Statistical Issues in Microbiome Data Analysis: Batch Effects and Multi-Omics Analysis

Weijia Fu

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

Master of Science

University of Washington

2019

Committee:

Michael Wu (Chair)

Timothy Thornton

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Statistical Issues in Microbiome Data Analysis: Batch Effects and Multi-Omics Analysis

Weijia Fu

Chair of the Supervisory Committee:
Michael Wu
Biostatistics

Progress in high throughput sequencing has facilitated the conduct of large scale microbiome profiling studies which have already begun to elucidate the role of microbes in many disorders and clinical outcomes. Despite the many successes, statistical analysis of data from these studies continues to pose a challenge. In the thesis, we proposed methods to study two specific challenges: batch effects and integrative analysis of microbiom and other -omics data. Both issues are increasingly relevant problems. As studies get larger, batching becomes inevitable and integrative analysis is imperative for gaining clues as to the mechanisms underlying discovered associations. The thesis is composed of two projects. In the first project, we compared six existing batch correction methods for microarray data when applied to microbiome data. Two real microbiome data sets were used to evaluate the performance using data visualization and several evaluation metrics. Our results suggest that an empirical bayes approach (ComBat), when applied appropriately, can outperform other methods. In the second project, we proposed a robust microbiome regression-based kernel association test (MiRKAT-R) to screen a large number of genomic markers for association with microbiome profiles. This approach utilizes a recently developed robust kernel machine test. We further propose to incorporate an omnibus test that simultaneously considers different models so as to allow for different relationships between the individual markers and microbiome composition. Systematic simulations and applications to real data show that the MiRKAT-R improves both type I error control and power.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor, Michael Wu, for his guidance, support and encouragement during my thesis project. I would also like to thank my committee member, Timothy Thornton, for serving as my committee member.

DEDICATION

to my family and friends

Chapter 1

INTRODUCTION

1.1 Microbiome Data

The human body is inhabited by a complex microbial community called the human microbiome [54]. Although primarily composed of bacteria, the microbiome also includes archaea, fungi, and viruses. Recently, profiling of the organisms comprising the human microbiome has been made possible through developments in next generation sequencing technology. High throughput microbiome profiling studies have now identified associations between microbiome composition and a wide range of human disease and traits including cancer, HIV, menopause, blood pressure, and others [50, 20, 43, 53].

Two technologies are commonly used for bacterial microbiome profiling: amplicon sequencing of the 16s rRNA gene (often referred to as 16s sequencing) and shotgun metagenomic sequencing [18, 29]. 16s sequencing involves amplifying and sequencing variable regions of the 16s rRNA gene which is ubiquitous across bacterial species, but hypervariable regions are heterogeneous across taxonomic categories allowing for identification of different bacterial taxa [27]. Traditionally, the 16S rRNA sequence reads are either mapped to an existing phylogenetic tree in a taxonomic-dependent way or clustered into operational taxonomic units (OTUs) at a certain similarity level — 97% similarity corresponds approximately with known species. More recently, an alternative has also been to use exact sequence variants as an alternative to OTUs [5]. References are used to assign taxonomic names related sequences. Shotgun metagenomic sequencing differs from 16s sequencing in that all DNA within a sample is sequenced rather than specific regions. Taxonomic profiling is again possible to identify the bacterial taxa present (possibly with higher resolution), but a separate advantage is that by mapping to appropriate references, the gene contents of the microbiomes are also made available, enabling functional analysis. This thesis focuses primarily on issues surrounding taxonomic profiling, i.e. abundances of bacterial taxa comprising the microbiome.

1.2 Statistical Analysis of Microbiome Data

Following sequencing and usual bioinformatics pipelines, a common starting point for down-stream analysis are bacterial count tables in which the total reads for each taxon are available for each sequenced sample. Subsequent down stream analyses include standard statistical objectives including exploratory analysis and visualization, correlative association studies, biomarker identification, and development of prediction and classification models.

In contrast to other areas of genomics, achievement of these objectives can be challenging due to the characteristics of microbiome data. Firstly, the microbiome data are heavily zero-inflated: the read counts are zero for many low frequency taxa. The sparsity reduces power, challenges distributional assumptions, and makes more sophisticated modeling hard. Secondly, in order to normalize for library size (different total numbers of reads per sample), the microbiome data are often compositional, which means there are constraints on the sum of the abundances for each individual. For example, to account for different total number of reads, one may choose to divide by the total read count for each sample such that the measurement for each taxon on an individual add to one (representing relative abundance). Thirdly, the microbiome data are high-dimensional. For each sample, there may be reads for hundreds to thousands of microbes. Finally, the data tend to be highly structured: microbial communities are highly connected from a functional perspective, a co-occurence perspective, and a phylogenetic perspective. Due to these collective challenges, standard statistical analysis methods may not be appropriate, particularly as there are also analytic objectives that are reflect objectives from the varied fields that have converged upon microbiome (epidemiology, microbiology, ecology, etc.). New approaches are constantly needed.

Currently, different groups and projects approach microbiome analysis very differently, depending on both the study objectives as well as conventions. On the correlative and association analysis front, common modes of analysis are individual taxon analysis. Within individual taxon analysis, the association between each taxon and a variable of interest (outcome or exposure) is assessed, one-by-one. The taxon measurement can be on the count scale, in which count-based methods are used [42, 37, 61]. More commonly, however, the data are normalized and/or transformed. The data can be normalized to many different quantities including total abundance, cumulative sums [45], or to the abundances of other taxa [39]. Compositional transformations such as center log-ratio transforms are often useful for mitigating compositional effects [1, 15]. Then one could apply standard statistical tools, e.g. t-test or Wilcoxon test, or more sophisticated tools. Different approaches have been

found to be optimal under different situations [55]. After obtaining a p-value for each taxon, multiple testing adjustments are applied. Common adjustments include false discovery rate control using Benjamini-Hochberg method [4] as well as tailored approaches [57, 22].

An alternative to individual taxon level analysis is community level analysis using beta diversity metrics, wherein overall microbiome composition is studied in relation to individual variables or outcomes [33, 53]. Focusing on overall composition provides a holistic view that emphasizes global community structure. Statistically, this facilitates identification of systematic community shifts, accommodates correlation among taxa, and harnesses phylogenetic relationships. The strategy often enjoys improved power over individual taxon analysis through reduced multiple testing burden and the ability to aggregate modest effects, e.g., when multiple taxa have individually modest, but concerted, effects [31]. Community level analysis often involves constructing a matrix of pairwise dissimilarities between communities [38, 7, 47]. Then, the top principal coordinates may be used for quantitative analysis, or the entire dissimilarity matrix can be tested for association with the outcome using methods like PERMANOVA [2]. Recently, the MiRKAT family of approaches was developed as a generalization of PERMANOVA that better allows for covariate/confounder adjustment, complex outcomes, and accommodation of multiple candidate dissimilarities [62, 59, 46].

A simpler alternative to beta-diversity analysis is alpha-diversity analysis which focuses on within-sample diversity. Then for each sample, a measure of the intrinsic diversity of the profiled community is obtained (e.g. richness, Shannon entropy, Simpson index, etc.). The alpha diversity is a univariate variable which can then be assessed for association with outcomes.

1.3 Outstanding Challenges: Batch Effects and Multi-Omic Analysis

Besides the statistical challenges for microbiome data analysis we mentioned above, there are also other challenges. In the following two part, we will talk about the batch effects and the problems we will fact for association test between microbiome composition and genomic data.

1.3.1 Project 1: Batch Effects

A central challenge of modern microbiome studies, particularly as sample sizes increase, is the issue of batch effects, which are systematic differences between samples in a study arising from differential sample or data processing [49]. Batch effects are ubiquitous across all areas of genomics and are induced when samples are processed under different settings.

Sources of batch effects include the different experiment times, processing sites, reagents and platforms. Since hundreds of samples may be needed for most studies, batch effects are almost inevitable because of technical and time constrains [11]. Batch effects can simultaneously lead to spurious findings or obfuscate true signals. Correction for batches is imperative for drawing accurate inference.

Despite the importance, little has been done on studying batch effects and computational corrections within the context of microbiome data. Therefore, in Project 1 of this thesis, we examine batch correction methods used within other contexts (particularly gene expression analysis) and assess their utility for microbiome sequencing data using real data sets.

1.3.2 Project 2: Multi-Omic Analysis

Despite the plethora of associations between microbiome and outcomes, the specific mechanisms by which the microbiome influences these conditions and health outcomes remain unclear. To this end, many studies are now interested in integrating other types of genomic data such as metabolomics [41], gene expression [44], and DNA methylation [12] into microbiome studies. These other genomic markers can serve as intermediaries between microbiome composition and outcomes and may provide clues as to the specific manner by which microbes drive subsequent processes and disease.

Recognizing that integrative analysis represents a broad, multi-faceted objective, we focus particularly on the problem of community level analysis. Specifically, in Project 2, we consider the problem of community level analysis of microbiome data with a large number of genomic outcomes. We adapt robust kernel-based strategies to rapidly screen a large number of individual genomic markers for association with microbiome composition at the community level.

Chapter 2

ASSESSMENT OF BATCH CORRECTION PROCEDURES FOR MICROBIOME PROFILING STUDIES

Batch effects are not unique to microbiome data and pose a challenge in everything from proteomics [3] and gene expression studies to DNA methylation. Correction procedures are commonly applied within these other -omics studies and have been found to be useful in mitigating batch effects [30]. The earliest approaches were based on scaling and centering data relative to batches – depending on the scale of the data being examined. For example, mean-centering (via PAMR) centers the mean of each batch to zero [51]. Extending these are approaches that involve estimating batch effects which are then regressed out using simple linear models (e.g. limma [52]). Later, more sophisticated approaches were developed, including the popular ComBat method [23] which represents the standard for gene expression and is also commonly used for DNA methylation data. ComBat also estimates the batch effect, but uses an empirical bayes approach which allows for estimation of the effect of batch on a particular feature while harnessing the other features under consideration. Ber is based upon an extension of the model used by ComBat for the estimation and removal of the batch effects using a two-stage regression procedure [19].

The limited work on microbiome data has focused on strategies that correct for batch effects at the association step [13] and in some cases require case-control designs [17]. Yet, it would be of considerable interest to develop a strategy that allows for general removal as association analysis is not necessarily the end goal. Additional objectives such as clustering, visualization, or prediction would benefit from such general batch adjustment. On the other hand, while little has been developed specifically for microbiome data, a range of batch correction procedures have been developed for other genomic data types. Whether these approaches are appropriate for microbiome data remains unclear and as the application of these approaches to microbiome data has not been systematically investigated.

The objective of this project is to fit this critical gap in the literature and systematically compare the performance of traditional batch effect removal methods when applied to microbiome data. We particularly choose to use two real data sets on which to benchmark

comparisons. In the first data set constitutes a set of samples from the Mid-Western Reference Panel (MWRP) [10]. These samples were re-processed and re-sequenced multiple times separately, effectively creating several batches. Thus, we apply different batch correction procedures and then assess the ability of the removal approaches in terms of maximizing reproducibility of abundances across batches for replicated samples. The second data set is from a large scale microbiome-cardiovascular health study which included three sequencing batches [53]. In addition to visualization, we assess variability resulting from batch effects and compare the performance of existing approaches for removing the batch induced variation though inclusion of outcome measurements.

2.1 Data normalization and Batch Correction Methods

In this section, we first describe two approaches for normalizing the data (prior to correction) followed by several batch correction approaches that we choose to compare. We then briefly describe the metrics for assessing batch correction performance.

2.1.1 Data normalization

Data normalization and transformation are necessary for accommodating different library size and variance stabilization.

Notationally, let $\widetilde{Y}_{ih}$ represent the raw count value for taxon h and sample i where $h \in 1, \ldots, H$ and $i \in 1, \ldots, n$.

Log relative abundance transformation (LRA)

The LRA transformation is given by

$$LRA(\widetilde{Y}_{ih}) = log\left(\frac{\widetilde{Y}_{ih}+1}{\sum_{h=1}^{H}\widetilde{Y}_{ih}+1}\right)$$

Under LRA, each taxon for each individual is scaled by the total count for the individual. Then the log transformation is used to reduce the impact of heavy tails. Since microbiome data is zero-inflated, we add one to every value in OTU table before taking LRA transform.

Centered log ratio transformation (CLR)

Although relative abundance is an interpretable quantity, it is subject to severe compositional effects. In short, a single highly abundant taxon could significantly drive the relative abundances of other taxa. As an alternative, instead of using the total abundance, the geometric mean can also be used in the denominator (preceding log-transformation). This is know as the CLR transformation [1] and is given by

$$CLR(\widetilde{Y}_{ih}) = log\widetilde{Y}_{ih} - \frac{1}{H}\sum_{h=1}^{H} log\widetilde{Y}_{ih}$$

Since microbiome data is zero-inflated, we added one to every value in OTU table before taking CLR transform. In addition to addressing some compositionality concerns, CLR transforms often encourage distributions to look more normal.

2.1.2 Batch effect correction methods

We considered 6 batch correction methods in our comparison study. Y_{ijg} and $Y_{ijg}*$ represents the normalized/transformed abundance of taxon g for sample j from batch i respectively.

Mean-centering Method

$$Y_{ijg}* = Y_{ijg} - \bar{Y_{ig}}$$

Mean-centering method (PAMR) [51] does a taxon-wise one-way ANOVA adjustment. By subtracting mean abundance of each taxon within each batch from all measurements in that batch, each batch have zero means after adjustment. It is implemented in the pamr R package.

Regression method

Limma [52] fits a linear regression model to the whole experiment data, including both batch and other covariates like treatment, then removes the component due to the batch effects by subtracting estimated batch coefficients. It is implemented in the limma R package.

Model-based location/scale adjustments

ComBat and ber are both based on location-scale model, which assume that the batch effects can be modeled by standardizing means and variances across batches, but they use different methods to estimate the location-scale (L/S) parameters. ComBat uses an Empirical Bayes method to borrow information between genes. The steps of implementing ComBat are as follows [23]:

1. Define an L/S model:
$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg}$$
where α_g is the overall abundance, X is a design matrix, and β_g is the vector of regression coefficients corresponding to X. γ_{ig} and δ_{ig} represent the additive and multiplicative batch effects of batch i for taxon g. ϵ_{ijg} follows a normal distribution with mean zero and variance σ_g^2.

2. Standardize the data: The standardized data Z_{ijg} is calculated by
$$Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha_g} - X\hat{\beta}_g}{\hat{\sigma_g}}$$

3. L/S parameters estimation: We assume
$$Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2), \gamma_{ig} \sim N(Y_i, \tau_i^2), \delta_{ig}^2 \sim InverseGamma(\lambda_i, \theta_i)$$
The hyperparameters are estimated empirically from standardized data using the method of moments. Non-parametric priors are used in ComBat non-parametric method. The EB estimates of $\gamma_{ig}*$ and δ_{ig}^2* are given by the conditional posterior means.

4. Adjust the data: The EB batch-adjusted data can be calculated by
$$Y_{ijg}* = \frac{\hat{\sigma_g}}{\hat{\delta_{ig}}*}(Z_{ijg} - \hat{\gamma_{ig}}*) + \hat{\alpha_g} + X\hat{\beta}_g$$

It has been shown that ComBat is more robust compared to other methods when sample size is small. ComBat is implemented in the sva R package.

Ber was developed as an extension of ComBat and uses a two-stage regression procedure to estimate the location and scale parameters of normalized data [19]. In the first stage, the location parameter is estimated using linear models on observed abundance levels including batches and other variables of interest as covariates. At second stage, scale parameter is estimated at second stage on the squared residuals produced in the first stage using linear model. In ber-bg method, bagging estimators are produced using bootstrap samples with replacement. Ber has low computational cost and also has computational advantange in high dimensional low sample size situation. Ber is implemented in the ber R package.

2.1.3 Evaluation of batch effect removal effectiveness

Here, we briefly describe different strategies for assessing the effectiveness of batch removal approaches.

Visualization and Principal Coordinates Analysis (PCoA)

Principal Coordinates Analysis (PCoA) was applied to raw and adjusted CLR, RA, LRA data of two data sets. We made multidimensional scaling (MDS) plot using the first two principle components to visualize the effectiveness of batch effect adjustments. We evaluated the performance of correcting means and scale by looking at the location and variance of each batch in MDS plot.

Correlation Between Replicate Pairs

When samples have been profiled more than once, across batches, a natural approach is to assess the correlation in the taxon abundances across the batches. Higher correlation suggests better reproducibility across batches. In many ways, these metrics are closer to the ideal, but require replicates across batches.

Principal Variance Component Analysis (PVCA)

Principal Variance Component Analysis (PVCA) integrates principal component analysis and variance components analysis. It can be used to quantify the proportion of variation attributable to each effect [32]. PVCA has four steps: 1) Perform PCA to get the first few principal components; 2) Fit a mixed model to each principal component with all variable of interest as random effects and any nuisance factors as fixed effects; (3) Average the estimated

variance components with their corresponding eigenvalues as weights; (4) Standardize the weighted average variance components estimates by dividing by sum. By conducting PVCA, we can get the proportion of variation induced by variable of interest, batch, the interaction between variable of interest and batch and residual. PVCA is implemented in the pvca R package.

Separation Score

Separation score measures the degree of separation between batches. Suppose there are two batches j and $j*$, for each observation in batch j, its k nearest neighbours are determined with respect to euclidean distance. Then the proportion of nearest neighbours belonging to batch j is calculated. MS_j is the average of n_j proportions. We define $S_j = |MS_j - n_{j*}/(n_j + n_{j*} - 1)|$, which measures the absolute difference between MS_j and its value expected in the absence of batch effects. Separation score is calculated as the average of S_j and the corresponding quantity when the roles of j and $j*$ are switched. When there are more than two batches, we first calculate the separation score between all pairs of batches, and then take the weighted mean with weights proportional to sample sizes. We'd expect adjusted data to have lower separation score [21]. Separation score is calculated in the bapred R package.

2.1.4 Data Set Description

Two real data sets with batch effects were used. The first was the MWRP data set, which is comprised of 255 samples with 3139 taxa from three batches. 118 of the samples were from batch 1. There were 40 samples in the second batch and 97 samples in the third batch. Many samples in MWRP data set which were processed more than one time. 234 out of 255 observations are replicates from 97 samples in this data set. We will calculate the correlations between the replicates later.

The second data set is from a microbiome sub-study of the CARDIA cohort (referred to as the CARDIA data set). In this sub-study of, 530 samples with 379 features from three batches. There were 87 samples in batch 1, 88 samples in batch 2, 89 samples in batch 3 and 4, 94 samples in batch 5 and 83 samples in batch 6. The indicator of hypertension was chosen as variable of interest. There're 185 samples with hypertension out of 530 samples.

2.2 Results

2.2.1 MWRP data set

Principal Coordinates Analysis

Based on MDS plots we produced, for raw data, we can see batch 2 and 3 were very similar while batch 1 was far away from batch 2 and 3. Also, the variance of batch 1 was much larger than batch 2 and 3. After adjustments, three batches came together. Basically, all batch effect removal methods removed batch effects more or less while ComBat-p, ComBat-n, ber and ber-bg outperformed other methods. When applied to CLR data, limma and PAMR were both good at correcting the mean but bad at correcting scale, which performed better than ratio-based methods. ComBat-p, ComBat-n, ber and ber-bg were good at correcting both the mean and scale. When applied to LRA data, limma and PAMR were both good at correcting the mean but bad at correcting scale. Other methods performed pretty well in LRA data.

Correlation of Replicates

transformation	CLR		LRA	
method	correlation	percentage(%)	correlation	percentage(%)
raw	0.8429		0.8429	
ComBat_p	0.9044	7.30	0.9028	7.11
ComBat_n	0.9016	6.96	0.9026	7.08
PAMR	0.5572	-33.89	0.7377	-1.25
limma	0.8702	3.24	0.8702	3.24
ber	0.9003	6.81	0.9024	7.06
ber_bg	0.9008	6.87	0.9027	7.09

Table 2.1: Correlations of replicates of raw CLR and LRA data and batch-corrected data using ComBat-p, ComBat-n, PAMR, limma, ber and ber-bg for MWRP data set.

Pairwise Pearson's correlations were calculated between the replicates. We'd expect replicates to have higher correlation after adjustment. Based on the medians of correlations of replicates, PAMR made the data worse after adjustments. ComBat-p, ComBat-n, ber and ber-bg outperformed other methods where ComBat-p adjusted data had the largest median

of correlations in both CLR and LRA data. Also, we could find the correlations of raw CLR and LRA data were the same but adjusted LRA data is a little larger than adjusted CLR data.

Separation Score

transformation	CLR		LRA	
method	separation score	percentage(%)	separation score	percentage(%)
raw	0.2430		0.3653	
ComBat_p	0.0252	-89.63	0.0697	-80.92
ComBat_n	0.0282	-88.40	0.0739	-79.77
PAMR	0.2955	21.60	0.3962	8.46
limma	0.4134	70.12	0.3962	8.46
ber	0.0391	-83.91	0.0805	-77.96
ber_bg	0.0402	-83.46	0.0762	-79.14

Table 2.2: Separation score of raw CLR and LRA data and batch-corrected data using ComBat-p, ComBat-n, PAMR, limma, ber and ber-bg for MWRP data set.

Based on separation score, we found ComBat-p, ComBat-p, ber and ber-bg performed much better when compared to other methods. ComBat-p had lowest separation score in both CLR and LRA data.

2.2.2 Cardia data set

Principal Coordinates Analysis

Based on MDS plots we produced, for raw data, we can see batch 6 is far away from other batches. After adjustments, all batches came together. The variances of six batches were similar. Basically, all batch effect removal methods remove batch effects more or less. Since the variance of six batches were similar, Limma and PAMR performed better when compared to data.obj data set. When applied to LRA data, all methods performed pretty well.

Proportion of variation induced by variable of interests

In Cardia data set, we regarded the indicator of hypertension as our variable of interest and then estimated the proportion of variance contributable to hypertension using PVCA. For

transformation	CLR		LRA	
method	PVCA	percentage(%)	PVCA	percentage(%)
raw	0.0256		0.0256	
ComBat_p	0.0294	14.84	0.0295	15.23
ComBat_n	0.0290	13.28	0.0290	13.28
PAMR	0.0279	8.98	0.0255	-3.91
limma	0.0279	8.98	0.0279	8.98
ber	0.0284	10.93	0.0284	10.93
ber_bg	0.0285	11.33	0.0285	11.33

Table 2.3: Proportion of variation induced by variable of interest of raw CLR and LRA data and batch-corrected data using ComBat-p, ComBat-n, PAMR, limma, ber and ber-bg for Cardia data set.

CLR data, all methods improved the data after adjustments. For LRA data, all methods except PAMR improved the data after adjustments. As a result, ComBat-p preseved more biological signal of interest in both CLR and LRA data.

Separation Score

transformation	CLR		LRA	
method	separation score	percentage(%)	separation score	percentage(%)
raw	0.2092		0.3653	
ComBat_p	0.0438	-79.06	0.0435	-88.09
ComBat_n	0.0452	-78.39	0.0448	-87.74
PAMR	0.0584	-72.08	0.1583	-56.67
limma	0.0584	-72.08	0.0588	-83.90
ber	0.0360	-82.79	0.0361	-90.12
ber_bg	0.0359	-82.84	0.0356	-90.25

Table 2.4: Separation Scores of raw CLR and LRA data and batch-corrected data using ComBat-p, ComBat-n, PAMR, limma, ber and ber-bg for Cardia data set.

Based on separation score, we found ComBat-p, ComBat-p, ber and ber-bg performed much better when compared to other methods in Cardia data set. Ber-bg had lowest separation score in both CLR and LRA data.

2.3 Remarks

We applied six batch effect removal methods to two real microbiome data sets to compare the performance of traditional microarray batch effect removal methods when applied to microbiome data. Although some methods performed well based on some evaluation metrics, ComBat and ber outperformed other methods overall. Thus, we recommened ComBat and ber for batch effect removal. Our study showed that the performance of batch effect removal methods could be affected by many factors such as number of batches, sample size and number of samples in each batch. More microbiome data sets with batch effect may be needed for future comparision study. Our study also showed different data normalization methods could affect the performance of batch effect removal methods. Since microbiome data has many different features compared to microarray data, we can also expect new batch effect removal methods tailored for microbiome data in the future.

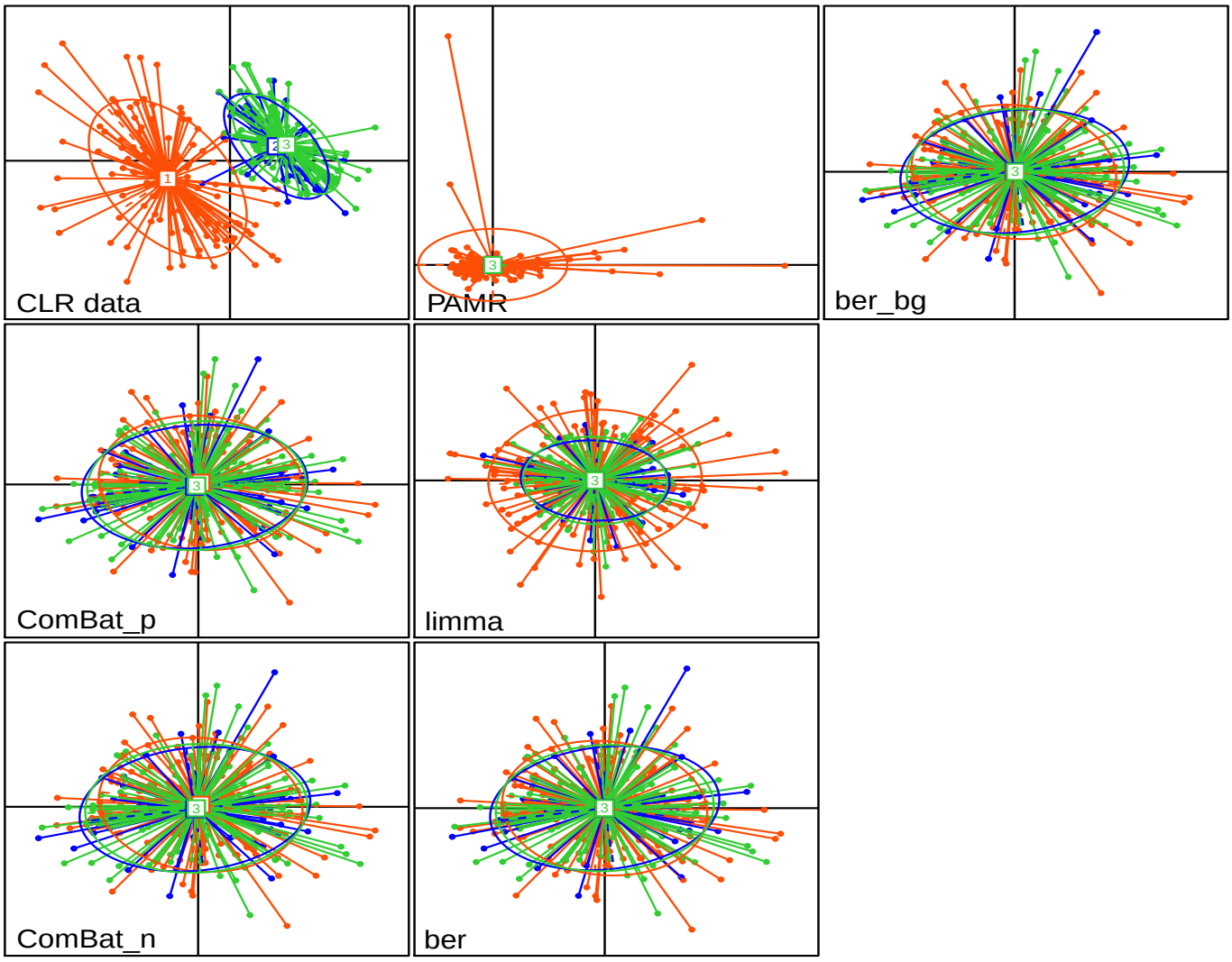

Figure 2.1: MDS plots of centered log ratio transformation (CLR) data and batch-corrected data using ComBat-p, ComBat-n, PAMR, limma, ber and ber-bg for MWRP data set.

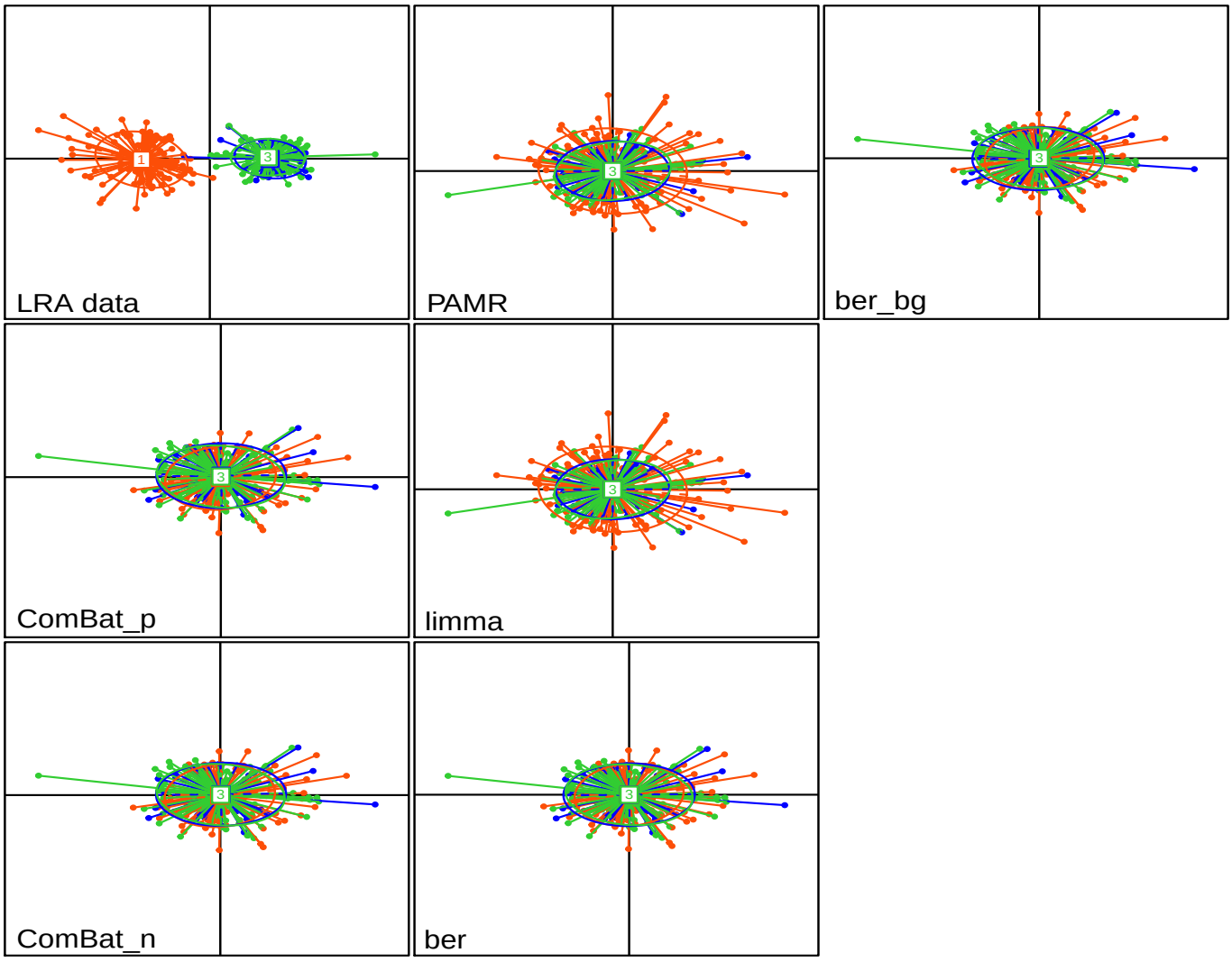

Figure 2.2: MDS plots of log relative abundance transformation (LRA) data and batch-corrected data using ComBat-p, ComBat-n, PAMR, limma, ber and ber-bg for MWRP data set.

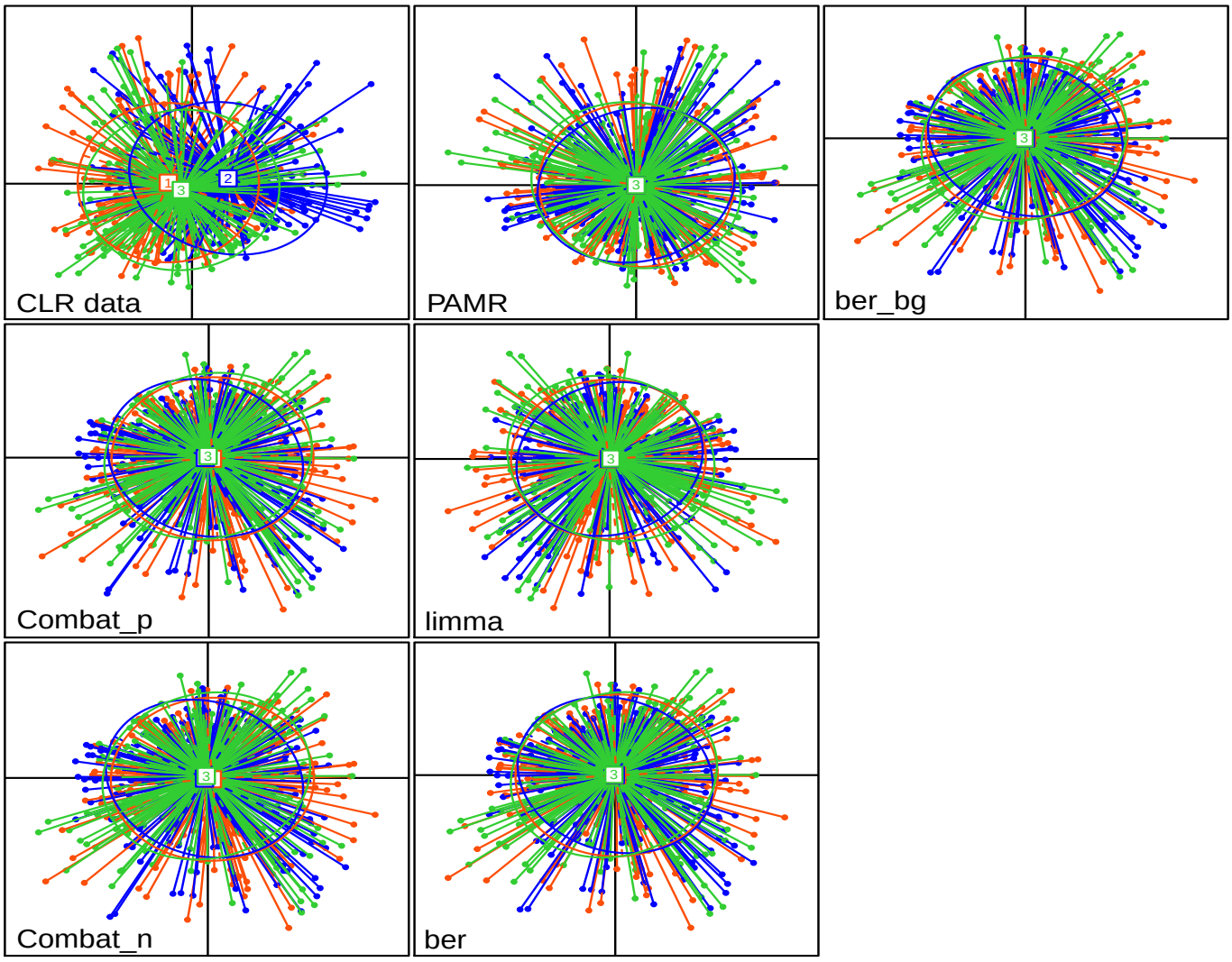

Figure 2.3: MDS plots of Centered log ratio transformation (LRA) data and batch-corrected data using ComBat-p, ComBat-n, PAMR, limma, ber and ber-bg for Cardia data set.

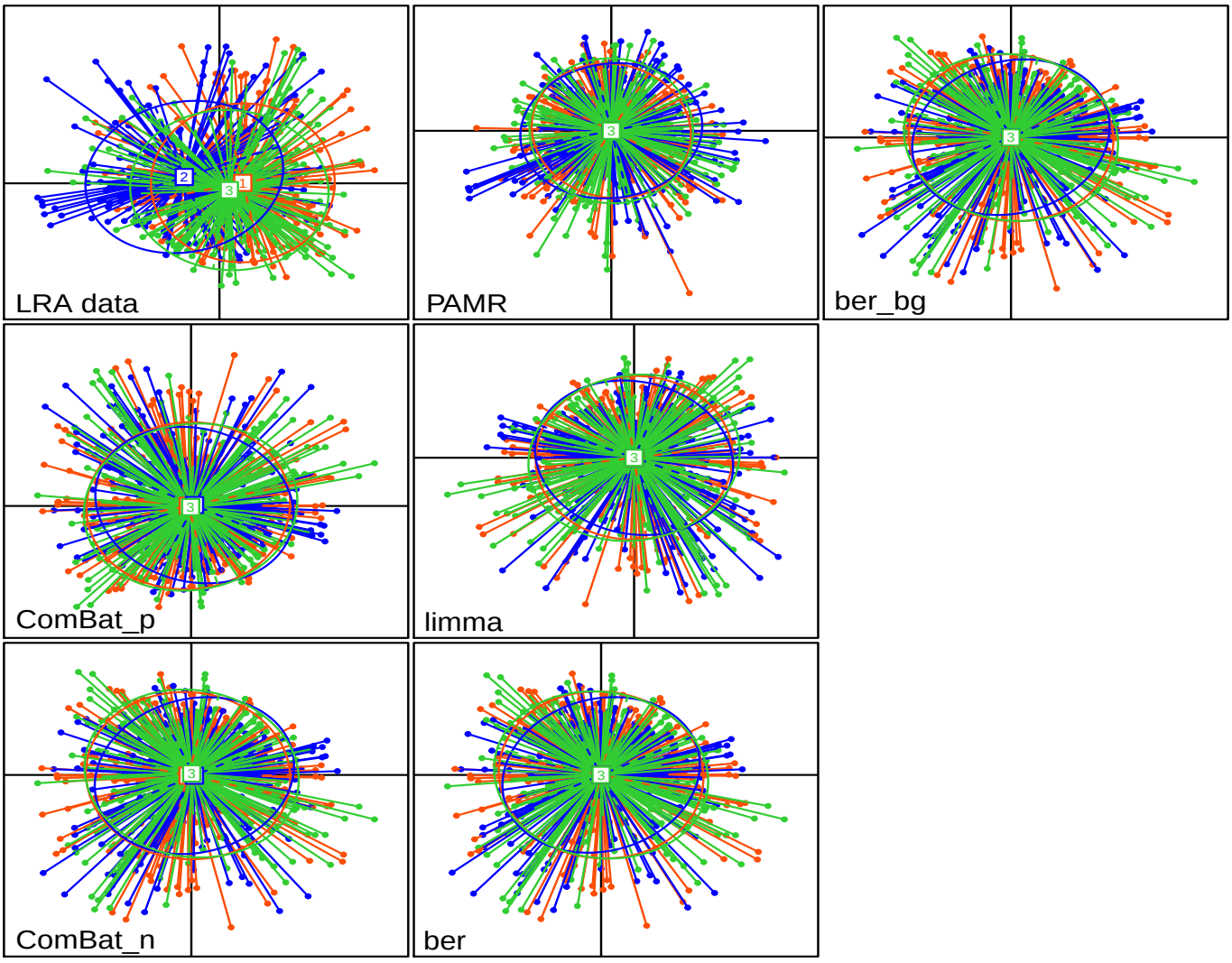

Figure 2.4: MDS plots of Log relative abundance transformation (LRA) data and batch-corrected data using ComBat-p, ComBat-n, PAMR, limma, ber and ber-bg for Cardia data set.

Chapter 3

SCREENING FOR ASSOCIATIONS BETWEEN MICROBIOME COMMUNITY PROFILES AND A LARGE NUMBER OF INDIVIDUAL GENOMIC OUTCOMES

Genomic markers can serve as important intermediaries between microbiome composition and outcomes and may serve as markers of effect or provide clues as to the specific manner by which microbes drive subsequent processes and disease. Consequently, identification of genomic markers related to microbiome composition represents an important problem for the field. Unfortunately, how to rapidly screen a large number of genomic markers for association with microbiome community composition remains unclear. Individual taxon analysis is simple, but is subject to multiple testing burden and also cannot capture important community level effects. On the other hand, while community level analysis offers a number of attractive features, existing approaches for community level analysis of the microbiome are not uniformly applicable to analyses focused on screening large numbers of genomic markers.

Screening for community level associations between microbiome composition and many genomic markers poses a grand challenge due to the high dimensionality and statistical irregularities of the markers. Permutation-based approaches are computationally expensive, often requiring hundreds of thousands of permutations to enable accurate p-value calculation in the tails. Regression-based approaches such as MiRKAT can allow for rapid assessment of associations, but the genomic markers often have poorly behaving distributions due to heteroskedasticity and presence of outliers [25], which can lead to both false negatives and false positives. Taking top principal coordinates for the community profiles for visualization or association analysis can mitigate some concerns, yet presumes that the top coordinates fully capture the variability of interest. These issues are often not of great concern when examining a single outcome, as investigators can identify optimal transformations of the data and conduct detailed analyses, but that is not possible when screening a large number of markers.

Beyond the sometimes poor distributional behavior, a separate challenge is that the manner in which a particular genomic marker is related to microbiome composition may differ

from marker to marker. Given that there are a range of different dissimilarities that can be used for capturing beta diversity, it becomes unclear which dissimilarity to use, since each may be optimal for different markers. The alternative approach, testing multiple dissimilarities for each marker, requires adjusting p-values to prevent inflated false positive rates; however, residual permutation, as used in MiRKAT, and perturbation [56] are computationally slow and therefore intractable in the present setting.

To facilitate rapid screening of individual genomic markers for association with microbiome composition, we propose the robust MiRKAT (MiRKAT-R) strategy, which is not a new method so much as a combination and translation of recent statistical developments to the context of microbiome analysis. MiRKAT-R works within the kernel machine testing framework [34], which underlies the MiRKAT family of methods, and incorporates a newly developed robust regression strategy [40]. As in the existing MiRKAT, a marker is regressed on covariates and a generally specified function of the microbiome profiles. This function, called a kernel, is fully determined based on a measure of similarity between individuals' microbiota. The kernel can be a transformation of existing distances and dissimilarities that capture important structure in microbiome data, including presence-absence effects and phylogeny. The regression framework enables rapid p-value calculation for the association between composition and outcomes while adjusting for confounders. However, whereas the usual MiRKAT and kernel machine framework uses a linear, least-squares, approach to regress markers on microbiome, MiRKAT-R adopts the recently developed robust kernel machine regression framework to regress the markers on beta diversity and obtain an analytic p-value for association. Robust regression allows for highly irregular distributions for the genomic markers while analytic p-value calculation ensures computational tractability. Furthermore, since the relationship between each marker and the microbiota may be different (and therefore better captured by different dis/similarity measures), we further develop an omnibus test based on combining results across different measures by incorporating a newly developed Cauchy-distribution transformation [36, 35] which precludes the need for permutation.

3.1 Robust Community-Level Testing for Microbiome-Marker Associations

We first briefly describe the MiRKAT framework [62] upon which our proposed methodology relies, then present our proposed robust approach.

3.1.1 MiRKAT Framework

Under the MiRKAT framework, we assume a microbiome study with n individuals. For each individual i, y_i denotes a particular genomic marker (we can consider each marker in turn), $\mathbf{X}_i$ denotes a vector of covariates (e.g. age, sex, education, etc.) and $\mathbf{Z}_i$ is the vector of microbial abundances. Then the MiRKAT framework operates under the model

$$y_i = \beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + h(\mathbf{Z}_i) + \varepsilon_i \tag{3.1}$$

where β_0 is an intercept, $\boldsymbol{\beta}$ is a vector of regression coefficients for additional covariates, and ε_i is an error term with mean zero and variance σ^2 (not necessarily normal). The relationship between the microbiota, $\mathbf{Z}_i$, and genomic marker, y_i, is encoded by the function $h(\cdot)$, which is assumed to be a member of a reproducing kernel Hilbert space defined by a positive definite kernel function $K(\cdot,\cdot)$. $K(\mathbf{Z}_i, \mathbf{Z}_{i'})$ is a measure of the similarity between subjects i and i' based on Z and, importantly, fully specifies the relationship between y and Z. Computing all pairwise similarities results in an $n \times n$ kernel matrix $\mathbf{K}$ with $(i,i')^{th}$ element equal to $K(\mathbf{Z}_i, \mathbf{Z}_{i'})$.

The matrix $\mathbf{K}$ may also be defined directly rather than through an explicit kernel function. Leveraging this within the context of microbiome studies, it is often convenient to define a matrix $\mathbf{D}$ of pairwise dissimilarities between individuals' microbial communities, then use the inverse relationship between similarity and dissimilarity to define the kernel matrix via $\mathbf{K} = (\mathbf{I} - \mathbf{1}\mathbf{1}'/n)\mathbf{D}^2(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)$. For example, $\mathbf{D}$ could be a matrix of Bray-Curtis dissimilarities or UniFrac distances.

Because $h(\mathbf{Z}_i)$ summarizes the association between the microbiome and the genomic marker, testing for no association is equivalent to testing $H_0 : h(\mathbf{Z}_i) = 0$. Through a connection between kernel methods and mixed models, this can be done by constructing the variance component score statistic

$$Q = \frac{(\mathbf{y} - \widehat{\mathbf{y}}_0)'\mathbf{K}(\mathbf{y} - \widehat{\mathbf{y}}_0)}{\widehat{\sigma}_0^2} \tag{3.2}$$

where $\widehat{\mathbf{y}}_0 = \widehat{\beta}_0 + \mathbf{X}\widehat{\boldsymbol{\beta}}$ with $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$ estimated under the null hypothesis, i.e. under the model where $\mathbf{h} = 0$. Under the null, Q follows a mixture of χ^2_1 distributions which can be easily calculated.

3.1.2 Robust MiRKAT Framework: MiRKAT-R

Although the usual MiRKAT framework has been successfully applied across a number of situations, genomic markers are frequently susceptible to problems such as outliers and heteroskedasticity [14, 25, 16], which can lead to both false positives and false negative findings. Therefore, we use the recently developed robust kernel machine approach [40] to enable more robust analysis.

We estimate the main effects of the covariates under the null hypothesis as

$$[\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}']' = \underset{\beta_0, \boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \phi(y_i - \beta_0 - \boldsymbol{\beta}\mathbf{X}_i') \quad (3.3)$$

where

$$\phi(x) = \begin{cases} \frac{1}{2}x^2 & |x| \leq k, \\ k(|x| - \frac{k}{2}) & |x > k \end{cases}$$

is the usual Huber's M-estimation loss function for achieving robust regression. With this $[\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}']'$, we can again calculate the residuals under the null as $\widehat{\boldsymbol{\varepsilon}}$. Then defining $u_i = \psi(\widehat{\varepsilon}_i/\widehat{s})$ where $\psi(x) = \frac{d}{dx}\phi(x)$ and $\widehat{s}$ is the estimated scale parameter, we can calculate the score statistic

$$\widetilde{Q} = \mathbf{u}'\mathbf{K}\mathbf{u}. \quad (3.4)$$

The form of the test statistic in (3.4) is essentially the same as (3.2) except the OLS residuals have been replaced by (scaled) residuals from the robust regression. In principle, under the null hypothesis, Q should also asymptotically follow a mixtures of χ^2 distribution. However, the mixing weights can be difficult to calculate, and when the same size is modest, the asymptotic distributions often do not work out well – the corresponding small sample corrections developed for MiRKAT [8] do not work for MiRKAT-R.

Therefore, following [40] and [60], we use an alternative strategy. Specifically, we assume that $\mathbf{K}$ is centered: if $\mathbf{K}_{orig}$ is a non-centered kernel calculated from a particular dissimilarity metric, the centered version is calculated as $\mathbf{K} = (\mathbf{I} - \mathbf{1}\mathbf{1}'/n)\mathbf{K}_{orig}(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)$. We can re-write (3.4) as $Q = \mathbf{u}'\mathbf{K}\mathbf{u} = tr(\mathbf{u}'\mathbf{K}\mathbf{u}) = tr([\mathbf{u}\mathbf{u}']\mathbf{K}) = tr(\mathbf{U}\mathbf{K})$. If we further standardize, then we have

$$\widetilde{Q} = \frac{Q}{\sqrt{tr(\mathbf{U}^2)tr(\mathbf{K}^2)}} = \frac{tr(\mathbf{U}\mathbf{K})}{\sqrt{tr(\mathbf{U}^2)tr(\mathbf{K}^2)}}, \quad (3.5)$$

which is equivalent to a kernel RV coefficient [48, 58]. We have previously generalized results

from the RV coefficient literature [26, 24] to facilitate testing in the kernel RV framework [58]. Specifically, we can calculate the moments of the finite sample permutation distribution of $\widetilde{Q}$ and analytically match these to the moments of a Pearson Type III distribution. Using the Pearson Type III distribution, we can analytically obtain a p-value for testing $H_0 : h(\mathbf{Z}) = 0$.

In this manner, for a given dissimilarity matrix $\mathbf{D}$ and corresponding kernel $\mathbf{K}$, we can test the association between each feature and the microbiota embedded with $\mathbf{K}$, followed by appropriate control for testing multiple genomic features. However, in practice, choosing a dissimilarity metric can be challenging.

3.1.3 Omnibus Testing Across Candidate Distances and Dissimilarities

Different dissimilarity and kernel metrics are optimal for detecting different types of relationships. For example, unweighted UniFrac focuses on presence/absence of taxa while incorporating phylogenetics, whereas the Aitchison distance focuses on quantitative abundance without regard for phylogeny. Unfortunately, the best dissimilarity to use depends on the true relationship between a particular marker and the microbiota, which is unknown *a priori*. Prior knowledge of the true relationship would preclude the need for analysis. Although permutation-based omnibus testing approaches have been proposed, these are impractical when screening a large number of genomic markers: permutation is inherently slow and the large multiple testing burden requires accurate p-values in the tails, which requires a large number of permutations.

Following [36], we use the Cauchy combination test (CCT) to get combined p-values based on individual p-values obtained from different dissimilarity and kernel matrices. Specifically, we assume that there are k candidate dissimilarities and kernels with p_i defined as the p-value from using MiRKAT-R with the i^{th} dissimilarity. To generate a final, combined p-value, the CCT uses a linear combination of transformed p-values as the test statistic, which is

$$T_{CCT} = \sum_{i=1}^{k} w_i tan\{(0.5 - p_i)\pi\} \tag{3.6}$$

where w_i's are nonnegative weights. Here we assign equal weight to each kernel metric. The transformed p-value $tan\{(0.5 - p_i)\pi\}$ is Cauchy distributed under the null hypothesis. Therefore, based on the cumulative function of the Cauchy distribution, the combined p-value

can be approximated by

$$p \approx 1/2 - arctan(T/w)/\pi \tag{3.7}$$

where $w = \sum_{i=1}^{k} w_i$. The p-value is exact in the situation where the individual p_i are uncorrelated, but in practice, the p-values can be quite highly correlated as they are generated from the same data. However, the remarkable result from using the Cauchy distribution is that the heavy tails make it such that the test is robust to the presence of correlation in the tails of the distribution (i.e., stringent significance levels). The CCT can poorly control false positives when the significance level is not low, but since our focus is on large scale screens, we are typically interested in much more stringent α levels.

3.2 Simulation Scenarios

Simulation studies were conducted under several simulation scenarios to verify that type I error would be well controlled using robust MiRKAT with different kernels and to investigate the power of omnibus robust MiRKAT. In addition to MiRKAT-R, we also consider application of the original MiRKAT procedure as well as a quantile kernel machine test which uses quantile instead of robust regression [28, 60], here refered to as MiRKAT-Q.

3.2.1 Simulations of Type I Error

We first simulated two covariates $\mathbf{X}_1$ and $\mathbf{X}_2$, where $\mathbf{X}_1$ is Bernoulli random variable with success probability 0.5 while $\mathbf{X}_2$ is continuous and generated from the normal distribution. Then a microbiome data set was simulated following Chen and Li's general approach [9], which bases simulation parameters on a real data set. Specifically, the dispersion parameters and proportion means were estimated from Charlson et al.'s real upper-respiratory-tract microbiome data set [6], which consists of 856 OTUs on 60 samples. We generated an OTU table for the same 856 OTUs in a simulated data set from a Dirichlet-multinomial distribution using the estimated parameters. Finally, we simulated outcomes from the standard normal distribution and the t distribution with 2 degrees of freedom. In total, there are four sets of simulated outcomes: normally-distributed outcomes with sample size 50 and 100, and t-distributed outcomes with sample size 50 and 100. The type I error rate was estimated as the proportion of p values less than significance level α, where α was set to be 0.05, 0.01 and 0.005.

3.2.2 Simulations of Empirical Power

For power simulations, we generated the covariates $\mathbf{X}_1$, $\mathbf{X}_2$ and microbiome data sets as before. Outcomes were generated according to one of two scenarios.

Under simulation scenario 1, the outcome was related to a cluster of taxa that depend on a phylogenetic tree. All the OTUs were partitioned into 20 clusters by performing the partitioning-around-medoids algorithm based on the cophenetic distances between taxa. A relatively abundant OTU cluster was chosen to be related to the outcome. The outcome was generated under the model:

$$y_i = 0.5X_{1i} + 0.5X_{2i} + \beta\, scale(\sum_{j\in A} Z_{ij}) + \epsilon_i$$

where ϵ_i follows a t distribution with 2 degrees of freedom. $\mathbf{Z}$ denotes a vector of microbial abundances. $\mathbf{A}$ denotes the indices of the OTUs in the selected cluster. The scale function standardizes the OTU abundance in the selected cluster to have mean 0 and SD 1. β is used to determine the effect size.

Under simulation scenario 2, the outcome was related to the ten most abundant OTUs in all samples. The outcome was simulated under the model:

$$y_i = 0.5X_{1i} + 0.5X_{2i} + \beta\, scale(\sum_{j\in A} \frac{Z_{ij}}{\bar{Z}_j}) + \epsilon_i$$

where ϵ_i, $\mathbf{Z}$, β, and the scale function are defined as before. $\bar{Z}_j$ denotes the average read count for the j-th OTU across samples, and $\mathbf{A}$ denotes the indices of the OTUs in the selected ten most abundant OTUs.

We conducted power simulations using a series of β values to compare the power of different tests under varied effect size. The empirical power of each test was estimated as the proportion of p values less than 0.05.

3.3 Data Illustration: Microbiome vs. Gene Expression Data

The IBD data set is from an inflammatory bowel disease (IBD) study that aimed to examine the association between host transcriptome and mucosal microbiome in patients with inflammatory bowel disease [44]. Paired host transcriptome and microbiome data were collected from 255 samples. Among the 255 samples, 196 were pre-pouch ileum (PPI) samples and 59 were

pouch samples. Expression levels for 19908 host transcripts and microbiome profiles with 7000 OTUs were measured for each sample. The IBD data set also contained other covariates including antibiotic use (yes/no), inflammation score, and disease outcome. Out of the 19908 host transcript expressions, we selected 7963 transcripts with higher variability than the remaining transcripts as outcomes. To test which of these were associated with the microbiome, we regressed the 7963 transcripts on microbiome compositions and two binary covariates including antibiotic use (yes/no) and PPI or pouch, separately. FDR control was used to adjust the p-values for multiple testing. The significance level α was set to be 0.05.

3.4 Results

3.4.1 Type I Error Results

The empirical type I error rates of MiRKAT, MiRKAT-R and MiRKAT-Q based on different kernels for t-distributed outcomes are shown in Table 1 for significance level $\alpha = 0.05$, 0.01 and 0.005. As is shown in Table 1, different kernels based on different distances will produce different type I error rates in the same test. When the outcomes are t-distributed, the type I error rates of MiRKAT based on Aitchison distance are highly inflated, resulting in highly inflated type I error of omnibus testing. Although the type I error rates of MiRKAT-R and MiRKAT-Q are slightly conservative for omnibus testing, both MiRKAT-R and MiRKAT-Q protect type I error rates much better compared to MiRKAT. The type I error rates of MiRKAT-R and MiRKAT-Q match closely.

3.4.2 Power Results

The power results under simulation scenario 1 and scenario 2 are shown in Figure 1 and 2, separately. From Figure 1.A and Figure 1.B, we can see the power produced by the same test based on different kernels differs greatly. For example, using MiRKAT-R, the kernel based on Bray-Curtis distance can produce highest power. However, the power produced by unweighted UniFrac is very low due to the moderately high abundance of the associated taxa (and resulting lack of information encoded in taxon presence/absence). The variations of power between approaches based on different kernels indicate that a proper choice of kernel is essential to have high power. A poor choice of kernel will lead to potentially substantial power loss. As different kernels will produce different power and we don't know which kernel will perform best, we use omnibus testing to consider multiple kernels simultaneously.

The power of MiRKAT and robust MiRKAT for omnibus testing under simulation scenario 1 is shown in Figure 1.B. Based on the plot, MiRKAT had the highest apparent power when effect size is small. However, this is not a fair comparison as MiRKAT cannot correctly control type I error with inflated false positive rate. Nonetheless, with increasing effect size, the power of MiRKAT was exceeded by MiRKAT-R and MiRKAT-Q. Overall, MiRKAT-R had highest power compared to the other approaches. MiRKAT-Q performed badly in power simulations and only had fair power when effect size is large. This is unsurprising, as quantile analysis effectively dichotomizes residuals whereas MiRKAT-R treats them continuously.

The results of power simulations under simulation scenario 2 are shown in Figure 2.B. As before, MiRKAT appears to have high power, though this is driven by inflated type I error rates and should be disregarded. For MiRKAT-R and MiRKAT-Q, which have correct size, MiRKAT-R is uniformly more powerful under this situation.

3.4.3 Data Application Results

Table 3.3 shows the results of analysis of the IBD data set. The number of significant transcripts found by each test can give insight into the power of each test. Among the 7963 transcripts, only 21 transcripts were showed significant by MiRKAT-Q, indicating the low power of MiRKAT-Q, which is in accordance with the results of the power simulations. The number of significant genes found by MiRKAT-R is 197, which is slightly larger than the number found by MiRKAT (184), showing that MiRKAT-R has higher power. Overall, looking at the number of significant associations for each method suggest that accommodation of outliers in genomic outcomes can increase the power of MiRKAT.

In observing the number of genes called significant by more then one method: generally, MiRKAT-R found most of the markers that were also called significant by MiRKAT and MiRKAT-Q. However, MiRKAT identified 42 genes that were not found by MiRKAT-R. These differences could be due to outliers that drive up false discoveries in addition to reducing power.

3.5 Remarks

In summary, both MiRKAT-R and MiRKAT-Q can control type I error well, while MiRKAT produces highly inflated type I error. MiRKAT-Q has the lowest power based on the results of the power simulation and data application, and thus is not a good choice for association testing. MiRKAT-R performed better under simulation scenario 1, while MiRKAT performed

better under simulation scenario 2. Although the performance of MiRKAT is acceptable with respect to the power simulations, it has the poorest performance in simulations of type I error. Combining the results of the type I error and power simulations, we can conclude that MiRKAT-R outperformed MiRKAT and MiRKAT-Q.

Although community level analysis using dis/similarities represents a powerful mode of analysis for microbiome data, the high dimensionality and irregular distributions of genomic data make existing approaches difficult to apply. Thus, in this paper, we have developed MiRKAT-R, which rapidly regresses individual genomic markers on overall microbiome composition summarized using dissimilarity measures. As a regression-based approach, MiRKAT-R is faster than permutation-based strategies and allows for easy adjustment of confounders. As a robust approach, MiRKAT-R accommodates outliers and heteroskedasticity present in genomic outcomes, reducing false negatives and accurately controlling false positives.

A separate mode of analysis is to assess the global association between large numbers of genomic markers, e.g., all transcripts in the transcriptome, and overall microbiome composition. This approach can be more powerful than analysis of individual markers for the same reasons that community level analysis can be more powerful than individual taxon level analysis. However, a limitation is that no individual markers are implicated, making it potentially difficult to interpret results and to design further studies interrogating the specific mechanisms by which the microbiome may be influencing outcomes. This is also a limitation of community level microbiome analysis, in that it is not possible to identify which individual taxa are associated with the outcome. After identifying associated genomic markers, community level microbiome analysis could be followed with individual taxon analysis to identify which taxa are driving the association.

We have focused on summarizing the microbiome using distances/dissimilarities or kernels, which do not consider the issue of compositionality, which is a central difficulty in analyzing microbiome data. On the one hand, community level analysis is focused on looking for global differences (whether there are any taxa related to the outcome), which theoretically mitigates some of the issues underlying compositionality; on the other hand, recent work suggests that compositionality may still be problematic. As our approach is largely generic, one could use recently developed distances/dissimilarities that accommodate compositionality directly.

Method	α	n	K_{BC}	K_{WU}	K_U	K_{AL}	Combined
MiRKAT	0.05	50	0.051	0.050	0.049	0.050	0.059
		100	0.051	0.050	0.050	0.050	0.058
	0.01	50	0.0105	0.0107	0.0097	0.0101	0.0110
		100	0.0104	0.0093	0.0100	0.0105	0.0110
	0.005	50	0.0052	0.0048	0.0053	0.0047	0.0054
		100	0.0053	0.0048	0.0047	0.0051	0.0055
MiRKAT-R	0.05	50	0.051	0.050	0.049	0.051	0.060
		100	0.051	0.047	0.052	0.050	0.058
	0.01	50	0.0110	0.0109	0.0103	0.0105	0.0124
		100	0.0111	0.0104	0.0096	0.0104	0.0121
	0.005	50	0.0055	0.0058	0.0051	0.0051	0.0065
		100	0.0058	0.0059	0.0050	0.0050	0.0064
MiRKAT-Q	0.05	50	0.051	0.049	0.049	0.049	0.058
		100	0.050	0.046	0.050	0.049	0.055
	0.01	50	0.0107	0.0100	0.0107	0.0103	0.0119
		100	0.0105	0.0100	0.0098	0.0094	0.0122
	0.005	50	0.0057	0.0055	0.0055	0.0051	0.0064
		100	0.0054	0.0057	0.0049	0.0048	0.0063

Table 3.1: Empirical type I errors for MiRKAT, MiRKAT-R and MiRKAT-Q with normally-distributed outcomes (n=50, 100). K_{BC} represents the Bray-Curtis kernel; K_{WU} represents the weighted UniFrac kernel; K_U represents the unweighted UniFrac kernel; K_{AL} represents the Aitchison linear kernel.

Method	α	n	K_{BC}	K_{WU}	K_U	K_{AL}	Combined
MiRKAT	0.05	50	0.046	0.052	0.047	0.090	0.075
		100	0.046	0.048	0.048	0.089	0.074
	0.01	50	0.0069	0.0091	0.0086	0.0310	0.0179
		100	0.0076	0.0076	0.0097	0.0324	0.0215
	0.005	50	0.0031	0.0040	0.0043	0.0197	0.0092
		100	0.0037	0.0036	0.0044	0.0220	0.0139
MiRKAT-R	0.05	50	0.048	0.049	0.050	0.051	0.056
		100	0.050	0.048	0.049	0.051	0.056
	0.01	50	0.0099	0.0101	0.0100	0.0098	0.0114
		100	0.0100	0.0103	0.0098	0.0103	0.0114
	0.005	50	0.0055	0.0053	0.0052	0.0052	0.0060
		100	0.0049	0.0058	0.0048	0.0057	0.0062
MiRKAT-Q	0.05	50	0.047	0.047	0.050	0.049	0.055
		100	0.049	0.047	0.049	0.051	0.056
	0.01	50	0.0105	0.0099	0.0104	0.0097	0.0113
		100	0.0106	0.0103	0.0095	0.0096	0.0121
	0.005	50	0.0052	0.0050	0.0051	0.0051	0.0056
		100	0.0055	0.0055	0.0052	0.0049	0.0061

Table 3.2: Empirical type I errors for MiRKAT, MiRKAT-R and MiRKAT-Q with t-distributed outcomes (n=50, 100). K_{BC} represents the Bray-Curtis kernel; K_{WU} represents the Weighted UniFrac kernel; K_U represents the Unweighted UniFrac kernel; K_{AL} represents the Aitchison linear kernel

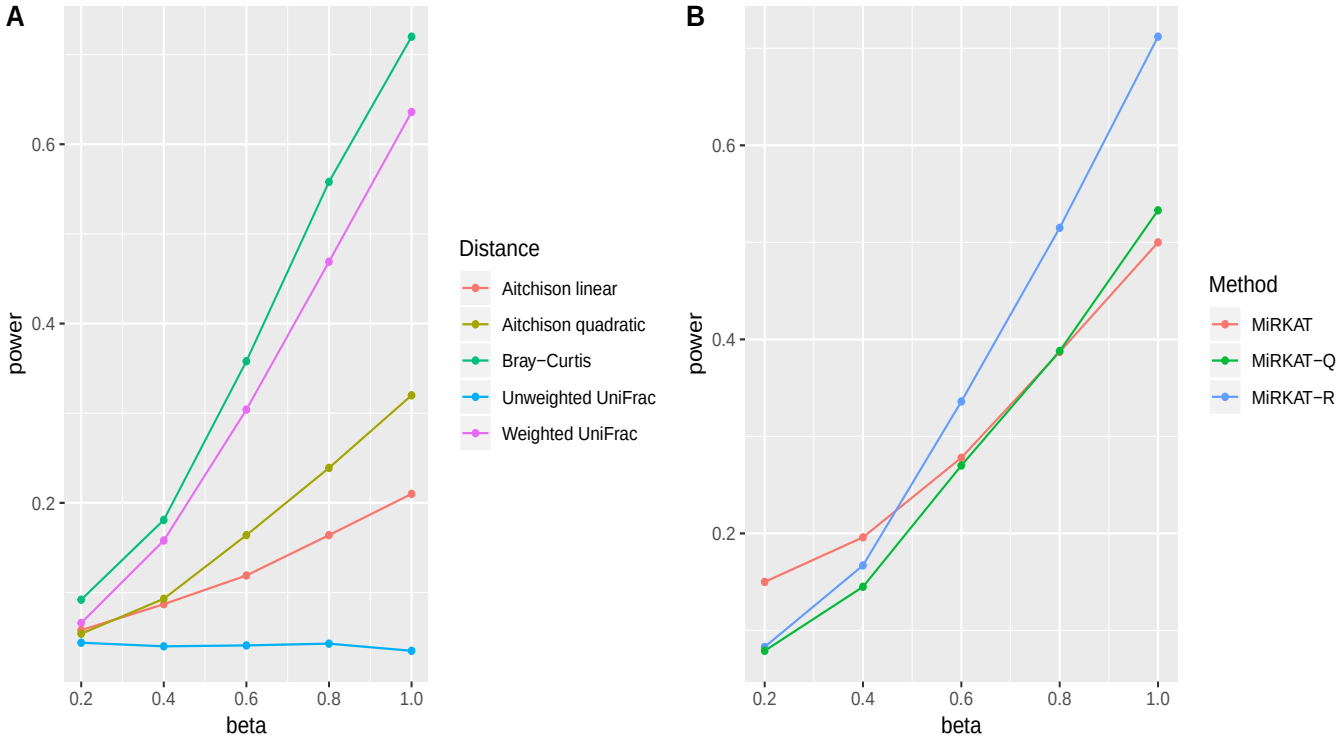

Figure 3.1: A: Power of MiRKAT-R based on different kernels under simulation scenario 1; B: Power of MiRKAT, MiRKAT-R and MiRKAT-Q for omnibus testing under simulation scenario 1

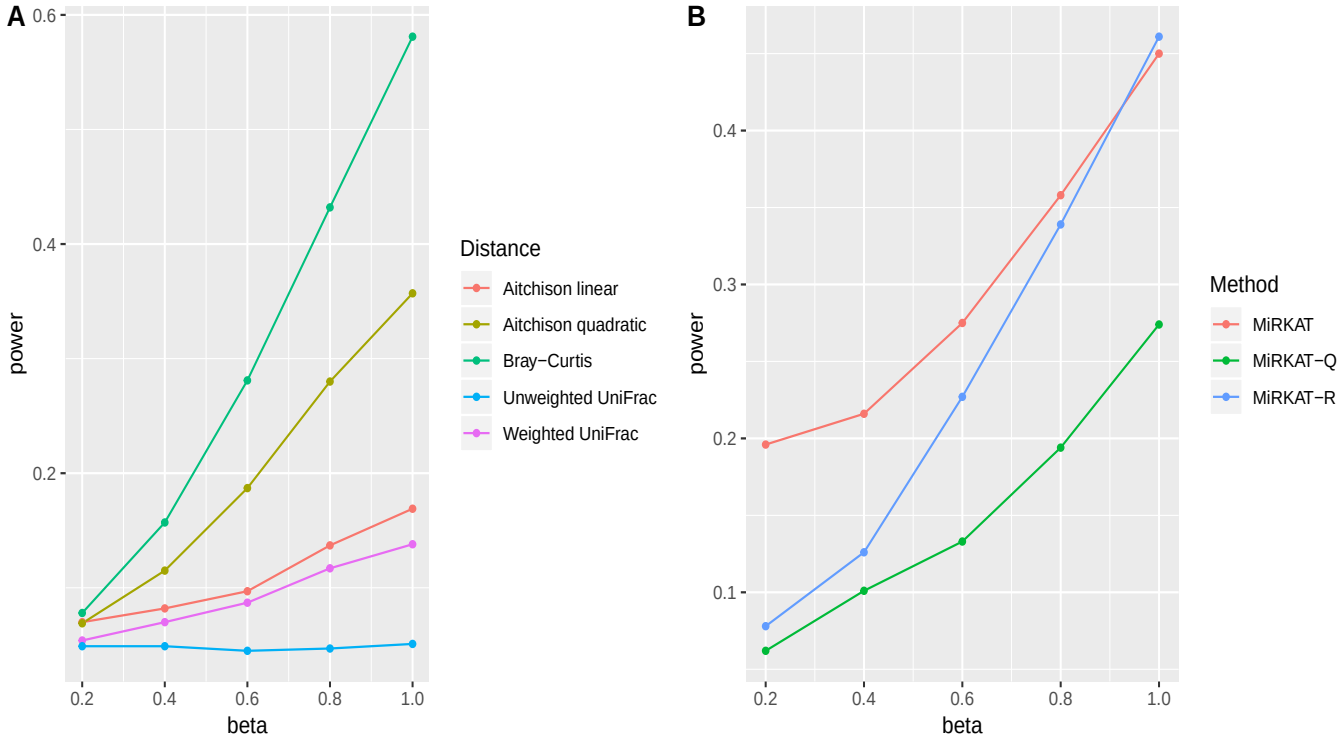

Figure 3.2: A: Power of MiRKAT-R based on different kernels under simulation scenario 2; B: Power of MiRKAT, MiRKAT-R and MiRKAT-Q for omnibus testing under simulation scenario 2

	MiRKAT	MiRKAT-R	MiRKAT-Q
MiRKAT	184	142	16
MiRKAT-R		197	17
MiRKAT-Q			21

Table 3.3: Number of significant expressions found by MiRKAT, MiRKAT-R and MiRKAT-Q in IBD data set. Each cell represents the number called significant by the method in the left and top of the table.

BIBLIOGRAPHY

[1] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society*, Series B, 1982.

[2] Marti J Anderson. A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1):32–46, 2001.

[3] Keith A Baggerly, Kevin R Coombes, and E Shannon Neeley. Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer. *Journal of Clinical Oncology*, 26(7):1186–1187, 2008.

[4] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[5] Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. Dada2: high-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7):581, 2016.

[6] 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.

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

[8] Jun Chen, Wenan Chen, Ni Zhao, Michael C Wu, and Daniel J Schaid. Small sample kernel association tests for human genetic and microbiome association studies. *Genetic epidemiology*, 40(1):5–19, 2016.

[9] Jun Chen and Hongzhe Li. Kernel methods for regression analysis of microbiome compositional data. In *Topics in Applied Statistics*, pages 191–201. Springer, 2013.

[10] Jun Chen, Euijung Ryu, Matthew Hathcock, Karla Ballman, Nicholas Chia, Janet E Olson, and Heidi Nelson. Impact of demographics on human gut microbial diversity in a us midwest population. *PeerJ*, 4:e1514, 2016.

[11] Kay Grennan Judith Badner Dandan Zhang Elliot Gershon Li Jin Chen, Chao and Chunyu Liu. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one*, 6(2), 2011.

[12] Rene Cortese, Lei Lu, Yueyue Yu, Douglas Ruden, and Erika C Claud. Epigenome-microbiome crosstalk: a potential new paradigm influencing neonatal susceptibility to disease. *Epigenetics*, 11(3):205–215, 2016.

[13] Hei Wong Jun Yu Yingying Wei Dai, Zhenwei and Inanc Birol. Batch effects correction for microbiome data with dirichlet-multinomial regression. *Bioinformatics*, 35(5), 2018.

[14] Z John Daye, Jinbo Chen, and Hongzhe Li. High-dimensional heteroscedastic regression with an application to eqtl data analysis. *Biometrics*, 68(1):316–326, 2012.

[15] Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

[16] Stephanie M Engel, Bonnie R Joubert, Michael C Wu, Andrew F Olshan, Siri E Håberg, Per Magne Ueland, Wenche Nystad, Roy M Nilsen, Stein Emil Vollset, Shyamal D Peddada, et al. Neonatal genome-wide methylation patterns in relation to birth weight in the norwegian mother and child cohort. *American journal of epidemiology*, 179(7):834–842, 2014.

[17] Sean M Gibbons, Claire Duvallet, and Eric J Alm. Correcting for batch effects in case-control microbiome studies. *PLoS computational biology*, 14(4):e1006102, 2018.

[18] Steven R Gill, Mihai Pop, Robert T DeBoy, Paul B Eckburg, Peter J Turnbaugh, Buck S Samuel, Jeffrey I Gordon, David A Relman, Claire M Fraser-Liggett, and Karen E Nelson. Metagenomic analysis of the human distal gut microbiome. *science*, 312(5778):1355–1359, 2006.

[19] Marco Giordan. A two-stage procedure for the removal of batch effects in microarray studies. *Statistics in Biosciences*, 6(1), 2014.

[20] Tiffany Hensley-McBain, Michael C Wu, Jennifer A Manuzak, Ryan K Cheu, Andrew Gustin, Connor B Driscoll, Alexander S Zevin, Charlene J Miller, Ernesto Coronado, Elise Smith, et al. Increased mucosal neutrophil survival is associated with altered microbiota in hiv infection. *PLoS pathogens*, 15(4):e1007672, 2019.

[21] Anne-Laure Boulesteix Hornung, Roman and David Causeur. Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment. *BMC bioinformatics*, 17(1), 2016.

[22] Lingjing Jiang, Amnon Amir, James T Morton, Ruth Heller, Ery Arias-Castro, and Rob Knight. Discrete false-discovery rate improves identification of differentially abundant microbes. *MSystems*, 2(6):e00092–17, 2017.

[23] Cheng Li Johnson, W. Evan and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1), 2007.

[24] Julie Josse, Jérôme Pagès, and François Husson. Testing the significance of the rv coefficient. *Computational Statistics & Data Analysis*, 53(1):82–91, 2008.

[25] Bonnie R Joubert, Siri E Håberg, Roy M Nilsen, Xuting Wang, Stein E Vollset, Susan K Murphy, Zhiqing Huang, Cathrine Hoyo, Øivind Midttun, Lea A Cupul-Uicab, et al. 450k epigenome-wide scan identifies differential dna methylation in newborns related to maternal smoking during pregnancy. *Environmental Health Perspectives*, 120(10):1425, 2012.

[26] Frédérique Kazi-Aoual, Simon Hitier, Robert Sabatier, and Jean-Dominique Lebreton. Refined approximations to permutation tests for multivariate inference. *Computational Statistics & Data Analysis*, 20(6):643–656, 1995.

[27] Christopher P Kolbert and David H Persing. Ribosomal dna sequencing as a tool for identification of bacterial pathogens. *Current opinion in microbiology*, 2(3):299–305, 1999.

[28] Dehan Kong, Arnab Maity, Fang-Chi Hsu, and Jung-Ying Tzeng. Testing and estimation in marker-set association study using semiparametric quantile regression kernel machine. *Biometrics*, 72(2):364–371, 2016.

[29] Justin Kuczynski, Christian L Lauber, William A Walters, Laura Wegener Parfrey, José C Clemente, Dirk Gevers, and Rob Knight. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, 13(1):47, 2012.

[30] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733, 2010.

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

[32] Pierre R. Bushel TzuMing Chu Li, Jianying and Russell D. Wolfinger. Principal variance components analysis: Estimating batch effects in microarray gene expression data. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*, 2009.

[33] Shih-Wen Lin, Neal D Freedman, Jianxin Shi, Mitchell H Gail, Emily Vogtmann, Guoqin Yu, Vanja Klepac-Ceraj, Bruce J Paster, Bruce A Dye, Guo-Qing Wang, et al. Beta-diversity metrics of the upper digestive tract microbiome are associated with body mass index. *Obesity*, 23(4):862–869, 2015.

[34] D. Liu, X. Lin, and D. Ghosh. Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics*, 63(4):1079–1088, 2007.

[35] Yaowu Liu, Sixing Chen, Zilin Li, Alanna C Morrison, Eric Boerwinkle, and Xihong Lin. Acat: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, 104(3):410–421, 2019.

[36] Yaowu Liu and Jun Xie. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, (just-accepted):1–29, 2018.

[37] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[38] Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, 71(12):8228–8235, 2005.

[39] Siddhartha Mandal, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*, 26(1):27663, 2015.

[40] Kara Martinez, Arnab Maity, Robert Yolken, Patrick Sullivan, and Jung-Ying Tzeng. The robust kernel association test. *arXiv preprint arXiv:1901.09419*, 2019.

[41] Ian H McHardy, Maryam Goudarzi, Maomeng Tong, Paul M Ruegger, Emma Schwager, John R Weger, Thomas G Graeber, Justin L Sonnenburg, Steve Horvath, Curtis Huttenhower, et al. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome*, 1(1):17, 2013.

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

[43] Caroline M Mitchell, Sujatha Srinivasan, Anna Plantinga, Michael C Wu, Susan D Reed, Katherine A Guthrie, Andrea Z LaCroix, Tina Fiedler, Matthew Munch, Congzhou Liu, et al. Associations between improvement in genitourinary symptoms of menopause and changes in the vaginal ecosystem. *Menopause*, 25(5):500–507, 2018.

[44] Xochitl C Morgan, Boyko Kabakchiev, Levi Waldron, Andrea D Tyler, Timothy L Tickle, Raquel Milgrom, Joanne M Stempak, Dirk Gevers, Ramnik J Xavier, Mark S Silverberg, et al. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome biology*, 16(1):67, 2015.

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

[46] Anna Plantinga, Xiang Zhan, Ni Zhao, Jun Chen, Robert R Jenq, and Michael C Wu. Mirkat-s: a community-level test of association between the microbiota and survival times. *Microbiome*, 5(1):17, 2017.

[47] Anna M Plantinga, Jun Chen, Robert R Jenq, and Michael C Wu. pldist: ecological dissimilarities for paired and longitudinal microbiome association analysis. *Bioinformatics*, 2019.

[48] Paul Robert and Yves Escoufier. A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Applied statistics*, pages 257–265, 1976.

[49] Andreas Scherer. Batch effects and noise in microarray experiments: sources and solutions. *John Wiley & Sons*, 868, 2009.

[50] Robert F Schwabe and Christian Jobin. The microbiome and cancer. *Nature Reviews Cancer*, 13(11):800, 2013.

[51] Graeme J. Smethurst Yvonne Hey Michal J. Okoniewski Stuart D. Pepper Anthony Howell Crispin J. Miller Sims, Andrew H. and Robert B. Clarke. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets-improving meta-analysis and prediction of prognosis. *BMC medical genomics*, 1(1), 2008.

[52] Gordon K. Smyth and Terry Speed. Normalization of cdna microarray data. *Methods*, 31(4), 2003.

[53] Shan Sun, Anju Lulla, Michael Sioda, Kathryn Winglee, Michael C Wu, David R Jacobs Jr, James M Shikany, Donald M Lloyd-Jones, Lenore J Launer, Anthony A Fodor, and K Meyer. Gut microbiota composition and blood pressure: The cardia study. *Hypertension*, 73:9981006, 2019.

[54] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804, 2007.

[55] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27, 2017.

[56] Michael C Wu, Arnab Maity, Seunggeun Lee, Elizabeth M Simmons, Quaker E Harmon, Xinyi Lin, Stephanie M Engel, Jeffrey J Molldrem, and Paul M Armistead. Kernel machine snp-set testing under multiple candidate kernels. *Genetic epidemiology*, 37(3):267–275, 2013.

[57] Jian Xiao, Hongyuan Cao, and Jun Chen. False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics*, 33(18):2873–2881, 2017.

[58] Xiang Zhan, A Plantinga, N Zhao, and Michael C Wu. A fast small-sample kernel independence test for microbiome community-level association analysis. *Biometrics*, page In press, 2017.

[59] Xiang Zhan, Xingwei Tong, Ni Zhao, Arnab Maity, Michael C Wu, and Jun Chen. A small-sample multivariate kernel machine test for microbiome association studies. *Genetic Epidemiology*, 2016.

[60] Xiang Zhan and Michael C Wu. Reader reaction: A note on testing and estimation in marker-set association study using semiparametric quantile regression kernel machine. *Biometrics*, 74(2):764–766, 2018.

[61] Xinyan Zhang, Himel Mallick, Zaixiang Tang, Lei Zhang, Xiangqin Cui, Andrew K Benson, and Nengjun Yi. Negative binomial mixed models for analyzing microbiome count data. *BMC bioinformatics*, 18(1):4, 2017.

[62] 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.