	0.0552	0.0068	0.0528	0.0084	0.0476	0.0084	0.0530

Table 2.1: Type I error rates for the $\alpha = (0.01, 0.05)$ and $n = (50, 100, 200, 300, 400, 500)$ based on outcome type and pathway.

2.3.3 Power

We assumed the following model for the alternative hypothesis.

$$g[(Y_i)] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + W_i\boldsymbol{\beta} + \gamma z_i \tag{2.16}$$

We set $\boldsymbol{\alpha} = (0.3, 0.3)$. For the first scenario, we assessed the performance of the metabolic potential while keeping the effects of the metabolic output constant. We set a non-zero constant value for $\approx 20\%$ of $\boldsymbol{\beta}$ while the remaining 80% of the coefficients were set to 0. We first tested the performance of the weighted VC method by varying the effects of γ. In the second scenario, we evaluated the performance of the method by keeping the metabolic potential constant while varying the effects of the metabolic output. We set a constant value for γ based on the first scenario and calculated the power of the method over a range of different constant values for $\boldsymbol{\beta}$. Although we only show the continuous outcome case for both scenarios, we observed similar results in the dichotomous outcome case. We compared the power of our proposed weighted VC test using both integrated methods (Min-p and

Cauchy) to methods that utilized only one dataset. Since the metagenomics data for each pathway contains only one value for each sample, we used a simple linear regression model in the continuous scenario:

$$g[(Y_i)] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \gamma z_i \tag{2.17}$$

For the metabolomics data, we modeled the association between the metabolites and outcome as:

$$g[(Y_i)] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + W_i \boldsymbol{\beta} \tag{2.18}$$

In contrast to the metagenomic data, there can be a modest to large number of metabolites in a pathway. To accommodate possible high-dimensionality of the metabolites, we used a standard VC test for the metabolomics-only analysis. We assumed that each β_j followed an arbitrary distribution with mean 0 and variance τ. Our test statistic for evaluating the null hypothesis: $\tau = 0$ was:

$$Q = \rho(Y - \hat{\boldsymbol{\mu}})^\top \boldsymbol{W}\boldsymbol{W}^\top (Y - \hat{\boldsymbol{\mu}}) \tag{2.19}$$

Q asymptotically follows a mixture of χ^2 distributions. Specifically $Q \sim \sum_{i=1}^{n} d_i \chi_1^2$ where d_i are the eigenvalues of $\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}_n \mathbf{W}\mathbf{W}^\top \boldsymbol{P}_n \boldsymbol{\Sigma}^{1/2}$ with $\boldsymbol{\Sigma} = \text{cov}(\boldsymbol{Y})$. The Davies method, which involves the numerical inversion of the characteristic function, can be used to obtain a p-value for the test statistic. In short, we can view our procedure for assessing the power of a metabolomics-only analysis as a simplified version of the proposed weighted VC test.

We present the results of Scenario 1 in Figure 2.1. For map00564, the metabolomics-only analysis only outperformed the integrated methods when there was no metagenomics effect, i.e. $\gamma = 0$. While the power of the metabolomics-only analysis decreased as the effect size of γ increased, the performance of the metagenomics-only analysis rapidly improved. For $\gamma > 0.1$, the integrated methods consistently outperformed the performance of the metagenomics-only

and metabolomics-only analysis for all sample sizes. With regards to the integrated methods, the Cauchy and Min-p approach had similar performance unless the sample size was small and the effect size of γ was large or small. Recall that the Min-p method relies on one $p-$value while the Cauchy method is an average of Cauchy-transformed $p-$values. As a result, the Min-p method will be especially advantageous when the contribution of each dataset is unequal, i.e. one data set has a dominating effect on the outcome. Similar results from the map00564 pathway were also observed in the map02010 pathway.

For the second scenario, we chose a value of γ that had moderate power in Scenario 1 and then varied the effects of $\boldsymbol{\beta}$. The results of Scenario 2 are outlined in Figure 2.2. Although the performances of the integrated methods increased as $\boldsymbol{\beta}$ increased, the metabolomics-only analysis had the greatest power at $n = 50$ since the contribution between the metabolomics data and metagenomics data becomes unequal; i.e. the metabolomics effect dominates the metagenomics effect. We note that the increase in the metabolomics-only performance is minimal and requires prior knowledge that the metabolites have a greater effect on the outcome than the metagenes. For an agnostic approach, it is still preferable to use either of the integrated methods. For $n = (200, 400)$, both integrated methods outperformed the other two methods across all values of $\boldsymbol{\beta}$ for both pathways.

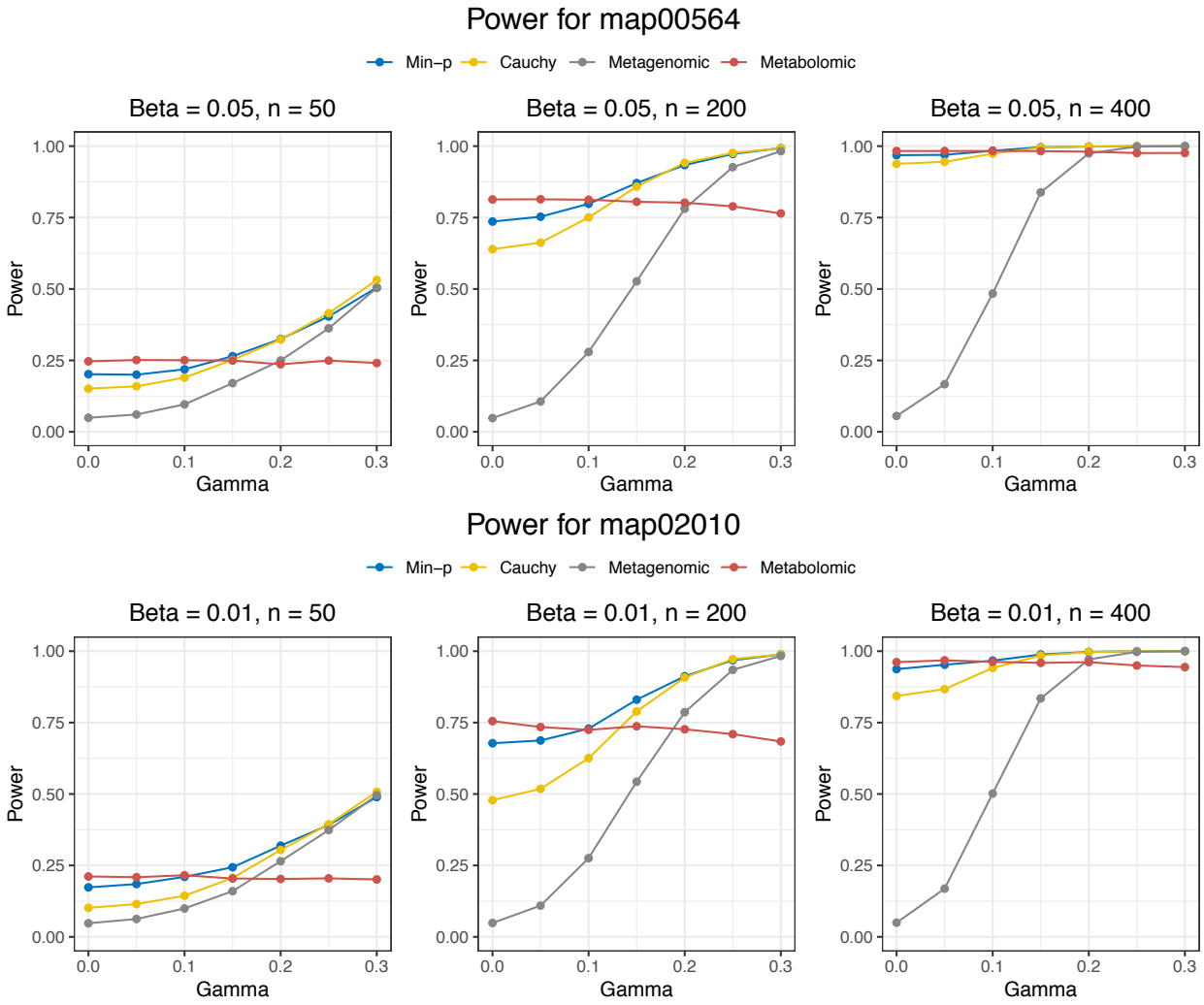


Figure 2.1: Simulation results for evaluating power in Scenario 1. The effects of the metabolomics input is kept constant as we evaluate the impact of increasing the effects of the metagenomics input for $n = (50, 200, 400)$ and for pathways map00564 and map02010. Power is defined as the proportion of $p-$values ≤ 0.05

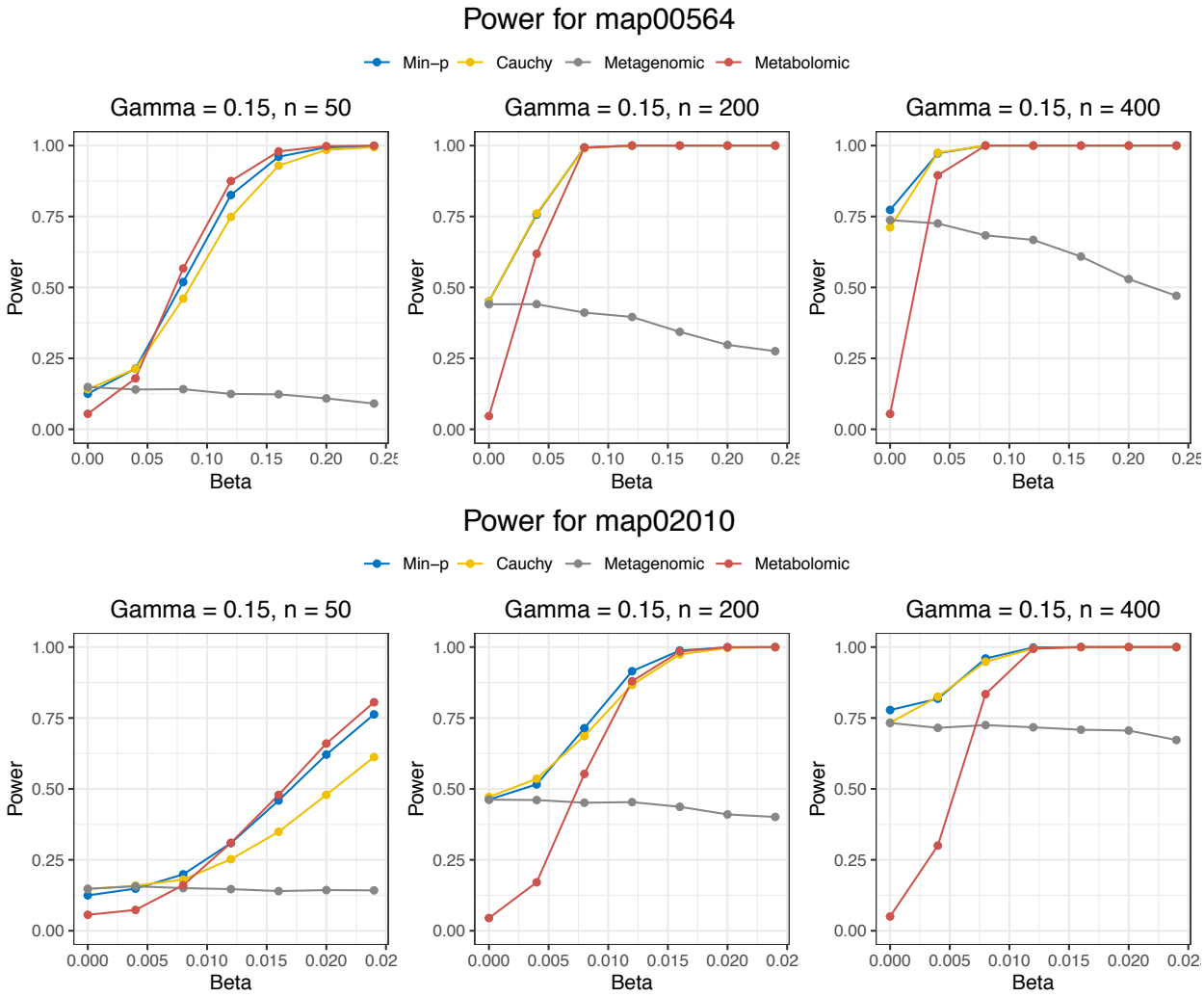


Figure 2.2: Simulation results for evaluating power in Scenario 2. The effects of the metabolomics input is kept constant as we evaluate the impact of increasing the effects of the metagenomics input for $n = (50, 200, 400)$ and for pathways map00564 and map02010. Power is defined as the proportion of $p-$values ≤ 0.05

2.4 Real Data Application

The Hispanic Community Health Study/Study of Latinos (HCHS/SOL) is a multicenter epidemiologic study aimed at assessing the risks and determinants of health in Hispanic/Latino populations across the United States. Approximately 16,000 participants across Miami, San Diego, Chicago and the Bronx area of New York are involved in the study. Cardiovascular disease (CVD) is particularly prevalent in Hispanics/Latinos in the United States and is the leading cause of mortality in this population. Several key risk factors for CVD have been identified and include hypertension, high cholesterol levels, diabetes and obesity. The goal of the HCHS/SOL study is to provide an objective and comprehensive overview of the prevalence of risk factors in the Hispanic/Latino population. Baseline findings of the HCHS/SOL study has shown that a large majority of this population, 71% in women and 80% in men, have at least one major risk factor for CVD [12].

Our work is motivated by a microbiome sub-study of HCHS/SOL. Recent developments in microbiome research have led to increased interest in understanding the crosstalk between the gut microbiome and disease risk. The Centers for Disease Control and Prevention estimate that more than 30 million people in the United States have type 2 diabetes (T2D), a major risk factor of CVD. Recent studies have shown the functional role of the gut microbiome in the pathophysiology of T2D. In particular, T2D is characterized by increased permeability of the gut. Reduced alpha diversity, as seen in patients with T2D, can lead to weaker tight junctions that result in increased insulin resistance and chronic inflammation [67]. Given that pathways disrupted by T2D can be contributed by both the host and microbiome, we applied the weighted VC method to these data sets to determine the subset of pathways significantly associated with T2D. Due to its status as a CVD risk factor, identifying significant pathways in T2D not only increases our understanding of this chronic disease but also helps aid in the development of comprehensive and effective preventive measures for CVD.

In our analysis, we used KEGG-based pathways in order to connect each metabolite to its metagenomic pathway. To encourage normality, we log-transformed both datasets. We

compared our method with a metagenomics-only and metabolomics-only analysis described earlier in the Simulations section. After data quality control, we had whole genome shotgun sequencing data and targeted serum metabolomic data for 46 pathways from 621 individuals. We applied our proposed method on each of the 46 pathways and adjusted our p-values using the false discovery rate (FDR) correction. Significant pathways were defined as having FDR ≤ 0.05.

	Min p-val	**Cauchy**	**Metagenomics**	**Metabolomics**
Min p	29	-	-	-
Cauchy	21	21	-	-
Metagenomics	0	0	0	-
Metabolomics	23	20	0	24

Table 2.2: Unique and overlapping significant pathways associated with Type 2 diabetes from the HCHS/SOL Study

The results of our real-data analysis are shown in Table 2.2. Cumulatively, this analysis identified many significant pathways associated with T2D. Quality of life is often broadly affected by T2D and there is substantive literature connecting T2D to renal function, coronary arterial disease, retinopathy, neuropathy, depression and dementia among many other conditions [74]. Since diabetes is a condition that affects so many biological processes, it is reasonable to see many pathways to be significant with diabetes status.

The Min-p method was able to detect more diabetes-relevant pathways compared to the other three methods. Additionally, as seen in Table 2.1, there were many overlapping pathways that were identified in Min-p method, Cauchy method and the metabolomics-only analysis; no significant pathways were detected in the metagenomics-only analysis. Based on this result and the large number of overlapping pathways between the Min-p method and metabolomics-only analysis, the differences in metabolic pathways between disease types seem to be driven primarily by the metabolites. However, it is still advantageous to consider

both data sets since the Min-p method was able to detect six unique pathways compared to the metabolomics-only analysis. Overall, our analysis has identified significant novel pathways that can be investigated in future studies to elucidate the mechanisms behind both T2D and CVD.

2.5 Discussion

We have proposed a weighted variance component framework for integrating metagenomics and metabolomics data when assessing the association between a pathway and an outcome of interest. In understanding the relationship between the host and the microbiome, metabolism can be represented in two parts: the metabolic potential (metagenomics) from the microbiome and the metabolic output (metabolomics) from the host. By employing a grid search strategy, we are able to agnostically evaluate the weighted average effect of both data sets. Compared to methods that only consider one data, we show through simulations that we actually increase power while also controlling for type 1 error if we consider both the metagenome and the metabolome in our analysis. Lastly, we applied our proposed approach to the HCHS/SOL study and identified several unique pathways significantly associated with T2D.

This approach may seem redundant since it is generally understood that the synthesis of metabolites occurs downstream of gene expression. Several theories may explain why we obtain higher power when we consider both data sets. First of all, metabolomic data can often often be noisy. In addition to measurement error from the equipment, levels of metabolites are highly variable on the time the sample was collected. A recent study examined the within-individual variability and between-individual variability in 385 metabolites from 60 individuals from baseline and one year into the Shanghai Physical Activity Study. Within-individual variability captured the majority of variation explained for 64% of the observed metabolites [64]. Furthermore, metagenomic data can capture shorter and more volatile metabolites that standard mass spectrometry procedures fail to detect. Lastly, in the context of our real data analysis, we had host metabolites and microbial genes. We can represent both

host and microbial mechanisms as a potential application of our proposed approach. Using a weighted VC framework to capture these two aspects of metabolism, we can better assess the association between an outcome and a metabolic pathway. As an extension, our method can be used to evaluate the cumulative effect of any two data sets (e.g. transcriptomics, proteomics, epigenomics, etc.).

We introduce two methods for obtaining p-values from our testing procedure: the minimum p-value method and the Cauchy combination test. The former method relies on an asymptotic result, specifically that for each ρ, the test statistic Q_ρ can be approximated as the sum of two independent chi-square distributions. Despite this asymptotic assumption, the Min-p method has been shown to have good performance in small sample sizes. The Cauchy combination test, on the other hand, is based on the result that the weighted sum of Cauchy-transformed p-values are still Cauchy under arbitrary correlation structures. Both the minimum p-value method and Cauchy combination test performed well in simulation studies, though we note the former method was able to detect more significant pathways in the real data analysis. Although both methods are suitable in the presence of correlated p-values, the main difference between the minimum p-value method and the Cauchy combination test is the representation of the final test statistic as a minimum or as an average, respectively. For example, the minimum p-value method is advantageous if there is a large difference between the minimum and the rest of the elements in the p-value vector. However, if there is low variability within the vector, the usage of the Cauchy combination test may lead to increased power. Additionally, we note that the Cauchy combination test is computationally faster and simpler to implement.

One limitation of our proposed method is that it does not provide the directional association of a microbial pathway. We note, however, that directionality is ambiguously defined in the field of metagenomics. Ostensibly, each pathway contains the abundance of all its respective genes. This does not mean, however, that each gene contributes to the increase activity of the pathway. For example, an increase in a gene involved in the tryptophan biosynthesis does not necessary result in an increase of its respective metabolite, tryptophan. One

purpose of our proposed method is to determine the subset of pathways associated with an outcome of interest. After properly controlling for type I error due to multiple testing, further downstream analyses can be performed on the set of significant pathways.

Through our simulation studies and real data analysis, we demonstrate that the integration of metabolic potential and metabolic output leads to increased power and the increased detection of significant pathways, respectively. As technology evolves, we now have the opportunity to collect multiple data sets to answer our scientific questions of interest. So far we have applied our method in the context of host and microbial contributions; our method can also be applied to host-only or microbiome-only data sets. Additionally, in our real-data example, we applied a *de novo* approach determining significant pathways, i.e. we used the weighted VC test on all available pathways. Our method may be powerful on a subset of more carefully selected pathways, like those with metabolites known to be relevant to bacterial production or function.

Chapter 3

VARIABLE ADJUSTMENT IN THE VISUALIZATION OF HIGH-DIMENSIONAL DATA

3.1 Introduction

The human microbiome consists of trillions of bacteria, viruses, and archaea. This vast array of living organisms within the human body can be thought of as a micro-ecosystem that is uniquely defined for each person. Although the initial colonization of the microbiome is strongly characterized by genetics, mode of birth and a variety of other factors, the microbiome remains dynamic throughout one's lifespan. Medication, environmental and/or lifestyle changes can lead to compositional differences in our microbiota, many of which result in profound functional effects in human health. In fact, the human microbiome has been implicated in several host diseases and conditions such inflammatory bowel disease (IBD), neurodevelopmental disorders, and certain cancers [18, 25]. This is primarily due to the microbiome's interactions with a number of host organs. Specifically, it has been shown that the gut microbiome interacts not only with proximal organs like the liver but also distal organs like the brain [10, 75]. Understanding the extensive and complex relationship between the microbiota and host can lead to the development of more effective strategies for treating diseases associated with the dysbiosis of the microbiome.

Although hypothesis-based approaches are popular in determining the association between microbial features and experimental conditions, methods for the visualization of microbiome data can be particularly powerful for demonstrating compositional changes in the microbiome as well. Data visualization tools are useful for summarizing and discovering new patterns from large amounts of microbiome data. Additionally, collaborators often prefer hypothesis-free methods because it allows for the full exploration of the data without

any prior assumptions. Data visualizations provide a data-driven approach for generating potential hypotheses for future studies [57].

Several challenges exist in the statistical analysis of microbiome data. Microbiome data is often high-dimensional. Because there are hundreds in not thousands of unique species in a single individual's microbiota, taxa can often outnumber sample size in any given study. Furthermore, microbiome data is compositional and subject to variations in library size. With respect to compositionality, microbiome data is should be viewed as proportions instead of raw counts due to limits in current sequencing technologies. As a result, species within the data matrix are not independent from each other. Failure to account for compositionality and zero values from rare species can lead to high false positive identification rates [19]. Due to the peculiarities of microbiome data, visualizing microbiome data using traditional methods, such as scatterplots or histograms, becomes ill-advised and ineffective. Instead, dimension-reduction methods are necessary in order to properly and efficiently interpret this data [41].

Instead of focusing on a species-by-species approach, we can instead utilize summary statistics for visualizing microbiome data. The most popular summary statistics in current literature are alpha diversity and beta diversity. Alpha diversity can be defined as either species richness or species evenness. The former refers to the raw number of species whereas the latter refers to the distribution of species in the sample. On the other hand, beta diversity, represented by distance metrics or metric-like measures, is used to quantify the pairwise relationships between samples based on high-dimensional data. Euclidean distances, which come from the most common metric, are sensitive to sample abundance and do not take species identity in consideration. Non-Euclidean measures, like the Bray-Curtis dissimilarity measure and the UniFrac distance, have been proposed for applications in microbiome data analysis [4]. For two microbiome samples with j species, the Bray-Curtis dissimilarity can be calculated by dividing the sum of the absolute differences between the j species by the total abundance of the two samples. The UniFrac distance, meanwhile, measures the distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that

leads to its descendants. The choice of measure is often context-specific; it has been shown that the Bray-Curtis measure performs poorly when there are rare but highly-abundant species in the data and the unweighted UniFrac distance performs poorly in the presence of highly prevalent but lowly abundant species [69].

Unsupervised learning methods can be used to identify patterns in the analysis and visualization of high-dimensional data. In regards to microbiome data, clustering methods such hierarchical clustering, partitioning around medoids (PAM) and Dirichlet multinomial mixtures (DMM) can determine if microbial composition is associated with a specific variable of interest such as treatment or disease type. Hierarchical clustering utilizes dissimilarity measures, such as the Bray-Curtis or UniFrac measure, in order to determine how clusters should be split. These splits can then be graphically represented by dendrograms, which depict clusters in a tree-like structure. Considered as an extension of k-means clustering, partitioning around medoids (PAM) is another popular machine learning method for identifying clusters. Instead of minimizing the sum of squared distances to cluster centers, as with k-means, PAM minimizes the sum of distances from the medoids and is more robust to the presence of outliers [69]. Lastly, Dirichlet multinomial mixture (DMM) generative models can be used to define metacommunities in microbiome data based on community structure and sample density and size. Each metacommunity can be represented by a vector of species probabilities, which arise from Dirichlet mixture components with different hyperparameters. Unlike the two previously described methods, DMM accounts for features commonly found in microbiome data such as rare taxa [26].

In addition to these unsupervised approaches, principal coordinates analysis (PCoA) is perhaps the most popular method for visualization and utilizes dissimilarity measures to depict clusters in the data[9]. Similar to principal components analysis (PCA), PCoA is based on eigenvalue equations that can be used to reduce the dimensionality of microbiome data to two or three dimensions that capture the most variance of the data. An important distinction of PCoA is its use of non-linear dissimiliarity measures (i.e. non-Euclidean distances). Unlike Euclidean distances used in PCA, the use of non-Euclidean measures are

particularly important in PCoA since they capture important features of microbiome data such as species identity and phylogenetic closeness. PCoA is used to visualize microbiome compositional differences based on beta diversity. The purpose of PCoA is often two-fold. First, PCoA plots can be used to identify clusters in the data, i.e. observations with similar microbiota compositions will be closer to each other in space and form clusters. Clusters are often delineated with color-coded points and 95% confidence interval ellipses based on experimental or disease conditions. Second, PCoA plots can also be used to evaluate microbiome variability within groups. For example, although the microbiota of infants have the lowest alpha diversity compared to older age groups, they have the highest interindividual variation; this corresponds to larger and more spread out points in the PCoA plot.

Although we have introduced various methods for the visualization of microbiome, they all fail to address the presence of covariates that obscure the true signal of the experiment. Advances in technology have allowed for more accessible processing and analysis of high-throughput data. Practical considerations, however, can limit the number of samples processed in a single sitting. Batch effects are often separate from biological factors and can be caused by a multitude of factors including time of day, laboratory conditions, and different reagent batches used in the sample preparation [40]. In the meta-analysis field, each individual study can also be considered as a batch effect. The ability to adjust for the differences in study design and data collection can be fruitful for merging insightful results for a specific disease or experimental condition. Overall, batch effects have a detrimental effect on the analysis of microbiome and other -omics data sets. They can produce both excessive false positives and false negatives and hinder inference between the microbiome and clinical variables of interest.

Several methods have been proposed for batch correction in genomics and microbiome data. Data with batch effects can be characterized as multilevel data, with each batch defined as a group. In this setup, we can represent each batch effect as a random effect in a linear mixed model. Although straightforward, this approach is often underpowered for high-dimensional data unless the effect size is large. Bayesian Dirichlet-multinomial

regression meta-analysis (BDMMA) is another method for batch correction that accounts for complex distributional traits found in microbiome data. By considering dependencies between taxa and incorporating batch effects in the Dirichlet-multinomial regression model, BDMMA can more precisely detect species associated with a disease condition across different studies. Unfortunately, BDMMA and other popular methods for batch effect correction are primarily used for association testing. It is unclear how these methods can be used to correct microbiome data for applications in data visualization. [30, 11].

We propose a two-step process for adjusting unwanted covariates from the principal coordinates (PCs) of microbiome data. In the first step, we perform a mean adjustment by projecting the PCs out of the unwanted covariate space using its hat matrix. In the second step, we adjust the second moment by assuming a linear relationship between the variance of the mean-adjusted PC and the unwanted covariate. We apply our proposed method to various visualization methods such as PCoA and hierarchical clustering across a wide variety of data sets ranging from an ecological study of invasive species to an international meta-analysis of colorectal cancer studies. Furthermore, we extend our two-step process to existing methods for evaluating multi-omics data (e.g. metabolomics and metagenomics).

The remainder of this chapter is organized as follows. In the next section, we provide a brief overview of multidimensonal scaling and the general procedure for obtaining PCs before outlining our proposed method of adjusting the mean and variance of the covariates out of the PCs. In Section 3, we apply our method to various visualization methods such as PCoA, hierarchical clustering, and permutational analysis of variance (PERMANOVA) across several real data sets. Finally we close the chapter with a brief discussion in Section 4.

3.2 Methods

First, we will review the general procedure for obtaining PCs using the classical multidimensional scaling method, also known as the Torgerson method. Our proposed method follows after obtaining the original PCs.

Let $\boldsymbol{Z}$ be the microbiome profiles for n individuals with m taxa. We can construct a matrix of pair-wise distances between individuals which we denote as $\boldsymbol{D}$. Non-Euclidean measures such as the Bray-Curtis dissimilarity matrix and the UniFrac distance can be used to represent these pairwise distances. Using $\boldsymbol{D}$, we can obtain the PCs by:

1. Computing the matrix of squared dissimilarities by squaring all elements of $\boldsymbol{D}$: $\boldsymbol{D}_{ij}^{(2)} = d_{ij}^2$

2. Double-centering $\boldsymbol{D}^{(2)}$: $\boldsymbol{M} = -\frac{1}{2}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{D}^{(2)}(\boldsymbol{I} - \boldsymbol{H})$, where $\boldsymbol{H} = \boldsymbol{1}(\boldsymbol{1}^\top\boldsymbol{1})^{-1}\boldsymbol{1}^\top$

3. Eigendecomposing $\boldsymbol{M}$ as $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$. We define $\boldsymbol{U}\boldsymbol{\Lambda}^{1/2}$ as the PCs. Generally, the first two columns of $\boldsymbol{U}\boldsymbol{\Lambda}^{1/2}$ are used for PCoA plots

For k covariates, let $\boldsymbol{X}$ be a $n \times k$ matrix of confounders we want to adjust out of our PCs. In the context of batch effects, $\boldsymbol{X}$ is the design matrix of dummy variables for each batch. A previous method proposed replacing $\boldsymbol{H}$ with the hat matrix, $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and taking the eigendecomposition of the covariate-adjusted Gower-centered matrix. Essentially, this corresponds to subtracting the mean of the covariate from each PC. As an extension of this method, we propose adjusting both the mean and variance of the unwanted covariates out of the PCs. Let $\boldsymbol{H}_{\text{covariate}} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, $\boldsymbol{\Phi} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}$ and $\boldsymbol{Z} = (\boldsymbol{I} - \boldsymbol{H}_{\text{covariate}})\boldsymbol{\Phi}$. We assume the following model for the variance of each column of $\boldsymbol{Z}$.

$$\text{Var}(z_{ij}) = \text{E}\left(z_{ij}^2\right) = \alpha + \boldsymbol{\beta}x_i^2 \qquad \text{where } \alpha, \beta_j \geq 0$$

Because $\text{Var}(z_{ij}) \geq 0$, we place a non-negative constraint on α and $\boldsymbol{\beta}$. Estimates of α and $\boldsymbol{\beta}$ can be obtained through non-negative least squares. We can adjust each column of the mean-adjusted PCs by using these estimates.

$$(Z_i) = \alpha + \beta x_i^2 \tag{3.1}$$

$$\frac{\sqrt{\alpha}}{\sqrt{\alpha + \beta x_i^2}}(Z_i) = \alpha \tag{3.2}$$

$$Z_{i,\text{adj}} = \frac{\sqrt{\hat{\alpha}}}{\sqrt{\hat{\alpha} + \hat{\beta} x_i^2}} Z_i \tag{3.3}$$

Note that we adjusted all the PCs even though only the first two or three PCs are typically used for data visualization. If the variance from batch effects were primarily captured by the first PC, the corresponding mean and variance batch adjusted PC would no longer capture the most variability in the data. As a result, we cannot use the adjusted PCs defined in Equation (3.3). Instead, let $\boldsymbol{M}^*_{\text{adj}} = \boldsymbol{Z}_{\text{adj}}\boldsymbol{Z}^\top_{\text{adj}}$. We take the eigendecomposition of $\boldsymbol{M}^*_{\text{adj}}$ to obtain $\boldsymbol{\Phi}_{\text{adj}} = \boldsymbol{U}_{\text{adj}}\boldsymbol{\Lambda}^{1/2}_{\text{adj}}$, the mean and variance-adjusted PCs. Although we have primarily focused on batch effects, which are categorical, our proposed method can be applied to continuous confounders or a mixture of continuous and categorical variables.

3.2.1 Extension to high-dimensional data

Due to the growing accessibility of high-throughput sequencing equipment, microbiome studies now collect several data sets from both the host and microbiome to have a more comprehensive understanding of their disease model or experimental condition. We extend our method to high-dimensional data sets to accommodate the increased availability of transcriptomics, proteomics and other -omics data sets. As described previously, the microbiome has been shown to interact with several host processes. In order to isolate the microbiome-only effects, we can let $\boldsymbol{X}$ instead be a $n \times m$ matrix where $m > n$. We modify our original two-step procedure in the following ways:

1. Using the ridge version of the hat matrix to obtain the mean-adjusted PCs

$$(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{\lambda}\boldsymbol{I})^{-1}\boldsymbol{X}\top)\boldsymbol{\phi} = \boldsymbol{Z}_{\text{high}}$$

2. Obtaining estimates of $\boldsymbol{\beta}$ through LASSO. Let $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_m\}$. For each jth column

of $\boldsymbol{Z}$, we let $\text{Var}(\boldsymbol{Z}_j) = Y$ and solve the following optimization problem:

$$\hat{\boldsymbol{\beta}} = \underset{\beta \in \mathbb{R}^m}{\arg\min} \|Y - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

We will use the selected estimates of $\hat{\boldsymbol{\beta}}$ to adjust $\boldsymbol{Z}_{\text{high}}$. We define $\boldsymbol{M}^*_{\text{adj,high}} = \boldsymbol{Z}_{\text{adj, high}} \boldsymbol{Z}^\top_{\text{adj, high}}$. Similar to the low-dimensional procedure, we can obtain the high-dimensional version of the mean and variance adjusted PCs by taking the eigendecomposition of $\boldsymbol{M}^*_{\text{adj,high}}$. If no variables are selected through LASSO for a specific column of $\boldsymbol{Z}$, the original PC is preserved.

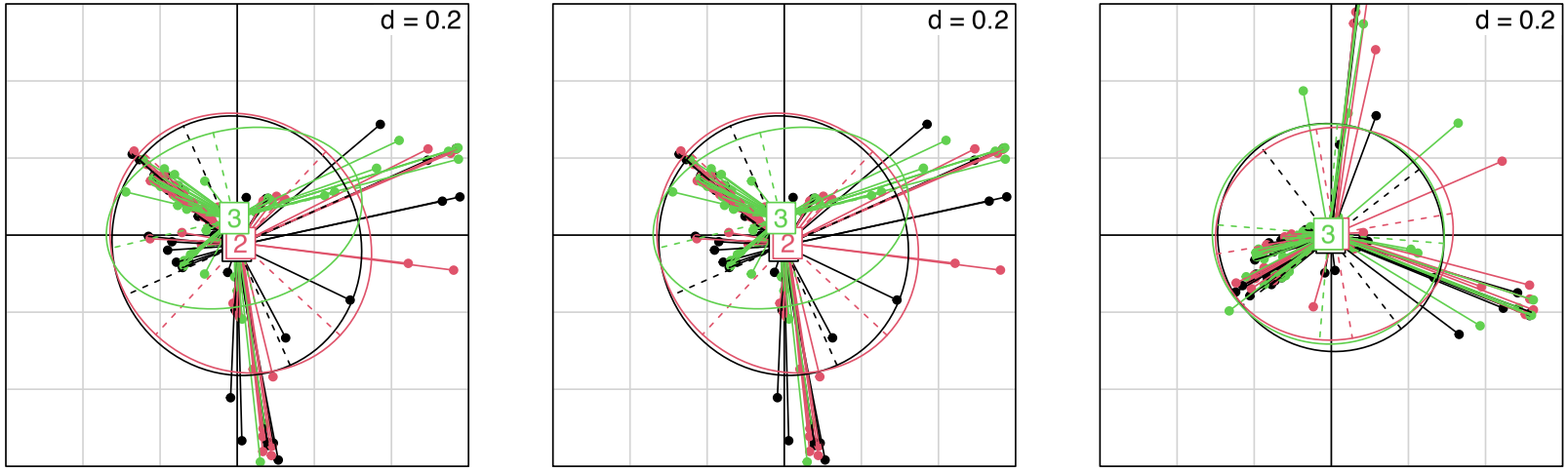
3.2.2 Mean and variance-adjusted dissimilarity matrices

The process of double centering the distance matrix results in its conversion to a similarity matrix, where larger values indicate increased similarity in microbial composition between two samples. Because $\boldsymbol{M}^*_{\text{adj}}$ is the mean and variance-adjusted similarity matrix, we can transform $\boldsymbol{M}^*_{\text{adj}}$ into a dissimilarity matrix. Let s_{ij} be the similarity measure between samples i and j and let d_{ij} be their corresponding distance. There are several generic methods to convert similarity matrices to dissimilarity matrices.

$$d_{ij} = \frac{1}{1 + s_{ij}}$$
$$d_{ij} = \max(s_{ij}) - s_{ij}$$
$$d_{ij} = \sqrt{1 - s_{ij}}$$

The resulting mean and variance-adjusted distance matrix can be used for additional analysis and visualization techniques. For example, we can perform hierarchical clustering and plot dendrograms to identify compositionally similar groups in our data. Another popular method used in microbiome data analysis is permutational multivariate analysis of variance (PERMANOVA). PERMANOVA is a semi-parametric method that uses a dissimilarity ma-

trix to construct a pseudo F-statistic that tests if there are differences in the location of the centroids between groups in the space of chosen dissimilarity measure [2]. By adjusting out unwanted covariates from our dissimilarity matrix, we can remove its effects and potentially observe stronger associations between the microbiome and our covariate of interest.



(a) Unadjusted (b) Mean adjusted (c) Both adjusted

Figure 3.1: PCoA plot from MsFLASH trial evaluating the use of vaginal estrodial tablet and/or vaginal moisterizing gel on the microbial composition of the vagina after one month of treatment. 1 - vaginal estradiol tablet plus placebo gel, 2 - vaginal moisturizing gel plus placebo tablet, 3 - Placebo gel and tablet.

3.3 Data application

We apply this method to several datasets to illustrate the utility of our proposed method. We considered the Bray-Curtis dissimilarity measure for all data examples. Although not shown, the UniFrac distance is also a popular measure and takes in phylogenetic information to calculate dissimilarities between pairwise samples.

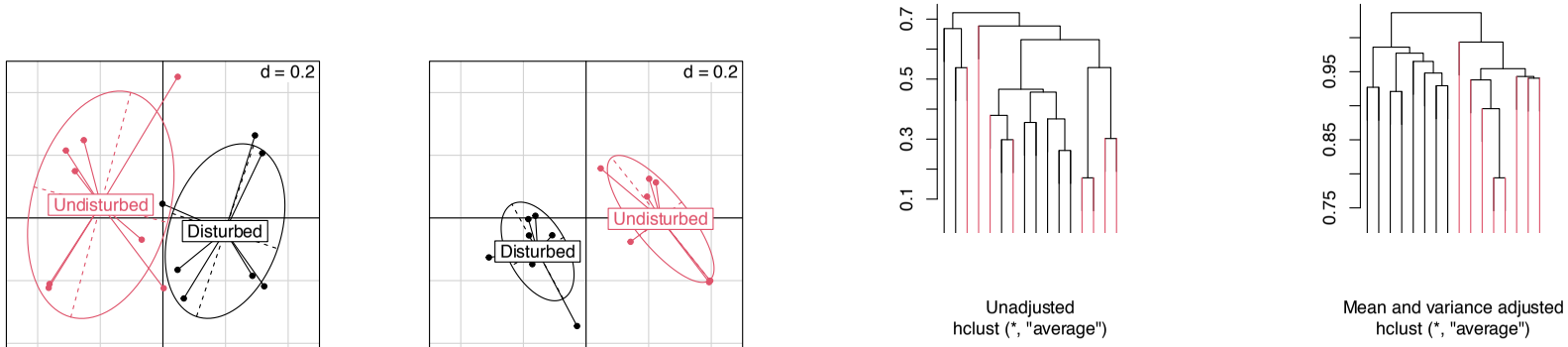
3.3.1 Finding Lasting Answers for Symptoms and Health Trials

First, we consider data from the Finding Lasting Answers for Symptoms and Health Trials (MsFLASH). Menopause occurs in the last third of a person's lifespan and is marked by the decreased secretion of estrogen and progesterone in the body. It is estimated that 1.5 million menstruating people undergo menopause each year world wide. Menopause is often associated with a wide-ranging array of symptoms such as lowered sexual libido, hot flashes, joint pain, and vaginal dryness. In fact, 86% of menopausal people reported consulting a clinician for symptom management in a population assessment of almost 400 individuals [65]. There is a dearth of research focused on this particular life stage despite the high prevalence of menopausal people reporting lower quality of life due to their symptoms. The purpose of MsFLASH is to develop effective treatments for the management of the most common and bothersome symptoms of menopause. Since its conception in 2008, this network has conducted five clinical trials totaling 1300 participants from the ages of 40 and 70.

Our data application is motivated by a microbiome sub-study of MsFLASH. It is believed that the vaginal microbiome may be associated with symptoms of vulvar, vaginal, or urinary discomfort in postmenopausal women [62]. Specifically, it has been shown that the *Lactobacillus* species is critical in the maintenance of vaginal homeostasis; unsurprisingly, the abundance of these species decreases during menopause due to decreased levels of estrogen and progesterone[55]. The severity of menopausal symptoms are also dependent on the stage of menopause. In study of over 1200 women in China, it was found that the degree of severity in fatigue, vaginal dryness, and joint pain were significantly different between

peri and postmenopausal people [63]. In this particular MsFLASH study, postmenopausal people with moderate-severe vulvovaginal discomfort were randomized to vaginal estradiol tablet (plus placebo gel), vaginal moisturizing gel (plus placebo tablet), or dual placebo for 12 weeks. After data quality control, the clinical study contained 410 participants with 381 taxa. We focused our analysis on week 4 of the trial to evaluate the changes in vaginal microbiome composition after one month on treatment. We treated age (continuous) as a potential confounder since microbial composition and severity of symptoms are associated with age.

Figure 3.1 shows a separation in microbiome composition by treatment in the unadjusted and mean-adjusted PCoA plot after a month of treatment, particularly with the placebo group. When we adjust for the mean and variance of age, however, the treatment effect disappears. This suggests that age may have a confounding effect on treatment and should be seriously considered during the development of effective treatments on mitigating the debilitating symptoms of menopause.



(a) Unadjusted (b) Adjusted (c) Dendrogram: unadjusted (d) Dendrogram: adjusted

Figure 3.2: Unadjusted and adjusted PCoA plots (a-b) and dendrograms (c-d) from the Tasmania ecological data set. Better separation is achieved by adjusting out location, which was observed to be a batch effect in the original data set.

3.3.2 Invasive species on meiobenthos species composition

We next consider an ecological dataset for the second illustration of our method. Although we have focused primarily on the microbiome, many of the current methods for microbiome analysis were originally developed for applications in ecology. In this example, we consider the role of invasive alien species (IAS) on the species composition of a local ecosystem. IAS are defined as introduced organisms that can overpopulate and cause harm to its new environment. After habitat loss, the introduction of IAS is considered to be the second greatest threat to diversity and affects almost 50% of all species globally [84]. The economic effects of IAS are also quite staggering; it is estimated that the effects of introducing up to 50,000 IAS in the United States have caused around 120 billion USD in financial losses. Invasive plants, such as yellow star thistle in California and the European cheatgrass in Idaho, comprise the majority of these species and have fundamentally altered the landscape of the United States [58].

Islands are highly susceptible to the effects of IAS due to their prevalence of highly unique species and the lack of competition to control invasive populations. The purpose of this study was to determine if disturbances from the soldier crab, *Mictyris platycheles*, led to disruptions in community structure of meiobenthos, which are small aquatic invertebrates that can inhabit both salt and freshwater habitats. Count data of 56 Meiobenthos species were collected from four locations across Eaglehawk Neck, Tasman Peninsula in southeastern Tasmania. Two of the locations were labeled as "Disturbed" (i.e. presence of *M. platycheles* detected) while the other two locations were labeled as "Undisturbed." Based on this experimental setup, species richness and species diversity were shown to be significantly reduced in nematode populations following disturbance from *M. platycheles*. On the other hand, the effects of this invasive species were much more subtle in the copepod populations [82]. Although this could be due to physiological differences between copepods and nematodes - nematodes are more sedentary and can be more susceptible to environmental changes - other environmental variables could account for this difference. We applied our variable adjust-

ment method on 56 species and 16 samples. We noted that high separation by location was observed in the PCoA plots (Figure B.1. Since species composition was highly dependent on its area, we treated location as a batch effect.

Our analysis produced mean and variance-adjusted PCs and distance matrices. The unadjusted and mean-and-variance adjusted PCoA plots are displayed in Figure 3.2. Some clustering was observed between disturbed and undisturbed samples [Figure 3.2(a)]. By adjusting out the mean and variance of location, which we treated as a batch effect, the difference between the two conditions became much more prominent [Figure 3.2(b)]. In regards to the distance matrices, we saw poor separation between Undisturbed and Disturbed samples in the unadjusted dendrogram (based on average agglomeration). By adjusting out the confounder, however, we achieved perfect separation, as represented in the dendrogram in Figure 3.2(d). This example demonstrates an important and highly useful feature of our method: by adjusting out the confounding variable, we are able to obtain better separation in the actual variable of interest.

We also performed a similar sub-analysis on copepods given the contrasting results between copepods and nematodes in the previous study by Warwick et. al. Based on the PCoA plot in Figure B.2(a), location heavily influenced the community structure of copepods, especially for Location 3. After adjusting out location, we saw slightly better clustering between Undisturbed and Disturbed populations (Figure B.2(b-c)), although not as striking as the analysis including all species of meiobenthos. In regards to Figure B.3(a-b), our hierarchical clustering analysis revealed near perfect separation between the two conditions after adjusting out the batch effect. Although there may still be important physiological difference between copepods and nematodes, our results, taken together, suggest that adjusting out location does improve the separation between Undisturbed and Disturbed copepod samples.

3.3.3 High variability between international colorectal cancer cohorts

So far, we have demonstrated the utility of our method on visualization techniques. In this colorectal cancer (CRC) data application, we will explore inferential methods for microbiome

data analysis using adjusted distance matrices. According to the World Health Organization, CRC is the third most common cancer worldwide and is the cause of almost a million deaths per year globally. Most incidences of CRC are characterized as sporadic CRC and are associated with age, obesity, diabetes, diet, and a variety of other lifestyle choices. Despite the vast majority ($> 85\%$) of cases arising as sporadic CRC, a small percent of cases have been linked to IBD [83]. Because IBD and several of these risk factors are also implicated in the dysbiosis of the microbiome, it is logical to link cases of CRC with an abnormal microbiota. It is well-established that diet is directly linked to the production of beneficial SCFAs, many of which are microbially produced. In the presence of a high-fat diet, levels of SCFAs become depleted and can result in chronic inflammation that jumpstarts carcinogenesis [61].

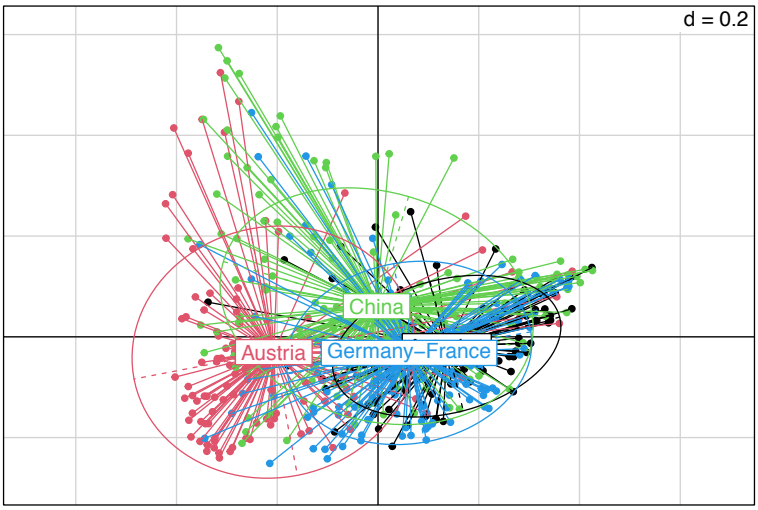


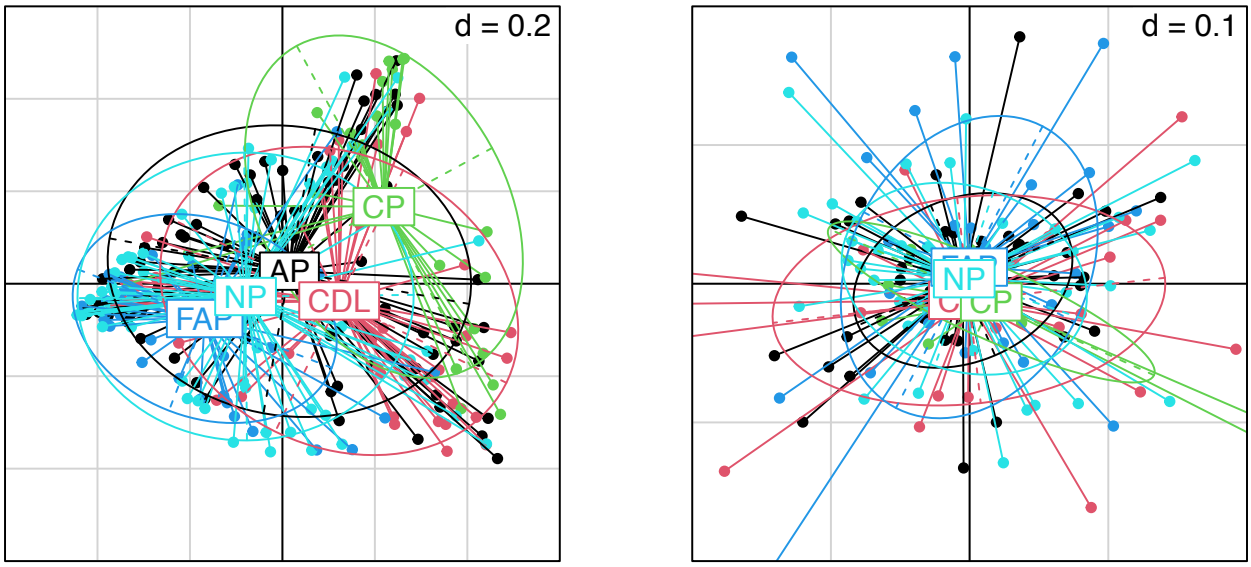
Figure 3.3: PCoA plot of four CRC cohorts from the United States (black), Austria (red), China (green), and Germany/France (blue). 10% of total variance was captured by cohort in the PERMANOVA analysis.

We focus on four CRC metagenomics studies from the United States, Austria, Germany/France, and China. The data processing procedure for all four data sets is outlined in Dai, et. al. [11]. In short, we had 323 microbiome categories and 526 individuals (United States: 100, Austria: 109, China: 165, Germany/France: 152) in our analysis. Summary

statistics of the demographics can be found in Supplementary Table B.1 for each cohort. Figure 3.3 revealed that while the microbiomes of individuals between Germany-France and the United States were similar to each other, they were quite different from the Austrian and Chinese samples. Several rationale, such as differences in culture, diet, study design, or data collection, may explain the some of the deviations we observe in the PCoA plot.

We treated cohort as a batch effect and adjusted its mean and variance out of the distance matrix. Using PERMANOVA, we assessed the variance captured by CRC diagnosis and cohort separately. The variance captured by CRC diagnosis decreased from 0.017 to 0.005 and the variance captured by cohort decreased from 0.105 to 0.01 when cohort effect was adjusted out. This analysis provides two key observations: 1) variation in compositional differences are captured more by cohort than by CRC diagnosis and 2) our adjustment method was successfully able to remove the cohort effect when it was treated as a confounder. This second observation is potentially profound since cohort differences can arise from a wide range of factors including variations in trial design and data collection to cultural and geographical differences between regions.

3.3.4 High-dimensional extension: classification of inflammatory bowel disease (IBD) subtypes



(a) Unadjusted (b) Adjusted

Figure 3.4: PCoA plots from the Pouchitis IBD dataset; AP: acute pouchitis, CP: chronic pouchitis, CDL: Crohn's disease-like phenotype, FAP: familial adenomatous polyposis, NP: no pouchitis

We demonstrate the high-dimensional extension of our method in this last data application. Inflammatory bowel disease (IBD) is a group of disorders characterized by chronic inflammation in the intestines. Marked by gastrointestinal pain among a multitude of other symptoms, IBD is chronic and may lead long term complications like irreversible bowel damage if left untreated. In regards to the microbiome, IBD was one of the first disorders to be characterized by decreased levels of beneficial organisms such as *Prevotella copri* and *Faecalibacterium prausnitzii* [22]. In addition to dysbiosis of the microbiome, patients with IBD also have increased gene expression in inflammation pathways such as autophagy and T-cell activation. Taken together, IBD can be characterized as the dysfunctional crosstalk between the gut microbiota and host mechanisms.

We wanted to evaluate the effect of host gene expression on microbial composition between different subtypes of pouchitis, or inflammation of the ileal pouch, in patients with IBD. We used 255 samples from the Mount Sinai Hospital in Toronto, Canada [54] who had all undergone proctocolectomy with ileal pouch-anal anastomosis at least one year prior to enrollment. In this data set, 76 patients had acute pouchitis (AP), 29 had chronic pouchitis (CP), 45 had Crohn's disease-like phenotype (CDL), 39 had familial adenomatous polyposis (FAP), and 66 had no pouchitis (NP). Both the microbiome and the host transcriptome have been implicated separately in the occurrence of pouchitis but it is unknown how these systems interact with each other. After data quality control, we had 244 samples, 263 OTUs, and 96 inflammation-related genes in the microbiome-host transcriptome study.

The PCoA plots for our analysis are in Figure 3.4. Clustering by subtype was apparent in the unadjusted PCoA plot. Using the high-dimensional extension of our proposed method, we observed no clustering after adjusting out host expression for inflammation genes in PCoA plot. In the PERMANOVA analysis, the p-value for association between subtype and microbial composition goes from 0.001 to 0.775 after adjusting for host inflammation genes. Although clustering by subtype was lost by our method after adjusting out the transcriptomics effects, it contradicts the key findings of the previous study, which stated that local transcriptional activity played a smaller role than early-life colonization on the composition of the microbiome. It is important to that the microbiome does not operate under a vacuum and understanding its interaction with host systems will be vital for developing effective therapies for preventing debilitating health issues in patients with IBD.

3.4 Discussion

Visualization of the microbiome data can be incredibly useful for depicting overall compositional differences of the microbiome between experimental factors or disease conditions. Unfortunately, confounding variables, like batch effects, can obstruct the true relationship between the microbiome and the variable of interest. Thus, we are motivated to develop methods that adjust out the effects of these covariates for the visual analysis of microbiome

data. Instead of a species-by-species approach, we work directly with summary statistics, specifically beta diversity, because of its effectiveness in the presence of high-dimensional data.

We demonstrated the utility of our method on four real data sets by adjusting the mean and variance of the confounding variable out of the PCs. In the MsFLASH example, we showed that treatment effect disappeared when we accounted for the participant's age in our analysis. In the meiobenthos example, we achieved perfect separation between the treatment variable (i.e. Undisturbed and Disturbed) after adjusting out location, which we defined to be a batch effect. We also observed better clustering after our two-step adjustment in the follow-up sub-analysis of the copepod data. As qualified by PERMANOVA,w e illustrated our method's ability to adjust out prominent differences in cohorts in the CRC meta-analysis example. Lastly, we demonstrated the high-dimensional extension of our method on an IBD study. Clustering based on pouchitis subtype decreased substantially after adjusting out host inflammation gene expression.

There are several reasons we worked directly on the distance metrics instead of the original data. 16S rRNA sequencing data, in particular, is discrete, highly zero-inflated, dispersed, and heterogeneous. Typical methods such as MMUPHin assume the data follows a zero-inflated Gaussian distribution, which fails to account for the complex distributional aspects of microbiome data [49]. More importantly, however, simply regressing out the effect of a confounder in the original data can prevent the construction of key statistics used in data visualization. For example, this straightforward approach can potentially change the data from discrete to continuous and introduce non-negative values, thus precluding the calculation of key beta diversity measures. Lastly, the purpose of this method is also field-contextual. Many methods exist for removing batch effects in association testing but not for purposes of visualization.

We note that our proposed method is a first pass attempt to account for unwanted covariates in our PCs. Methods like Covariate Adjusted Principal Coordinates Analysis (aPCoA) proposed a first-moment adjustment of the PCs [68]. As an extension, we wanted to evalu-

ate the performance of our method if we additionally adjusted for the second moment of the confounding variable. We demonstrate this improvement in the MsFLASH data application. By adjusting both the mean and variance effect of age, we were able to completely remove the treatment effect from the PCs. Although we could additionally adjust out the kurtosis as a simple extension of our method, a non-parametric take of this method could instead be preferable. Quantile regression, in particular, can be used to correct the entire conditional distribution of each taxon for each sample through quantile-quantile matching. One particular method, ConQuR (Conditional Quantile Regression) has adapted this framework and serves as a non-parametric extension to approaches that use two-part models to account for zero-inflated outcomes. Unlike our proposed method, ConQuR is computationally intensive and models each taxon individually. Since microbiome data is compositional, this approach may fundamentally change the overall microbiome profile in unintended ways. As demonstrated in Section 4, we still observe substantial results with our parametric method on numerous data applications.

We have presented a method for controlling covariates in the visualization of microbiome data. Batch effects, an example of a potential confounder, can introduce technical variation that obscures the true relationship between the microbiome and our covariate of interest. We have outlined a two-step process for adjusting the PCs of a microbial data set. In the first step, we adjust the mean of the potential confounder out of the PC. In the second step, we assume a linear model between the unwanted covariate and the second moment of each PC. We can adjust out the variance by estimating its coefficient with non-negative least squares. In addition to its applicability in visualization techniques, this method also can be used in inferential approaches such as PERMANOVA. By adjusting out unwanted covariates out of our data, we have the opportunity to accurately visualize the relationship between the microbiome and our variable of interest.

Chapter 4

EVALUATING EFFECT MODIFICATION IN HIGH-DIMENSIONAL DATA

4.1 Introduction

The microbiome plays an integral role in human health and physiology. The colonization of the human microbiome occurs immediately after birth and its resulting composition is dependent of a number of factors such as mode of birth, host genetics, and maternal diet [78]. The environment and the microbiome continue to play interconnecting roles post-infancy. For example, while there is strong evidence connecting the composition of the gut microbiome with the onset of inflammatory bowel disease (IBD), smoking status and age have been associated with more aggressive forms of Crohn's disease [90]. Treatment efficacy can also be affected by the species that reside in our microbiota. A striking example of this occurred during a PrEP study evaluating the efficacy of a vaginal gel in the prevention of HIV infection. The researchers found a specific organism within the vaginal microbiome in participants for whom treatment was ineffective; it was later discovered that this species actually had the ability to metabolize the treatment and render it ineffective [79]. Progress in high-throughput biotechnology has culminated in the development of large scale metagenomics studies wherein a large number of metagenomic markers are screened for association with outcomes of interest. We can gain a more comprehensive understanding of these complex diseases by developing methods for interaction testing between the microbiome and host traits.

Variance component (VC) score test based multi-marker analyses represent a powerful and commonly used class of statistical methods for analyzing high-throughput genomic data. Under this mode of analysis, the joint effect of multiple genomic markers is treated as a

random effect which is then assessed for association with an outcome variable [46]. Assessing the joint effect of multiple markers is often more powerful due to reduced multiple testing burden and ability to aggregate individually modest effects [**?**]. VC models have been shown to be particularly attractive since they make few assumptions regarding the relationship between markers and adapt to correlation in the data. These approaches were first developed for the analysis of gene expression data [21, 46] but have been extended to many other data types including genetics [36, 86], microbiome/metagenomics [96], methylation [80], and others [93]. In fact, VC testing represents the *de facto* standard approach in many analytic pipelines and has been responsible for successfully identifying thousands of associations across a wide range of disease and genomic data types.

Despite the popularity of VC testing within the context of genomics, the majority of VC testing approaches focus on testing main effects of microbial markers. It is often of interest to assess whether the microbial markers interact with an environmental covariate (broadly defined as non-microbiome covariates), i.e. effect modification. An example occurred in a recent microbiome study within the broader Coronary Artery Risk Development in Young Adults (CARDIA) study. In this study, fecal specimens were collected with the objective of understanding how the microbiome is related to systolic blood pressure, a major cardiovascular disease endpoint. After profiling the specimens and pre-processing, 221 bacterial taxa were measured on each individual. Then in the primary analysis, a multi-marker test identified an association between overall microbiome composition and systolic blood pressure, an important factor in cardiovascular disease. Within the microbiome field, this multi-marker testing approach is called β-diversity analysis [98]. Beyond the primary analysis, however, investigators were also interested in understanding whether the demographic factors (Age, Sex and Race) interacted with microbiome profiles. This knowledge could provide clues as to future risk prevention strategies. Unfortunately, the majority of approaches for assessing multi-marker effects have focused on main effects, and existing approaches for interaction analysis may not be appropriate in this context.

A few VC and other tests for interactions between multiple markers and other covariates

have been developed. They often fall into two categories. Under the first category, dimension reduction tools are used to summarize the multiple markers, e.g. using the top principal component(s) [42, 47]. Under the second category, ridge regression is used to control the main effects of the markers (or treat them as a vector of random effects – often equivalent to ridge regression) and then the interaction effect can be assessed using a generalization of VC score tests or alternative approaches [48, 44, 43, 50, 45, 95, 52]. However, a central challenge for all of these approaches is the need to correctly model the main effects of the markers in an unbiased fashion. In particular, in the absence of main effects, all of these approaches are valid with regard to type I error control, but even modest main effects violate the implicit assumptions that many dimension reduction based approaches make regarding the effects of the markers. This results in biased estimation of the main effects of the markers and inflated type I error from incorrectly estimating the null model [44]. On the other hand, ridge regression also fails to provide an unbiased estimate and may also lead to type I error inflation, particularly if the main effects are modest or large as we would expect in the CARDIA study. Many interaction tests developed for SNP data have none or negligible main effects, but for other types of genomic data, such as microbiome data, new methods are necessary.

In this paper, we propose a new marker-environment testing approach that offers improved type I error control while maintaining power. As an alternative to ridge regression based approaches, it has been recently demonstrated that fitting a least-squares model from LASSO (least absolute shrinkage and selection operator)-selected variables can lead to valid inference [97]. With this in mind, we propose a two-step testing procedure for evaluating marker-environment interaction effects on an outcome. In the first step, we perform variable selection on the microbiome data. Because LASSO is guaranteed to select $p < n$ variables, we can proceed with ordinary least squares (OLS) to estimate the remaining coefficients and use the standard VC framework to estimate the effect modification of the microbiome. We will evaluate the performance of our method with more complex kernels, such as the UniFrac kernel and Bray-Curtis kernel. Simulations show that our improved VC score test correctly

controls for type I error whereas existing approaches often have significantly inflated type I error. Importantly, we apply our proposed approach to the CARDIA data set and identify microbiome by covariate interactions that have important biological and eventual translational implications. Although our we apply our approach to microbiome data, we emphasize that the approach is general and can be applied for other, non-microbiome based studies as well. The improved method is also shown to be highly valued in applications for real data testing interactions between microbiome markers and environment variables but is also adaptable for testing gene-environment interactions.

The remainder of this chapter is organized as follows. In the next section, we introduce notation and briefly review existing VC score tests for interactions. Then in Section 3, we present our proposed two-step procedure. We describe the approach and discuss several possible microbiome-specific kernels. In Section 4, we illustrate the advantages of our proposed approach through simulations. The lasso based VC test is then applied to the CARDIA data set in Section 5, and we conclude with a brief discussion in Section 6.

4.2 Variance Component Tests for Interactions

Here, we briefly introduce notation and describe previous VC interaction tests. We assume that we have data from a study with sample size n, where each row of data is independent and identically distributed random vectors $(Y_i, \tilde{\boldsymbol{X}}_i)$ for $i = 1, \ldots, n$. Y_i represents a continuous outcome (BMI in the CARDIA study) lying in the real number space $\mathbb{R}$, and $\tilde{\boldsymbol{X}}_i$ consists of two parts as $\tilde{\boldsymbol{X}}_i = (\boldsymbol{X}_i, \boldsymbol{G}_i)$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{im})$ is a vector of m non-genomic environmental exposures, and $\boldsymbol{G}_i = (G_{i1}, \ldots, G_{ip})$ is a vector of the p genomic markers that we are interested in studying (221 bacterial taxa in CARDIA). We further define $\boldsymbol{Z}_i = \boldsymbol{X}_i \boldsymbol{G}_i = (X_{i1}G_{i1}, \ldots, X_{im}G_{i1}, \ldots, X_{im}G_{ip})$ to be vector marker-environment interaction terms of length $m \times p$ for the i-th sample.

For a simple continuous outcome, we can relate the outcome to the main effects of the environmental variables, the genomic markers, and their interactions using the usual linear

model

$$Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta}_1 + \boldsymbol{G}_i^T\boldsymbol{\beta}_2 + \boldsymbol{Z}_i^T\boldsymbol{\gamma} + \epsilon_i = \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} + \boldsymbol{Z}_i^T\boldsymbol{\gamma} + \epsilon_i \tag{4.1}$$

where $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{im})^T$ is the vector of regression coefficients for the environmental covariates, $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2p})^T$ is the vector of regression coefficients for the genomic markers, and $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{1p}, \ldots, \gamma_{mp})$ is the vector of regression coefficients of $m \times p$ marker-environment interaction parameters. We set $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T]^T$. The error terms ϵ_i follows an independent identically distributions with mean zero and variance σ^2. For clarity, we omit the intercept as well as other variables that might be included in the model (e.g. demographics or other confounders), but these are trivially included within our frameworks.

Since the objective is to test for a marker-environment interaction, this corresponds to assessing H_0: $\boldsymbol{\gamma} = \mathbf{0}$, i.e. for all $j = 1, 2, \ldots, mp$, $\gamma_j = 0$. While this could proceed via a usual p-DF likelihood ratio or other classical test, such tests lose power especially for moderate to large numbers of interaction terms, and even fail due to difficulties in deriving the likelihood when the total number of parameters is larger than the sample size of the data or when there is strong correlation in the data. To overcome this difficulty and to increase power, VC tests assume that each γ_j follows some finite independent distribution with mean 0 and variance τ^2. Hence testing the null hypothesis H_0: $\boldsymbol{\gamma} = \mathbf{0}$ is equivalent to H_0: $\tau^2 = 0$. This null can be assessed by constructing the score statistic

$$Q = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{G}\hat{\boldsymbol{\beta}}_2)^T\boldsymbol{Z}\boldsymbol{Z}^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{G}\hat{\boldsymbol{\beta}}_2) \tag{4.2}$$

where $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are estimates for the main effects of the environment and markers under the null model $Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta}_1 + \boldsymbol{G}_i^T\boldsymbol{\beta}_2 + \epsilon_i$. Under H_0, Q asymptotically follows a mixture of χ^2 distribution.

In early VC tests for main effects, $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ would be the usual least squares estimates. Unfortunately, when the dimensionality of $\boldsymbol{X}$ or $\boldsymbol{G}$ is modest to large or when there is strong

correlation in the data, least squares fails, making it difficult to estimate $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ and obtain Q. Existing approaches use ridge regression estimates for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ or equivalently treat the main effects as vectors of subject specific random effects[48, 44, 43, 45]. While this mitigates some of the challenges associated with estimating the null model, these approaches unfortunately fail to control type I error when p is not small and when there are substantial main effects.

The fundamental difficulty is that ridge regression provides inherently biased estimation (at the trade-off of reduced variance). This bias is negligible in situations where there are no main effects or very small main effects, as in many genetic association studies, but in other types of genomic data, including the motivating microbiome study, main effects can be large, leading to substantially inflated type I error.

4.3 Variable selection via LASSO

4.3.1 Procedure and test statistic

We propose a two-step procedure where we first fit a LASSO model with all the microbiome and environmental markers under the null model. Since we are primarily concerned with the high-dimensionality of the microbiome markers, we place no penalty on the environmental parameters, i.e. the environmental variables are always selected. In the second step, we fit an OLS model with the environmental markers and LASSO-selected microbiome markers. Let $\hat{\boldsymbol{\beta}}_2^{(q)}$ be the estimated coefficients for the LASSO-selected microbiome parameters. Our test statistic takes on a similar form to previous methods:

$$Q = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{G}\hat{\boldsymbol{\beta}}_2^{(q)})^T \boldsymbol{Z}\boldsymbol{Z}^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{G}\hat{\boldsymbol{\beta}}_2^{(q)}) \tag{4.3}$$

Q approximately follows a mixture of chi-square distributions, i.e. $Q \sim \sum_j \lambda_j \chi_1^2$ where $\{\lambda_1, \ldots, \lambda_j\}$ are the eigenvalues of $\mathbf{P}_0^{1/2}\mathbf{Z}\mathbf{Z}^T\mathbf{P}_0^{1/2}$ and
$P_0 = \widehat{\mathbf{V}}^{-1} - \widehat{\mathbf{V}}\mathbf{X}(\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\widehat{\mathbf{V}}^{-1}$; $\mathbf{V} = diag(\sigma^2)$. The p-value for Q can be calculated using Davies method.

4.3.2 Microbiome-specific kernels

Thus far, we have expressed our testing procedure using the linear kernel, $\boldsymbol{Z}\boldsymbol{Z}^T$. In the context of microbiome data analysis, however, it may be more powerful to utilize microbiome-specific kernels derived from beta diversity measures. It is also important to note that microbiome data carry unique features such as sparsity and compositionality. Many microbial species are rare and are only observed in a handful of species. Furthermore, the assumption of true independence between taxa does not hold because of the limited capacity of next generation sequencing equipment; the abundance of one taxa subsequently depends on the abundance of all the other taxa. Simulating from standard distributions such the Normal or Poisson distribution, may not be appropriate for modeling microbiome data.

In our proposed method, we will focus on several popular beta diversity measures: the Bray-Curtis dissimiliarity and UniFrac distance. The choice of kernel is often context-specific and its performance depends on the specific relationship between the taxa and the outcome. The Bray-Curtis dissimilarity is a function of the shared number of species between two samples whereas the UniFrac distance integrates phylogenetic data in its calculation. The UniFrac distance can subsequently be divided into two submeasures. The weighted UniFrac distance takes abundance of the taxa into account, therefore the effect of low abundance species are diminished compared to the unweighted UniFrac distance. In order to construct the kernels, we use the linear kernel for the environmental data, i.e. $\boldsymbol{K}_E = \boldsymbol{X}\boldsymbol{X}^\top$. The microbiome kernel is defined as $\boldsymbol{K}_E$, and is derived by double-centering the beta diversity matrix. We construct the interaction kernel by using Hadamard product of $\boldsymbol{K}_E$ and $\boldsymbol{K}_G$. Using $\boldsymbol{K}$ as a generic kernel function, we can rewrite our test statistic as

$$Q = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{G}\hat{\boldsymbol{\beta}}_2^{(q)})^T \boldsymbol{K} (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{G}\hat{\boldsymbol{\beta}}_2^{(q)}) \tag{4.4}$$

Like with the linear kernel, our test statistic asymptotically follows a mixture of chi-square distributions.

4.3.3 Assumptions

It has been established that data dredging can lead to inflated Type 1 error and narrower confidence intervals [60]. The previously-described method appears to be a clear example of data-dredging since we first fit a LASSO model in order to choose a subset of the coefficients to use for the VC test statistic. However, it has been shown that LASSO-selected variables are deterministic and non-data-dependent under certain assumptions. Zhao et. al. demonstrated that this result yielded asymptotically valid inference through the development of the naive confidence interval and naive score test [97].

Similar to LASSO, this method assumes the irrepresentable condition and the beta-min condition. The beta-min condition requires that all nonzero regression coefficients must be sufficiently large. In the context of our method, this refers to the moderate to large main effects of our taxa. The irrepresentable condition, on the other hand, states that variables that are important, i.e. non-zero and sufficiently large, are uncorrelated with the other variables. Although these assumptions are unverifiable, our goal is not to produce unbiased estimates of the main effects in VC-based interaction tests and we are not interested in inference of the specific $\boldsymbol{\beta}_{2j}$. With regards to the second assumption, microbiome data is inherently compositional due to limitations in current sequencing technologies. Compositionality induces correlation in the data, which is a violation of the irrepresentable condition. However, as the number of species increases, correlation should also decrease. Although, the results from our proposed approach should be interpreted carefully, we show through empirical results in the next section that our method works.

4.4 Simulations

4.4.1 Type 1 Error

We conducted simulations to evaluate type I error of the proposed VC interaction test as well as existing methods. Specifically, we considered our proposed method, the existing ridge regression VC test for interactions approach, as well as a VC test using OLS to estimate the

main effects of the markers — noting that the latter approach cannot be used when the number of markers and environmental variables exceeds the sample size.

For each simulation, we generated data sets with n samples from the linear model where the outcome for each sample was generated as:

$$y_i = X_{1i}\beta_1 + X_{2i}\beta_2 + \boldsymbol{G}_i\boldsymbol{\beta}_2 + \boldsymbol{Z}_i\boldsymbol{\gamma} + \epsilon_i \tag{4.5}$$

The objective is to test for an interaction between the environmental covariates, $\boldsymbol{X}_i$, and the genomic markers, $\boldsymbol{G}_i$. The error terms, ϵ_i, are generated from independent and identical standard normal distribution. For the environmental variables, X_{1i} follows a standard normal distribution and X_{2i} follows a Binomial distribution where $p = 0.5$. Under the null distribution, $\boldsymbol{\gamma} = 0$.

For Simulation Setting 1, we set $p < n$ and allowed n to vary from 50 to 400 and for p to vary from 10 to 40. For each value of n, we conducted 5,000 simulations where 5% of the genomic markers had a true large effect on the outcome. We applied the LASSO based VC interaction test, ridge regression based VC interaction test, and OLS based VC interaction tests to each data set. Type I error was defined as the proportion of p-values less than $\alpha = 0.05$. Results for Simulation Setting 1 for $p = 20$ are presented in Figure 4.1. We see inflated type I error when ridge regression is used to control for the main effects of the genomic markers and this continues to persist even at large sample sizes. Specifically, at $n = 400$, the type 1 error inflation is at 8 times larger than the nominal level. We see similar results of inflated type 1 error for the OLS based VC interaction tests, though to a lesser degree than the ridge regression based tests. As sample size increases, type 1 error from the OLS based tests does appear to become reasonable. On the other hand, our proposed LASSO based method correctly controls from type 1 error even at smaller sample sizes. We do observe type 1 error inflation when $n = 50$ but this deviation is relatively modest, especially compared to the inflation we see in the ridge regression based and OLS based tests.

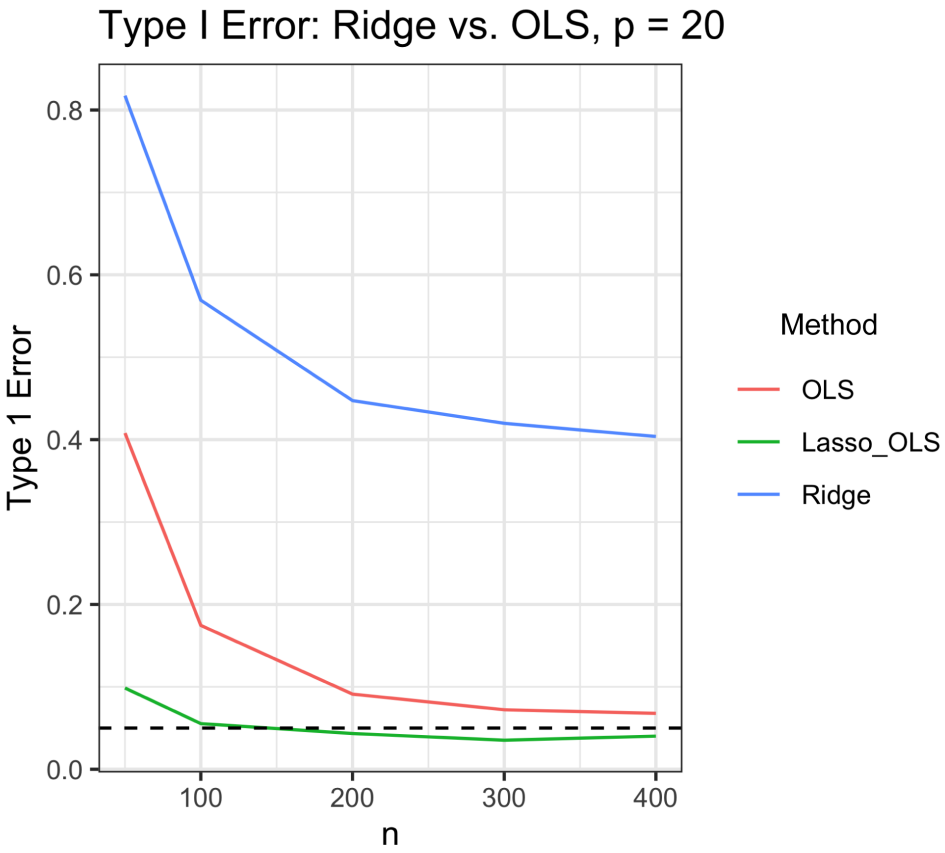


Figure 4.1: Type 1 error rates between ridge regression based VC tests, OLS based VC tests and LASSO based VC tests.

In many study settings, we will often observe $p > n$, especially in microbiome studies where there are often hundreds if not thousands of taxa per subject. For Simulation Setting 2, we considered scenarios where $p > n$. We generated $\boldsymbol{G}_i$ from independent multivariate normal distributions and n independent Poisson distributions for continuous and count data respectively. Under Simulation Setting 2, we set $p = 850$ and varied n from 100 to 250. All other conditions were the same as Simulation Setting 1. Because OLS based approaches were not appropriate under Simulation Setting 2, we only evaluated the performances of the ridge regression based VC interaction tests and the LASSO based VC interaction tests.

n	Normal	Poisson
100	0.0624	0.0492
150	0.0482	0.0486
200	0.0564	0.047
250	0.0448	0.0516

Table 4.1: Type 1 error rates for the LASSO based VC tests on data generated from normal and Poisson distributions

Results for Simulation Setting 2 are presented in Table 4.1. We excluded results from the ridge regression based VC interaction tests since type 1 error was near 1 for all sample sizes and distributions. In the LASSO based VC interaction tests, there is slightly inflated type 1 error for normally distributed data at $n = 100$, although this effect dissipates as sample size increases. We see that our proposed method correctly maintains type 1 error for all sample sizes for the Poisson distributed data.

Thus far, we have extended our LASSO based approach to consider high-dimensional settings. In Simulation Setting 3, we further extend our evaluation by generating data that better mimics data collected from microbiome studies. We simulated microbiome data sets using the general approach outlined by Chen and Li [27]. To account for overdispersion, we simulated taxa data by drawing from a Dirichlet multinomial distribution for each individual. We used upper-respiratory tract microbiome data in order to obtain realistic dispersion parameters and proportion means for each species. Specifically, as outlined by Charlson et al., this data consists of 856 taxa across 60 samples and contains phylogenetic tree information necessary for calculating UniFrac distances [7].

In addition to using the linear kernel for our LASSO based VC interaction test statistic, we also considered more complex kernels that captured nonlinear taxa effects as well as phylogenetic relationships between species. In order to construct the kernels, we defined the environmental kernel as $\boldsymbol{K}_E = \boldsymbol{X}\boldsymbol{X}^\top$, i.e. the linear kernel of the environmental variables,

and the microbiome kernel as $\boldsymbol{K}_E$, which was derived from the UniFrac distance (unweighted and weighted) and the Bray-Curtis dissimilarity measure. The interaction kernel was defined as the Hadamard product of $\boldsymbol{K}_E$ and $\boldsymbol{K}_G$. We simulated 5000 microbiome datasets for $n = 200, 300, 400$ and $p = 856$. Because $n < p$ and the type 1 error rate from the ridge regression based VC interaction tests was close to 1 in Simulation Setting 2, we only evaluated the type 1 error performance of the LASSO based VC interaction tests under the linear kernel, the unweighted UniFrac kernel, the weighted UniFrac kernel, and the Bray-Curtis kernel. The results of Simulation Setting 3 are presented in Table 4.2 and show that all kernels correctly control for type 1 error across all samples.

n	**Linear**	**Weighted**	**Unweighted**	**Bray-Curtis**
200	0.054	0.051	0.057	0.049
300	0.057	0.052	0.056	0.053
400	0.058	0.051	0.049	0.053

Table 4.2: Type 1 error rates for the LASSO based VC tests with the linear, weighted UniFrac (Weighted), Unweighted UniFrac (Unweighted) and Bray-Curtis kernel.

4.4.2 Power

We conducted additional power simulations to show that our proposed method could still maintain reasonable power whilst controlling type I error. We evaluated the performance of our method on $n = 200, 300, 400$ across 5000 simulations. For each simulation, we generated data sets with n samples from the linear model where the outcome for each sample was generated as:

$$y_i = X_{1i}\beta_1 + X_{2i}\beta_2 + \boldsymbol{G}_i\boldsymbol{\beta}_2 + \boldsymbol{Z}_i\boldsymbol{\gamma} + \epsilon_i \tag{4.6}$$

$\boldsymbol{X}$ and $\boldsymbol{G}$ were generated under the same conditions as Simulation Setting 3 in the type 1 error simulations. We focused on this setting so that we could evaluate power across four

different kernels. We hypothesized that microbiome-specific kernels could have greater power because they more accurately captured the potentially complex relationships between taxa. In order to test this hypothesis, we used the partition across medoids (PAM) algorithm to divide the OTUs into 20 clusters. Because PAM uses the cophenetic distance between taxa in its analysis, each cluster corresponded to a potential bacterial lineage. We chose a relatively abundant lineage containing 19% of the OTU counts to be associated with the outcome. Specifically, all species in that lineage had a nonzero value for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. For each nonzero entry j, we set $\beta_j = 1$ and varied values of γ_j from 0.01 to 0.08. Power was estimated as the proportion of p-values less than $\alpha = 0.05$.

Results for the power simulations are presented in Figure 4.2. As expected, power increased with greater sample sizes and values of γ_j. For $n = 200$, the weighted Unifrac kernel performed just as well as the linear kernel while the performance of the other three kernels fell short. With increased sample sizes, as seen in $n = 300, 400$, the weighted UniFrac kernel had the best performance and had greater power over the linear kernel. Additionally and although outperformed by the weighted UniFrac kernel, the Bray-Curtis kernel also had greater power over the linear kernel. Because the simulated outcome was driven by both lineage and abundance of OTUs, the unweighted UniFrac kernel failed to capture these characteristics and had the worst performance out of the four kernels.

4.5 Real Data Application

Our work is motivated by a microbiome sub-study of the broader Coronary Artery Risk Development in Young Adults (CARDIA) study, a cohort with more than 30 years of follow up. Recall that within the primary analysis, an overall association between microbiome composition and systolic blood pressure (SBP) was identified, though no individual bacterial taxa were implicated. This finding was important as the microbiota are modifiable risk factors representing a potential means for reducing cardiovascular risk. However, an unrealized secondary objective was to assess whether the association between microbiome composition and SBP was heterogeneous by the major demographic variables of Age, Sex, and Race.

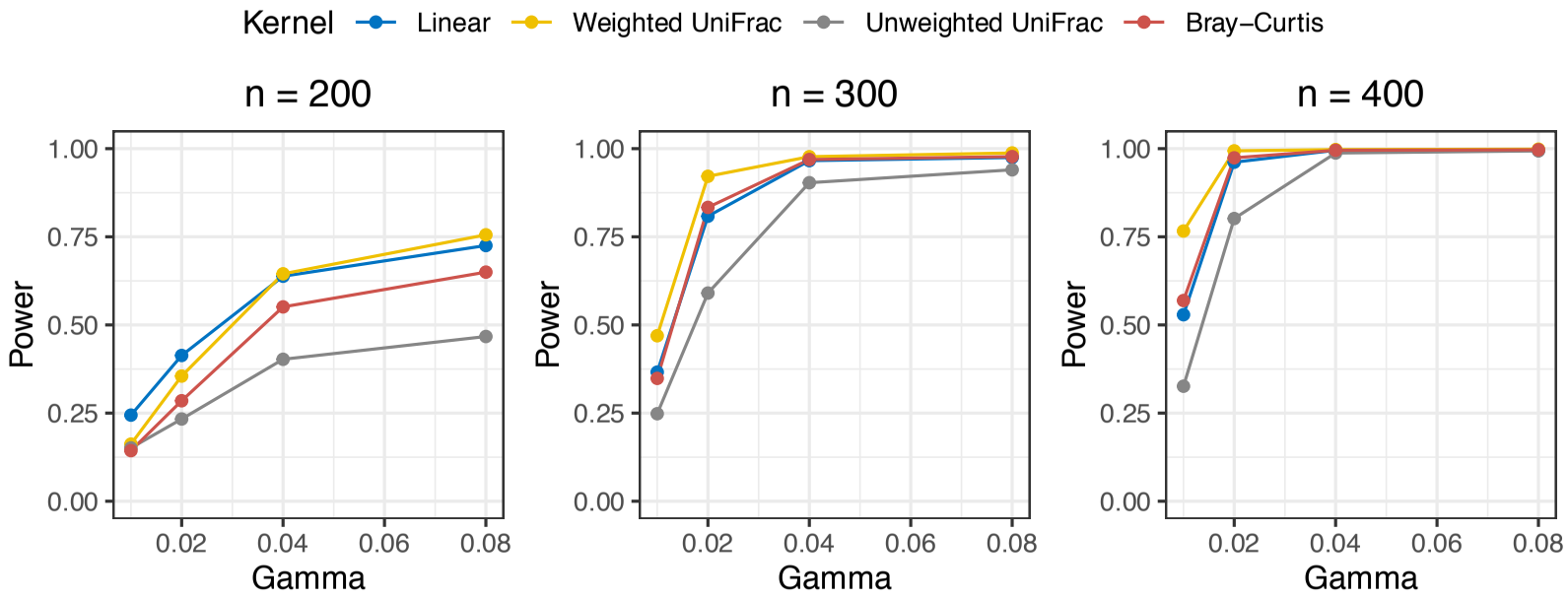


Figure 4.2: Simulation results evaluating the power of the LASSO based VC tests for $n = (200, 300, 400)$ across different interaction effect sizes (Gamma).

Therefore, we applied our proposed method to specifically test for the interaction between bacterial taxa and each demographic variable.

After data quality control, the microbiome sub-study consisted of $p = 221$ bacterial taxa each measured on $n = 633$ individuals. Using the linear kernel for our proposed approach, we found the p-values for the interaction between Age (continuous), Sex, and Race (White vs. Non-White) with the microbiome to be 0.05, 0.27, and 0.03, respectively. For the Bray-Curtis kernel, the p-values for Age, Sex, and Race was 0.01, 0.35, and 0.02 respectively. This indicates that the association between microbiome composition and SBP is different based on age and race, which reflects broader views in the field that age and race specific CVD prevention strategies are needed.

4.6 Discussion

We propose an improved VC score test for multi-marker by environment interaction through a two-step process involving variable selection with LASSO. Compared to existing, classical VC interaction tests based on using OLS or ridge regression to control for main effects, our method maintains type 1 error rates even as dimensionality increases. In the CARDIA study, we were able to identify associations between overall microbiome profiles and key demographic variables on systolic blood pressure. Although we focus on a microbiome studies with continuous outcomes in our simulations and real data application, our approach can be easily extended to many other studies and data types.

While the linear kernel achieves reasonable power, we can increase power of our LASSO based VC test by considering more complex kernels, such as the UniFrac or Bray-Curtis kernel. The choice of kernel is often scenario specific. A recent study compared the performance of several beta diversity measures on four published datasets [69]. Based on this analysis, the Bray-Curtis dissimilarity measure performed well in the presence of highly prevalent and abundant OTUs. On the other hand. the performance of the unweighted UniFrac suffered substantially when an increasing number of zeros in the data were substituted with low count values. Although the overall composition changes minimally, the unweighted UniFrac

distance relies solely on the presence or absence of a species and does not take its abundance into consideration.

The choice of beta diversity measure relies on *a priori* information of how the microbiome relates to the outcome. Knowledge of this, however, would preclude the need for analysis. Several methods have been developed to address this pitfall. Shi et. al [69] proposed a weighted average of the Bray-Curtis dissimilarity and the unweighted UniFrac distance. Under a flat weight, the combined metric outperformed its individual components in a clinical setting where global separation between experimental groups was not expected. Although the authors suggested a flat weight on the Bray-Curtis and unweighted UniFrac measures, prior knowledge on the microbiome data in relation to its outcome is still necessary for best performance. Optimal microbiome regression-based kernel association test (MiRKAT) was developed as an extension to MiRKAT in order to evaluate all possible beta diversity kernels. Under optimal MiRKAT, the test statistic is the minimum p-value from the vector of p-values obtained by each individual kernels. To avoid issues of multiple comparison, residual permutation is then used to generate the null distribution of the minimum p-value. Since all kernels are simultaneously considered, *a priori* knowledge of the data is not necessary for good performance [96]. We can apply the same residual permutation approach to our LASSO based VC test in order to agnostically assess all possible interaction kernels.

So far our proposed method has shown good performance on main effects with large influence on the outcome. However, our performance may potentially suffer in the presence of microbiome markers with small but non-zero effect size. As with our choice of kernel, we do not have *a priori* knowledge on how each taxa influences the outcome. We can slightly modify our two-step approach in order to address this issue. The first step of the method remains the same; LASSO is fit on all the data to select the taxa with large influence. In the second step, we can then fit ridge regression on the same data with a modified penalty matrix, i.e. no penalty terms are placed on taxa chosen by LASSO. This approach may allow us to estimate across a range of different effect sizes for each taxa.

The popularity of microbiome studies has exploded in recent decades due to its impact

on human health. Innovations in sequencing technology has increased the accessibility of collecting microbiome data, thus adding to the demands of these conducting these studies. Because of its interactions with host and environmental factors, it is now pertinent to integrate microbiome data into clinical studies. Previous methods for interaction testing have failed due to inflated type 1 error rates resulting from taxa with modest to large effect sizes. Through simulations and a real data application using the CARDIA study, we show that our LASSO-based VC test maintains type 1 error with good power. By applying this method to current clinical studies, we have the potential to develop more effective treatment strategies for complex human diseases and conditions.

Chapter 5

DISCUSSION AND FUTURE DIRECTIONS

5.1 Summary

There is a mountain of evidence pointing to the microbiome's integral role in human health. As we have outlined in Chapter 1, the composition of the microbiome has been connected to several wide-ranging diseases and conditions. Although improvements in sequencing technology have increased the accessibility of human microbiome studies, several statistical challenges remain in the analysis of this type of data.

In Chapter 2, we proposed a weighted variance component (VC) test for testing the cumulative effects of a metabolic pathway on an outcome of interest. Within the sphere of metabolism, we defined metagenomic data as the metabolic potential and metabolomic data as the metabolic output in this framework. Because we do not know *a priori* the effect each component has on the outcome, we used a grid search approach to calculate a p-value for each weight and presented two methods, the Min-p and Cauchy method, to obtain an overall p-value. Through simulations, we showed that our weighted VC test procedure had greater power than methods that only used one data set while also maintaining type 1 error rate. With regards to the real data application, the Min-p method was able to detect more diabetes-relevant pathways in the HCHS/SOL study compared to approaches that only considered either metagenomic data or metabolomic data.

In Chapter 3, we introduced a two-step procedure for adjusting covariates in the visualization of microbiome data. We focused primarily on principal coordinates (PCs), which are derived from beta diversity measures. In the first step, we projected the mean effect of the covariates out of the PCs. In the second step, we used linear regression to model the association between the variance of the PC and the unwanted covariates. By estimating its

coefficient with non-negative least squares, we adjusted the variance of the unwanted covariate out. Although we focused on adjusting out batch effects, our method could be applied on continuous variables as well. We demonstrated the utility of our method on a menopause trial, an ecological study on invasive species, a meta-analysis of CRC cohorts, and an IBD sub-study on pouchitis.

Finally in Chapter 4, we presented a LASSO based variance component (VC) test for evaluating the interaction between the microbiome and an environmental/host trait on an outcome of interest. The biggest challenge for this type of analysis is estimating the main effects because microbiome data is often high-dimensional. Previous methods employed ridge regression to estimate the taxa coefficients which led to inflated type 1 error when effect sizes were moderate to large. To address this issue, we proposed a two-step procedure where we applied the typical VC test framework on LASSO-selected variables. Through simulations, we showed that our method maintains type 1 error rate compared to OLS and ridge regression based VC tests. Additionally, using non-linear interaction kernels, such as those derived from the UniFrac measure, led to greater power. By applying this method to the CARDIA data set, we found that the interaction between the microbiome and age and race separately were significantly associated with systolic blood pressure.

5.2 Future Directions

The field of microbiome research is rapidly expanding. Consequently, the demand for statistical method development for microbiome applications will increase as experimental designs and clinical trials become more complicated. In this section, we will give an overview on several statistical avenues with great potential for discovering insightful findings in real world applications.

Host genetics and microbiome composition are deeply intertwined [88, 77]. In a cohort of over 1500 healthy individuals, it was estimated that almost a third of all fecal bacterial taxa were heritable. Despite this large proportion, only six SNPs were found to be significantly associated with the relative abundance of specific taxa [77]. Additionally, although

several similar studies have also identified sets of loci associated with microbes, there are few replicated results between each study. A recent *Nature Genetics* editorial exemplified the need for better analytical tools for microbiome genome-wide association studies (mbGWAS). Current statistical tools in mbGWAS are typically underpowered and cannot accommodate heterogeneous microbial traits [1]. For increased power, we can apply the general VC framework to identify individual taxa associated with rare variants or groups of variants. We can then modify this approach with a zero-inflated quantile model in order to address the heterogeneity of microbial abundance.

In addition to finding loci associated with microbial species, it may be of interest to account for relatedness in our data. Many of the current statistical tools used in microbiome studies assume independence between subjects. However, because we know that host genetics are associated with microbial composition, independence may not hold when individuals are related to one another. Methods that account for relatedness in kernel-based genetic association tests can be extended for applications in microbiome studies. GLMM-MiRKAT, an extension of MiRKAT, includes a random effect that models the within-cluster correlation in responses [33]. Although the authors only used a random intercept to represent relatedness between twins in their real data application, we can model genetic relatedness more completely by reformulating the GLMM model. Specifically, it has been suggested that the covariance between random effects can be modeled by the additive and dominant genetic variance components. In the cases of unequal family sizes and genetically diverse relatedness, we can apply a Cholesky decomposition based re-parameterization of the genetic effects in order to achieve identifiable variance components [81].

It has been well-established that the microbiome is dynamic and often fluctuates with time. Longitudinal studies for the microbiome can be advantageous since they can properly capture the variability of host and microbial markers over time. Although we previously referenced GLMM-MiRKAT as a method for controlling genetic relatedness, it can also be applied to longitudinal data. The main purpose of GLMM-MiRKAT is to determine if the microbiome, a single data set, is associated with the outcome of interest while accounting

for repeated measurements. We have increasing access to multiple types of data as sequencing technologies evolve. As a result, it may advantageous to combine the frameworks of GLMM-MiRKAT and the weighted VC testing procedure (Chapter 2) in order to test for the cumulative effects of a metabolic pathway on an outcome in a longitudinal study.

In the previous paragraph, we have described methods where repeated measures are accounted for but are not necessarily the main purpose of the analysis. Some of these compositional fluctuations, however, are important and have been found to be associated with disease outcomes such as bacterial vaginosis and neonate late onset sepsis [37, 71]. In a longitudinal study of IBD, it was found that case subjects had greater volatility of the microbiome, which was defined the as the intra-individual taxa ratios between consecutive time points, than control subjects [8]. Although volatility was defined as the ratio between individual species, it may be more comprehensive to define volatility as the ratio of alpha diversity measures between time points. Furthermore, volatility is a commonly studied subject in the field of economics; it may be useful to extend methods used to detect market volatility to applications relating to the microbiome.

So far, we have only discussed approaches for the inference and visualization of microbiome data. There are several motivations for methods that can predict future microbiomes given the current compositional profile. This incentive is especially important for fecal matter transplants (FMT), in which fecal samples from healthy donors are given to a recipient in order to modify the ecosystem of their gut microbiome. FMT have been shown to be widely effective in treating *Clostridioides difficile* infections but its efficacy in other diseases remains unexplored. Despite its success, the molecular mechanism behind FMT remains relatively unexplored and it is unknown exactly which subset of species are responsible for treating C. difficile infection. FMT is often an unpleasant procedure and isolating these species in a probiotic cocktail may increase the accessibility of this treatment. Furthermore, additional research is necessary to determine if other factors, such as donor-recipient compatibility, are associated with greater improvement of FMT treatment [87].

Predicting microbiome profiles is challenging given the complex chemical and physical

processes that occur within and between both host and microbiome systems. To overcome this obstacle, several methods in machine learning have been proposed for predicting future metabolomic and metagenomic profiles. Deep learning approaches are particularly powerful because they do not necessarily require comprehensive knowledge about species-species interactions, the availability of chemical resources, or the spatial structure of the microbiome. Michel-Mata, et. al., proposed a deep learning framework that uses compositional Neural Ordinary Differential Equation (cNODE) to predict future assemblages based on prior microbial compositions while accounting for zero values in the data. Despite this advantage, cNODE permits only one steady state; this is an unrealistic assumption given the complexity and diversity of the human gut microbiome [53]. Improvements in deep learning methods, such as those with more relaxed assumptions, will be necessary in order to accurately predict the success of modifying the microbiome's ecosystem through far-reaching treatments like FMT.

We have only outlined a small subset of the potential paths for new method development in microbiome studies. The formation of said methods will be timely and important for the future health and safety of our population. In addition to its popularity in the research field, an increasing amount of products are being advertised as ways to improve our microbiome. Despite countless celebrity promotions, probiotic yogurt and supplements have had limited effectiveness in drastically changing the composition of our microbiome. Statistical methods that can accurately assess the impact of clinical treatments and consumer products will be pertinent in order to provide unbiased results that protect our population. In other words, we anticipate that these avenues for new statistical methods will be invaluable in increasing our understanding of the collection of complex ecologic, chemical and physical processes that make up our microbiome.

BIBLIOGRAPHY

[1] Our genes, our microbes. *Nature Genetics*, 54(2):95–95, 2022.

[2] Marti J. Anderson. *Permutational Multivariate Analysis of Variance (PERMANOVA)*, pages 1–15. John Wiley Sons, Ltd, 2017.

[3] S. Ardizzone, S. Bollani, P. Bettica, M. Bevilacqua, P. Molteni, and G. Bianchi Porro. Altered bone metabolism in inflammatory bowel disease: there is a difference between Crohn's disease and ulcerative colitis. *Journal of Internal Medicine*, 247(1):63–70, Jan 2000.

[4] E. W. Beals. Bray-Curtis Ordination: An Effective Strategy for Analysis of Multivariate Ecological Data. *Advances in Ecological Research*, 14:1–55, 1984.

[5] Robert H Berk and Douglas H Jones. Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(1):47–59, 1979.

[6] M. Carabotti, A. Scirocco, M. A. Maselli, and C. Severi. The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. *Annals of Gastroenterology*, 28(2):203–209, 2015.

[7] 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.

[8] A. G. Clooney, J. Eckenberger, E. Laserna-Mendieta, K. A. Sexton, M. T. Bernstein, K. Vagianos, M. Sargent, F. J. Ryan, C. Moran, D. Sheehan, R. D. Sleator, L. E. Targownik, C. N. Bernstein, F. Shanahan, and M. J. Claesson. Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. *Gut*, 70(3):499–510, 03 2021.

[9] M. Cox and T. Cox. *Handbook of Data Visualization*, chapter Multidimensional Scaling, pages 315–347. Springer Handbooks Comp.Statistics, Berlin Heidelberg, 2008.

[10] J. F. Cryan, K. J. O'Riordan, C. S. M. Cowan, et al. The Microbiota-Gut-Brain Axis. *Physiological Reviews*, 99(4):1877–2013, 10 2019.

[11] Z. Dai, S. H. Wong, J. Yu, and Y. Wei. Batch effects correction for microbiome data with Dirichlet-multinomial regression. *Bioinformatics*, 35(5):807–814, 03 2019.

[12] Martha L. Daviglus, Amber Pirzada, and Gregory A. Talavera. Cardiovascular Disease Risk Factors in the Hispanic/Latino Population: Lessons From the Hispanic Community Health Study/Study of Latinos (hchs/sol). *Progress in Cardiovascular Diseases*, 57(3):230–236, 2014. Cardiovascular Diseases in Hispanics.

[13] R. J. DeBerardinis and N. S. Chandel. Fundamentals of cancer metabolism. *Science Advances*, 2(5):e1600200, 05 2016.

[14] M. V. DiLeo, G. D. Strahan, M. den Bakker, and O. A. Hoekenga. Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. *PLoS One*, 6(10):e26683, 2011.

[15] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

[16] Eric A. Franzosa, Katherine Huang, James F. Meadow, Dirk Gevers, Katherine P. Lemon, Brendan J. M. Bohannan, and Curtis Huttenhower. Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences*, 112(22):E2930–E2938, 2015.

[17] S. M. Gibbons, C. Duvallet, and E. J. Alm. Correcting for batch effects in case-control microbiome studies. *PLoS Computational Biology*, 14(4):e1006102, 04 2018.

[18] J. A. Gilbert, M. J. Blaser, J. G. Caporaso, J. K. Jansson, S. V. Lynch, and R. Knight. Current understanding of the human microbiome. *Nature Medicine*, 24(4):392–400, 04 2018.

[19] Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8:2224, 2017.

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

[21] Jelle J Goeman, Sara A Van De Geer, Floor De Kort, and Hans C Van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[22] J. Halfvarson, C. J. Brislawn, R. Lamendella, Y. Vázquez-Baeza, W. A. Walters, L. M. Bramer, M. D'Amato, F. Bonfiglio, D. McDonald, A. Gonzalez, E. E. McClure, M. F. Dunklebarger, R. Knight, and J. K. Jansson. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature Microbiology*, 2:17004, Feb 2017.

[23] J. Han, J. Meng, S. Chen, and C. Li. Integrative analysis of the gut microbiota and metabolome in rats treated with rice straw biochar by 16S rRNA gene sequencing and LC/MS-based metabolomics. *Sci Rep*, 9(1):17860, 11 2019.

[24] S. Hawinkel, F. Mattiello, L. Bijnens, and O. Thas. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief Bioinform*, 20(1):210–221, 01 2019.

[25] B. A. Helmink, M. A. W. Khan, A. Hermann, V. Gopalakrishnan, and J. A. Wargo. The microbiome, cancer, and cancer therapy. *Nature Medicine*, 25(3):377–388, 03 2019.

[26] I. Holmes, K. Harris, and C. Quince. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One*, 7(2):e30126, 2012.

[27] Mingxiu Hu, Yi Liu, and Jianchang Lin. *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, volume 55. Springer Science & Business Media, 2013.

[28] D. Jiang, C. R. Armour, C. Hu, M. Mei, C. Tian, T. J. Sharpton, and Y. Jiang. Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. *Frontier Genetics*, 10:995, 2019.

[29] D. Jiang, C. R. Armour, C. Hu, M. Mei, C. Tian, T. J. Sharpton, and Y. Jiang. Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. *Front Genet*, 10:995, 2019.

[30] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, Jan 2007.

[31] R. R. Kalyani and J. M. Egan. Diabetes and altered glucose metabolism with aging. *Endocrinology & Metabolism Clinics of North America*, 42(2):333–347, Jun 2013.

[32] S. Kim, K. A. Sohn, and E. P. Xing. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25(12):i204–212, Jun 2009.

[33] H. Koh, Y. Li, X. Zhan, J. Chen, and N. Zhao. A Distance-Based Kernel Association Test Based on the Generalized Linear Mixed Model for Correlated Microbiome Studies. *Frontier Genetics*, 10:458, 2019.

[34] A. D. Kostic, R. J. Xavier, and D. Gevers. The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology*, 146(6):1489–1499, May 2014.

[35] James A. Koziol and Michael D. Perlman. Combining independent chi-squared tests. *Journal of the American Statistical Association*, 73(364):753–763, 1978.

[36] Lydia Coulter Kwee, Dawei Liu, Xihong Lin, Debashis Ghosh, and Michael P Epstein. A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2):386–397, 2008.

[37] J. A. Lambert, S. John, J. D. Sobel, and R. A. Akins. Longitudinal analysis of vaginal microbiome dynamics in women with recurrent bacterial vaginosis: recognition of the conversion process. *PLoS One*, 8(12):e82599, 2013.

[38] M. A. E. Lawson, I. J. O'Neill, M. Kujawska, S. Gowrinadh Javvadi, A. Wijeyesekera, Z. Flegg, L. Chalklen, and L. J. Hall. Breast milk-derived human milk oligosaccharides promote Bifidobacterium interactions within a single ecosystem. *ISME Journal*, 14(2):635–648, 02 2020.

[39] S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, D. C. Christiani, M. M. Wurfel, and X. Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, 91(2):224–237, Aug 2012.

[40] J. T. Leek, R. B. Scharpf, H. C. Bravo, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 10 2010.

[41] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

[42] Jia Li, Rui Tang, Joanna M Biernacka, and Mariza De Andrade. Identification of gene-gene interaction using principal components. In *BMC proceedings*, volume 3, page S78. BioMed Central, 2009.

[43] Shaoyu Li, Yuehua Cui, et al. Gene-centric gene–gene interaction: A model-based kernel machine method. *The Annals of Applied Statistics*, 6(3):1134–1161, 2012.

[44] 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.

[45] Xinyi Lin, Seunggeun Lee, Michael C Wu, Chaolong Wang, Han Chen, Zilin Li, and Xihong Lin. Test for rare variants by environment interactions in sequencing association studies. *Biometrics*, 72(1):156–164, 2016.

[46] Dawei Liu, Debashis Ghosh, and Xihong Lin. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9(1):292, 2008.

[47] Meng Lu, Hye-Seung Lee, David Hadley, Jianhua Z Huang, and Xiaoning Qian. Logistic principal component analysis for rare variants in gene-environment interaction analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 11(6):1020–1028, 2014.

[48] Shujie Ma, Lijian Yang, Roberto Romero, and Yuehua Cui. Varying coefficient model for gene–environment interaction: a non-linear look. *Bioinformatics*, 27(15):2119–2126, 2011.

[49] Siyuan Ma, Dmitry Shungin, Himel Mallick, Melanie Schirmer, Long H. Nguyen, Raivo Kolde, Eric Franzosa, Hera Vlamakis, Ramnik Xavier, and Curtis Huttenhower. Population structure discovery in meta-analyzed microbial communities and inflammatory bowel disease. *bioRxiv*, 2020.

[50] Arnab Maity and Xihong Lin. Powerful tests for detecting a gene effect in the presence of possible gene–gene interactions using garrote kernel machines. *Biometrics*, 67(4):1271–1284, 2011.

[51] F. Maleki, K. Ovens, D. J. Hogan, and A. J. Kusalik. Gene Set Analysis: Challenges, Opportunities, and Future Research. *Front Genet*, 11:654, 2020.

[52] Rachel Marceau, Wenbin Lu, Shannon Holloway, Michèle M Sale, Bradford B Worrall, Stephen R Williams, Fang-Chi Hsu, and Jung-Ying Tzeng. A fast multiple-kernel method with applications to detect gene-environment interaction. *Genetic epidemiology*, 39(6):456–468, 2015.

[53] Sebastian Michel-Mata, Xu-Wen Wang, Yang-Yu Liu, and Marco Tulio Angulo. Predicting microbiome compositions from species assemblages through deep learning. *iMeta*, 1(1):e3, 2022.

[54] X. C. Morgan, B. Kabakchiev, L. Waldron, A. D. Tyler, T. L. Tickle, R. Milgrom, J. M. Stempak, D. Gevers, R. J. Xavier, M. S. Silverberg, and C. Huttenhower. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biology*, 16:67, Apr 2015.

[55] A. L. Muhleisen and M. M. Herbst-Kralovetz. Menopause and the vaginal microbiome. *Maturitas*, 91:42–50, Sep 2016.

[56] J. K. Nicholson, E. Holmes, J. Kinross, R. Burcelin, G. Gibson, W. Jia, and S. Pettersson. Host-gut microbiota metabolic interactions. *Science*, 336(6086):1262–1267, Jun 2012.

[57] Jannes Peeters, Olivier Thas, Ziv Shkedy, Leyla Kodalci, Connie Musisi, Olajumoke Evangelina Owokotomo, Aleksandra Dyczko, Ibrahim Hamad, Jaco Vangronsveld, Markus Kleinewietfeld, et al. Exploring the microbiome analysis and visualization landscape. *Frontiers in Bioinformatics*, page 69, 2021.

[58] David Pimentel, Rodolfo Zuniga, and Doug Morrison. Update on the environmental and economic costs associated with alien-invasive species in the united states. *Ecological Economics*, 52(3):273–288, 2005. Integrating Ecology and Economics in Control Bioinvasions.

[59] S. L. Prescott, D. L. Larcombe, A. C. Logan, C. West, W. Burks, L. Caraballo, M. Levin, E. V. Etten, P. Horwitz, A. Kozyrskyj, and D. E. Campbell. The skin microbiome: impact of modern environments on skin ecology, barrier integrity, and systemic immune programming. *World Allergy Organ J*, 10(1):29, 2017.

[60] P. Ranganathan, C. S. Pramesh, and M. Buyse. Common pitfalls in statistical analysis: The perils of multiple testing. *Perspectives in Clinical Research*, 7(2):106–107, 2016.

[61] M. Rebersek. Gut microbiome and its role in colorectal cancer. *BMC Cancer*, 21(1):1325, Dec 2021.

[62] S. D. Reed, A. Z. LaCroix, G. L. Anderson, et al. Lights on MsFLASH: a review of contributions. *Menopause*, 27(4):473–484, 04 2020.

[63] X. Ruan, Y. Cui, J. Du, F. Jin, and A. O. Mueck. Prevalence of climacteric symptoms comparing perimenopausal and postmenopausal Chinese women. *J Psychosom Obstet Gynaecol*, 38(3):161–169, 09 2017.

[64] J. N. Sampson, S. M. Boca, X. O. Shu, R. Z. Stolzenberg-Solomon, C. E. Matthews, A. W. Hsing, Y. T. Tan, B. T. Ji, W. H. Chow, Q. Cai, D. K. Liu, G. Yang, Y. B. Xiang, W. Zheng, R. Sinha, A. J. Cross, and S. C. Moore. Metabolomics in epidemiology: sources of variability in metabolite measurements and implications. *Cancer Epidemiology, Biomarkers Prevention*, 22(4):631–640, Apr 2013.

[65] N. Santoro, C. N. Epperson, and S. B. Mathews. Menopausal Symptoms and Their Management. *Endocrinol Metab Clin North Am*, 44(3):497–515, Sep 2015.

[66] R. Sender, S. Fuchs, and R. Milo. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*, 14(8):e1002533, 08 2016.

[67] Sapna Sharma and Prabhanshu Tripathi. Gut microbiome and type 2 diabetes: where we are and where to go? *The Journal of Nutritional Biochemistry*, 63:101–108, 2019.

[68] Y. Shi, L. Zhang, K. Do, C. B. Peterson, and R. R. Jenq. aPCoA: covariate adjusted principal coordinates analysis. *Bioinformatics*, 36(11):4099—4101, 2020.

[69] Y. Shi, L. Zhang, C. B. Peterson, K. A. Do, and R. R. Jenq. Performance determinants of unsupervised clustering methods for microbiome data. *Microbiome*, 10(1):25, 02 2022.

[70] K. Skowron, J. Bauza-Kaszewska, Z. Kraszewska, N. Wiktorczyk-Kapischke, K. Grudlewska-Buda, J. Kwiecińska-Piróg, E. Wałecka-Zacharska, L. Radtke, and E. Gospodarek-Komkowska. Human Skin Microbiome: Impact of Intrinsic and Extrinsic Factors on Skin Microbiota. *Microorganisms*, 9(3), Mar 2021.

[71] C. J. Stewart, N. D. Embleton, E. C. L. Marrs, D. P. Smith, T. Fofanova, A. Nelson, T. Skeath, J. D. Perry, J. F. Petrosino, J. E. Berrington, and S. P. Cummings. Longitudinal development of the gut microbiome and metabolome in preterm neonates with late onset sepsis and healthy controls. *Microbiome*, 5(1):75, 07 2017.

[72] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, Oct 2005.

[73] B. T. Tierney, Z. Yang, J. M. Luber, M. Beaudin, M. C. Wibowo, C. Baek, E. Mehlenbacher, C. J. Patel, and A. D. Kostic. The Landscape of Genetic Content in the Gut and Oral Human Microbiome. *Cell Host Microbe*, 26(2):283–295, 08 2019.

[74] A. Trikkalinou, A. K. Papazafiropoulou, and A. Melidonis. Type 2 diabetes and quality of life. *World Journal of Diabetes*, 8(4):120–129, Apr 2017.

[75] A. Tripathi, J. Debelius, D. A. Brenner, M. Karin, R. Loomba, B. Schnabl, and R. Knight. The gut-liver axis and the intersection with the microbiome. *Nature Reviews Gastroenterology & Hepatology*, 15(7):397–411, 07 2018.

[76] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The human microbiome project. *Nature*, 449(7164):804–810, Oct 2007.

[77] W. Turpin, O. Espin-Garcia, W. Xu, M. S. Silverberg, D. Kevans, M. I. Smith, D. S. Guttman, A. Griffiths, R. Panaccione, A. Otley, L. Xu, K. Shestopaloff, G. Moreno-Hagelsieb, A. D. Paterson, K. Croitoru, M. Abreu, P. Beck, C. Bernstein, L. Dieleman, B. Feagan, K. Jacobson, G. Kaplan, D. O. Krause, K. Madsen, J. Marshall, P. Moayyedi, M. Ropeleski, E. Seidman, S. Snapper, A. Stadnyk, H. Steinhart, M. Surette, D. Turner, T. Walters, B. Vallance, G. Aumais, A. Bitton, M. Cino, J. Critch, L. Denson, C. Deslandres, W. El-Matary, H. Herfarth, P. Higgins, H. Huynh, J. Hyams, D. Mack, and J. McGrath. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat Genet*, 48(11):1413–1417, 11 2016.

[78] Y. Vandenplas, V. P. Carnielli, J. Ksiazyk, M. S. Luna, N. Migacheva, J. M. Mosselmans, J. C. Picaud, M. Possner, A. Singhal, and M. Wabitsch. Factors affecting early-life intestinal microbiota development. *Nutrition*, 78:110812, 10 2020.

[79] J. Velloza and R. Heffron. The Vaginal Microbiome and its Potential to Impact Efficacy of HIV Pre-exposure Prophylaxis for Women. *Current HIV/AIDS Reports*, 14(5):153–160, 10 2017.

[80] Biqi Wang, Anita L DeStefano, and Honghuang Lin. Integrative methylation score to identify epigenetic modifications associated with lipid changes resulting from fenofibrate treatment in families. In *BMC Proceedings*, volume 12, page 28. BioMed Central, 2018.

[81] Tao Wang, Peng He, Kwang Woo Ahn, Xujing Wang, Soumitra Ghosh, and Purushottam Laud. A re-formulation of generalized linear mixed models to fit family data in genetic association studies. *Frontiers in Genetics*, 6, 2015.

[82] R.M. Warwick, K.R. Clarke, and J.M. Gee. The effect of disturbance by soldier crabs mictyris platycheles h. milne edwards on meiobenthic community structure. *Journal of Experimental Marine Biology and Ecology*, 135(1):19–33, 1990.

[83] Jürgen Weitz, Moritz Koch, Jürgen Debus, Thomas Höhler, Peter R Galle, and Markus W Büchler. Colorectal cancer. *The Lancet*, 365(9454):153–165, 2005.

[84] David S. Wilcove, David Rothstein, Jason Dubow, Ali Phillips, and Elizabeth Losos. Quantifying Threats to Imperiled Species in the United States: Assessing the relative importance of habitat destruction, alien species, pollution, overexploitation, and disease. *BioScience*, 48(8):607–615, 08 1998.

[85] M. C. Wu and X. Lin. Prior biological knowledge-based approaches for the analysis of genome-wide expression profiles using gene sets and pathways. *Statistical Methods in Medical Research*, 18(6):577–593, Dec 2009.

[86] 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.

[87] Y. Xiao, M. T. Angulo, S. Lao, S. T. Weiss, and Y. Y. Liu. An ecological framework to understand the efficacy of fecal microbiota transplantation. *Nature Communications*, 11(1):3329, 07 2020.

[88] Hailiang Xie, Ruijin Guo, Huanzi Zhong, Qiang Feng, Zhou Lan, Bingcai Qin, Kirsten J. Ward, Matthew A. Jackson, Yan Xia, Xu Chen, Bing Chen, Huihua Xia, Changlu Xu, Fei Li, Xun Xu, Jumana Yousuf Al-Aama, Huanming Yang, Jian Wang, Karsten Kristiansen, Jun Wang, Claire J. Steves, Jordana T. Bell, Junhua Li, Timothy D. Spector, and Huijue Jia. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Systems*, 3(6):572–584.e3, 2016.

[89] H. Yamamoto, T. Fujimori, H. Sato, G. Ishikawa, K. Kami, and Y. Ohashi. Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis. *BMC Bioinformatics*, 15:51, Feb 2014.

[90] A. J. Yarur, S. G. Strobel, A. R. Deshpande, and M. T. Abreu. Predictors of aggressive inflammatory bowel disease. *Gastroenterol Hepatol (N Y)*, 7(10):652–659, Oct 2011.

[91] H. Zeng, S. Umar, B. Rust, D. Lazarova, and M. Bordonaro. Secondary Bile Acids and Short Chain Fatty Acids in the Colon: A Focus on Colonic Microbiome, Cell Proliferation, Inflammation, and Cancer. *Int J Mol Sci*, 20(5), Mar 2019.

[92] X. Zhan, A. D. Patterson, and D. Ghosh. Kernel approaches for differential expression analysis of mass spectrometry-based metabolomics data. *BMC Bioinformatics*, 16:77, Mar 2015.

[93] Xiang Zhan, Andrew D Patterson, and Debashis Ghosh. Kernel approaches for differential expression analysis of mass spectrometry-based metabolomics data. *BMC Bioinformatics*, 16(1):77, 2015.

[94] A. Zhang, M. Matsushita, L. Zhang, H. Wang, X. Shi, H. Gu, Z. Xia, and J. Y. Cui. Cadmium exposure modulates the gut-liver axis in an Alzheimer's disease mouse model. *Commun Biol*, 4(1):1398, 12 2021.

[95] G Zhao, R Marceau, D Zhang, and JY Tzeng. Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. *Genetics*, 199(3):695–710, 2015.

[96] 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.

[97] Sen Zhao, Daniela Witten, and Ali Shojaie. In Defense of the Indefensible: A Very Naive Approach to High-Dimensional Inference. *Statistical Science*, 36(4):562 – 577, 2021.

[98] Xuan Zhu, Jian Wang, Cielito Reyes-Gibby, and Sanjay Shete. Processing and analyzing human microbiome data. In *Statistical Human Genetics*, pages 649–677. Springer, 2017.

Appendix A

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

In this section, we show the power simulation results for the weighted VC method when we have a dichotomous outcome. This section is adapted from the continuous example in the main text.

$$\text{logit}[\text{P}(Y_i - 1)] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + W_i\boldsymbol{\beta} + \gamma z_i \tag{A.1}$$

For the first scenario, we assessed the performance of the metabolic potential while keeping the effects of the metabolic output constant. We set a constant value for $\approx 20\%$ of $\boldsymbol{\beta}$ while the remaining 80% of coefficients were set to 0. We first tested the performance of the weighted VC method by varying the effects of γ. In the second scenario, we evaluated the performance of the method by keeping the metabolic potential constant and varying the effects of the metabolic output. We set a constant value for γ based on the first scenario and calculated the power of the method over a range of different constant $\boldsymbol{\beta}$ values. We compared the power of the weighted VC test, using both integrated methods (Min-p and Cauchy) to methods that just used one dataset. Since the metagenomics data for each pathway contains just value for each sample, we used a simple logistic regression model for the metagenomic data in the continuous scenario:

$$\text{logit}[\text{P}(Y_i - 1)] = \alpha_0 + \alpha_1 X_1 + \gamma z_i \tag{A.2}$$

We modeled the association between the metabolites and outcome as:

$$g[(Y_i)] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + W_i\boldsymbol{\beta} \tag{A.3}$$

In contrast of the metagenomics-only analysis, we used a variance component test for the metabolomics-only analysis since there can be a modest to large number of metabolites in a pathway. We assumed that each β_j followed an arbitrary distribution with mean 0 and variance τ. Our test statistic for evaluating the null hypothesis: $\tau = 0$ was:

$$Q = \rho(Y - \hat{\boldsymbol{\mu}})^T \mathbf{W}\mathbf{W}^\top (Y - \hat{\boldsymbol{\mu}}) \tag{A.4}$$

Q asymptotically follows a mixture of χ^2 distributions. Specifically $Q \sim \sum_{i=1}^{n} d_i \chi_1^2$ where d_i are the eigenvalues of $\boldsymbol{\Sigma}^{1/2}\boldsymbol{P}_n\mathbf{W}\mathbf{W}^\top\boldsymbol{P}_n\boldsymbol{\Sigma}^{1/2}$ with $\boldsymbol{\Sigma} = \text{cov}(\boldsymbol{Y})$. The Davies method, which involves the numerical inversion of the characteristic function, can be used to obtain a p-value for the test statistic. In short, our procedure for assessing the power of a metabolomics-only analysis is a simplified version of the proposed weighted VC test.

As shown in Figure A.1, our results from Scenario 1 are largely the same as the continuous outcome example. For map00564, the metabolomics-only analysis only outperformed the integrated methods when there was no metagenomics effect, i.e. $\gamma = 0$. However, for $\gamma > 0.2$ the integrated methods consistently outperformed the performance of the metagenomics-only and metabolomics-only analysis for all sample sizes. With regards to the integrated methods, the Cauchy and Min-p approach had similar performance unless the sample size was small and the effect size of γ was large or small. Recall that the Min-p method relies on one $p-$value while the Cauchy method is an average of $p-$values. Similar results were also observed in the map02010 pathway.

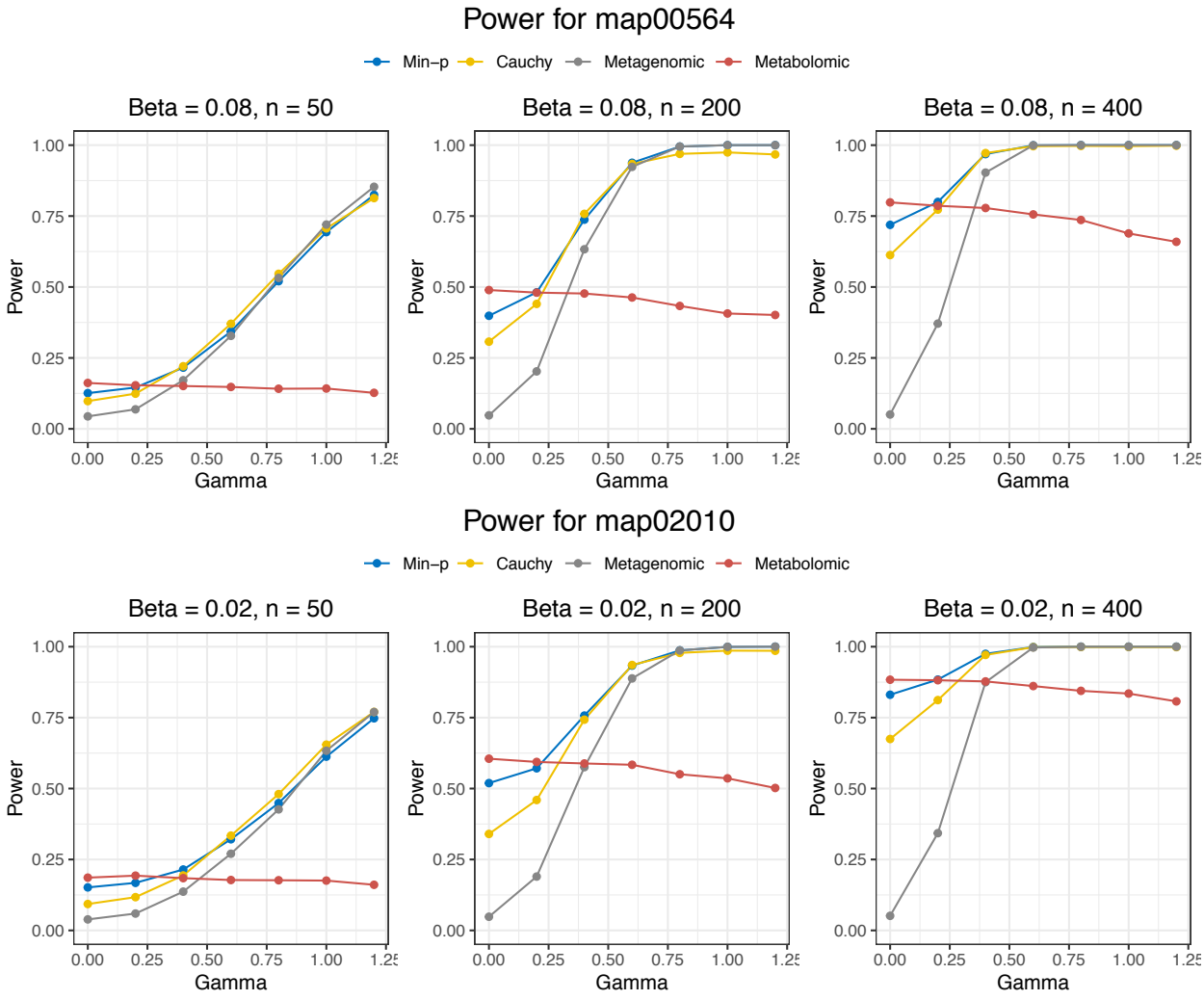


Figure A.1: Simulation results for evaluating power in Scenario 1 with a dichotomous outcome. The effects of the metabolomics input is kept constant as we evaluate the impact of increasing the effects of the metagenomics input for $n = (50, 200, 400)$ and for pathways map00564 and map02010. Power is defined as the proportion of $p-$values ≤ 0.05

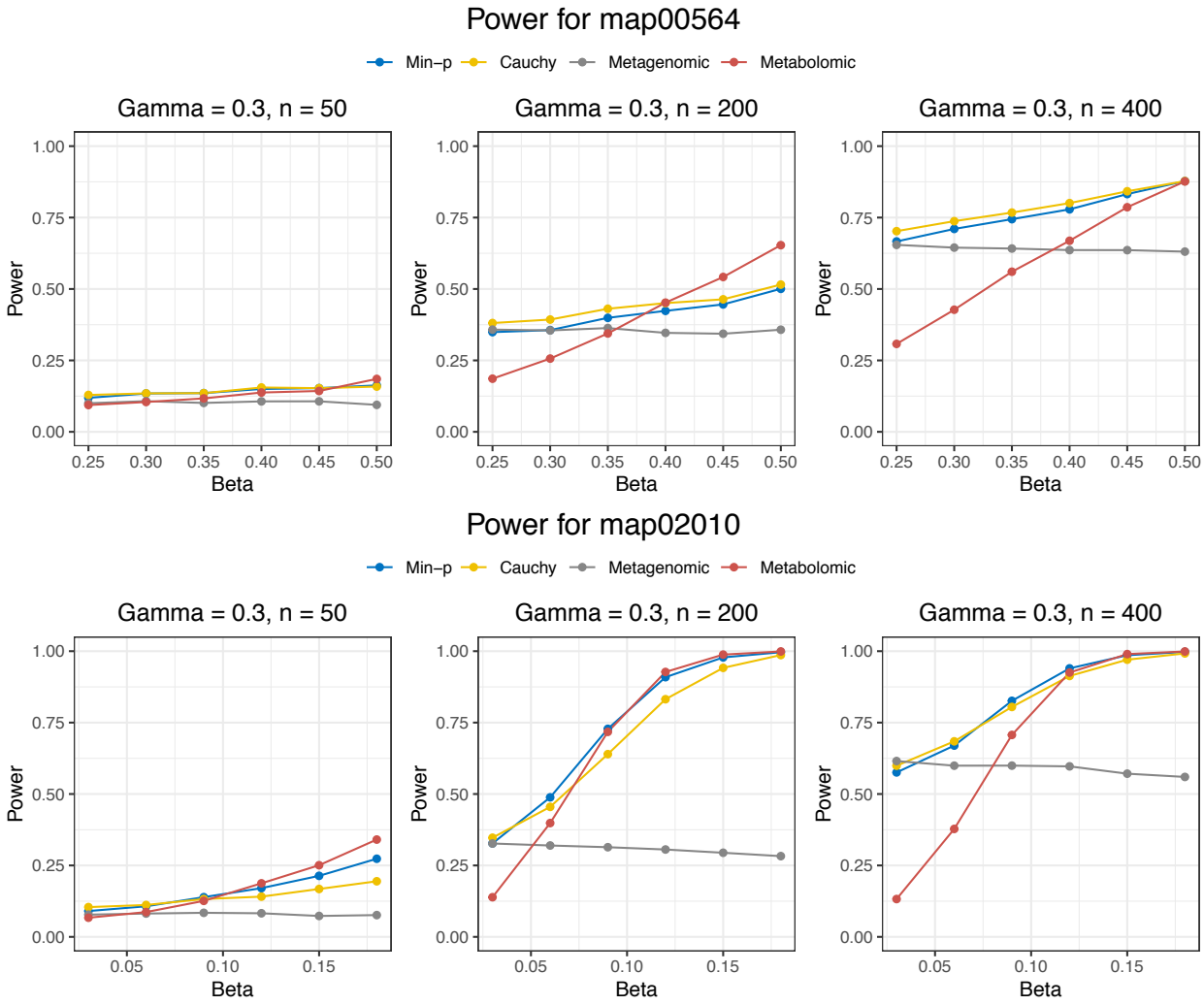


Figure A.2: Simulation results for evaluating power in Scenario 1 with a dichotomous outcome. The effects of the metagenomics input is kept constant as we evaluate the impact of increasing the effects of the metabolomics input for $n = (50, 200, 400)$ and for pathways map00564 and map02010. Power is defined as the proportion of $p-$values ≤ 0.05

For the second scenario, we chose a value of γ that had moderate power in Scenario 1 and then varied the effects of $\boldsymbol{\beta}$. The results of Scenario 2 are outlined in Figure A.2. Although the performances of the integrated methods increase as $\boldsymbol{\beta}$ increases, the metabolomics-only analysis has the largest power at when $\boldsymbol{\beta}$ is large since the contribution between the metabolomics

data and metagenomics data is unequal. We note that the increase in performance in the metabolomics-only approach is minimal and requires prior knowledge that the metabolites have a greater effect on the outcome than the metagenes. For an agnostic approach, it is still preferable to use either of the integrated methods.

Appendix B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

B.1 Additional figures and tables from the Meiobenthos study in Tasmania

Sixteen samples were collected from four locations around Tasmania in order to determine if the presence of *M. platycheles* disturbed the community structure of Meiobenthos species. Although the covariate of interest in this analysis was presence of *M. platycheles*, the location of the sample strongly influences its species composition (Figure B.1)

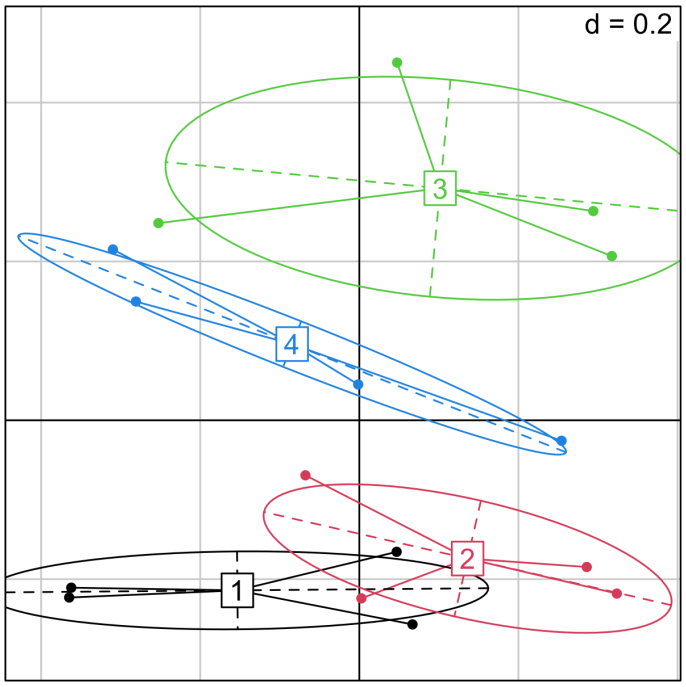
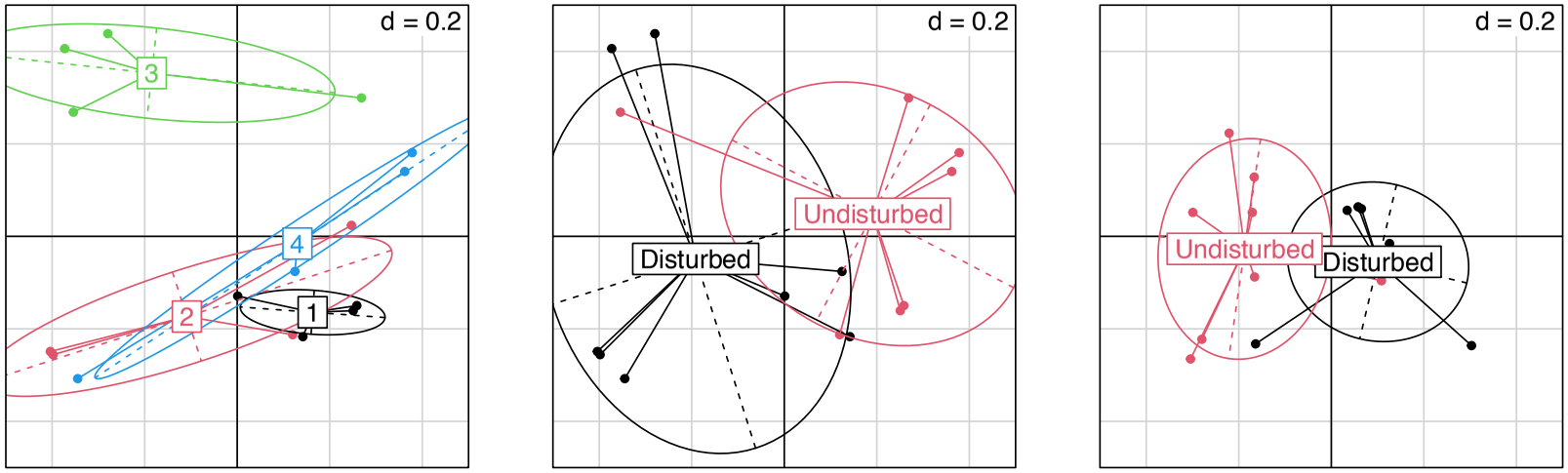
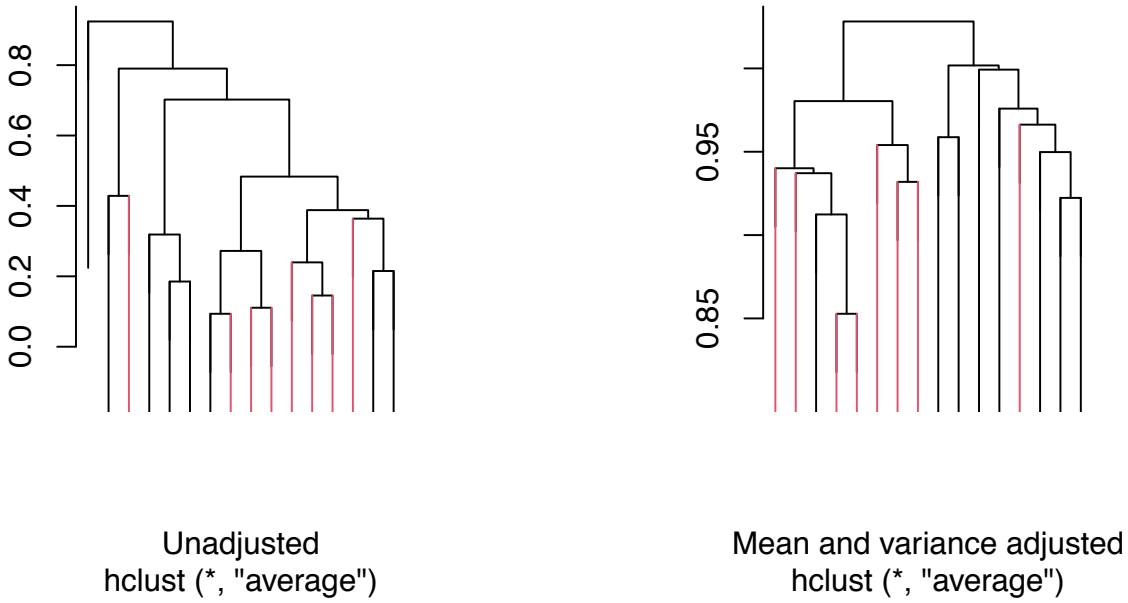


Figure B.1: PCoA plot by location in the Tasmania ecological data set as described in Warwick, et. al. [82]. Since clustering is observed by location, we will treat it as a batch effect.



(a) Batch effect of location (b) Unadjusted (c) Adjusted

Figure B.2: PCoA plots depicting the separation between copepods in Undisturbed and Disturbed locations. This data is from the Tasmania ecological data set as described in Warwick, et. al. [82]



(a) Unadjusted

(b) Adjusted

Figure B.3: Dendrograms depicting the separation between copepods in Undisturbed and Disturbed locations. This data is from the Tasmania ecological data set as described in Warwick, et. al. [82]

We performed a subanalysis on copepod data. As seen in Figure B.2(a), location also strongly influenced copepod composition. We saw better clustering by disturbance, our variable of interest, after adjusting out location (Figure B.2(b-c), Figure B.3).

B.2 Additional tables from the CRC meta-analysis

n	United States		Austria		China		Germany/France	
	Control	CRC	Control	CRC	Control	CRC	Control	CRC
Sample Size	52	48	63	46	92	73	64	88
Age	61.2 (11.0)	61.0 (13.5)	67.1 (6.4)	67.1 (10.9)	58.5 (7.6)	65.9 (10.6)	58.8 (13.0)	64.4 (12.2)
Gender (% Female)	28.8%	46.4%	41.2%	39.1%	44.6%	35.6%	50%	39.8%
BMI	25.4 (4.3)	24.9 (4.3)	27.6 (3.8)	26.5 (3.5)	23.9 (3.3)	24.1 (3.2)	24.7 (3.2)	25.9 (4.3)

Table B.1: Demographic summary statistics of the CRC cohorts. Continuous variables are presented as: Mean (standard deviation)