1-1-2012

Statistical Methods for Human Microbiome Data Analysis

Jun Chen
University of Pennsylvania, jchen1981@gmail.com

Statistical Methods for Human Microbiome Data Analysis

Abstract
The human microbiome is the totality of the microbes, their genetic elements and the interactions they have with surrounding environments throughout the human body. Studies have implicated the human microbiome in health and disease. Two central themes of human microbiome studies are to identify potential factors influencing the microbiome composition, and to define the relationship between microbiome features and biological or clinical outcomes. With the development of next generation sequencing technologies, the human microbiome composition can be interrogated using high-throughput DNA sequencing. One strategy sequences the bacterial 16S ribosomal RNA gene for species identification. These 16S sequences are usually clustered into Operational Taxonomic Units (OTUs). Analysis of such OTU data raises several important statistical challenges, including taking into account the phylogenetic relationship among OTUs and modeling high-dimensional overdispersed count data. This dissertation presents three statistical methods developed specifically for 16S data analysis centering around the two themes. To test the association between overall microbiome composition and a covariate/an outcome, a testing procedure based on a generalized UniFrac distance was developed. The generalized UniFrac distance corrects the unduly weighting of classic UniFrac distances on either highly abundant or rare lineages, and was shown to be more powerful than the classic UniFracs. Under the framework of canonical correlation analysis (CCA), a structure-constrained sparse CCA was proposed to select the OTUs and their correlated covariates. A phylogenetic structure-constrained penalty function was imposed to induce certain smoothness on the linear coefficients according to the OTU phylogenetic relationship. Structure-constrained sparse CCA performed much better than sparse CCA in selecting relevant OTUs. Finally, a sparse Dirichlet-multinomial regression (SDMR) model was developed to link the microbiome composition to environmental covariates and to select the most important covariates and their affected OTUs. SDMR accounts for the overdispersion of OTU counts and uses a sparse group L1 penalty function to facilitate selection of covariates and OTUs simultaneously. These methods were illustrated using simulations as well as a real human gut microbiome data set from a study of dietary effects on gut microbiome composition.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
Hongzhe Li

Keywords
High-dimensional statistics, Metagenomics, Microbiome, Variable selection

STATISTICAL METHODS FOR HUMAN MICROBIOME DATA ANALYSIS

Jun Chen

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2012

Supervisor of Dissertation

Hongzhe Li, PhD, Professor of Biostatistics and Statistics

Graduate Group Chairperson

Maja Bucan, PhD, Professor of Genetics

Dissertation Committee

Mingyao Li, Assistant Professor of Biostatistics

Frederic Bushman, Professor of Microbiology

Nancy Zhang, Associate Professor of Statistics

Li-San Wang, Assistant Professor of Pathology and Laboratory Medicine

STATISTICAL METHODS FOR HUMAN MICROBIOME DATA ANALYSIS

Dedicated to my family, my friends and everyone I love.

ACKNOWLEDGEMENT

I would like to thank my advisor Hongzhe Li for his meticulous supervision of my research during the last five years. I would have never been able to finish my dissertation without his guidance. It is him who cultivates my deep interest in statistics and brings me to the forefront of statistical methodological research. I have learned from him the essential skills to be a good statistician, from writing a statistics paper to giving a clear presentation. Most importantly, I have learned how to think statistically, which will prepare me well for future independent research. I am very certain that his influence will carry over into my future career. I would like to thank my collaborators Rick Bushman, Gary Wu and Jim Lewis for providing me the microbiome data. I really enjoyed the collaborative experiences, which not only helped me master existing statistical tools but also motivated me to develop new statistical methods. These methods comprise the main part of this dissertation. My heartfelt thanks also go to my committee members Mingyao Li, Rick Bushman, Nancy Zhang, Li-san Wang and Carlo Maley for their time and efforts to serve on my committee. Their valuable advices during committee meetings have helped me to improve my dissertation.

This research is partially supported by NIH grants CA127334 and DK083981.

ABSTRACT

STATISTICAL METHODS FOR HUMAN MICROBIOME DATA ANALYSIS

Jun Chen

Hongzhe Li, PhD

The human microbiome is the totality of the microbes, their genetic elements and the interactions they have with surrounding environments throughout the human body. Studies have implicated the human microbiome in health and disease. Two central themes of human microbiome studies are to identify potential factors influencing the microbiome composition, and to define the relationship between microbiome features and biological or clinical outcomes. With the development of next generation sequencing technologies, the human microbiome composition can be interrogated using high-throughput DNA sequencing. One strategy sequences the bacterial 16S ribosomal RNA gene for species identification. These 16S sequences are usually clustered into Operational Taxonomic Units (OTUs). Analysis of such OTU data raises several important statistical challenges, including taking into account the phylogenetic relationship among OTUs and modeling high-dimensional overdispersed count data. This dissertation presents three statistical methods developed specifically for 16S data analysis centering around the two themes. To test the association between overall microbiome composition and a covariate/an outcome, a testing procedure based on a generalized UniFrac distance was developed. The generalized UniFrac distance corrects the unduly weighting of classic UniFrac distances on either highly abundant or rare lineages, and was shown to be more powerful than the classic UniFracs. Under the framework of canonical correlation analysis (CCA), a structure-constrained sparse CCA was proposed to select the OTUs and their correlated covariates. A phylogenetic structure-constrained penalty function was imposed to induce certain smoothness on the linear coefficients according to the OTU phylogenetic relationship. Structure-constrained sparse CCA performed much better than sparse CCA in selecting relevant OTUs. Finally, a sparse Dirichlet-multinomial

regression (SDMR) model was developed to link the microbiome composition to environmental covariates and to select the most important covariates and their affected OTUs. SDMR accounts for the overdispersion of OTU counts and uses a sparse group l_1 penalty function to facilitate selection of covariates and OTUs simultaneously. These methods were illustrated using simulations as well as a real human gut microbiome data set from a study of dietary effects on gut microbiome composition.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

1.1. Human microbiome in health and disease

We are not living alone. The human body is home to 10 trillion (10^{14}) microbial cells, exceeding at least 10-fold the number of human cells (Whitman *et al.*, 1998). The totality of the microbes (*microbiota*), their genomes (*metagenome*) and the environment in which they interact constitutes the human *microbiome* (Cho and Blaser, 2012). The human microbiome contains taxa from across the tree of life including bacteria, viruses, micro-eukaryotes, and archaea, that interact with one another and with the host, greatly impacting the human health and physiology (Clemente *et al.*, 2012). The human microbiome encodes 100 times more genes than the human genome, providing traits that humans did not need to evolve on their own (Qin *et al.*, 2010). The emerging concept of human "supra-organism" views humans as a composite of microbial and human cells with human genetic landscape as an aggregate of the genes in the human genome and the microbiome, and the human metabolic features as a blend of human and microbial traits (Turnbaugh *et al.*, 2007). In contrast to the human genome, the human microbiome is highly variable. It displays substantial intra-individual variation at different body sites (gut, skin, lung, vagina, oral cavity *etc.*), inter-individual variation at the same body sites and intra-individual variation at different times (Costello *et al.*, 2009).

The human microbiome plays an important role in promoting human health. For example, the human gut microbiome can harvest otherwise inaccessible nutrients, synthesize certain vitamins, promote the proper development of the immune system and protect us from pathogens (Turnbaugh *et al.*, 2007). Increasingly more human microbiome studies have implicated the human microbiome in the pathogenesis of many human diseases such as obesity, diabetes, inflammatory bowel disease (IBD), irritable bowel syndrome (IBS), vaginosis and even cancers (Cho and Blaser, 2012; Plottel and Blaser, 2011; Pflughoeft and Versalovic, 2011; Littman and Honda, 2012; Holmes *et al.*, 2011; Kinross *et al.*, 2011).

Higher Firmicutes to Bacteroidetes ratios and reduced species diversity have been observed in obese humans (Ley *et al.*, 2005, 2006). Two recent studies found that the abundance of phylum Fusobacteria increased significantly in the colon of colorectal cancer patients (Castellarin *et al.*, 2012; Kostic *et al.*, 2012). These findings have profound implications. If the microbiome effect is causal, new therapeutic strategies can be designed to treat diseases by modulating the microbiome composition (Virgin and Todd, 2011; Collison *et al.*, 2012). Even if the microbiome alteration is a result of disease process, the affected taxa in the microbiome can still serve as biomarkers for disease prevention and early diagnosis (Segata *et al.*, 2011; Knights *et al.*, 2011).

Many factors can influence the human microbiome composition (Turnbaugh *et al.*, 2007). These factors include the host genotype (Spor *et al.*, 2011), host physiological status such as aging (Biagi *et al.*, 2010), host pathophysiological status (Turnbaugh *et al.*, 2009), host lifestyle such as dietary habit (De Filippo *et al.*, 2010; Wu *et al.*, 2011a) and host environment (Dominguez-Bello *et al.*, 2010). The genotypic effect on the microbiome may explain the missing link between genetics and disease. A disease-susceptibility genotype may affect the disease outcome through the alteration of the microbiome composition (Virgin and Todd, 2011; Spor *et al.*, 2011).

Fueled by technological advancement, large-scale endeavors such as the Human Microbiome Project (HMP) (Peterson *et al.*, 2009) by the US National Institutes of Health and the European Metagenomics of the Human Intestinal Tract (MetaHIT) (Ehrlich, 2011) have been undertaken to characterize the compositional range of the "healthy" microbiome, to define the relationship between microbiome features and biological or clinical outcomes, and to identify potential factors influencing the microbiome composition. To achieve these goals, powerful statistical methods need to be developed to make full use of the data structure and to guard against false discoveries in this ultra high-dimensional setting.

1.2. Metagenomic approaches for human microbiome studies

Prior to the era of high-throughput DNA sequencing, researchers study the microbiome by cultivating the microbes from collected environmental samples, which is very laborious and time-consuming, and yet the majority of the microbes can not be cultivated, blinding us to see the global picture of the real microbial world. With the development of next generation sequencing such as Roche/454 pyrosequencing and Illumina Solexa sequencing, the human microbiome can now be studied by direct DNA sequencing. The DNA sequencing based approach to study the microbiome is called *metagenomics*. There are basically two metagenomic approaches to sequence the microbiome (Kuczynski *et al.*, 2011). The first approach is 16S ribosomal RNA (rRNA) gene targeted amplicon sequencing, where part of the 16S rRNA gene of the bacterial genome (1.5kb) is sequenced (Andersson *et al.*, 2008). This approach is used exclusively for determining the taxonomic composition and species diversity of the bacterial community. One advantage of the 16S rRNA gene is its taxonomic coverage: 16S rRNA gene is present in all bacteria. Furthermore, 16S rRNA gene contains both conserved region that can be used to design PCR primers to amplify regions of interest, and variable regions (V1-V9) that can be used for fine level taxonomic classification. Another advantage is the availability of several large databases of 16S rRNA gene reference sequences and taxonomies, such as Ribosomal Database Project (RDP), Greengenes and SILVA (Cole *et al.*, 2009; DeSantis *et al.*, 2006; Pruesse *et al.*, 2007). As with any PCR-based approach, there are problems of PCR bias and chimeric reads associated with PCR amplification. However, by choosing appropriate primer set according to the studied microbial community, the problem of PCR bias can be alleviated (Kuczynski *et al.*, 2011). By using efficient computational algorithms, the chimeric reads can be readily detected (Haas, 2011). Due to its simplicity, relatively low cost and availability of mature analysis pipelines, 16S rDNA sequencing is routinely employed to profile the taxonomic content of the community.

The second approach to sequence the microbiome is shotgun metagenomic sequencing (sim-

ply referred to as metagenomic sequencing), which involves randomly sequencing all the genomic DNA in the samples (Tringe *et al.*, 2005; Gill *et al.*, 2006; von Mering *et al.*, 2007; Dinsdale *et al.*, 2008; Turnbaugh *et al.*, 2009; Qin *et al.*, 2010; Arumugam *et al.*, 2011; Iverson *et al.*, 2012) . This approach can reveal the gene content of the microbiome as well as the taxonomic content. It has been reported that the taxonomic content of the microbiome varies tremendously across individuals but the gene content remains similar, indicating the importance of studying the gene content (Turnbaugh *et al.*, 2009). The shotgun approach is potentially unbiased and can be used to study other communities such as the viral community (Minot *et al.*, 2011). The bottleneck of this approach is the development of efficient computational tools (read mapping, binning and assembly) to process the massive amount of short reads produced (Wooley and Ye, 2010). The ambiguity of the reads poses a great challenge since each read can come from any region of any microbial genome of unknown genome size and abundance with some regions being more divergent than others. Many databases and software packages are being developed to analyzing the shotgun metagenomic data (Huson *et al.*, 2007; Meyer *et al.*, 2008; Markowitz *et al.*, 2008; Seshadri *et al.*, 2007; Goll *et al.*, 2010; Angiuoli *et al.*, 2011) .

The statistical methods presented in this dissertation are developed specifically for 16S rDNA sequence data, though they can also be adapted for analyzing shotgun metagenomic data. The processing of 16S data can be taxonomy-dependent, where 16S sequences are compared to existing 16S databases (Matsen *et al.*, 2010), or taxonomy-independent, where 16S sequences are clustered based on their divergence. The taxonomy-independent approach is more prevalent and many tools such as QIIME (Caporaso *et al.*, 2010b), mothur (Schloss *et al.*, 2009) and VAMPS use this approach to process 16S sequences. Fig. 1 shows the pipeline of 16S data generation and processing (QIIME pipeline). DNA is first extracted from the environmental samples. Some variable region of the 16S rRNA gene such as V1-V2 region is PCR amplified using barcoded primer set. Barcoding enables high-throughput multiplex sequencing (Hamady *et al.*, 2008). The barcoded PCR amplicons are pooled and subject to Roche/454 pyrosequencing using the GS FLX platform. The average read length

of Roche/454 Genome Sequencer FLX Titanium system can be up to 500bp or more. The Roche/454 platform produces sequence reads in SFF format (Standard Flowgram Format). This *.sff file contains the original flowgrams (light signal strength) and quality scores for each read in addition to other information. The platform also converts *.sff file into a sequence (*.fna) file and quality (*.qual) file.

After obtaining the raw reads, samples are assigned to the multiplex reads based on barcodes, and low-quality and ambiguous reads are removed. The filtered sequences are clustered into sequence clusters called *Operational Taxonomic Units* (OTUs) based on sequence similarity. OTUs are intended to represent some degree of taxonomic relatedness. For example, when sequences are clustered at 97% sequence similarity, each resulting cluster is typically thought of as representing a biological species. OTU picking is a critical step of 16S data processing and has a large effect on downstream analysis based on OTU data. The QIIME pipeline uses uclust algorithm (Edgar, 2010) to form OTUs as default. Currently there are a number of competing algorithms for OTU picking (Sun *et al.*, 2012). Determining the optimal way of picking OTUs is an active research area. Each OTU has an associated representative sequence. RDP classifier (Wang *et al.*, 2007) can be used to assign a bacterial lineage to the representative sequence. The RDP classifier is a naive Bayes classifier, which provides taxonomic assignments from domain to genus, with confidence estimates for each assignment. The OTU representative sequences are further aligned using template guided alignment method (e.g. PyNAST) or de novo alignment method (e.g. MUSCLE). For large data sets, PyNAST is preferred for its computational efficiency (Caporaso *et al.*, 2010a). Chimeric reads are removed based on the aligned sequences(Haas, 2011). A phylogenetic tree is then built on the aligned sequences using a tree-building algorithm (e.g. FastTree, Price *et al.* 2009) possibly after lanemasking the hypervariable regions. Optionally, a denoising procedure based on the flowgram data (*.sff) can be performed prior to the pipeline to reduce sequencing errors due to homopolymers(Quince *et al.*, 2009). The final output of the pipeline is an OTU table recording the counts for each OTU in each sample and a phylogenetic tree of the OTUs. The species-level OTUs can be aggregated into higher

taxonomic levels based on their assigned taxonomic lineages.

1.3. Motivation: The Penn gut microbiome project

As part of the Human Microbiome Demonstration Projects (UH2/UH3), the principal investigators here at Penn study the relationship between diet, genetic factors, and the gut microbiome in Crohn's disease. This is a collaborative project involving the PIs from Microbiology department (Rick Bushman) and Gastroenterology division(Gary Wu and James Lewis). We propose to investigate the hypothesis that consistent changes in the human gut microbiome are associated with Crohn's disease, a form of inflammatory bowel disease, and that altered microbiota contributes to pathogenesis. Analysis of this problem is greatly complicated by the fact that multiple factors influence the composition of the gut microbiome, including diet, host genotype, and disease state. Sequencing data alone cannot yield a useful picture of the role of the microbiome in disease if samples are confounded with uncontrolled variables. To untangle the major confounding variables, we first conduct a Cross-sectional Study of Diet and Stool Microbiome Composition (COMBO), to evaluate the association between dietary intake and the composition of the gut microbiome in healthy subjects in the outpatient setting. About 100 human subjects were enrolled in this study. The long-term dietary intake was determined by food frequency questionnaire (FFQ). Based on the FFQ, the intake values of 214 nutrients were calculated by nutritionists. Demographic data such as body mass index (BMI), age and sex were also available. For these subjects, stool samples were collected and the V1-V2 region of the 16S rRNA gene was sequenced by Roche/454 GS FLX Titanium system. Pyrosequencing produced about one million reads with an average read length of about 350bp.

The statistical methods developed in this dissertation are mainly motivated by analysis of the data from the COMBO project. Specifically, we develop new statistical methods to address the following problems:

- Given an outcome/a covariate such as BMI and nutrient intake, we want to test the as-

sociation between the covariate and the overall microbiome composition characterized by the OTU abundances and their phylogenetic constraint.

- If there are a large number of covariates as in the COMBO data set, where we have 214 nutrients, we want to select the most important covariates/nutrients.

- Finally, we want to perform more detailed analysis and select not only the covariates but also their associated OTUs.

The statistical methods to address these questions should take into account the characteristics of the OTU data discussed in the next section.

1.4. Characteristics of OTU/taxa data and statistical challenges

Fig. 2A shows the OTU count table for the 98 COMBO samples. From the count table, we can see five major characteristics of the OTU data.

First, the OTU count data are high-dimensional. The number of OTUs usually exceeds the number of samples. For example, at the species level (97% similarity), the COMBO data have 17,303 OTUs. Even consolidating the species-level OTUs into genera, we still have 127 genera. Variable selection becomes important for analysis of high-dimensional data.

Second, the OTUs are related by a phylogenetic tree (Fig. 2AB). The phylogenetic tree provides an important prior knowledge on the evolutionary relationship among OTUs. It can be useful in at least three ways:

The tree is informative to define a biologically meaningful distance measure. Consider a scenario, where we have two microbial communities with each having a unique set of OTUs. If the two sets of OTUs are closely related and interleaved on the tree, the two microbial communities share much of their evolutionary history and intuitively their distance should be small. On the other extreme, if these two sets of OTUs are located on different clades of the tree, each community has their own unique evolutionary history and their distance should be much larger than the first situation. Without the phylogenetic tree, we can not

distinguish between these two situations.

The tree can guide OTU selection. Closely related OTUs are genetically more similar and they are expected to have similar biological functions and respond to the environmental perturbation in a similar way. They have a natural tendency to be selected together. This is an important prior knowledge that we should exploit in OTU selection.

The tree also provides a hierarchical grouping of the OTUs. Environmental factor tends to affect one/several OTU lineages (OTUs that share a common ancestor) of different depths. Typically, microbiome data analysis is often performed at different taxonomic levels (genus, family, order, class, phylum). However, the taxonomic classification is usually arbitrary. Using the grouping structure implied by the OTU tree is more natural.

Third, the OTU counts are overdispersed, meaning that the variance of the counts is much larger than what would be predicted by assuming a common multinomial model. Fig. 4B shows that the counts generated by the multinomial model lacks variation compared to the real counts (Fig. 4A). Models allowing for overdispersion should be used to model the counts.

Fourth, the distribution of OTU abundance is very skewed (Fig. 2C). The majority of the OTUs have very low abundance with a large fraction being singletons (OTUs with only one sequence). The OTU count data are usually dominated by a small number of highly abundant OTUs. However, it is hard to say the low-abundance OTUs are not important. Statistical methods that place too much weight on the abundant lineages may miss important findings if the biologically important change occurs in less abundant lineages.

Finally, the distribution of the OTU occurrence probability is also skewed (Fig. 2D). Only a few OTUs are shared across samples, and the rest are seen in only a small percentage of the samples. This results in excessive 0's in the OTU data. Excessive 0's may be a result of count overdispersion or due to other mechanisms. Adequately modeling excessive 0's is important for OTU data.

The original OTU data are high-dimensional count data on the phylogenetic tree. Since the sequencing depth varies from sample to sample, the count data are usually normalized into proportions. As a way of further summarization, the data can be represented as pairwise distances between the samples, which can be thought as projecting the original data into lower dimensional space. In principle, we can develop statistical methods on any level of data summarization (counts, proportions and distances). However, from counts to distances, a large amount of information is lost, and the statistical power will be reduced accordingly. For proportion data, the variability associated with multinomial sampling is lost. Pairwise distances only capture certain features of the data.

In summary, the statistical challenges of OTU data analysis include incorporation of the OTU phylogenetic tree information, treatment of rare OTUs, modeling high-dimensional counts with overdispersion and excessive 0's, modeling high-dimensional composition/proportion data, and defining distance measures that capture a variety of microbiome differences.

1.5. Organization of the dissertation

The rest of the dissertation is organized as follows. Chapter 2 presents a distance based method for testing the association between overall microbiome composition and environmental/biological covariates. A generalized UniFrac distance (GUniFrac), which extends the classic UniFrac distances (Lozupone and Knight, 2005; Lozupone *et al.*, 2007), is defined between two microbiomes. The power of GUniFrac based test is compared to other UniFrac variants using simulations as well as the COMBO data set and an oropharyngeal microbiome data set(Charlson *et al.*, 2010).

When there are a large number of covariates, variable selection becomes important. Chapter 3 and Chapter 4 provide two methods for selecting the most important covariates under different frameworks. Chapter 3 proposes a method for structure-constrained sparse canonical correlation analysis (ssCCA), taking into account the phylogenetic relationship among OTUs. This method does not assume a specific probability model and can be regarded

as an exploratory analysis method. It takes OTU proportion data and outputs the most correlated OTUs and covariates. An efficient coordinate descent algorithm is implemented to obtain the ssCCA solution. The performance of ssCCA is compared to sparse CCA using simulations and the COMBO data set.

ssCCA takes the OTU proportion data, which does not consider the variation associated with multinomial sampling, and the result does not show detailed individual OTU-covariate associations. To deal with these limitations, a sparse Dirichlet-multinomial regression (SDMR) method, which links the OTU counts to covariates under a regression setting, is proposed. SDMR takes the OTU count data and outputs all identified associations. The OTU counts are modeled using Dirichlet-multinomial distribution to account for overdispersion, and a sparse group l_1 penalty function is imposed to achieve desired sparsity. A block-coordinate descent algorithm is implemented to obtain the maximum penalized likelihood estimate. The selection performance of SDMR is also evaluated using simulations and the COMBO data set in comparison to other possible models.

Chapter 5 concludes the dissertation with future research directions. Chapter 2-4 are self-contained and can be read independently.

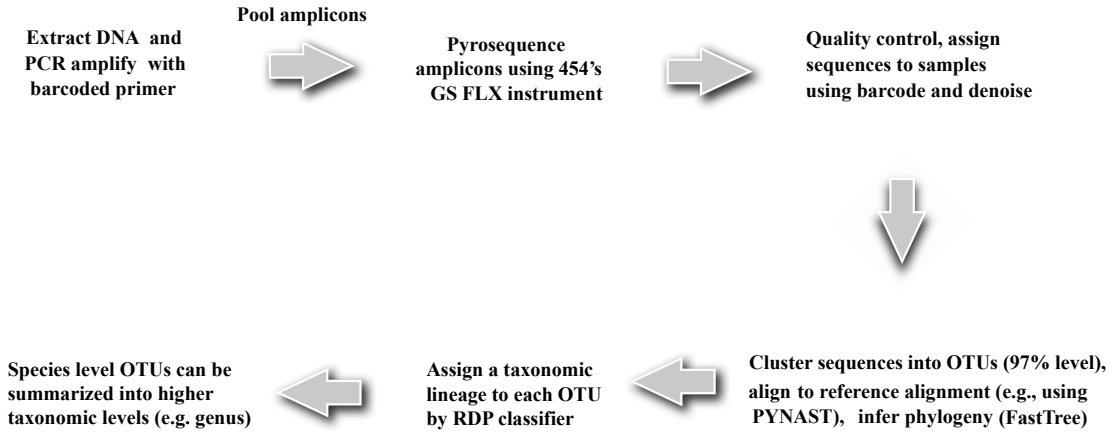

Figure 1: **Pipeline for 16S sequence data generation and processing.**

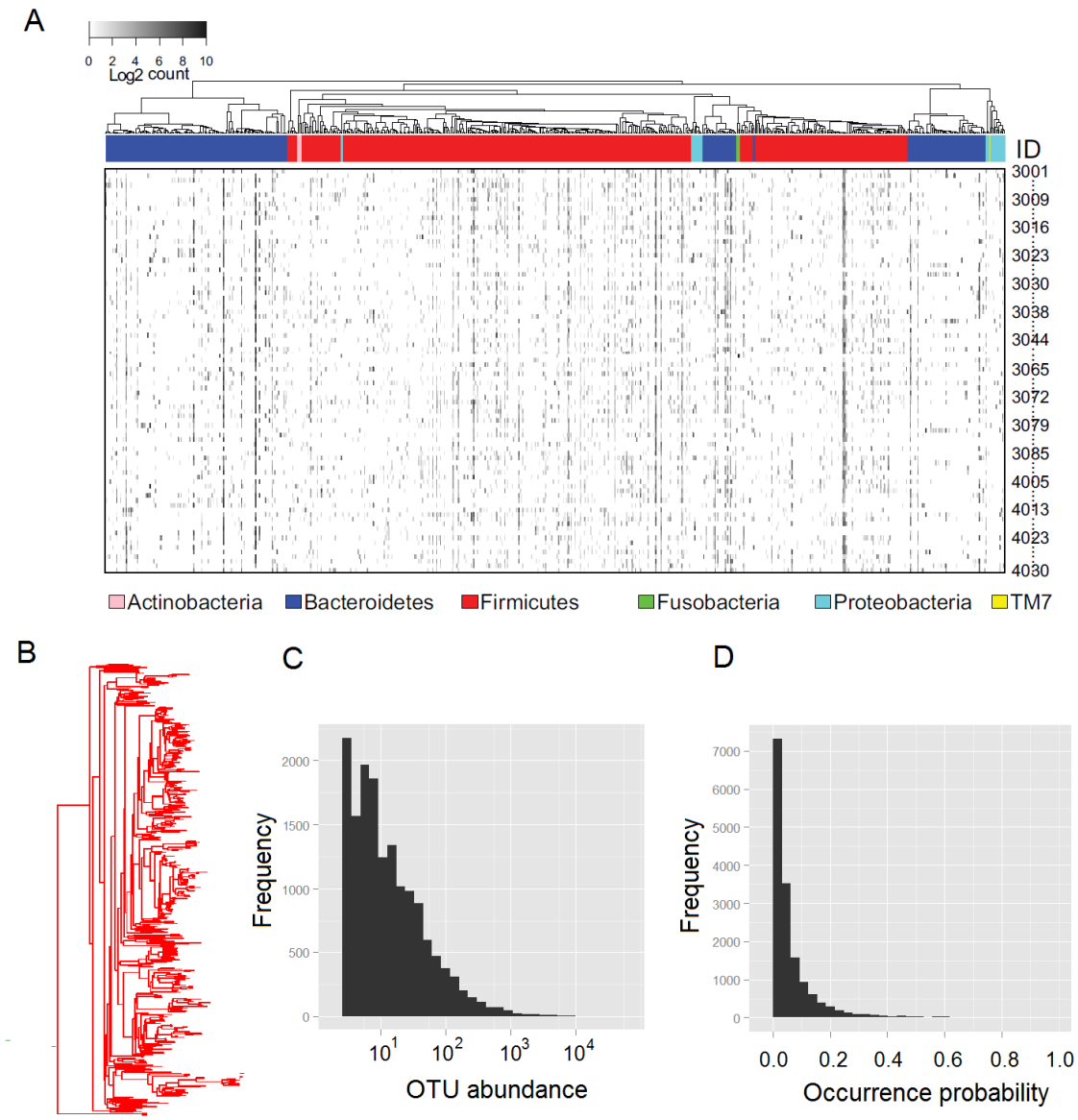

Figure 2: **Characteristics of OTU data illustrated using the COMBO data.** (A) The heatmap shows the OTU counts for the 98 COMBO samples. Rows represent samples and columns correspond to OTUs. These OTUs are related by a phylogenetic tree colored by phyla. The gray scale indicates the level of abundance on a log scale with white meaning zero counts (see legend). (B) The phylogenetic tree of OTUs. (C) The histogram shows the OTU abundance distribution. The OTU abundance (x-axis) is on the log scale. (D) The histogram shows the distribution of the OTU occurrence probability.

CHAPTER 2 : Associating Microbiome Composition with Environmental Covariates using Generalized UniFrac Distance

In this chapter, we propose a new distance for characterizing the difference between two microbial communities(microbiomes). Distance based statistical tests have been applied to test the association of microbiome composition with environmental/biological covariates. The unweighted and weighted UniFrac distances are the most widely used distance measures. However, these two measures assign too much weight either to rare lineages or to highly abundant lineages, which can lead to loss of power when the important composition change occurs in moderately abundant lineages. We develop generalized UniFrac distance that extends weighted and unweighted UniFrac distances for detecting a much wider range of biologically relevant changes. We evaluate the use of generalized UniFrac distance in associating microbiome composition with environmental covariates using extensive Monte Carlo simulations. Our results show that tests using the unweighted and weighted UniFrac distances are less powerful in detecting abundance change in moderately abundant lineages. In contrast, the generalized UniFrac distance is most powerful in detecting such changes, yet it retains nearly all its power for detecting rare or highly abundant lineages. The generalized UniFrac distance also has an overall better power than the joint use of unweighted/weighted UniFracs distances. Application to two real microbiome data sets have demonstrated gains in power in testing the associations between human microbiomes and dietary intake and smoking. An R package has been developed for generalized UniFrac distance and is available at **http://cran.r-project.org/web/packages/GUniFrac**.

2.1. Introduction

Understanding the compositional differences of microbial communities is essential in microbial ecology. With the development of next generation sequencing technologies, microbiome composition can now be determined by direct DNA sequencing without the need for laborious cultivation. There has been great interest in human microbiome studies in different

body sites, ranging from skin (Grice *et al.*, 2009) to gut (Qin *et al.*, 2010; Arumugam *et al.*, 2011; Muegge *et al.*, 2011; Wu *et al.*, 2011a) and respiratory tract (Charlson *et al.*, 2010, 2011; Sze *et al.*, 2012). Important insights have been gained from analysis of large-scale human microbiome data, including the discovery of enterotypes (Arumugam *et al.*, 2011) and discovery of the link between diet and these enterotypes (Wu *et al.*, 2011a).

Two recurring themes in human microbiome studies are to identify potential environmental factors that are associated with microbiome composition, and to define the relationship between microbiome features and biological or clinical outcomes. The goal is to provide a better understanding of the factors that shape our microbiome and, potentially, contribute to the development of new therapeutic strategies to modulate the microbiome composition (Spor *et al.*, 2011; Virgin and Todd, 2011) and affect the human health. Testing the association of microbiome composition with potential environmental factors using OTU abundances directly is difficult due to high dimensionality, non-normality and phylogenetic structure of the OTU data. Instead, distance based non-parametric test, in which a distance measure is defined between any two microbiome samples, is usually used to achieve this goal (Fukuyama *et al.*, 2012; Evans and Matsen, 2012; Kuczynski *et al.*, 2010a; Wu *et al.*, 2011a, 2010; Charlson *et al.*, 2010). The power of the distance based test depends on a proper choice of a distance measure. Numerous distance measures have been proposed to compare microbial communities (Kuczynski *et al.*, 2010b; Swenson, 2011). Phylogenetic distance measures, which account for the evolutionary relationship among species, provide far more power because they exploit the degree of divergence between different sequences. Among these, the UniFrac distances are the most popular ones (Lozupone and Knight, 2005; Lozupone *et al.*, 2007). There are two versions of UniFrac distances: an unweighted UniFrac distance that considers only species presence and absence information and counts the fraction of branch length unique to either community, and a weighted UniFrac distance that uses species abundance information and weights the branch length with abundance difference. Unweighted UniFrac distance is most efficient in detecting abundance change in rare lineages. When the abundance of a rare lineage falls below a certain threshold,

the sequencing machine may not be able to pick it up and it will appear absent in the final data set. On the other hand, weighted UniFrac distance is most sensitive to detect change in abundant lineages since it uses absolute abundance difference in its definition. However, Unweighted/weighted UniFrac distances may not be very powerful in detecting change in moderately abundant lineages. Recently, a variance adjusted weighted UniFrac distance (VAW-UniFrac), which moderates the branch proportion difference by its variance, was developed to account for the fact that weighted UniFrac distance does not consider the variation of the weights under random sampling (Chang *et al.*, 2011). VAW-UniFrac was shown to increase the power over weighted UniFrac distance for detecting the difference between two microbial communities.

In this chapter, we introduce generalized UniFrac distance that unifies weighted UniFrac and unweighted UniFrac distances. The new generalized UniFrac distance covers a series of distances ranging from weighted to unweighted UniFrac by adjusting the weight on the branches. The generalized UniFrac distance is designed to provide a robust and powerful tool for detecting a wider range of biologically relevant changes in microbiome composition. We conduct extensive Monte Carlo simulation studies under various conditions to evaluate their power in detecting environmental influence on microbiome composition using PERMANOVA (McArdle, 2001), a distance based non-parametric test. Although each distance in the series can perform the best in certain scenarios, none has optimal performance under all conditions considered. However, analyses based on the generalized UniFrac distance are shown to be more robust and has overall the best performances across a range of possible scenarios. We demonstrate the power gain of using this distance in detecting the microbiome differences by analysis of two real human gut microbiome data sets related to linking human gut microbiome composition to long-term diet (Wu *et al.*, 2011a) and testing oropharyngeal microbiome difference between smokers and non-smokers (Charlson *et al.*, 2010).

2.2. Generalized UniFrac distance between two microbial communities

Consider two microbiome communities A and B and suppose that we have a rooted phylogenetic tree with n branches. Let b_i be the length of the branch i and p_i^A, p_i^B are the taxa proportions descending from the branch i for community A and B, respectively. The unique fraction metric, or UniFrac, measures the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from either one environment or the other, but not both. The original definition refers to unweighted UniFrac (Lozupone and Knight, 2005), which is mathematically defined as

$$d^U = \sum_{i=1}^{n} \frac{b_i \left| I(p_i^A > 0) - I(p_i^B > 0) \right|}{\sum_{i=1}^{n} b_i},$$

where $I(.)$ is the indicator function and only presence/absence of species of branch i, $I(p_i^A > 0)$ and $I(p_i^B > 0)$, are used in the definition. The distance definition d^U completely ignores the taxa abundance information. In contrast, the (normalized) weighted UniFrac distance (Lozupone *et al.*, 2007) weights the branch length with abundance difference and is defined as

$$d^W = \frac{\sum_{i=1}^{n} b_i \left| p_i^A - p_i^B \right|}{\sum_{i=1}^{n} b_i (p_i^A + p_i^B)}.$$

Note that d^W can not be reduced to d^U even if we convert abundance data into presence/absence data. Also note that d^W uses the absolute proportion difference $\left| p_i^A - p_i^B \right|$ in its formulation. The consequence of using the absolute difference is that the value of d^W is determined mainly by branches with large proportions and is less sensitive to the abundance changes on the branches with small proportions. To attenuate the weight on branches with large proportions, we may instead use the relative difference $\left| p_i^A - p_i^B \right| / (p_i^A + p_i^B) (\in [0, 1])$

in the formulation. We denote this distance measure as

$$d^{(0)} = \frac{\sum_{i=1}^{n} b_i \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{n} b_i},$$

where $\sum_{i=1}^{n} b_i$ in the denominator is the normalizing factor so that $d^{(0)} \in [0, 1]$. Now if we dichotomize the abundance data using the indication function $I(.)$, $d^{(0)}$ is reduced to d^U. So $d^{(0)}$ can be seen as the "weighted version" of d^U. Using the relative differences, we place equal emphasis on every branch and the distance is not dominated by the branches with large proportions, since the relative difference does not depend on the magnitude of p_i^A, p_i^B. However, the low-abundance branches may be more noisy and the relative difference may amplify such noises. To strike a balance between relative difference and absolute difference, we weight the branch length both by the relative difference and its importance indicated by the branch proportion. We propose the following generalized UniFrac distance

$$d^{(\alpha)} = \frac{\sum_{i=1}^{n} b_i (p_i^A + p_i^B)^\alpha \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{n} b_i (p_i^A + p_i^B)^\alpha},$$

where $\alpha \in [0, 1]$ controls the contribution from high-abundance branches, and $\sum_{i=1}^{n} b_i (p_i^A + p_i^B)^\alpha$ is the normalizing factor so that $d^{(\alpha)} \in [0, 1]$. Branches with zero proportions for both communities will not be included in the calculation. As α changes from 0 to 1, more emphasis is placed on high-abundance branches. When $\alpha = 1$, $d^{(\alpha)}$ is reduced to d^W. When $\alpha = 0$, we get $d^{(0)}$ defined above.

Therefore, by varying α from 1 to 0 , we achieve a series of distances ranging from d^W to $d^{(0)}$. Note $d^{(0)}$ is obtained by dichotomizing the abundance in $d^{(0)}$, but is different from $d^{(0)}$.

We are particularly interested in $d^{(0.5)}$, the distance in the middle of the distance series

$$d^{(0.5)} = \frac{\sum_{i=1}^{n} b_i \sqrt{p_i^A + p_i^B} \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{n} b_i \sqrt{p_i^A + p_i^B}}.$$

We also compare $d^W, d^{(0.5)}, d^{(0)}$ and d^U to VAW-UniFrac distance d^{VAW}, which is defined as:

$$d^{VAW} = \frac{\sum_{i=1}^{n} b_i \frac{|p_i^A - p_i^B|}{m(m - m_i)}}{\sum_{i=1}^{n} b_i \frac{p_i^A + p_i^B}{m(m - m_i)}},$$

where m_i is the total number of sequences from both communities on the ith branch, and m is total number of sequences.

2.3. Statistical test based on UniFrac distances

We study the power of generalized UniFrac distance using the distance-based non-parametric test for association of microbiome composition with environmental covariates. Suppose we have a set of m environmental covariates. We assume that we have collected microbiome data and the m-dimensional covariates data $\mathbf{X}$ on n samples. We apply the PERMANOVA procedure (McArdle, 2001) (Permutational Multivariate Analysis of Variance Using Distance Matrices, "adonis" function from R package "vegan"), which partitions the distance matrix among sources of variation, fits linear models to distance matrices and uses a permutation test with pseudo-F ratios to obtain the p-values. The pseudo-F statistic is defined as:

$$F = \frac{\mathrm{tr}(\mathbf{HGH})/(m-1)}{\mathrm{tr}[(\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H})]/(n-m)},$$

where tr(.) is the trace function of a matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the hat (projection) matrix of the design matrix $\mathbf{X}$, $\mathbf{G}$ is Gower's centered matrix and n, m is the number of samples and the number of predictors respectively. Let d_{ij} be the distance between

community i and j, and denote $\mathbf{A} = (a_{ij}) = (-\frac{1}{2}d_{ij}^2)$. The Gower's matrix is defined as

$$\mathbf{G} = (\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n})\mathbf{A}(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}),$$

where $\mathbf{1}$ is a vector of 1's.

Since d^U and d^W reflect the abundance change in either rare lineages or abundant lineages, combining d^U and d^W may potentially increase the overall power. Instead of applying Bonferroni correction to the p values from separate PERMANOVA tests using d^U or d^W to control the family-wise type I error rate, a more powerful approach is to take the maximum of pseudo-F statistics for d^U and d^W as a new test statistic. The significance of the pseudo-F statistics is assessed based on permutations.

2.4. Simulation studies

2.4.1. Simulation strategies

We use two simulation strategies to evaluate the power of the generalized UniFrac distance under various conditions. The first strategy is a modification of the simulation method proposed by Schloss (2008), where we draw points (16S rDNA sequences) from a 2D circle with known densities (Fig. 3A). This strategy facilitates simulations of different community characteristics such as species evenness and richness. The Euclidean distance between points is analogous of the genetic distance between the sequences. The diameter of the circle represents the maximum genetic divergence between any pair of sequences within a sample. The area of the circle is proportional to the richness and the density distribution of the circle is proportional to the evenness. By varying the centroid positions (o) and their radius (r), it is possible to vary the fraction of shared membership and species richness within each sample (Fig. 3B,D). By varying the point distribution on the circle (density proportional to r^α, where α controls the degree of evenness and $\alpha = 0.5$ for uniformly distribution), it is possible to change the species evenness (Fig. 3C). We also simulate scenarios where lineages of different abundance levels change by a k fold (Fig. 3E-G). These are achieved

by simulating the community with point mass concentrated at the circle center ($r^{1.0}$) and varying the point density in different regions of the 2D circle corresponding to abundant lineages ($0-0.2r$ from the center, Fig. 3E), moderately abundant lineages ($0.4r-0.8r$ from the center, Fig. 3F), and rare lineages ($0.8r-1.0r$ from the center, Fig. 3G). We further bin the sampled points into small hexagons as "OTU"s before calculating the UniFrac distances ("hexbin" function from the R package "hexbin"). The phylogenetic tree of these "OTU"s is built using NJ algorithm (Neighbor Joining, "nj" function in R) and rooted by midpoint rooting method. UniFrac distances are then calculated based on the NJ tree and "OTU" abundances. Each replication consists of drawing 400 points from each community, a bin size of 0.015 units to form "OTUs" ($\sim$ 300 OTUs per sample), and the maximum distance between any two points is 0.3 units ($r = 0.15$), corresponding to typical phylum level divergence of 30% for 16S rRNA gene. These conditions allow us to simulate the sampling intensity and biodiversity found within a typical 16S rRNA gene targeted sequencing experiment (Schloss, 2008).

The second set of simulations utilizes a real oropharyngeal microbiome data set consisting of 60 samples and 856 OTUs from Charlson *et al.* (2010) (Fig. 3H). A common way of modeling multivariate count data is to use the multinomial model. However, the multinomial model assumes fixed underlying proportions for each sample, which does not hold for real microbiome data due to high degree of heterogeneity among the samples. The real OTU count distribution (Fig. 4A) exhibits more variance than expected from a multinomial model (Fig. 4B). To realistically simulate the data, it is important to model extra-variation or overdispersion of the OTU counts. This can be achieved by using the Dirichlet-multinomial (DM) model (Mosimann, 1962), which assumes the underlying proportions of the multinomial model come from a Dirichlet distribution. The density function of a DM random

variable $\mathbf{N}$ is given as

$$P(\mathbf{N}=\mathbf{n}) = \binom{n}{\mathbf{n}} \frac{\prod_{j=1}^{k}\prod_{r=1}^{n_j}\{\pi_j(1-\theta)+(r-1)\theta\}}{\prod_{r=1}^{n}\{1-\theta+(r-1)\theta\}},$$

where $n=\sum_j n_j$ is total count, k is the OTU number, and proportion mean $\pi=(\pi_1,\pi_2,\cdots,\pi_k)$ and dispersion θ are parameters. When $\theta=0$, it is reduced to multinomial model. We estimate the DM parameters π,θ using maximum likelihood method ("dirmult" function from R package "dirmult"). We then generate OTU counts using the DM model with the estimated parameters and $1,000$ counts per sample. Fig. 4C shows an OTU heatmap generated by the DM model, in which the overdispersion is similar to that of the real data. To study the power of UniFrac variants for identifying potential environmental factors, we let the abundance of a certain OTU cluster change in response to environment. We use UPGMA tree of OTUs ("hclust" function in R) based on the OTU distance matrix calculated under the K80 nucleotide substitution model (Felsenstein, 2003), and partition the 856 OTUs into 20 clusters using Partitioning Around Medoids (PAM) ("pam" function from R package "cluster") based on patristic distances (the length of the shortest path linking two OTUs on the tree). These OTU clusters are highlighted in different colors in Fig. 3H.

We call the first strategy 2D circle based simulation and the second tree based simulation. For power calculation, we use $2,000$ replications.

2.4.2. Comparison of the power of different UniFrac variants using 2D circle based simulations

We use PERMANOVA to test for environmental effects and compare the power of $d^W, d^{(0.5)}, d^{(0)}$, d^U and d^{VAW} . Specifically, we simulate two environmental conditions (e.g. smoking vs non-smoking) under which we draw 10 samples each. We then vary the degree of community difference under these two conditions and produce the power curve over a grid of 10 for each

UniFrac distance. We investigate six scenarios, where the environmental factor affects the community membership, species evenness, species richness, most abundant lineages, moderately abundant lineages, and rare lineages respectively (Fig. 3B-G). For each scenario, we vary one community characteristic (Table 1).

Suppose x_1 and x_2 are the mean values of the community characteristic for condition 1 and 2. We simulate 10 communities for each condition with community characteristic value $x_{ij} \sim \text{Uniform}(x_j - s, x_j + s)$ for $i = 1 \cdots 10$ and $j = 1, 2$, where s controls the variation within each condition. Each community is sampled once. Initially, we let $x_1 = x_2$ (no difference) and then increase the difference between x_1 and x_2 to simulate stronger environmental effects. PERMANOVA is then performed on the distance matrices and the power curve is created over a grid of 10 using type I error $\alpha = 0.05$. Fig. 5 shows the power curves for different UniFrac distances under the six scenarios considered. When the environmental factor has no effect ($x_1 = x_2$), PERMANOVA controls the type I error at the nominal level of 0.05 for all five UniFrac distances. As the environmental effects become stronger, all the distances have better power. When the environmental factor affects the community membership or richness (Panel 1, 3), all the distances give a similar power and their power curves are nearly identical. For the evenness change scenario (Panel 2), the power of d^W and $d^{(0.5)}$ is very close and is more powerful than $d^{(0)}$ and d^U. d^W is the most powerful for detecting change in most abundant lineages (Panel 4) but is much less powerful for change in rare lineages (Panel 6). d^U shows an opposite trend: it is the most powerful for detecting change in rare lineages (Panel 6) but has almost no power for change in most abundant lineages (Panel 4). In contrast, $d^{(0.5)}$ is the most powerful for detecting change in moderately abundant lineages (Panel 5). They are also the most robust among the distances investigated: its power is close to the best UniFrac distance under all scenarios. The performance of $d^{(0)}$ lies between $d^{(0.5)}$ and d^U, and is also very robust. Finally, the performance of d^{VAW} is almost identical to $d^{(0.5)}$ under this simulation setting.

In the above simulations, we use a bin size of 0.015 to form "OTU"s ($\sim$ 300 OTUs per

sample). To study the effect of bin size, we compare the power curves of UniFrac distances using a smaller bin size of 0.01 ($\sim$ 700 OTUs per sample) or a larger bin size of 0.03 ($\sim$ 80 OTUs per sample). The bin size does not change the general conclusion (Appendix Fig. A1). To study the effect of tree construction methods, we also construct the phylogenetic tree using UPGMA. The general conclusions still hold (Appendix Fig. A2).

2.4.3. Comparison of the power of different UniFrac variants using tree based simulations

We also compare the power of different UniFrac distances for detecting environmental effects using tree based simulations that mimic the oropharyngeal microbiome data (see Section 2.5.2 for details). The phylogenetic tree of the 856 OTUs is partitioned into 20 clusters (Fig. 3H). The mean OTU proportions and the dispersion parameter are estimated from the real data by fitting a Dirichlet-multinomial (DM) model. We assume that the environmental factor causes an increase of the abundance of a particular OTU cluster. Specifically, suppose that the proportion of the ith OTU cluster under condition 1 is p_i. For condition 2, the proportion of ith OTU cluster is increased by k fold where k varies from 1 (no difference) to $1/\sqrt{p_i}$ (strong effect) on a grid of 10. The proportion vector is re-normalized to sum to 1. Next, 10 samples are simulated for each condition with their OTU counts generated by the DM model with the corresponding proportion vector and the common dispersion parameter. As expected, the five UniFrac distances differ in their power for detecting environmental effects for the 20 OTU clusters tested. Except for $d^{(0)}$, all the UniFrac distances have their best-performance scenarios. d^W, $d^{(0.5)}$, d^U and d^{VAW} achieve the highest power in 7, 6, 3 and 1 cases respectively. For the remaining 3 cases, d^W and $d^{(0.5)}$ are equally the most powerful (Appendix Fig. A3). The results are consistent with the 2D-circle based simulation: d^W is most powerful for detecting the environmental effects on most abundant lineages, $d^{(0.5)}$ for moderately abundant lineages and d^U for rare lineages. In contrast, performance of the test with $d^{(0)}$ and d^{VAW} is generally between d^U and $d^{(0.5)}$. The power of d^W and d^U has a reciprocal relationship and neither of them is as robust as $d^{(0.5)}$. Fig. 6A shows the power curves of four representative cases. As the proportion of the affected

cluster decreases from 19.7% to 0.9%, d^W becomes less powerful and the power of d^U has the opposite trend.

In the simulations presented above, the power is calculated assuming we know the cluster effected. Since the cluster affected can be abundant or rare, we randomly choose an affected OTU cluster in each replication and calculate the power over $2,000$ replications. We also report the power for the test combining d^W and d^U by taking the maximum of their pseudo-F statistics. We denote this method as d^{MAX}. Fig. 6B (left plot) demonstrates that d^U and d^{VAW} have the lowest overall power than the other distances, and $d^{(0.5)}$ and d^{MAX} have the best power indicating combining d^U and d^W can increase power. In contrast, $d^{(0)}$and d^W are in between and as the environmental effect becomes stronger, $d^{(0)}$ becomes as powerful as $d^{(0.5)}$ and d^{MAX}. Lastly we assume that the environmental factor affects a random set of 40 OTUs from the phylogenetic tree instead of a random OTU cluster. At this extreme, where phylogenetic relationship is no longer important, $d^{(0.5)}$ has even higher power than the other distances, followed by $d^{(0)}$, d^{MAX}, d^W, d^U and d^{VAW} (see Fig. 6B, right plot). Overall, $d^{(0.5)}$ has a better power than any other UniFrac distances including the one that combines d^W and d^U.

2.5. Application to real data analysis

2.5.1. Results from analysis of a data set linking long-term diet to gut microbiome composition

Diet strongly affects the human health, partly by modulating gut microbiome composition. Wu *et al.* (2011a) studied the long term diet effect on the human gut microbiome, where the diet information was converted into a vector of micro-nutrient intakes. A cross-sectional analysis of 98 healthy volunteers were enrolled in this study. Diet information was collected using food frequency questionnaire (FFQ). The questionnaires were converted to intake amounts of 214 micro-nutrients. Nutrient intake was further normalized using the residual method to standardize for caloric intake. Stool samples were collected and V1-V2 region

of the 16S rRNA gene was sequenced by 454/Roche GS FLX Titanium system. The 16S pyrosequences were denoised (Quince *et al.*, 2009) prior to taxonomic assignment yielding an average of $9,265 \pm 3,864(SD)$ reads per sample. The denoised sequences were then analyzed by the QIIME pipeline (Caporaso *et al.*, 2010b) with the default parameter setting. The OTU table contains 3068 OTUs after discarding the singleton OTUs. We use the phylogenetic tree generated by QIIME (FastTree algorithm, Price *et al.* 2009) to construct the UniFrac distances. One objective of the study is to identify nutrients that have a significant impact on the gut microbiome composition. We use PERMANOVA to test for association of microbiome composition with nutrient intake based on different UniFrac distance matrices. We compare $d^{(0.5)}$ with d^U, d^W, their combination d^{MAX} and d^{VAW}. We plot the number of selected nutrients against different p-value cutoffs to create a ROC-like curve (Fig. 7). Clearly the curve for $d^{(0.5)}$ is above all the other four curves. Wilcoxon signed-rank tests show that $d^{(0.5)}$ results in smaller p-values than other distances ($p < 0.05$), indicating that$d^{(0.5)}$ is most powerful in selecting the relevant microbiome-associated nutrients. Using d^W or d^U only could miss important associations. Power of d^{VAW} is the second best. Interestingly, d^{MAX}, the joint use of d^W and d^U, does not increase the power over d^W, indicating most associations can be recovered by d^W alone.

2.5.2. Results from analysis of an oropharyngeal microbiome data set smokers and non-smokers

Cigarette smokers have an increased risk of multiple diseases, including upper respiratory tract infections. Previous studies had linked smoking to specific respiratory tract bacteria but the consequences of smoking for global airway microbial community composition had not been fully clarified. Charlson *et al.* investigated the smoking effect on the oropharyngeal and nasopharyngeal bacterial communities using 454 pyrosequencing of 16S sequence tags Charlson *et al.* (2010). Specifically, a total of 291 swab samples from the right and left nasopharynx and oropharynx of 29 smoking and 33 nonsmoking healthy asymptomatic adults were collected. V1-V2 region of the bacterial 16S rRNA gene was PCR-amplified using

individually barcoded primer set and subject to multiplexed pryosequencing by 454/Roche GS FLX Titanium system. The pyrosequences were denoised (Quince *et al.*, 2009) prior to taxonomic assignment and yielded an average of $1,335 \pm 603(SD)$ reads per airway sample. The denoised sequences were then analyzed using the QIIME pipeline (Caporaso *et al.*, 2010b) with default parameter setting. We use the left oropharyngeal samples in this study. After removing two samples with read number less than 500 and discarding singleton OTUs, we finally have an OTU table of 60 samples (28 smokers vs 32 nonsmokers) and 856 OTUs. The phylogenetic tree produced by QIIME was used to construct the distances.

We test the smoking effect on the oropharyngeal microbial community composition by applying PERMANOVA (10,000 permutations). All the five UniFrac distances achieve statistical significance at $\alpha = 0.05$ level, indicating smoking alters the community composition. However, test using $d^{(0.5)}$ produces the smallest p-value of 0.006, followed by 0.008 from $d^{(0)}$. The p-values based on d^W, d^U and d^{VAW} are 0.012, 0.019 and 0.043 respectively. We also perform a principle coordinate analysis on the distance matrices, and plot the samples on the first two principle coordinates (Fig. 8). $d^{(0.5)}$ separates the samples better than the other three distance measures. This indicates that smoking might affect not only the predominant lineages but also these less abundant lineages in the oropharyngeal microbial community. We then performed Wilcox rank-sum test or Fisher's exact test to select the differential OTUs. At $\alpha = 0.05$ level, we identify 32 OTUs (Appendix Table A1). These OTUs belong to genera Prevotella (8), Lachnospiraceae (5), Veillonella (3), Streptococcus (2), Fusobacterium (2), Treponema (2), Neisseria (1), Haemophilus (1), Megasphaera (1), Dialister (1), Moryella (1), Erysipelotrichaceae (1) and four genera from Actinobacteria. Most of the selected OTUs are moderately abundant or rare, so we expect $d^{(0.5)}$ and $d^{(0)}$ to have better power.

Finally, we study the effect of tree constructing methods on generalized UniFrac distance. Besides using the tree from QIIME, we also construct the phylogenetic tree by NJ, UPGMA, parsimony and maximum likelihood methods. NJ and UPGMA are based on the

distance matrix generated by the R function "dist.dna" from the "ape" package (pairwise.deletion=T) under the K80 nucleotide substitution model Felsenstein (2003). The parsimony and maximum likelihood methods are implemented using the DNAPARS and DNAML program with default parameter setting in PHYLIP 3.69. All the unrooted trees are rooted using midpoint rooting method. We observe that different tree constructing methods produce similar results (Appendix Fig. A4).

2.6. Discussion

Microbiome data are multivariate count data in its original form and are statistically challenging to analyze due to their high dimensionality, phylogenetic constraints among species/OTUs, overdispersion and excessive zeros. To circumvent the difficulty, the data are often summarized in the form of distance matrix. Testing association of microbiome composition with environmental covariates is performed using the distance matrix. We have demonstrated in simulations that the weighted and unweighted UniFrac impose large weight either to abundant lineages or to rare lineages, they can be underpowered in detecting change in moderately abundant lineages. Since microbiome composition change could occur in any lineages, our generalized UniFrac distance, which unifies the weighted and unweighted UniFrac in a common framework enable us to detect a much wider range of biologically relevant changes. Our simulation studies have clearly demonstrated that the generalized UniFrac distance $d^{(0.5)}$ is more robust than d^W or d^U, and its performances are in general comparable to the best UniFrac distances among the scenarios we considered. In addition, the generalized UniFrac distance is very robust to tree constructing methods. We suggest the use of $d^{(0.5)}$ for testing association of microbiome composition with environmental covariates to avoid missing important findings.

Both weighted and unweighted UniFrac distances are sensitive to sampling depth (Lozupone *et al.*, 2010). Inflated distances at lower sampling depth are caused by sampling variation especially for these rare lineages. The generalized UniFrac distance is also sensitive to sampling depth (Fig. 9). However, as the sampling effort increases, the distance stabilizes.

For the gut microbiome data set, we found a sequencing depth of ~ 1000 reads is sufficient to stabilize the generalized UniFrac distance. To overcome potential adverse effects of uneven sampling, rarefaction is usually employed to subsample the samples to the same depth. When the sampling depth varies greatly across the samples, rarefaction will throw away a significant portion of the 16S reads and increase the sampling variation artificially. We found that rarefaction is not necessary, at least, in the context of testing the association of the microbiome composition with covariates (Fig. 10).

The VAW-UniFrac (Chang *et al.*, 2011) also up-weights the differences on less abundant lineages by adjusting the variance of the weights. If we assume the number of reads from the two communities are the same and divide the weights in $d^{(0.5)}$ by $\sqrt{(2 - p_i^A - p_i^B)}$, then $d^{(0.5)}$ becomes VAW-UniFrac. Usually the majority of the branches have low proportions, so $\sqrt{(2 - p_i^A - p_i^B)}$ for majority of the branches are similar ($\sim \sqrt{2}$). This accounts for the similarity of $d^{(0.5)}$ and d^{VAW} in the 2D circle based simulations. When the phylogenetic tree is constructed based on the OTUs of multiple samples (>2), the lowest common ancestor (LCA) of any two communities is not necessarily the same as the root of the whole tree. This is frequently seen in real data, where some samples occupy only a subtree. These common branches between the LCA and the root are not included in the calculation of VAW-UniFrac distance due to division by 0 for these branches. Ignoring these common branches will inflate the distance. In contrast, $d^{(0.5)}$ does not have this limitation. This accounts for the superiority of $d^{(0.5)}$ over d^{VAW} in tree-based simulation as well as in real data analysis.

The power of UniFrac variants can also be compared in the context of testing whether two microbial communities differ significantly as in (Schloss, 2008; Chang *et al.*, 2011). Instead of comparing power for detecting the difference between two communities, we focus our evaluations on the performance of UniFrac distances for associating microbiome composition to environmental covariates by collecting multiple independent samples. The rational is that as the sequence depth increases, two sample comparison will have increased power

to detect differences due to sources that we are not interested in (random noises), such as the individual-to-individual variability, day-to-day variability, sampling location variability or even technical variability (e.g. sample preparation). Multiple samples from a population coupled with multivariate statistical methods such as the distance based PERMANOVA, provide powerful design and analysis methods to overcome these potential random noises (Lozupone *et al.*, 2010). As more and more large-scale microbiome data sets are being collected, we expect that our generalized UniFrac distance can help to identify important covariates that are associated with the microbiomes that could be missed using the commonly used UniFrac distances. In addition to identifying environmental covariates that may be determinants of microbiome composition, our approach would be equally suited to identifying microbiome features associated with biological or clinical outcomes, which is needed to begin to understand the impact of the microbiome on health.

Table 1: **Parameter values used in power study for 2D circle-based simulation**

Parameter	Condition 1 (x_1)	Condition 2 (x_2)	Value of s
Centroid position (o)	$0 \sim 0.016$	$0 \sim -0.016$	0.03
Evenness (α)	$1 \sim 0.7$	$1 \sim 1.3$	0.3
Radius (r)	$0.15 \sim 0.132$	$0.15 \sim 0.168$	0.03
Scenario 4 Fold change (k)	1 (Fixed)	$1 \sim 2.5$	1
Scenario 5 Fold change (k)	1 (Fixed)	$1 \sim 2.8$	1
Scenario 6 Fold change (k)	1 (Fixed)	$1 \sim 6$	2

The values of x_1, x_2 are evenly spaced on a grid of 10. For Scenario 6, we decreased the abundances.

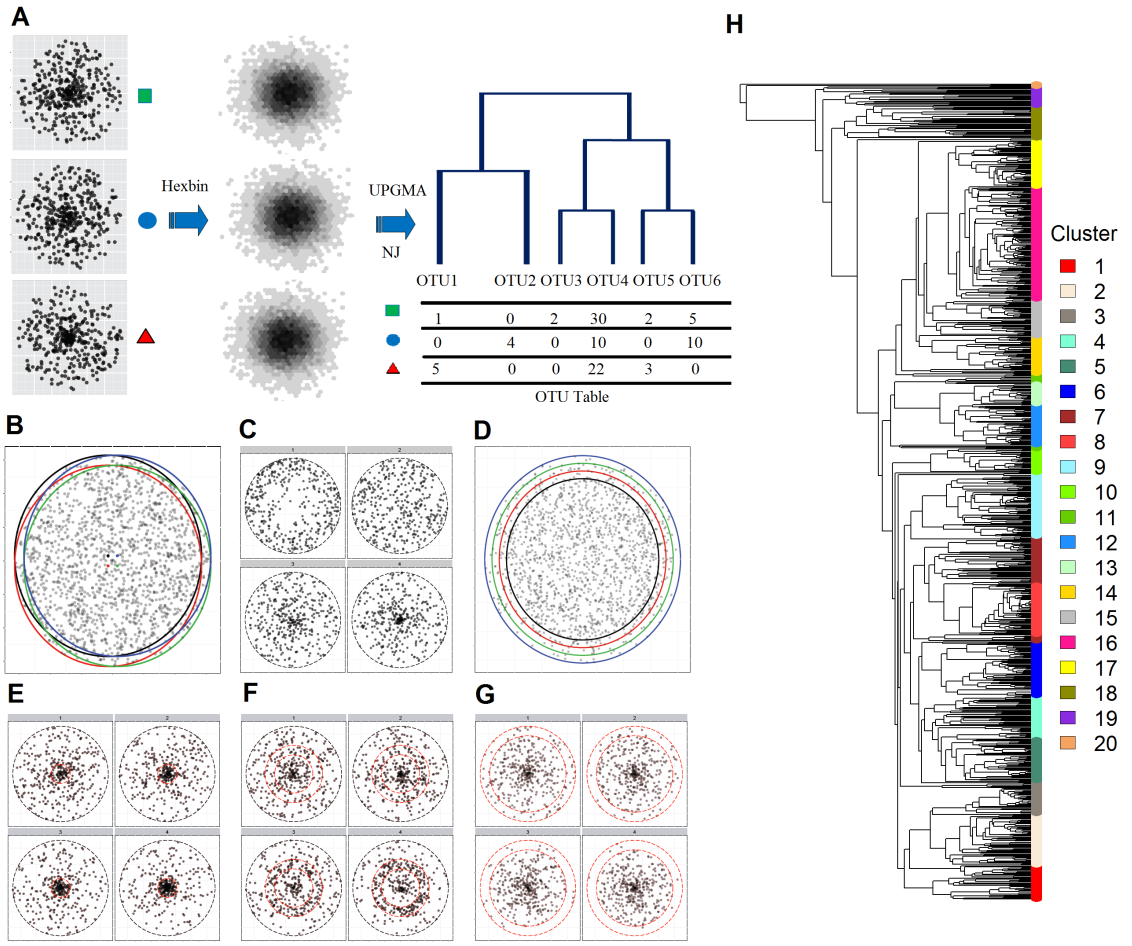

Figure 3: **Two simulation strategies to evaluate the generalized UniFrac distance.** A-G, 2D circle based simulation of microbial communities with different characteristics. (A) The microbial community is represented by a 2D circle. Points are drawn from the circle to simulate the 16S based sampling process. These points are further binned into small hexagons as OTUs. UPGMA or NJ method is used to build the OTU phylogenetic tree. Six scenarios are investigated, where the difference occurs in: community membership (B), evenness (C), richness (D), most abundant lineages (E), moderately abundant lineages (F) and rare lineages (G). The affected lineages are indicated by a red circle or ring. H, tree based simulation of microbial communities based on the phylogenetic tree and Dirichlet-multinomial model. A real OTU phylogenetic tree from an oropharyngeal microbial community data set is used. These OTUs are roughly divided into 20 clusters (lineages) by performing PAM method using the OTU patristic distance matrix. Each cluster is subjected to abundance change in response to the environment. Counts are generated from a Dirichlet-multinomial model.

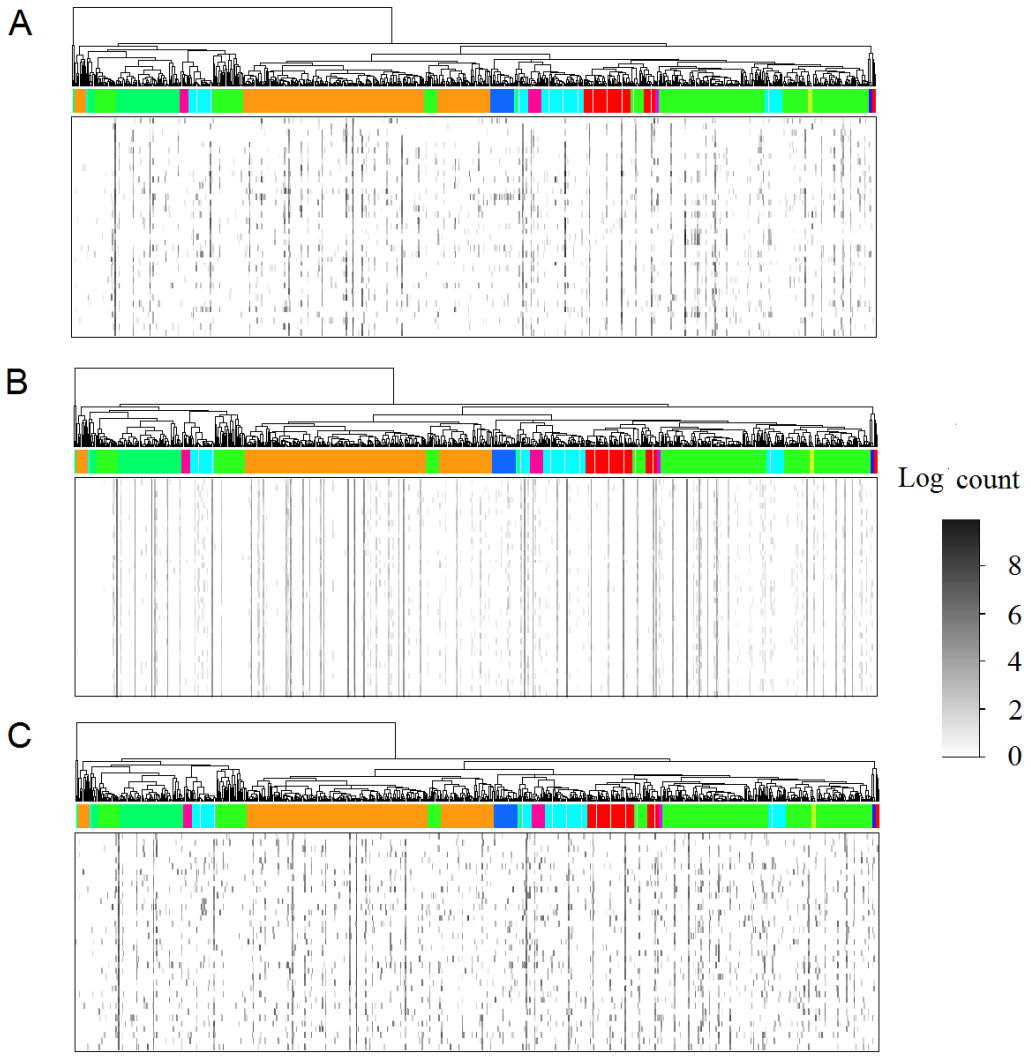

Figure 4: **Comparison of multinomial model and Dirichlet-multinomial model for simulating OTU counts for an oropharyngeal microbial community.** (A) The heatmap shows the OTU abundance distribution from a real oropharyngeal microbial community of 60 samples. Rows represent samples while columns correspond to OTUs. These OTUs are related by a phylogenetic tree colored by phyla. The gray scale indicates the level of abundance on a log scale with white meaning zero counts (see legend). (B) The OTU counts are generated by assuming a multinomial model, where the parameters are estimated from (A). (C) The OTU counts are generated by assuming a Dirichlet multinomial (DM) model, where the parameters are estimated from (A). The DM models overdispersion better than the multinomial model.

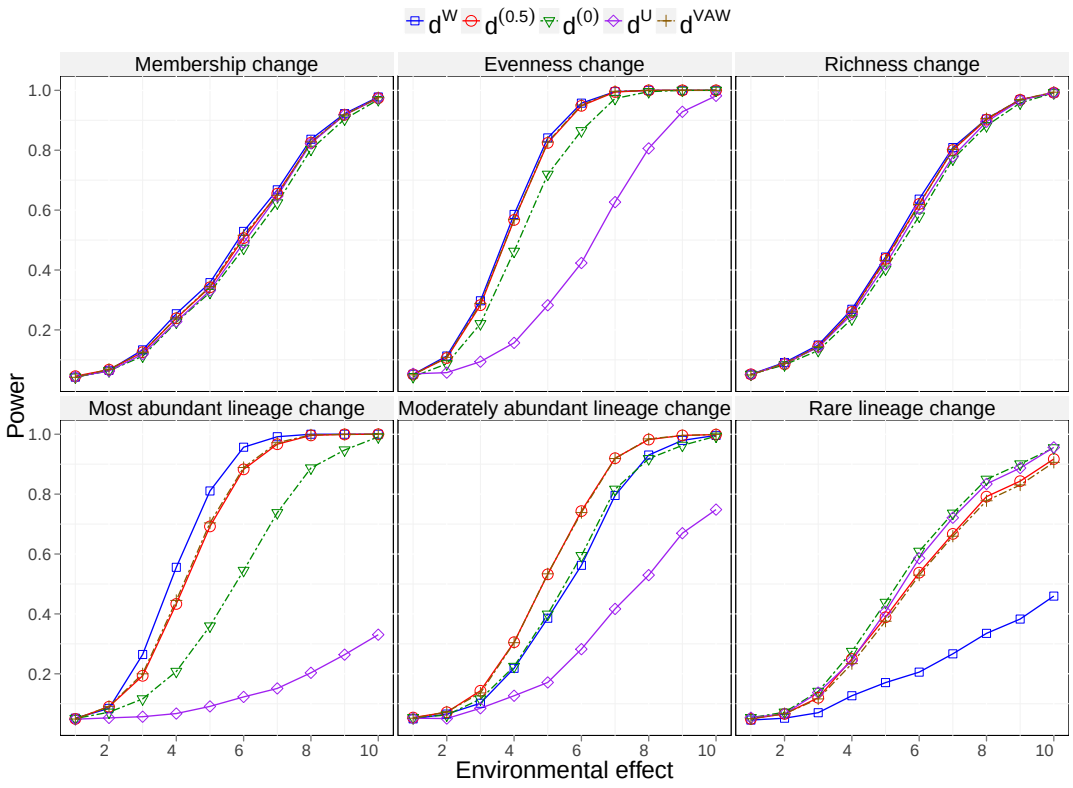

Figure 5: **Power comparison of different UniFrac variants for detecting environmental effects using 2D circle based simulation.**PERMANOVA is used for testing hypotheses. The specific community difference caused by different environmental conditions is indicated in the panel title. The power curves are created by varying the degree of environmental effect. The initial point of the power curve is the power when there is no environmental effect.

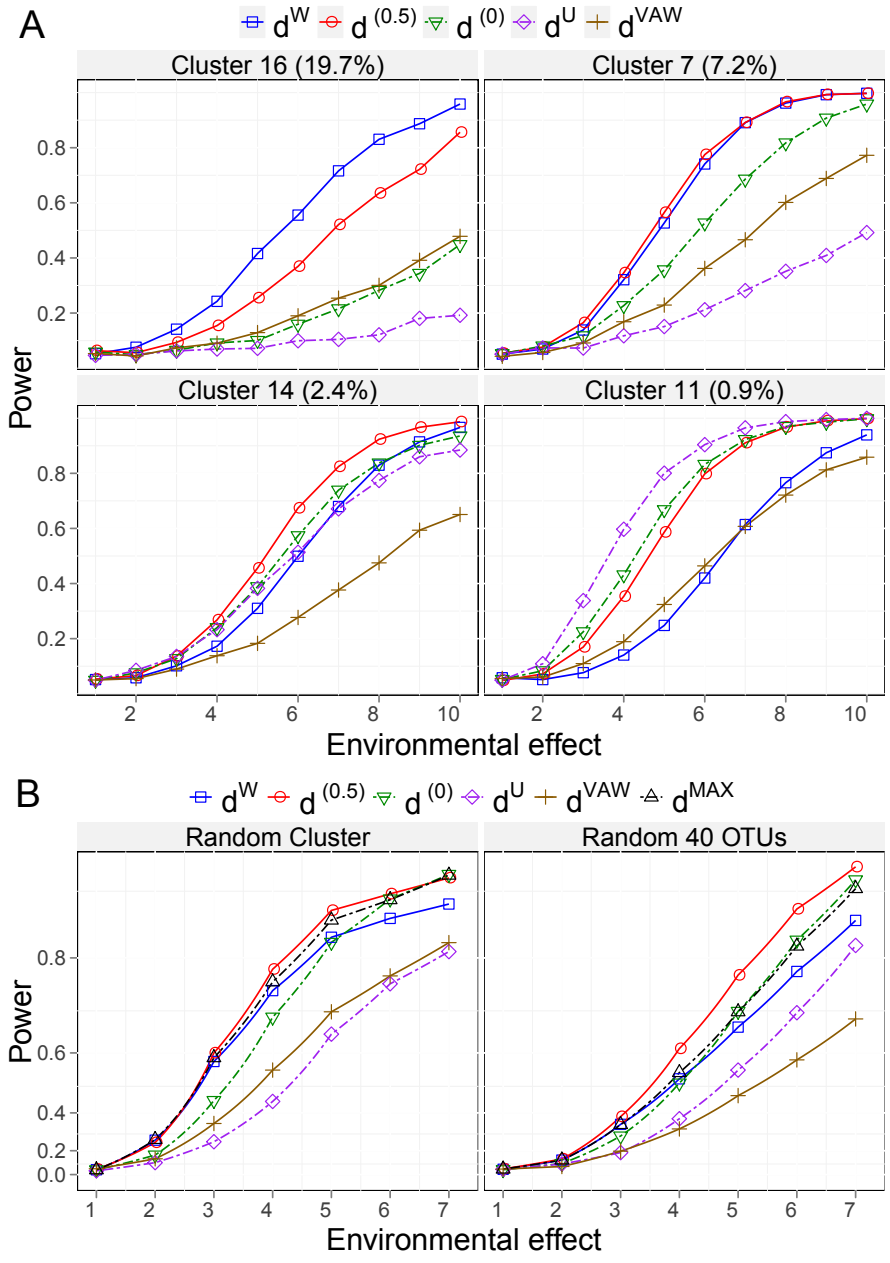

Figure 6: **Power comparison of different UniFrac variants for detecting environmental effects using tree based simulation.** PERMANOVA is used for testing hypotheses. The power curves are created by varying the degree of environmental effect. (A) the environmental factor affects a particular lineage (OTU cluster). Four example lineages of different abundance levels that are affected by environment are given. The lineage abundance is given in parentheses in the panel title. (B) the environmental factor affects a random lineage (left panel) or a random subset of 40 OTUs (right panel). The initial point of the power curve is the power when there is no environmental effect.

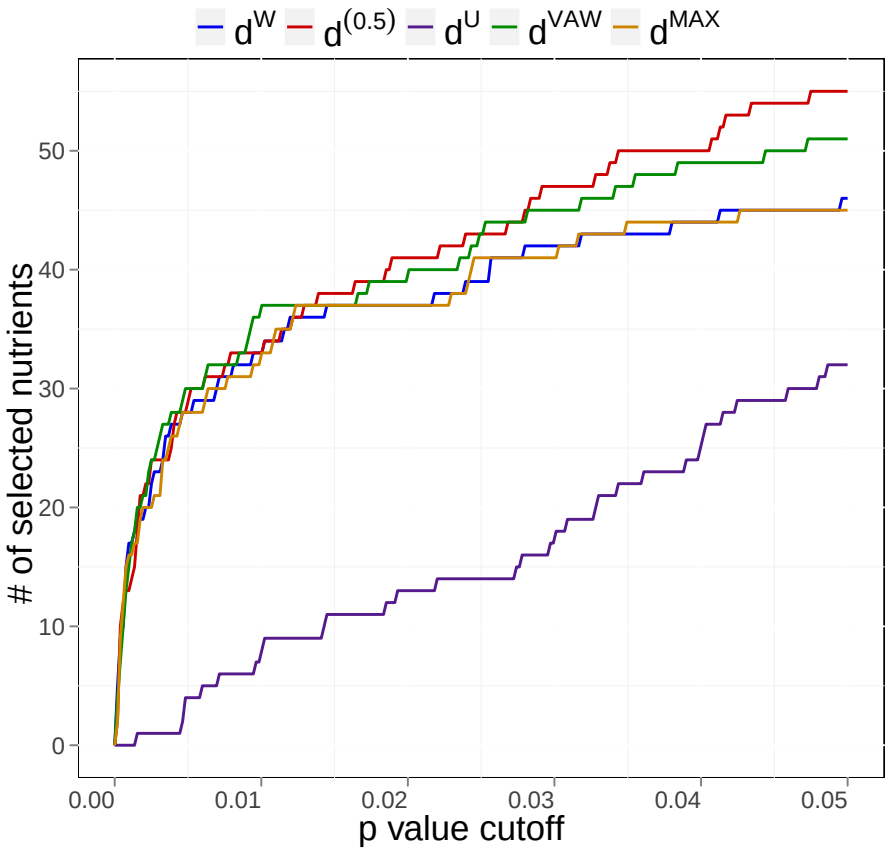

Figure 7: **Comparison of different UniFrac variants for detecting nutrient effects on gut microbiome composition.**PERMANOVA is used for testing hypotheses. 214 nutrients are included in the testing. The curves are generated by varying the p-value cutoffs.

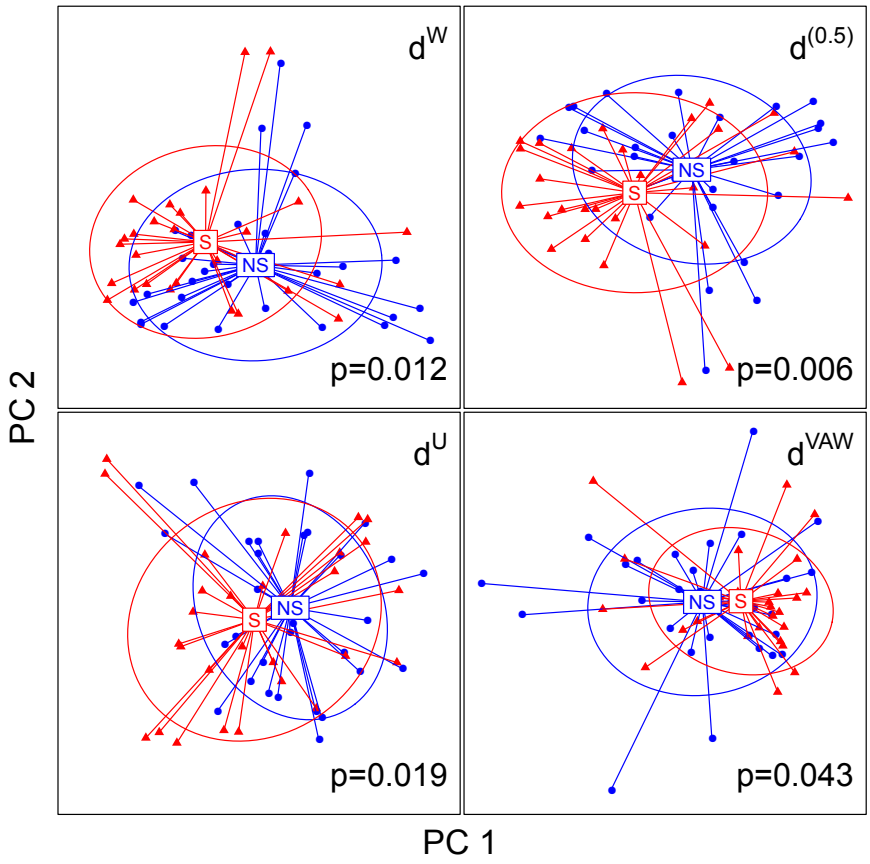

Figure 8: **Comparison of different UniFrac variants for clustering samples from smokers and nonsmokers.** Principle coordinate analysis is performed on the distance matrices of d^W, $d^{(0.5)}$, d^U and d^{VAW}. The samples are plotted on the first two principle coordinates. The PERMANOVA p values are also indicated in the figure. The ellipse center indicates groups means, its main axis corresponds to the first two principle components from principle component analysis (PCA), and the height and width are variances on that direction.

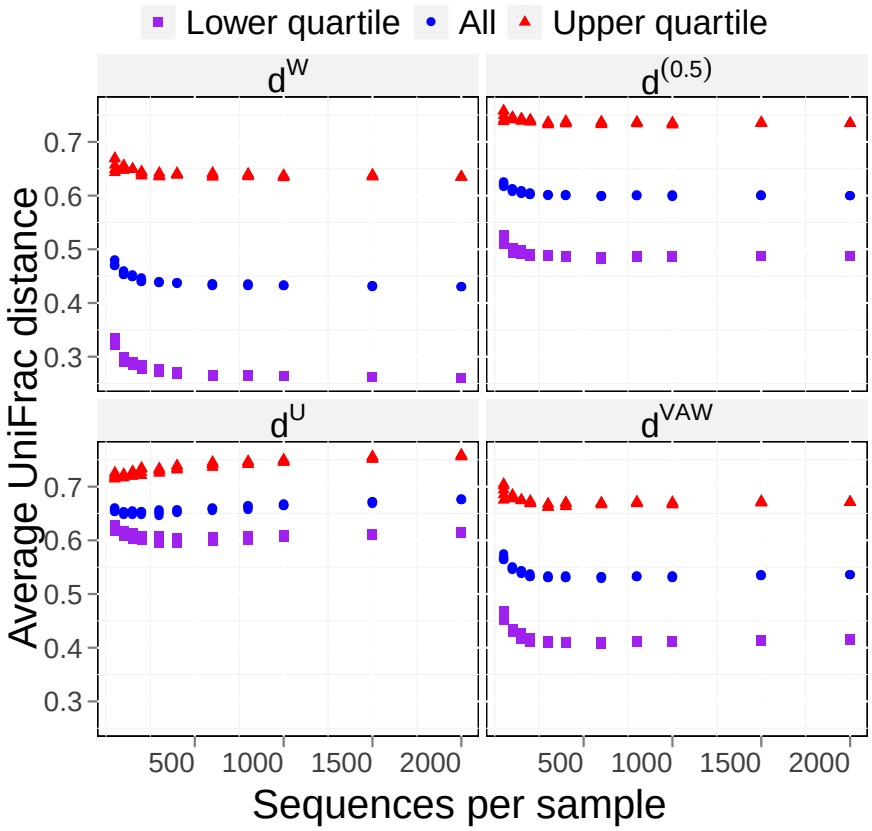

Figure 9: **Sensitivity of generalized UniFrac distance to sampling depth by rarefaction of data from a study of diet effect on the gut microbiome.** This study produced $\sim$ 1 million reads from the V12 region of 16S rRNA using pyrosequencing. The samples with less than 2408 sequences were first excluded (leaving 98 samples). For five replications, sequences from 98 samples were subsampled to different depth (between 50 to 2000). Pairwise distances were calculated for the four UniFrac variants (d^W, $d^{(0.5)}$, $d^{(U)}$ and d^{VAW}). To assess the the effects of community divergence on the sensitivity to sampling, the most similar and most different pairs of samples were identified from un-subsampled samples (2408 sequences) as those in the upper and lower quartile of UniFrac values calculated separately for all UniFrac variants. The points represent the average UniFrac value at each sampling depth for all pairs ('All') and the pairs that were in the upper and lower quartiles. Each point represents one replicate.

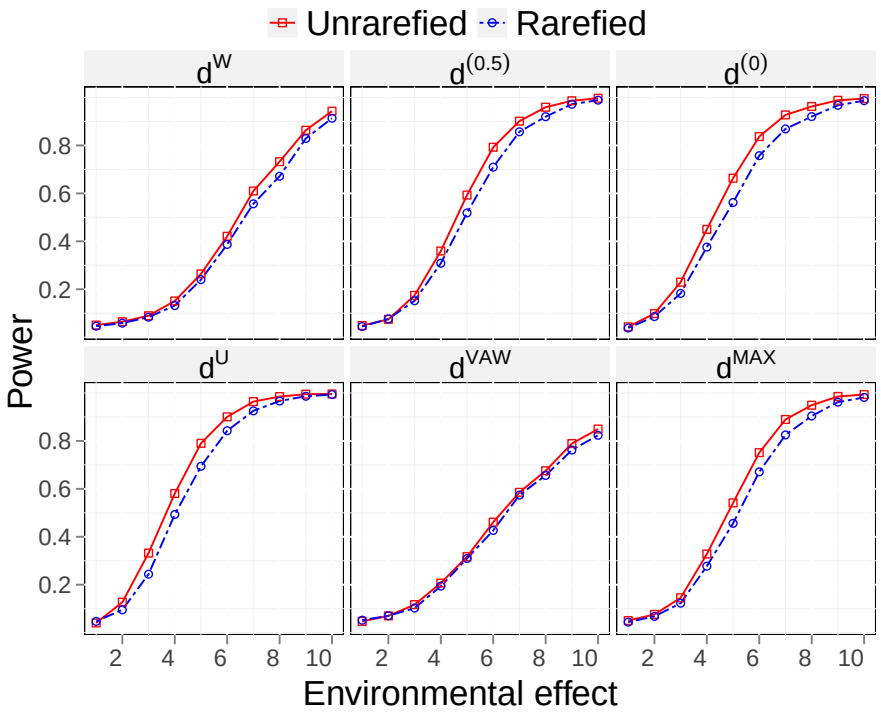

Figure 10: **Effect of rarefaction on the power of testing the association of microbiome composition with covariates.** The tree based simulation approach is used to investigate the effect of rarefaction on the power of PERMANOVA test. Two conditions are simulated with 10 samples under each condition. We let the environmental factor increases the abundance of OTU Cluster 11 of the tree for illustration purpose. Negative binomial (NB) model is used to generate the sampling depth for each sample. The parameters of the NB model is adjusted to have mean 1,000 and standard deviation (SD) of 300 so the sampling depth can range from 400 to 1600 (2SD) sequences per sample. We then compare the power of the test before and after rarefaction calculated over 2,000 replications. For the rarefaction case, we rarefied all the samples to the lowest sampling depth seen in the samples. We found that rarefaction decreases the power for all the UniFrac variants tested.

CHAPTER 3 : Structure-Constrained Sparse Canonical Correlation Analysis for Microbiome Data

In this chapter, we develop a method for structure-constrained sparse canonical correlation analysis (ssCCA) in high dimensional setting, motivated by studying the association between nutrient intakes and the human gut microbiome composition. Compared to sparse canonical correlation analysis, ssCCA takes into account the phylogenetic relationship among the bacteria when selecting the most correlated bacterial taxa and covariates. Our ssCCA formulation utilizes a phylogenetic structure-constrained penalty function to impose certain smoothness on the linear coefficients according to the phylogenetic relationship among the taxa. An efficient coordinate descent algorithm is developed for optimization. A human gut microbiome data set is used to illustrate the method. Both simulations and real data application show that ssCCA performs better than the standard sparse CCA in identifying meaningful variables when there are structures in the data.

3.1. Introduction

Bacterial taxa are not independent of each other and are related evolutionarily by a phylogenetic tree. Taxa that are phylogenetically close usually behave similarly or have similar biological functions. Such phylogenetic tree information has been effectively utilized in the commonly used UniFrac distance between two microbiome samples (Lozupone and Knight, 2005). In an attempt to visualize the human gut microbiomes from different samples, Purdom (2011) proposed a phylogenetic tree-based principle component analysis (PCA) on the 16S data set. This phylogenetic PCA was shown to separate the environmental samples in a biologically more sensible way than the standard PCA.

In this chapter, we consider another commonly used dimension-reduction method, canonical correlation analysis (CCA), that can be used to relate the bacteria taxa with environmental covariates when the number of covariates is large. Our motivating example is a data set generated from a human gut microbiome study at the University of Pennsylvania, where

we aim to associate nutrient intakes to the bacterial composition in the human gut. Here we have both the nutrient intake data and the bacterial abundance data measured on the same individual (see Chapter 1 and Section 3.6 for details). We are interested in selecting the bacterial taxa and nutrients that are mostly correlated. CCA aims to identify the linear combinations of two sets of variables that are maximally correlated with each other and provides an important tool to summarize the overall dependency structures between two sets of variables. It has been applied to linking two sets of high dimensional genomic data measured on the same set of samples.

The standard CCA however does not perform variable selection and hence usually lacks biological interpretability especially when the dimension of variables is high. When the number of variables exceeds the number of observations, CCA can not be applied directly due to the singularity of the covariance matrix. To overcome these two major limitations, various types of sparse CCA (sCCA) have been proposed and developed and applied to genomic data analysis (Waaijenborg *et al.*, 2008; Witten *et al.*, 2009). In sCCA, a sparsity penalty function such as the l_1 penalty is often imposed on the linear coefficients in order to explain the correlation between two data sets using the least number of variables. The sparsity constraint in sCCA not only makes the computation feasible but also increases the biological interpretability of the selected variables.

Available approaches to sCCA do not, however, exploit the prior structure information among the variables. In many applications, there exists some structure among the set of variables in the CCA analysis. These structures can be some simple group structure such as gene sets or graphical structure such as gene networks in genomic studies. By including this prior structure information of the data, one can gain better biological insights from the analysis. This has been clearly demonstrated in sparse regression analysis (Li and Li, 2008, 2010).

We consider particularly to utilize the phylogenetic structure of data from human microbiome studies in CCA analysis. The phylogenetic information of the bacterial taxa could

guide us to select relevant taxa in the context of CCA by inducing a tendency to select closely related taxa together, since these taxa are very likely to be associated with the covariates in a similar fashion. In order to effectively utilize the phylogenetic information, we propose to develop a structure-constrained sparse CCA (ssCCA), where we impose an additional structure-constrained penalty function based on the phylogenetic tree structure. The ssCCA extends the sparse CCA formulation of Witten *et al.* (2009) by imposing a smoothness penalty for the loading coefficients of the taxa based on their closeness on the phylogenetic tree. We also develop an efficient coordinate descent algorithm to implement the ssCCA. Our simulations that mimic real microbiome data demonstrate that ssCCA can result in much better performance in selecting the bacteria that are associated with other environmental variables. Our analysis of the microbiome and nutrient data has identified that fat-related nutrients are closely related to human gut microbiome composition, a conclusion that agrees with a previous analysis of the data set (Wu *et al.*, 2011a).

The rest of the chapter is organized as follows. The idea of using phylogenetic tree-structure is presented in Section 3.2. A brief review of CCA and the formulation of ssCCA are given in Section 3.3. Details of the coordinate descent algorithm is presented in Section 3.4. Results from simulation studies to evaluate our method are given in Section 3.5. An application to a real human microbiome study to associate nutrient intakes with bacterial abundances is presented in Section 3.6. Finally, a brief discussion of the method and results is given in Section 3.7.

3.2. Construction of the phylogenetic tree and Laplacian matrix

The method proposed in this chapter is mainly applied to OTU-based 16S data. This means that each of the N 16S sequences belongs to one of p OTUs/taxa. Each OTU is characterized by a representative DNA sequence and can be assigned a taxonomic lineage by comparing to known bacterial 16S rRNA database (see Chapter 1 for details). Most species level OTUs are in extremely low abundances with a large proportion of OTUs being simply singletons possibly due to sequencing error. We can further aggregate the OTUs

from the same genus to form genus level OTUs and perform analysis on genus level, which is more robust to sequencing error and can reduce the number of variables significantly. A distance between any two OTUs can be computed using the OTU representative sequences based on some evolution model such as Jukes-Cantor, Kimura and Felsenstein model and a phylogenetic tree for the OTUs can be built based on these distances (Felsenstein, 2003).

Let $\mathbf{x} = (x_1, x_2, \cdots, x_p)^T$ represent the vector of the relative abundances of p OTUs obtained from the 16S sequencing, where each OTU is a leaf node of a phylogenetic tree of all the OTUs. We first construct an adjacency matrix using a pairwise distance matrix between any two OTUs. With the given phylogenetic tree, we can use the patristic distance, which is the sum of the branch lengths linking the two OTUs, or we can use the genetic distance between sequences without explicitly constructing the tree. The distance is usually normalized to the scale of [0,1], with 0 for identity and 1 for complete difference. Denote d_{jk} the distance between OTU j and OTU k. We then form a $p \times p$ adjacency matrix $\mathbf{A}$ with the diagonal elements of 1 and the jkth element between OTU j and k defined as

$$a_{jk} = 1/d_{jk}^2 \quad \text{for} \quad j \neq k. \tag{3.1}$$

By taking the square of d_{jk}, large edge weight is given to closely related OTUs. At the same time, the edge weights for distantly related OTUs are made small. Other ways to construct the adjacency are possible. The simplest way is to define a simple thresholding function

$$a_{jk} = \begin{cases} 1 & \text{if } d_{jk} \leq r \\ 0 & \text{if } d_{jk} > r. \end{cases}$$

Or we can also define a continuous measure. Several possible measures are given by

$$\begin{aligned} a_{jk} &= (1 - d_{jk}^m)^n \\ a_{jk} &= \exp(-d_{jk}^m) \\ a_{jk} &= 1/d_{jk}^m, \end{aligned}$$

where the power $m, n > 0$, the value of m, n determines how much weight to put on the edges between close OTUs and we can simply use $m = n = 1$. Finally, the discrete and continuous measures can be combined. The flexibility of constructing a adjacency matrix provides us a powerful means to incorporate prior information into the analysis.

Note that phylogenetic tree is a special case of general undirected graphs and the adjacency matrix is related to the Laplacian matrix associated with the graph. For a given adjacency matrix $\mathbf{A}$, define $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_p)$, where $d_j = \sum_{k=1}^{p} a_{jk}$. The associated Laplacian matrix is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$ (Chung, 1997). The Laplacian matrix $\mathbf{L}$ is associated with a labeled weighted graph $\mathcal{G} = (V, \mathcal{E}, w)$ with vertex set $V = 1, \ldots, p$ and edge set $\mathcal{E} = \{(j, k) : (j, k) \in V \times V\}$. Here a_{jk} is the weight of edge (j, k) and d_j is the degree of vertex j. For a given vector $\mathbf{u}$, it is easy to show that

$$\mathbf{u}^T \mathbf{L} \mathbf{u} = \sum_{1 \le j < k \le p} a_{jk} (u_j - u_k)^2, \tag{3.2}$$

which measures the smoothness of the vector $\mathbf{u}$ with respect to the labeled weighted graph $\mathcal{G}$. Based on this interpretation, Li and Li (2008, 2010) proposed a smoothness penalty of the form $\mathbf{u}^T \mathbf{L} \mathbf{u}$ in high dimensional regression setting. The structure constraint has a local smoothing effect by encouraging the variables that are linked on the prior graphical structure to have similar coefficients. In the next section, we extend sCCA to include this smoothness penalty to further encourage some smoothness of the coefficients in linear projections.

3.3. Structure-constrained sparse canonical correlation analysis

We consider CCA between random vectors $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p)^T$ and $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_q)^T$, where vector $\mathbf{x}$ contains the abundances of the p OTUs on a given phylogenetic tree and $\mathbf{y}$ is the q-dimensional vector of the environmental covariates. Let $\mathbf{A}$ be the adjacency matrix defined in previous section based on the phylogenetic tree structure and $\mathbf{L}$ be the corresponding Laplacian matrix.

3.3.1. Problem setup and standard CCA

Given two column vectors of random variables $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p)^T$ and $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_q)^T$, CCA aims to find two projection directions $\mathbf{u}_1 \in \mathbb{R}^p$ and $\mathbf{v}_1 \in \mathbb{R}^q$ so that

$$\begin{aligned}(\mathbf{u}_1, \mathbf{v}_1) &= \arg\max_{\mathbf{u},\mathbf{v}} \mathrm{Corr}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y}) \\ &= \arg\max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^T\Sigma_{\mathbf{xy}}\mathbf{v}}{\sqrt{(\mathbf{u}^T\Sigma_{\mathbf{xx}}\mathbf{u})(\mathbf{v}^T\Sigma_{\mathbf{yy}}\mathbf{v})}},\end{aligned}$$

where $\Sigma_{\mathbf{xx}}, \Sigma_{\mathbf{yy}}$ and $\Sigma_{\mathbf{xy}}$ are covariance and cross-covariance matrices. This maximization is equivalent to

$$\max_{\mathbf{u},\mathbf{v}} \ \mathbf{u}^T\Sigma_{\mathbf{xy}}\mathbf{v} \tag{3.3}$$

$$\text{subject to} \quad \mathbf{u}^T\Sigma_{\mathbf{xx}}\mathbf{u} = 1, \quad \mathbf{v}^T\Sigma_{\mathbf{yy}}\mathbf{v} = 1. \tag{3.4}$$

Here $\mathbf{u}_1, \mathbf{v}_1$ are called the first pair of canonical vectors while the new variables $\eta_1 = \mathbf{u}_1{}^T\mathbf{x}, \xi_1 = \mathbf{v}_1{}^T\mathbf{y}$ are called the first pair of canonical variables or latent variables and $\rho_1 = \mathrm{Corr}(\eta_1, \xi_1)$ is referred as the first canonical correlation.

Higher order canonical variables and canonical correlations can be obtained in a stepwise fashion. Let $k = \min(p, q)$. For $s = 2, \cdots, k$, we can successively find canonical vectors $(\mathbf{u}_2, \mathbf{v}_2), \cdots, (\mathbf{u}_k, \mathbf{v}_k)$ by

$$\begin{aligned}&(\mathbf{u}_s, \mathbf{v}_s) = \arg\max_{\mathbf{u},\mathbf{v}} \mathrm{Corr}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y}) \\ \text{subject to} \quad &\mathbf{u}^T\Sigma_{\mathbf{xx}}\mathbf{u} = \mathbf{v}^T\Sigma_{\mathbf{yy}}\mathbf{v} = 1, \mathbf{u}^T\Sigma_{\mathbf{xx}}\mathbf{u}_t = \mathbf{v}^T\Sigma_{\mathbf{yy}}\mathbf{v}_t = 0, \text{for } 1 \le t < s.\end{aligned}$$

The solution of the above maximization problem can be obtained by performing singular value decomposition (SVD) on the matrix $\mathbf{K}$:

$$\begin{aligned}\mathbf{K} &= \Sigma_{\mathbf{xx}}^{-1/2}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1/2} \\ &= \textstyle\sum_{i=1}^{k} d_i \mathbf{u}_i^* \mathbf{v}_i^*,\end{aligned}$$

then $\mathbf{u}_s = \Sigma_{\mathbf{xx}}^{-1/2}\mathbf{u}_s^*$ and $\mathbf{v}_s = \Sigma_{\mathbf{yy}}^{-1/2}\mathbf{v}_s^*$ for $s = 1, \cdots, k$.

Suppose that we have collected *i.i.d* n samples of $\mathbf{x}$ and $\mathbf{y}$, denoted by $\mathbf{X}$ and $\mathbf{Y}$. Assume both are column-standardized to have mean 0 and unit l_2 norm. When data are available, one estimates $\mathbf{u}$ and $\mathbf{v}$ by replacing $\Sigma_{\mathbf{xy}}, \Sigma_{\mathbf{xx}}$ and $\Sigma_{\mathbf{yy}}$ by the observed sample covariance and variance matrices $\mathbf{X}^T\mathbf{Y}$, $\mathbf{X}^T\mathbf{X}$ and $\mathbf{Y}^T\mathbf{Y}$. As in most applications, we focus on the first canonical vector pair, which captures most of the correlation between the two data sets.

3.3.2. Formulation of ssCCA

When the dimensions p and q are high, regularization is required in order to obtain a unique solution to the optimization problem (3.3). Given the tuning parameters $c_1 > 0, c_2 > 0, c_3 > 0$, we propose the following ssCCA criterion that extends the sCCA of Witten *et al.* (2009):

$$\begin{aligned} & \max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} \\ \text{subject to} \quad & \mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u} \leq 1, \mathbf{v}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v} \leq 1, \\ & pen(\mathbf{u}) \leq c_1, pen(\mathbf{v}) \leq c_2, \mathbf{u}^T\mathbf{L}\mathbf{u} \leq c_3. \end{aligned} \tag{3.5}$$

where $pen(\mathbf{u})$ and $pen(\mathbf{v})$ are sparsity penalty functions such as the l_1 penalty function that is defined as $pen(\mathbf{u}) = \sum_{i=1}^{p} |u_i|$. Different from the sCCA formulation, we impose another structure constraint on the coefficient vector $\mathbf{u}$ through the quadratic Laplacian quantity defined in (3.2), $\mathbf{u}^T\mathbf{L}\mathbf{u} \leq c_3$. This constraint encourages smoothness of the estimated coefficients of the OTUs that are closely related on the phylogenetic tree. Smaller value of the tuning parameter c_3 results in smoother estimate of the coefficient vector $\mathbf{u}$ over the phylogenetic tree.

It has been shown that in other high-dimensional problems, treating the covariance matrix as diagonal can yield good results (Tibshirani *et al.*, 2003; Dudoit *et al.*, 2001; Witten *et al.*, 2009). For this reason, rather than using (3.5) as our ssCCA criterion, following the same strategy adopted by many of the existing sCCA algorithms (Waaijenborg *et al.*, 2008;

Parkhomenko *et al.*, 2009; Witten *et al.*, 2009), we substitute in the identity matrix I for $\mathbf{X}^T\mathbf{X}$ and $\mathbf{Y}^T\mathbf{Y}$, which gives the ssCCA formulation that we use in this chapter:

$$\begin{aligned} &\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} \\ \text{subject to} \quad &\|\mathbf{u}\|_2^2 \leq 1, \|\mathbf{v}\|_2^2 \leq 1, pen(\mathbf{u}) \leq c_1, pen(\mathbf{v}) \leq c_2, \mathbf{u}^T\mathbf{L}\mathbf{u} \leq c_3. \end{aligned} \tag{3.6}$$

3.4. Coordinate descent algorithm for ssCCA

3.4.1. Algorithm to obtain the first ssCCA factor

Using l_1 penalty, the ssCCA criterion (3.6) can be written as:

$$\begin{aligned} &\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} \\ \text{subject to} \quad &\|\mathbf{u}\|_2^2 \leq 1, \|\mathbf{v}\|_2^2 \leq 1, \|\mathbf{u}\|_1 \leq c_1, \|\mathbf{v}\|_1 \leq c_2, \mathbf{u}^T\mathbf{L}\mathbf{u} \leq c_3, \end{aligned}$$

where $\|\mathbf{u}\|_1$ and $\|\mathbf{v}\|_1$ are the $l_1$norm of the vectors $\mathbf{u}$ and $\mathbf{v}$. To facilitate computation, we write constraints on $\mathbf{u}$ in Lagrangian form and the ssCCA criterion becomes:

$$\begin{aligned} &\min_{\mathbf{u},\mathbf{v}} \left\{ -\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} + \frac{1}{2}\|\mathbf{u}\|_2^2 + \lambda_1\|\mathbf{u}\|_1 + \frac{\lambda_2}{2}\mathbf{u}^T\mathbf{L}\mathbf{u} \right\} \\ \text{subject to} \quad &\|\mathbf{v}\|_2^2 \leq 1, \|\mathbf{v}\|_1 \leq c_2, \end{aligned} \tag{3.7}$$

where $\lambda_1 \geq 0, \lambda_2 \geq 0$ and $c_2 > 0$ are tuning parameters. If the coefficient $1/2$ in the criterion (3.7) is changed to $k/2$ for some constant $k > 0$, then the solution of the new criterion with tuning parameter $(\lambda_1, \lambda_2, c_2)$ will correspond to the solution of original problem with tuning parameter $(\lambda_1, \lambda_2/k, c_2)$ up to a scaling of $\mathbf{u}$. Also note that when $\lambda_2 = 0$, ssCCA is reduced to sCCA. Since the Laplacian penalty function $\frac{\lambda_2}{2}\mathbf{u}^T\mathbf{L}\mathbf{u}$ is convex in $\mathbf{u}$, so the criterion (3.7) remains biconvex in $\mathbf{u}$ and $\mathbf{v}$, so we can still use an iterative method to solve this optimization problem:

Algorithm to Obtain the First ssCCA Factor

1 Initialize $\mathbf{v}$ as the first right singular vector with an unity l_2 norm from the singular value decomposition of $\mathbf{X}^T\mathbf{Y}$.

2 Iterate until convergence:

- (a) $\mathbf{u} \leftarrow \arg\min_{\mathbf{u}} \left\{ -\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} + \frac{1}{2}\|\mathbf{u}\|_2^2 + \lambda_1\|\mathbf{u}\|_1 + \frac{\lambda_2}{2}\mathbf{u}^T\mathbf{L}\mathbf{u} \right\}$, which can be solved by a graph-constrained regression problem Li and Li (2010) :

$$\mathbf{u} \leftarrow \arg\min_{\mathbf{u}} \left\{ \frac{1}{2}\left\| \mathbf{X}^T\mathbf{Y}\mathbf{v} - \mathbf{u} \right\|_2^2 + \lambda_1\|\mathbf{u}\|_1 + \frac{\lambda_2}{2}\mathbf{u}^T\mathbf{L}\mathbf{u} \right\}.$$

- (b) $\mathbf{v} \leftarrow \arg\min_{\mathbf{v}} -\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v}$ subject to $\|\mathbf{v}\|_2^2 \leq 1, \|\mathbf{v}\|_1 \leq c_2$, which is given by

$$\mathbf{v} \leftarrow \frac{S((\mathbf{u}^T\mathbf{X}^T\mathbf{Y})^T, \delta)}{\left\| S((\mathbf{u}^T\mathbf{X}^T\mathbf{Y})^T, \delta) \right\|_2},$$

where $S(.,.)$ the soft-thresholding function, i.e.,

$$S(a,b) = \begin{cases} \text{sgn}(a)(|a| - b) & \text{if } |a| > b \\ 0 & \text{Otherwise,} \end{cases}$$

and $\delta = 0$ if this results in $\|\mathbf{v}\|_1 \leq c_2$; otherwise, δ is chosen so that $\|\mathbf{v}\|_1 = c_2$. The choice of δ can be determined using a binary search (Witten *et al.*, 2009).

Let $\mathbf{L} = \mathbf{U}\Gamma\mathbf{U}^T$ and $\mathbf{S} = \mathbf{U}\Gamma^{1/2}$, then Step 2(a) can be converted into a simple Lasso problem as in Li and Li (2008) :

$$\mathbf{u} \leftarrow \arg\min_{\mathbf{u}} \left\{ \frac{1}{2}\|\mathbf{A}^*\mathbf{u} - \mathbf{b}^*\|_2^2 + \lambda_1\|\mathbf{u}\|_1 \right\},$$

where $\mathbf{A}^*_{2p\times p} = \begin{pmatrix} \mathbf{I}_{p\times p} \\ \sqrt{\lambda_2}\mathbf{S}^T \end{pmatrix}$, $\mathbf{b}^*_{2p} = \begin{pmatrix} \mathbf{X}^T\mathbf{Y}\mathbf{v} \\ \mathbf{0}_p \end{pmatrix}$, $\mathbf{I}_{p\times p}$ is $p \times p$ identity matrix and $\mathbf{0}_p$ is a p-dimensional vector of 0's. Note that no intercept is included in this Lasso problem and coordinate descent algorithm can be implemented to obtain the solution at given λ_1

(Friedman *et al.*, 2007).

Though the objective function is biconvex, i.e., is convex in either $\mathbf{u}$ or $\mathbf{v}$, it is not convex in $(\mathbf{u}^T, \mathbf{v}^T)^T$, so the coordinate descent algorithm does not necessarily converge to the global optimum; however, by using the first right singular vector of the covariance matrix as the initial starting point, it does converge to a stationary point (Tseng and Yun, 2009) and interpretable solutions.

3.4.2. Choosing tuning parameters

The tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, c_2)$ control the model complexity and have to be tuned. We use a M-fold two-stage cross-validation (2CV) method to choose $\boldsymbol{\lambda}$. First, we divide all the samples into M disjoint subgroups, also known as folds, and denote the index of samples in the mth fold by I_m for $m = 1, \cdots, M$. The M-fold cross-validated function is defined as

$$CV(\boldsymbol{\lambda}) = \frac{1}{M} \sum_{m=1}^{M} \text{Corr}(\mathbf{X}_m^T \widehat{\mathbf{u}}_{-m}(\boldsymbol{\lambda}), \mathbf{Y}_m^T \widehat{\mathbf{v}}_{-m}(\boldsymbol{\lambda})) \tag{3.8}$$

where $\text{Corr}(.,.)$ is the correlation function and $\widehat{\mathbf{u}}_{-m}(\boldsymbol{\lambda}), \widehat{\mathbf{v}}_{-m}(\boldsymbol{\lambda})$ is the estimate of $\mathbf{u}, \mathbf{v}$ based on the samples $(\cup_{m=1}^{M} I_m) \backslash I_m$ with $\boldsymbol{\lambda}$ as the tuning parameter. It is well known that cross validation can perform poorly on model selection problems involving l_1 penalties (Meinshausen and Bühlmann, 2006) due to shrinkage in the values of the non-zero elements of the projection coefficients. To reduce the shrinkage problem, 2CV re-estimates the non-zero coefficients without penalization by performing the singular value decomposition on the training data set excluding the variables with zero coefficients in the penalized procedure. Specifically, for a given tuning parameter $\boldsymbol{\lambda}$, we re-calculate the loading coefficients using the variables that are selected by ssCCA and use these coefficients in the CV score (3.8). This avoids the bias of the estimates due to penalization. We then choose $\boldsymbol{\lambda}^* = \text{argmax}_{\boldsymbol{\lambda}} CV(\boldsymbol{\lambda})$ as the best tuning parameters. From our simulations, we observe that the 2CV procedure almost always performs better than standard CV without re-estimating the parameters.

We report all our results based on the 2CV procedure. This two-stage approach was also used for tuning parameter selection in other settings when l_1 penalization is used (James *et al.*, 2010).

3.5. Simulation studies

We present Monte Carlo simulations to evaluate ssCCA in identifying the relevant variables that explain the correlation between two multivariate vectors. The solution of sCCA is obtained by setting $\lambda_2 = 0$ in ssCCA. The simulations are carried out to mimic an association study between nutrient intakes and genus level OTU abundances that is presented in Section 3.6. Since the phylogenetic tree implies distances between the OTUs, we simulate the distance matrix directly. Specifically, since OTUs are often clustered on the phylogenetic tree, we generate random OTU clusters of size from 1 to 15. If two OTUs are from the same cluster (e.g. from the same taxonomic rank *family*), then their distance is drawn from a uniform distribution on $(0.1, 0.2)$; if two OTUs are from different clusters, then their distance is drawn from a uniform distribution on $(0.2, 1)$. We then construct the adjacency matrix $\mathbf{A}$ using the method (3.1) based on the distances.

3.5.1. Simulation based on a latent variable model

We use a latent variable model to generate the data matrices $\mathbf{X}$ and $\mathbf{Y}$ where the dependency between these two sets of variables are induced by a latent random variable ζ and the variances in $\mathbf{x}, \mathbf{y}$ can be explained in part by ζ. We assume $\mathbf{x} = \zeta \mathbf{w_x} + \boldsymbol{\epsilon}_\mathbf{x}$ and $\mathbf{y} = \zeta \mathbf{w_y} + \boldsymbol{\epsilon}_\mathbf{y}$, where $\zeta \sim N(0, \sigma_\zeta^2)$, $\boldsymbol{\epsilon}_\mathbf{x}, \boldsymbol{\epsilon}_\mathbf{y}$ are random noise vectors that follow $\boldsymbol{\epsilon}_\mathbf{x}, \boldsymbol{\epsilon}_\mathbf{y} \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$, and $\mathbf{w_x} \in \mathbb{R}^p, \mathbf{w_y} \in \mathbb{R}^q$ are column vectors of pre-set weights. The $\sigma_\epsilon / \sigma_\zeta$ ratio controls the overall association strength between $\mathbf{x}$ and $\mathbf{y}$, with small value indicating strong association. The coefficients $\mathbf{w_x}$ and $\mathbf{w_y}$ control the relative contributions of individual variables to the overall association. We assume that only the first $p_\mathbf{x}$ elements of $\mathbf{w_x}$ and the first $q_\mathbf{y}$ elements of $\mathbf{w_y}$ are nonzero. Table 2 shows the parameters used in simulation. If we let $\sum_{i=1}^{p_\mathbf{x}} |w_{\mathbf{x},i}| = \sum_{i=1}^{p_\mathbf{y}} |w_{\mathbf{y},i}| = 1$, then the highest correlation between linear combinations of

$\mathbf{x}$ and $\mathbf{y}$ is given by Parkhomenko *et al.* (2009):

$$\rho_{max} = \frac{\sigma_\zeta^2}{\sqrt{(\sigma_\zeta^2 + p_{\mathbf{x}}\sigma_\epsilon^2)(\sigma_\zeta^2 + p_{\mathbf{y}}\sigma_\epsilon^2)}}. \tag{3.9}$$

We fix $\sigma_\epsilon^2 = 1$ and vary σ_ζ^2 to control the strength of the canonical correlation. When $\sigma_\zeta = 5$, $\rho_{max} \approx 0.7$.

3.5.2. Evaluation of the selection performance

We evaluate the performance of our methods in terms of selecting the relevant variables that lead to correlation between random vectors $\mathbf{x}$ and $\mathbf{y}$ by considering models with various combinations of the parameters. For each simulated data set, we use five-fold 2CV to select the tuning parameter values and then compute true positive rate (TPR), false positive rate (FPR) and Matthew's correlation coefficient (MCC) to measure the selection performance for both $\mathbf{x}$ and $\mathbf{y}$. These three measures are defined as

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN},$$
$$MCC = \frac{TP \times TN - FP \times FN}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)},$$

where TP, FP, TN, FN are true positives, false positives, true negatives and false negatives, respectively. For each model, we generate the observed data set $\mathbf{X}$ and $\mathbf{Y}$ 100 times and summarize TPR, FPR and MCC as averages over 100 runs. Results from 10-fold 2CV are very similar and are omitted here.

We also compare the performance of different methods using the ROC curve (FPR against TPR) for identifying the relevant taxa OTUs by varying the tuning parameters. Specifically, the three tuning parameters are searched over a $10 \times 10 \times 10$ grid for a total of 1000 tuning parameter combinations. For each combination, we obtain FPR and TPR, which represents one point in the ROC plot. The ROC curve is then obtained by joining these points for each run. We then average the ROC curves over 100 runs to produce an average

ROC curve.

3.5.3. Comparison of ssCCA and sCCA under one latent variable model

We consider models with various combinations of the parameters (labeled **A1** to **D2**), including the number of relevant OTU clusters, the signal strength as measured by σ_ζ^2 and the dimensions p and q and present the results in Table 3 and Figure 11. We observe that the advantage of ssCCA over sCCA is more obvious under weak association (Model **A1**). As the signal becomes stronger, the performance of sCCA becomes closer to ssCCA (Model **A2**). This agrees with our intuition: the advantage of ssCCA lies in borrowing information from closely related OTUs and when the association is weak, pooling information across closely related OTUs can really help the OTU selection. Another interesting observation is that better selection of OTUs can lead to better selection of nutrients, which is best shown in the weak association case by obtaining a higher MCC. We also observe that that as the dimension increases, both ssCCA and sCCA become less efficient in selecting relevant OTUs and nutrients (Models **B1** and **B2**). However, ssCCA performs consistently better than sCCA in all dimensions considered. Finally, as the cluster size decreases, we do not see a significant deterioration of the selection performance of ssCCA (Models **C1** and **C2**). ssCCA still performs better than sCCA. As long as the cluster contains more than one OTU, using structure information always improves variable selection.

Since the smoothness penalty encourages that the variables that are close on the phylogenetic tree to have similar linear projection coefficients, we evaluate the sensitivity of ssCCA when this assumption does not hold. We investigate the performance of ssCCA when data and prior contradict each other. We consider the model where the first 10 elements of $\mathbf{w_x}$ have different coefficients but with the same signs and take values that are equally spaced on $[0.08, 0.12]$ (model **D1**). The performance of ssCCA is still much better than sCCA. The Model **D2** considers the scenario when the first 5 and the second 5 elements of $\mathbf{w_x}$ are 0.1s and -0.1s respectively, where the coefficients are different and have different signs. This scenario violates our model assumption that closely linked OTUs have similar coeffi-

cients, the structure-constrained penalty now has an adverse effect. This is clearly seen in the ROC plot (Figure 11 **D2**). However, when the 2CV procedure is applied to select the tuning parameters and the corresponding OTUs and nutrients, the performance of ssCCA and sCCA is very similar (Table 3). This is because that if the prior structure information is not useful, 2CV procedure tends to select $\lambda_2 = 0$, which reduces ssCCA to sCCA. Therefore, the selection performance of ssCCA should be at least as good as sCCA, but ssCCA performs better when the prior knowledge is correct.

3.5.4. Comparison of ssCCA and sCCA under complex models

We compare the performance of ssCCA and sCCA under several complex models and also present the results in Table 3 and Figure 11. Under Model **E**, we consider the scenario when the noises are correlated with correlation $0.4^{|i-j|}$ for ϵ_i and ϵ_j for both $\mathbf{x}$ and $\mathbf{y}$. The performances of ssCCA and sCCA are both slightly worse when compared to Model **A2** when the noises are independent. ssCCA still outperforms sCCA.

We then consider Model **F** where we simulate count data with zeros. Specifically, we first generate the data matrix $\mathbf{X}$ as previously. We then convert it into a proportion matrix $\mathbf{P}$ and generate the counts based on $\mathbf{P}$. For the jth column X_j, we first map the column values into the range of $[0, p_j^{max}]$ by a linear transformation $p_{ij} = \frac{x_{ij} - \min_i(x_{ij})}{\max_i(x_{ij}) - \min_i(x_{ij})} p_j^{max}$, where p_j^{max} is sampled from [0.01, 0.1], so the maximum OTU abundance can vary by 10 folds. Rows of $\mathbf{P}$ are further scaled to sum up to 1. Given the OTU proportions for each sample, we generate the counts using a Dirichlet-multinomial model with a total count of 1000 and an overdispersion of 0.01. Since we introduce extra variation by simulating counts, we increase the first 10 components of $\mathbf{w_x}$ to 0.4 to achieve moderate association ($\rho_1 \approx 0.7$). Under this parameter setting, the data matrix contains about 20% 0's. To apply ssCCA and sCCA, we convert the simulated count matrix into a proportion matrix. Table 3 **F** and Figure 11 **F** again show ssCCA outperforms sCCA in selecting the relevant variables.

Finally, we consider two models where two orthogonal directions induce the correlation

between two sets of random vectors. We assume $\mathbf{x} = \zeta_1 \mathbf{w}_{\mathbf{x}}^1 + \zeta_2 \mathbf{w}_{\mathbf{x}}^2 + \boldsymbol{\epsilon}_{\mathbf{x}}$ and $\mathbf{y} = \zeta_1 \mathbf{w}_{\mathbf{y}}^1 + \zeta_2 \mathbf{w}_{\mathbf{y}}^2 + \boldsymbol{\epsilon}_{\mathbf{y}}$, where under Model **G**, the two directions are given by

$$\mathbf{w}_{\mathbf{x}}^1 = (\underbrace{0.1, \cdots, 0.1}_{5}, \underbrace{0.1, \cdots, 0.1}_{5}, \underbrace{0, \cdots, 0}_{90})^T,$$

and

$$\mathbf{w}_{\mathbf{x}}^2 = 0.5(\underbrace{0.1, \cdots, 0.1}_{5}, \underbrace{-0.1, \cdots, -0.1}_{5} \underbrace{0, \cdots, 0}_{90})^T.$$

We assume that $\mathbf{w}_{\mathbf{y}}^1, \mathbf{w}_{\mathbf{y}}^2$ are the same as $\mathbf{w}_{\mathbf{x}}^1, \mathbf{w}_{\mathbf{x}}^2$, and the OTUs from the same cluster have the same coefficients on on the first direction. Under Model **H**, we consider model misspecification where the two directions are given by

$$\mathbf{w}_{\mathbf{x}}^1 = (\underbrace{0.1, 0.1, -0.1, -0.1, 0.1}_{5}, \underbrace{0.1, 0.1, -0.1, -0.1, -0.1}_{5}, \underbrace{0, \cdots, 0}_{90})^T$$

and

$$\mathbf{w}_{\mathbf{x}}^2 = 0.5(\underbrace{0.1, \cdots, 0.1}_{5}, \underbrace{0.1, \cdots, 0.1}_{5} \underbrace{0, \cdots, 0}_{90})^T,$$

and $\mathbf{w}_{\mathbf{y}}^1, \mathbf{w}_{\mathbf{y}}^2$ are the same as $\mathbf{w}_{\mathbf{x}}^1, \mathbf{w}_{\mathbf{x}}^2$. OTUs from the same cluster have coefficients of different signs on the first direction. ssCCA has higher true positive and lower false positive rates and higher area under the ROC curve (see Table 3 **G** and Figure 11 **G**). Under the model misspecification (Model **H**), the performances of ssCCA and sCCA are comparable.

3.6. Application to real data analysis

We apply ssCCA to a microbiome study on association between the nutrient intake and bacterial abundances in the human gut conducted at the University of Pennsylvania. One goal of the study is to investigate the relationship between diet and microbiome composition and to identify a short list of potential nutrients and their associated bacteria in human gut. For this study, both gut microbiome 16S data and the nutrient intake data are available for 99 healthy subjects. Fecal samples were obtained from these 99 subjects and bacterial

DNA was extracted using standard protocol. After multiplexed 454 pyrosequencing, about $900,000$ high quality, partial ($\sim$370bp) 16S rRNA gene sequences were generated. These sequences were analyzed using the Qiime pipeline* (Caporaso *et al.*, 2010b), where the sequences were clustered at 97% sequence identity into OTUs and assigned a taxonomic identity using the RDP classifier (Wang *et al.*, 2007). We consolidated these species level OTUs into 119 genera (genus level OTUs) and used the representative sequence from the most abundant species level OTU as the genus level representative sequence for distance calculation and for construction of the phylogenetic tree. In our analysis, we further excluded the uncommon genera that occurred in less than 1/4 of the samples so we only considered $p = 40$ relatively common genera (See Figure 12). These 99 subjects also completed a carefully designed Food Frequency Questionnaire (FFQ). Based on the FFQ, the daily intake for $q = 214$ nutrients were calculated for each subject by nutritionists. Because the nutrient intake is clearly dependent on the overall energy consumption, we regressed the nutrient intake on the total energy consumption and took residuals as the normalized nutrient intake. Our final data set can be summarized as the OTU abundance matrix $\mathbf{X}_{99\times 40}$ and the nutrient intake matrix $\mathbf{Y}_{99\times 214}$. Since the sampling depths are very different for different samples, we normalize the counts into proportions and standardize the columns to have mean 0 and variance 1.

The goal of our analysis is to investigate the overall association between gut bacteria abundances and nutrient intakes. We used the method presented in equation (3.1) to construct the adjacency matrix $\mathbf{A}$ and the distances between any two OTUs were calculated using the "K80" model (R "ape" package, "dist.dna" function). Five-fold 2CV was performed to search the optimal tuning parameters on a grid of $20 \times 20 \times 20$ and the range of the tuning parameters were set to explore all possible models: from the most dense to the most sparse model. We applied ssCCA to the data set and identified 24 nutrients and 14 genera whose linear combinations gave a cross-validated correlation of 0.42 between gut bacterial

*Denoising and rarefaction were not performed for this analysis, hence one more sample and different number of genera compared to the other two chapters.

abundances and nutrients. Figure 12 shows the heatmap of pair-wise correlations between these selected nutrients and OTUs, where the estimated loading coefficients are given in the parentheses. The signs of the estimated loading coefficients correspond very well to the pair-wise correlations. The nutrients related to fats are clustered together while the other nutrients show an opposite direction of association.

The selected microbiome-associated nutrients are biologically interpretable. More than half of the selected nutrients are related to fat. It has been experimentally shown that fats can change the gut microbiome composition independent of obesity in mouse study (Hildebrandt *et al.*, 2009). There are also 4 selected nutrients related to Choline and it was found by a recent human microbiome study that the composition of the gastrointestinal microbiome changed with the choline levels of diets (Spencer *et al.*, 2011). The selected nutrients are also consistent with the candidate nutrients we identified using a distance-based testing procedure (Wu *et al.*, 2011a). This procedure utilized the overall UniFrac distances (Lozupone and Knight, 2005) between microbiomes of any two subjects computed using both the OTU abundances and the phylogenetic relationship among them. 20 out of 24 nutrients selected by ssCCA were in the nutrients selected by the distance-based individual testing method at the false discovery rate of 25%.

The pattern of selected OTUs is also interesting. The selected OTUs are marked with red circles in the phylogenetic tree of Figure 12. We see that the closely related OTUs tend to be selected together, for example, the genus *Parabacteroides* and *Marinilabilia*, *Butyrivibrio* and *Coprococcus* and *Anaerostipes* and *Lachnospiraceae Incertae Sedis* are all close relatives on the tree. ssCCA tends to select closely related OTUs together by making the coefficients of neighbors similar through imposing phylogenetic tree-constrained smoothness penalty. This feature of ssCCA can also be viewed as borrowing information from nearby OTUs, that is, if several neighbors all exhibit similar weak association, ssCCA amplifies the signal strength and selects them together. On the other hand, if some OTU exhibits low-level association but all its neighbors show the opposite evidence, ssCCA will not select that

OTU.

As a comparison, sCCA that does not account for the phylogenetic relationship among the OTUs only selects one OTU, the *Firmucute Lachnospira*, which was also selected by ssCCA, but a total of 122 nutrients. The interpretation of the result is not as clear as that from ssCCA. The resulting combinations gave a cross-validated correlation of 0.39, smaller than that obtained from ssCCA.

3.7. Discussion

In this chapter, we have extended the sparse CCA to incorporate the graphical structure among the variables in canonical correlation analysis. When the number of variables exceeds the number of samples, using prior structure information to guide variable selection is very important. The prior knowledge could lead to a solution that is biologically more interpretable. The sparse sCCA utilizes the phylogenetic information to select the bacterial OTUs that are associated with covariates. The power of the ssCCA method has been demonstrated in the simulation studies and its performance is unanimously better than sCCA in all the simulated scenarios when there are structures in the data. Even when the prior information is not completely accurate, our method still performs comparably to sCCA due to selection of the tuning parameter by cross-validation.

Our method could also be applied to analysis of other types of genomic data. Due to development of high throughput sequencing methods, it has become common for researchers to obtain two or more genomic measurements on the same samples such as the gene expression data and genotyping data. For example, in eQTL study, we want to associate genotype vector to gene expression vector. sCCA has been applied to this problem and produced some encouraging results. However, genes are also related by gene networks, which in turn introduces some structures to both gene expression and genotype data. We can apply ssCCA to such data by incorporating the prior network information. We expect to gain certain insights in identifying the genetic variants that are associated with genetic pathways.

One limitation of the ssCCA formulation is that it assumes a linear relationship among the variables, which may not always hold for OTU compositional/abundance data. Our analysis of the gut microbiome data did not indicate too much deviation from the linearity between OTU abundances and nutrient intakes. One interesting future research is to develop structure-constrained nonlinear measures of association and sparse nonlinear CCA. One promising idea is to incorporate the phylogenetic tree information of the OTU abundances into kernel CCA (Dauxois and Nkiet, 1997).

Table 2: **Parameters used in simulation studies.** For parameters with multiple entries, the first value is used as the baseline simulations. When one parameter is varied, the baseline parameters are used for other parameters.

Parameters	Value
Sample size (n)	100
OTU No. (p)	100, 200, 400
Relevant OTU No. ($p_{\mathbf{x}}$)	10
Weights for relevant OTUs ($\mathbf{w}_{\mathbf{x}}$)	0.1,0.1,. . . ,0.1
Nutrient No. (q)	100, 200, 400
Relevant nutrient No. ($q_{\mathbf{y}}$)	10
Weights for relevant nutrients ($\mathbf{w}_{\mathbf{y}}$)	Equally spaced on $[0.08, 0.12]$
Relevant OTU cluster No. (Size)	1(10), 2 (5,5), 3(3,3,4)
Latent variable SD (σ_ζ)	5, 4
Random error SD (σ_ϵ)	1

Table 3: **Simulation results to evaluate ssCCA under models of different association signals, dimension sizes, cluster sizes, model misspecification and complexity.** Five-fold 2CV is used to select the tuning parameters. As a comparison, results from sCCA are also presented. Each column represents a measure of selection performance for OTU ($\mathbf{x}$) or nutrient ($\mathbf{y}$). TPR: true positive rate, FPR: false positive rate, MCC: Matthew's correlation coefficient. The results are averaged over 100 replications with SD indicated in the parenthesis.

	Selection of $\mathbf{x}$ variables			Selection of $\mathbf{y}$ variables		
Method	TPR-$\mathbf{x}$	FPR-$\mathbf{x}$	MCC-$\mathbf{x}$	TPR-$\mathbf{y}$	FPR-$\mathbf{y}$	MCC-$\mathbf{y}$
A1 - 1 cluster, $\sigma_\zeta = 4, p, q = 100$						
ssCCA	0.91(0.20)	0.07(0.10)	0.76(0.22)	0.78(0.22)	0.12(0.12)	0.58(0.18)
sCCA	0.70(0.31)	0.09(0.12)	0.56(0.22)	0.75(0.24)	0.12(0.12)	0.54(0.21)
A2 - 1 cluster, $\sigma_\zeta = 5, p, q = 100$						
ssCCA	0.96(0.10)	0.03(0.08)	0.89(0.16)	0.87(0.17)	0.05(0.09)	0.78(0.16)
sCCA	0.89(0.17)	0.05(0.08)	0.79(0.17)	0.87(0.16)	0.05(0.09)	0.77(0.17)
B1 - 1 cluster, $\sigma_\zeta = 5$, p, q=200						
ssCCA	0.98(0.08)	0.05(0.11)	0.87(0.19)	0.87(0.16)	0.07(0.11)	0.75(0.18)
sCCA	0.89(0.17)	0.09(0.15)	0.74(0.22)	0.87(0.16)	0.08(0.11)	0.72(0.20)
B2 - 1 cluster, $\sigma_\zeta = 5$, p, q=400						
ssCCA	0.89(0.30)	0.06(0.11)	0.74(0.33)	0.81(0.28)	0.12(0.13)	0.60(0.30)
sCCA	0.77(0.32)	0.09(0.23)	0.66(0.32)	0.78(0.31)	0.11(0.12)	0.57(0.32)
C1 - 2 clusters, $\sigma_\zeta = 5, p, q = 100$						
ssCCA	0.93(0.14)	0.03(0.07)	0.88(0.15)	0.83(0.16)	0.05(0.09)	0.76(0.16)
sCCA	0.87(0.16)	0.05(0.08)	0.78(0.16)	0.85(0.16)	0.06(0.10)	0.76(0.17)
C2 - 3 clusters, $\sigma_\zeta = 5, p, q = 100$						
ssCCA	0.94(0.11)	0.03(0.07)	0.88(0.15)	0.88(0.15)	0.07(0.11)	0.75(0.18)
sCCA	0.89(0.16)	0.05(0.10)	0.80(0.18)	0.88(0.16)	0.07(0.10)	0.76(0.18)
D1 - 1 cluster, $\sigma_\zeta = 5, p, q = 100$, variable coefficients of the same signs						
ssCCA	0.95(0.11)	0.02(0.05)	0.90(0.13)	0.86(0.19)	0.06(0.10)	0.76(0.17)
sCCA	0.87(0.15)	0.04(0.08)	0.79(0.15)	0.88(0.16)	0.07(0.10)	0.75(0.18)
D2 - 1 cluster, $\sigma_\zeta = 5, p, q = 100$, variable coefficient of opposite signs						
ssCCA	0.89(0.14)	0.05(0.09)	0.81(0.17)	0.89(0.15)	0.08(0.11)	0.75(0.20)
sCCA	0.90(0.15)	0.04(0.09)	0.82(0.17)	0.90(0.14)	0.07(0.11)	0.76(0.19)
E - correlated noise, 1 cluster, $\sigma_\zeta = 5, p, q = 100$						
ssCCA	0.92(0.18)	0.04(0.07)	0.84(0.19)	0.78(0.21)	0.05(0.08)	0.72(0.17)
sCCA	0.85(0.20)	0.05(0.10)	0.77(0.18)	0.82(0.21)	0.06(0.10)	0.73(0.18)
F - count data, 1 cluster, $\sigma_\zeta = 5, p, q = 100$						
ssCCA	0.92(0.16)	0.04(0.11)	0.84(0.17)	0.72(0.26)	0.06(0.14)	0.71(0.20)
sCCA	0.72(0.22)	0.09(0.15)	0.62(0.18)	0.80(0.24)	0.08(0.16)	0.75(0.23)
G - two directions, 1 cluster, $\sigma_\zeta = 5, p, q = 100$						
ssCCA	0.95(0.13)	0.03(0.08)	0.87(0.17)	0.85(0.17)	0.05(0.09)	0.76(0.16)
sCCA	0.85(0.19)	0.07(0.10)	0.73(0.17)	0.82(0.19)	0.06(0.09)	0.72(0.16)
H - two directions, 2 clusters, model misspecification, $\sigma_\zeta = 5, p, q = 100$						
ssCCA	0.83(0.20)	0.05(0.09)	0.74(0.19)	0.88(0.19)	0.11(0.13)	0.67(0.20)
sCCA	0.87(0.18)	0.06(0.10)	0.76(0.18)	0.89(0.17)	0.10(0.13)	0.69(0.20)

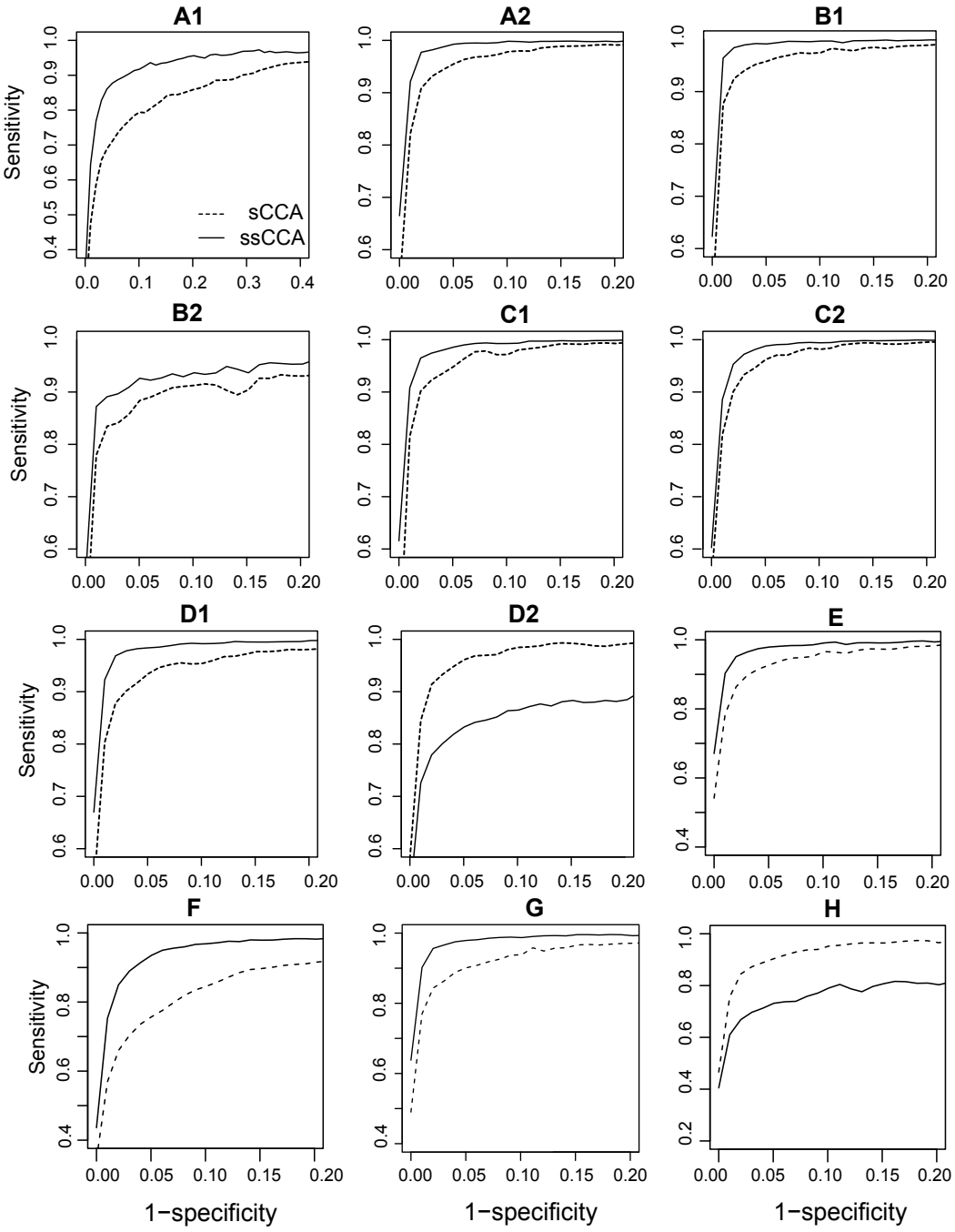

Figure 11: **ROC curves for OTU selection using ssCCA and sCCA for Models A1 - H.** The corresponding model parameters are given in Table 3.

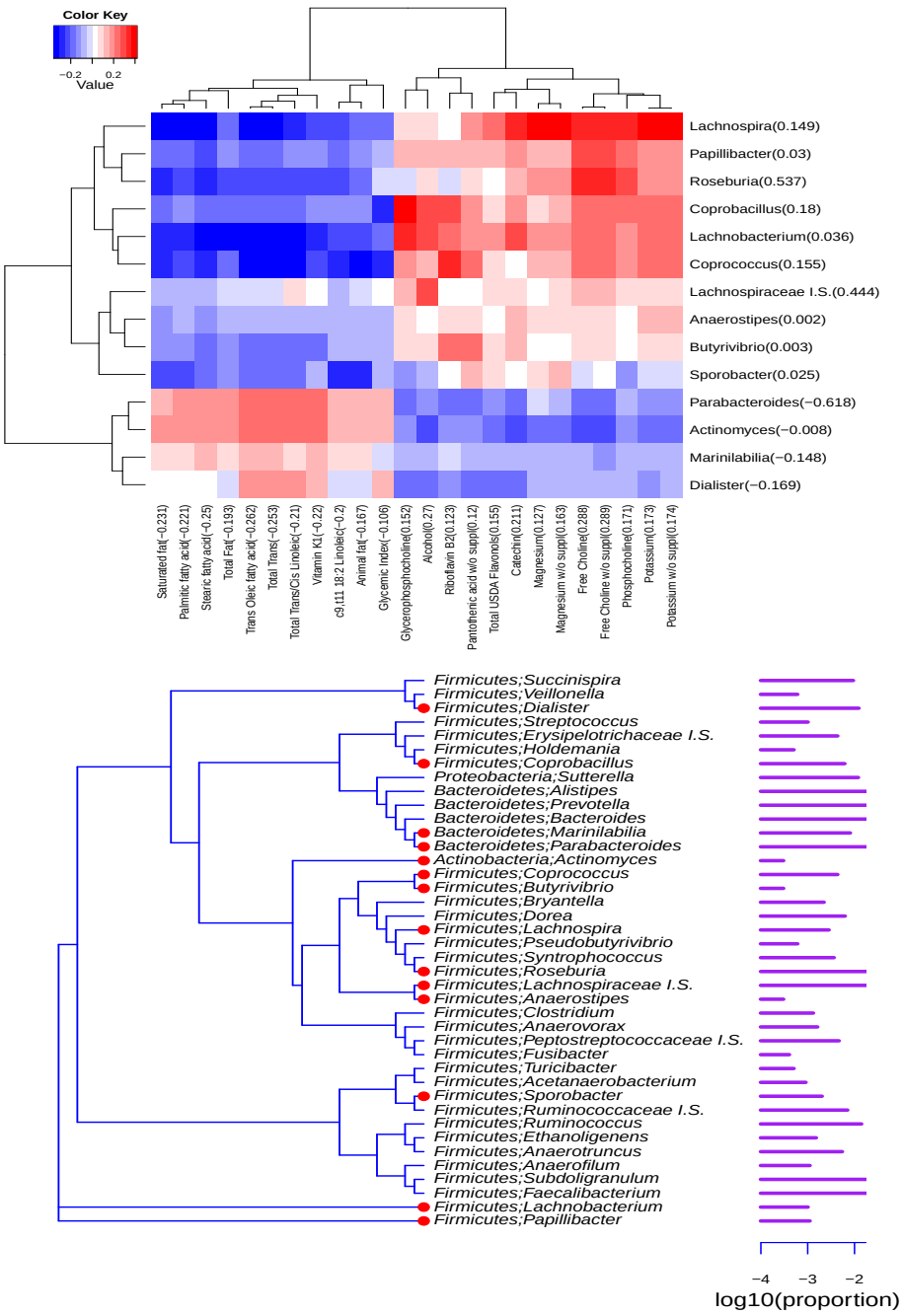

Figure 12: **Associating gut microbiome composition with dietary nutrient intakes using ssCCA.** Top: heatmap that shows the correlations between the selected genera and nutrients. The number in parenthesis of each variable is the estimated loading coefficient. Red and blue colors indicate positive and negative correlations respectively. Bottom: Phylogenetic tree of the 40 genera used in the analysis. The genera selected by ssCCA are marked with red circles. The bars on the right side indicate the average relative abundances of these genera on log 10 scale.

CHAPTER 4 : Variable Selection for Sparse Dirichlet-Multinomial Regression with Applications to Microbiome Data Analysis

In this chapter, we present a model-based regression method to link the microbiome composition with environmental covariates. We propose to model the OTU counts using a Dirichlet-multinomial (DM) regression model in order to account for overdispersion of observed counts. The DM regression model can be used for testing the association between microbiome composition and covariates using the likelihood ratio test. However, when the number of the covariates is large, multiple testing can lead to loss of power. To deal with the high dimensionality of the problem, we develop a penalized likelihood approach to estimate the regression coefficients and to select the variables by imposing a sparse group ℓ_1 penalty to encourage both group-level and within-group sparsity. Such a variable selection procedure can lead to selection of the relevant covariates and their associated OTUs. An efficient block-coordinate descent algorithm is developed to solve the optimization problem. We present extensive simulations to demonstrate that the sparse DM regression can result in better identification of the microbiome-associated covariates than models that ignore overdispersion or only consider the proportions. We demonstrate the power of our method in an analysis of a data set evaluating the effects of nutrient intake on human gut microbiome composition.

4.1. Introduction

Recent studies have linked the microbiome with human diseases including obesity and inflammatory bowel disease (Virgin and Todd, 2011). It is therefore important to understand how genetic or environmental factors shape the human microbiome in order to gain insight into etiology of many microbiome-related diseases and to develop therapeutical measures to modulate the microbiome composition. Benson *et al.* (2010) demonstrated that genetic variants are associated with the mouse gut microbiome. Wu *et al.* (2011a) showed that dietary nutrients are associated with the human gut microbiome. Both studies have consid-

ered a large number of genetic loci or nutrients and aimed to identify the genetic variants or nutrients that are associated with the gut microbiome. When there are a large number of possible covariates affecting the microbiome composition, variable selection becomes necessary. Variable selection can not only increase biological interpretability but also provide researchers with a short list of top candidates for biological validation. The methods we develop in this chapter are particularly motivated by an ongoing study at the University of Pennsylvania to link the nutrient intake to the human gut microbiome. In this study, gut microbiome data were collected on 98 normal volunteers. In addition, food frequency questionnaire (FFQ) were filled out by these individuals. The questionnaires were scored and the quantitative measurements of 214 micronutrients were obtained. Details of the study and the data set can be found in Chapter 1 and Section 4.6. Our goal is to identify the nutrients that are associated with the gut microbiome and also their associated OTUs.

Most of the microbiome studies used distance-based methods to link the microbiome and environmental covariates, where a distance metric such as the generalized UniFrac distance we presented in Chapter 2 was defined between two microbiome samples and statistical analysis was then performed using the distances. However, the choice of distance metric is sometimes subjective and different distances vary in their power of identifying relevant environmental factors. Another limitation of distance-based methods is its inefficiency for detecting subtle changes since distances summarize the overall relationship. In addition, such distance-based approaches do not provide information on how covariates affect the microbiome compositions and which OTUs are affected. Therefore, it is desirable to model the composition directly instead of summarizing the data as distances.

In this chapter, we consider the sparse Dirichlet-multinomial (DM) regression (Mosimann, 1962) to link high-dimensional covariates to OTU counts from microbiome data. The DM regression model is chosen in order to model the overdispersed OTU counts. The observed OTU count variance is much larger than that predicted by a multinomial model that assumes fixed underlying OTU proportions, an assumption that is hardly met for real microbiome

data. Uncontrollable sources of variation such as individual-to-individual variability, day-to-day variability, sampling location variability or even technical variability such as sample preparation lead to enormous variability in the underlying proportions. In contrast, the DM model assumes that the underlying OTU proportions come from a Dirichlet distribution. We use a log linear link function to associate the mean OTU proportions with covariates. In this DM modeling framework, the effects of the covariates on OTU proportions can be tested using the likelihood ratio test.

When the number of the covariates is large, we propose a sparse group ℓ_1 penalized likelihood approach for variable selection and parameter estimation. The sparse group ℓ_1 penalty function (Friedman *et al.*, 2010) consists of a group ℓ_1 penalty and an overall ℓ_1 penalty, which induce both group-level sparsity and within-group sparsity. This is particularly relevant in our setting. For the nutrient-microbiome association example, we have p nutrients and q OTUs, so the fully parameterized model has $(p+1) \times q$ coefficients including the intercepts, since each nutrient-OTU association is characterized by one coefficient. The q coefficients for each nutrient constitute a group. If we assume many nutrients have no or ignorable effect on the microbiome composition, the groups of coefficients associated with these irrelevant nutrients should be zero altogether, which is a group-level sparsity that is achieved by imposing a group ℓ_1 penalty. However, the group ℓ_1 penalty does not perform within-group selection, meaning that if one group is selected, all the coefficients in that group are non-zeros. In the case of nutrient-microbiome association, we are also interested in knowing which OTUs are associated with a selected nutrient. By imposing an overall ℓ_1 penalty, within-group selection becomes possible. Therefore, we impose a sparse group ℓ_1 penalty not only to select these important nutrients but also to recover relevant nutrient-OTU associations.

Section 4.2 reviews the Dirichlet-multinomial model for count data. Section 4.3 introduces the Dirichlet-multinomial regression framework for incorporating covariate effects and proposes a likelihood ratio statistic for testing the covariate effect. Section 4.4 proposes a

sparse group ℓ_1 penalized likelihood procedure for variable selection for the DM models followed by a detailed description of a block-coordinate descent algorithm in Section 4.4.1. Section 4.5 shows simulation results and Section 4.6 demonstrates the proposed method on a real human gut microbiome data set to associate the nutrient intake with the human gut microbiome composition.

4.2. Dirichlet-multinomial model for microbiome composition data

Suppose we have q OTUs/taxa and their counts $Y = (Y_1, Y_2, ..., Y_q)$ are random variables. Denote $\mathbf{y} = (y_1, y_2, ..., y_q)$ as the observed counts. The simplest model for count data is the multinomial model and its probability function is given as:

$$f_M(y_1, y_2, \cdots, y_q; \boldsymbol{\phi}) = \binom{y_+}{\mathbf{y}} \prod_{j=1}^{q} \phi_j^{y_j},$$

where $y_+ = \sum_{j=1}^{q} y_j$ and $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_q)$ are underlying species proportions with $\sum_{j=1}^{q} \phi_j = 1$. Here the total OTU count y_+ is determined by the sequencing depth and is treated as an ancillary statistic since its distribution does not depend on the parameters in the model. The mean and variance of the multinomial component Y_j $(j = 1 \cdots q)$ are:

$$\mathrm{E}(Y_j) = y_+\phi_j, \qquad \mathrm{Var}(Y_j) = y_+\phi_j(1 - \phi_j). \tag{4.1}$$

For microbiome composition data, the actual variation is usually larger than what would be predicted by the multinomial model, which assumes fixed underlying proportions. This increased variation is due to the heterogeneity of the microbiome samples and the underlying proportions vary among samples. To account for the extra-variation or overdispersion, we assume the underlying proportions $(\phi_1, \phi_2, \cdots, \phi_q)$ are themselves positive random variables $(\Phi_1, \Phi_2, \cdots, \Phi_q)$ subject to the constraint $\sum_{j=1}^{q} \Phi_j = 1$. One commonly used distribution

is the Dirichlet distribution (Mosimann, 1962) with the probability function given by

$$f_D(\phi_1, \phi_2, \cdots, \phi_q; \boldsymbol{\gamma}) = \frac{\Gamma(\gamma_+)}{\prod_{j=1}^{q} \Gamma(\gamma_j)} \prod_{j=1}^{q} \phi_j^{\gamma_j - 1},$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_q)$ are positive parameters, $\gamma_+ = \sum_{j=1}^{q} \gamma_j$ and $\Gamma(\cdot)$ is the Gamma function. The mean and variance of the Dirichlet component Φ_j $(j = 1 \cdots q)$ are:

$$\mathrm{E}(\Phi_j) = \frac{\gamma_j}{\gamma_+}, \qquad \mathrm{Var}(\Phi_j) = \frac{\gamma_j(\gamma_+ - \gamma_j)}{(1+\gamma_+)\gamma_+^2}.$$

The mean is proportional to γ_j and the variance is controlled by γ_+, which can be regarded as a "precision parameter". As γ_+ becomes larger, the proportions are more concentrated around the means.

The Dirichlet-multinomial (DM) distribution (Mosimann, 1962) results from a compound multinomial distribution with weights from the Dirichlet distribution (Parameterization I):

$$\begin{aligned} f_{DM}(y_1, y_2, \cdots, y_q; \boldsymbol{\gamma}) &= \int f_M(y_1, y_2, \cdots, y_q; \boldsymbol{\phi}) f_D(\boldsymbol{\phi}; \boldsymbol{\gamma}) d\boldsymbol{\phi} \\ &= \binom{y_+}{\mathbf{y}} \frac{\Gamma(y_+ + 1)\Gamma(\gamma_+)}{\Gamma(y_+ + \gamma_+)} \prod_{j=1}^{q} \frac{\Gamma(y_j + \gamma_j)}{\Gamma(\gamma_j)\Gamma(y_j + 1)}. \end{aligned} \tag{4.2}$$

The mean and variance of the DM distribution for each component Y_j $(j = 1, \cdots, q)$ is given by

$$\mathrm{E}(Y_j) = y_+ \,\mathrm{E}(\Phi_j), \qquad \mathrm{Var}(Y_j) = y_+ \,\mathrm{E}(\Phi_j)\{1 - \mathrm{E}(\Phi_j)\} \left(\frac{y_+ + \gamma_+}{1 + \gamma_+} \right). \tag{4.3}$$

Comparing (4.3) with (4.1), we see that the variation of the DM component is increased by a factor of $(y_+ + \gamma_+)/(1 + \gamma_+)$, where γ_+ controls the degree of overdispersion with a larger value indicating less overdispersion. Using an alternative parameterization, the probability

function can be written as (Parameterization II):

$$f^*_{DM}(y_1, y_2, ..., y_q; \boldsymbol{\phi}, \theta) = \binom{y_+}{\mathbf{y}} \frac{\prod_{j=1}^{q} \prod_{k=1}^{y_j} \{\phi_j(1-\theta) + (k-1)\theta\}}{\prod_{k=1}^{y_+} \{1 - \theta + (k-1)\theta\}}, \tag{4.4}$$

where $\phi_j = \gamma_j/\gamma_+$ is the mean and $\theta = 1/(1+\gamma_+)$ is the dispersion parameter. When $\theta = 0$, it is easy to verify (4.4) is reduced to the multinomial distribution.

4.3. Dirichlet-multinomial regression for incorporating the covariate effects

When there is no covariate effect, the DM model can be used to produce more accurate estimates of OTU proportions of a given microbiome sample than the simple multinomial model, due to its ability to model the overdispersion. Beyond proportion estimation, microbial ecologists are more interested in associating the microbiome composition with some environmental covariates. Suppose we have n microbiome samples and q species. Let $\mathbf{Y} = (y_{ij})_{n\times q}$ be the observed count matrix for the n samples. Let $\mathbf{X} = (x_{ij})_{n\times p}$ be the design matrix of p covariates for n samples. We assume the parameters γ_j $(j = 1, \cdots, q)$ in the DM model (Parameterization I) depends on the covariate via the following log-linear model

$$\gamma_j(\mathbf{x}^i) = \exp(\alpha_j + \sum_{k=1}^{p} \beta_{jk} x_{ik}), \tag{4.5}$$

where $\mathbf{x}^i$ is the ith row vector of $\mathbf{X}$ and β_{jk} is the coefficient for jth OTU with respect to kth covariate, whose sign and magnitude measures the effect of kth covariate on the jth OTU. From (4.3), we see that $\mathrm{E}(Y_{ij}) \propto \exp(\alpha_j) \prod_{k=1}^{p} \exp(\beta_{jk} x_{ik})$, where $\exp(\alpha_j)$ can be interpreted as the baseline abundance level for species j and the coefficient β_{jk} indicates the magnitude of the kth covariate effect on species j. Though the log linear link is assumed mainly for ease of computation, it is biologically consistent, in that microorganisms usually exhibit exponential growth in favorable environment.

For notational simplicity, we denote β_{j0} as α_j and augment $\mathbf{X}$ with an n-vector of 1's as its

first column. We number the columns from 0 to p. The link function becomes

$$\gamma_j(\mathbf{x}^i) = \exp(\sum_{k=0}^{p} \beta_{jk} x_{ik}). \tag{4.6}$$

Let $\boldsymbol{\beta}$ be the $q \times (p+1)$ regression coefficient matrix, $\boldsymbol{\beta}^j = (\beta_{j0}, \cdots, \beta_{jp})^T$ be the vector of coefficients for the jth OTU ($j = 1, \cdots, q$) and $\boldsymbol{\beta}_k = (\beta_{1k}, \cdots, \beta_{qk})^T$ be the vector of coefficients for the kth covariate ($k = 0, \cdots, p$). We also use $\boldsymbol{\beta}$ to denote the $q(p+1)$ vector that contains all the coefficients. Substituting (4.5) into DM probability function (4.2) and ignoring the part that does not involve the parameters, the log likelihood function given the covariates is given by

$$\begin{aligned} l(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{n} \Bigg[& \tilde{\Gamma}\left(\sum_{j=1}^{q} \gamma_j(\mathbf{x}^i; \boldsymbol{\beta}^j)\right) - \tilde{\Gamma}\left(\sum_{j=1}^{q} y_{ij} + \sum_{j=1}^{q} \gamma_j(\mathbf{x}^i; \boldsymbol{\beta}^j)\right) \\ & + \sum_{j=1}^{q} \left\{ \tilde{\Gamma}\left(y_{ij} + \gamma_j(\mathbf{x}^i; \boldsymbol{\beta}^j)\right) - \tilde{\Gamma}\left(\gamma_j(\mathbf{x}^i; \boldsymbol{\beta}^j)\right) \right\} \Bigg]. \end{aligned} \tag{4.7}$$

where $\tilde{\Gamma}(.)$ is the log gamma function.

Based on the likelihood function (4.7), one can test the effect of a given covariate or the joint effects of all the covariates on the microbiome composition using the standard likelihood ratio test (LRT). To solve the maximization problem, we implemented the Newton-Raphson algorithm, since the gradient and Hessian matrix of the log likelihood can be calculated analytically. Alternatively, we can use general-purpose optimization algorithm such as nlm in R, which computes the gradient and Hessian numerically. By selecting an appropriate starting point (*e.g.* $\boldsymbol{\alpha} = \boldsymbol{\beta} = \mathbf{0}$), for moderate-size problems in the dimensions p and q, the algorithm converges to a stationary point sufficiently fast.

With a large number of covariates in the DM regression model, direct maximization of the likelihood function becomes infeasible or unstable. When each covariate is tested separately using the LRT, adjustment for multiple testing is required. In addition, when the number

of OTUs q is large, the null distribution of the LRT has large degrees of freedom and therefore reduced power. It is also desirable to select the individual OTUs associated with the covariate. Although one can test the null hypothesis $H_0 : \beta_{jk} = 0$ for each (j, k) pair by the LRT, adjustment of multiple comparisons can lead to a loss of power. In next section, we present a sparse group ℓ_1 penalized estimation for variable selection and parameter estimation for sparse DM regression models.

4.4. Variable selection for sparse Dirichlet-multinomial regression

To perform variable selection, we estimate the regression coefficient vector $\boldsymbol{\beta}$ in model (4.6) by minimizing the following sparse group ℓ_1 penalized negative log-likelihood function,

$$pl(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}, \lambda_1, \lambda_2) = -l(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}) + \lambda_1 \sum_{k=1}^{p} \|\boldsymbol{\beta}_k\|_2 + \lambda_2 \sum_{k=1}^{p} \|\boldsymbol{\beta}_k\|_1 , \tag{4.8}$$

where $l(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X})$ is the log-likelihood function defined as in (4.7), λ_1 and λ_2 are the tuning parameters and $\|\boldsymbol{\beta}_k\|_1 = \sum_{j=1}^{q} |\beta_{ik}|$ is the ℓ_1 norm and $\|\boldsymbol{\beta}_k\|_2 = \sqrt{\sum_{j=1}^{q} \beta_{ik}^2}$ is the group ℓ_1 norm of the coefficient vector $\boldsymbol{\beta}_k$, respectively. We do not penalize the intercept vector $\boldsymbol{\beta}_0$. The first part of the sparse group ℓ_1 penalty is the group ℓ_1 penalty that induces group-level sparsity, which facilitates selection of the covariates that are associated with OTU counts. The second ℓ_1 penalty on all the coefficients facilitates the within-group selection, which is important for interpretability of the resulting model. A similar penalty involving both group ℓ_1 and ℓ_1 terms is discussed in (Peng *et al.*, 2009) and (Friedman *et al.*, 2010) for regularized multivariate linear regression. When $\lambda_2 = 0$, criterion (4.8) reduces to the group lasso.

4.4.1. A block-coordinate gradient descent algorithm for sparse group ℓ_1 penalized DM regression

The sparse group ℓ_1 estimates of $\boldsymbol{\beta}$ can be obtained by minimizing the penalized negative log-likelihood function (4.8):

$$\hat{\boldsymbol{\beta}}_{\lambda_1,\lambda_2} = \arg\min_{\boldsymbol{\beta}} \left\{ -l(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}) + \lambda_1 \sum_{k=1}^{p} \|\boldsymbol{\beta}_k\|_2 + \lambda_2 \sum_{k=1}^{p} \|\boldsymbol{\beta}_k\|_1 \right\}.$$

Using the general block coordinate gradient descent algorithm of Tseng and Yun (2009), we develop in the following an efficient algorithm to solve this optimization problem. Meier *et al.* (2008) present a block coordinate gradient descent algorithm for group lasso for logistic regression that includes only the group ℓ_1 penalty (i.e., $\lambda_2 = 0$). In contrast, our optimization problem (4.8) has two non-differentiable parts, both at the individual β_{jk} and at the group $\boldsymbol{\beta}_k$ levels.

The key idea of the algorithm is to combine a quadratic approximation of the log-likelihood function with an additional line search. First we expand (4.7) at current estimate $\hat{\boldsymbol{\beta}}^{(t)}$ to a second-order Taylor series. The Hessian matrix is then replaced by a suitable matrix $\mathbf{H}^{(t)}$. We define

$$l_Q^{(t)}(\mathbf{d}) = l(\hat{\boldsymbol{\beta}}^{(t)}) + \mathbf{d}^T \nabla l(\hat{\boldsymbol{\beta}}^{(t)}) + \frac{1}{2}\mathbf{d}^T \mathbf{H}^{(t)} \mathbf{d}, \tag{4.9}$$

where $\mathbf{d} \in \mathbb{R}^{q(p+1)}$. Also denote $\nabla l(\hat{\boldsymbol{\beta}}^{(t)})_k$ and $\mathbf{d}_k$ the gradient and increment with respect to $\hat{\boldsymbol{\beta}}_k^{(t)}$ for the kth group, and $\nabla l(\hat{\boldsymbol{\beta}}^{(t)})_{sk}$ and $\mathbf{d}_{sk}$ with respect to $\hat{\beta}_{sk}^{(t)}$. We then minimize the following function $pl_Q^{(t)}(\mathbf{d})$ with respect to the kth penalized parameter group:

$$\begin{aligned} pl_Q^{(t)}(\mathbf{d}) &= -l_Q^{(t)}(\mathbf{d}) + \lambda_1 \sum_{k=1}^{p} \left\| \hat{\boldsymbol{\beta}}_k^{(t)} + \mathbf{d}_k \right\|_2 + \lambda_2 \sum_{k=1}^{p} \left\| \hat{\boldsymbol{\beta}}_k^{(t)} + \mathbf{d}_k \right\|_1 \\ &\approx pl(\hat{\boldsymbol{\beta}}^{(t)} + \mathbf{d}; \mathbf{Y}, \mathbf{X}, \lambda_1, \lambda_2). \end{aligned} \tag{4.10}$$

We restrict ourselves to vectors $\mathbf{d}$ with $\mathbf{d}_j = \mathbf{0}$ for $j \neq k$ and the corresponding $q \times q$

submatrix $\mathbf{H}_{kk}^{(t)}$ for kth group is a diagonal matrix of the form $\mathbf{H}_{kk}^{(t)} = h_k^{(t)}\mathbf{I}_q$ for some scalar $h_k^{(t)} \in \mathbb{R}$.

The solution to the general optimization problem of the form (4.10) is given by Theorem 1 and its Corollary in the Appendix. Let $S = \{s \mid |\nabla l(\hat{\boldsymbol{\beta}}^{(t)})_{sk} - h_k^{(t)}\hat{\beta}_{sk}^{(t)}| < \lambda_2\}$ and $\bar{S}$ be the set $\{1, \cdots, q\}\backslash S$. Denote $\mathbf{d}_{Sk}$ the subvector of $\mathbf{d}_k$ with indices in S and $\mathbf{d}_{\bar{S}k}$ in $\bar{S}$. The minimizer of (4.10) can be decomposed into two parts: The first part $\mathbf{d}_{Sk}^{(t)}$ can be obtained by

$$\mathbf{d}_{Sk}^{(t)} = -\hat{\boldsymbol{\beta}}_{Sk}^{(t)}.$$

The second part $\mathbf{d}_{\bar{S}k}^{(t)}$ can be computed by minimizing:

$$f^{(t)}(\mathbf{d}_k) \quad = \quad -\left\{\mathbf{d}_k^T\mathbf{u}_k^{(t)} + \frac{1}{2}\mathbf{d}_k^T\mathbf{H}_{kk}^{(t)}\mathbf{d}_k\right\} + \lambda_1 \left\|\hat{\boldsymbol{\beta}}_k^{(t)} + \mathbf{d}_k\right\|_2 \tag{4.11}$$

with respect to $\mathbf{d}_{\bar{S}k}$ (set components other than $\mathbf{d}_{\bar{S}k}$ to be 0), where

$$\mathbf{u}_k^{(t)} = \left[\nabla l(\hat{\boldsymbol{\beta}}^{(t)})_k - \lambda_2 \text{sgn}\left\{\nabla l(\hat{\boldsymbol{\beta}}^{(t)})_k - h_k^{(t)}\hat{\boldsymbol{\beta}}_k^{(t)}\right\}\right]$$

and $\text{sgn}(\cdot)$ is the sign function.

Minimization of (4.11) with respective to $\mathbf{d}_{\bar{S}k}$ can be performed in a similar fashion as in Meier *et al.* (2008) for the group ℓ_1 penalty. Specifically, if $\left\|\mathbf{u}_{\bar{S}k}^{(t)} - h_k^{(t)}\boldsymbol{\beta}_{\bar{S}k}^{(t)}\right\|_2 < \lambda_1$, the minimizer of equation (4.11) for $\mathbf{d}_{\bar{S}k}$ is

$$\mathbf{d}_{\bar{S}k}^{(t)} = -\hat{\boldsymbol{\beta}}_{\bar{S}k}^{(t)}.$$

Otherwise

$$\mathbf{d}_{\bar{S}k}^{(t)} = -\frac{1}{h_k^{(t)}}\left\{\mathbf{u}_{\bar{S}k}^{(t)} - \lambda_1 \frac{\mathbf{u}_{\bar{S}k}^{(t)} - h_k^{(t)}\hat{\boldsymbol{\beta}}_{\bar{S}k}^{(t)}}{\left\|\mathbf{u}_{\bar{S}k}^{(t)} - h_k^{(t)}\hat{\boldsymbol{\beta}}_{\bar{S}k}^{(t)}\right\|_2}\right\}.$$

For the unpenalized intercept, the solution can be directly computed:

$$\mathbf{d}_0^{(t)} = -\frac{1}{h_0^{(t)}} \nabla l(\hat{\boldsymbol{\beta}}^{(t)})_0.$$

If $\mathbf{d}^{(t)} \neq \mathbf{0}$, an inexact line search using the Armijo rule will be performed. Let $\alpha^{(t)}$ be the largest value in $\{\alpha_0 \delta^l\}_{l \geq 0}$ such that

$$pl(\hat{\boldsymbol{\beta}}^{(t)} + \alpha^{(t)} \mathbf{d}^{(t)}) - pl(\hat{\boldsymbol{\beta}}^{(t)}) \leq \alpha^{(t)} \sigma \Delta^{(t)},$$

where $0 < \delta < 1, 0 < \sigma < 1, \alpha_0 > 0$, and $\Delta^{(t)}$ is the improvement in the objective function $pl(\boldsymbol{\beta})$ using a linear approximation, *i.e.*,

$$\begin{aligned} \Delta^{(t)} = -\mathbf{d}^{(t)T} \nabla l(\hat{\boldsymbol{\beta}}^{(t)}) &+ \lambda_1 \left\{ \sum_{k=1}^{p} \left\| \hat{\boldsymbol{\beta}}_k^{(t)} + \mathbf{d}_k^{(t)} \right\|_2 - \sum_{k=1}^{p} \left\| \hat{\boldsymbol{\beta}}_k^{(t)} \right\|_2 \right\} \\ &+ \lambda_2 \left\{ \sum_{k=1}^{p} \left\| \hat{\boldsymbol{\beta}}_k^{(t)} + \mathbf{d}_k^{(t)} \right\|_1 - \sum_{k=1}^{p} \left\| \hat{\boldsymbol{\beta}}_k^{(t)} \right\|_1 \right\}. \end{aligned}$$

Finally, we update the current estimate by

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + \alpha^{(t)} \mathbf{d}^{(t)}.$$

For $\mathbf{H}_{kk}^{(t)}$, we use the same choice as in Meier *et al.* (2008), that is,

$$h_k^{(t)} = -\max \left[\text{diag}\{-\nabla^2 l(\hat{\boldsymbol{\beta}}^{(t)})_{kk}\}, c^* \right],$$

where $c^* > 0$ is a lower bound to ensure convergence. In this chapter, we use the standard choices for the parameters: $\alpha_0 = 1, \delta = 0.5, \sigma = 0.1$ and $c^* = 0.001$ (Tseng and Yun, 2009) in the block coordinate descent algorithm to ensure the convergence of the algorithm.

Remark: In each iteration of the algorithm detailed above, when estimating the kth column

of the $q \times p$ coefficient matrix $\boldsymbol{\beta}$ with all other columns fixed, the algorithm first identifies the coefficients with zero estimates, denoted by set S in the algorithm. For the coefficients in set S, $d_{Sk}^{(t)} = -\hat{\boldsymbol{\beta}}_{Sk}^{(t)}$, and therefore when $\alpha^t = 1$, $\hat{\boldsymbol{\beta}}_{Sk}^{(t+1)} = \hat{\boldsymbol{\beta}}_{Sk}^{(t)} + \alpha^t d_{Sk}^{(t)} = 0$ and the coefficients in S are shrank to zero. Based on its definition, the set S depends on the turning parameter λ_2 and a larger value of λ_2 leads to fewer non-zero coefficients. The algorithm then performs a group shrinkage of the non-zero estimates of the coefficients in the complementary set $\bar{S}$. These non-zeros coefficients can further be shrunk to zero as a group if the condition $||\mathbf{u}_{\bar{S}k}^{(t)} - h_k^{(t)}\boldsymbol{\beta}^{(t)}||_2 < \lambda_1$ is met, in which case $d_{\bar{S}k}^{(t)} = -\hat{\boldsymbol{\beta}}_{\bar{S}k}^{(t)}$ and therefore $\hat{\boldsymbol{\beta}}_{\bar{S}k}^{(t+1)} = \hat{\boldsymbol{\beta}}_{\bar{S}k^{(t)}} + d_{\bar{S}k}^{(t)} = 0$. Clearly, this group shrinkage depends on the tuning parameter λ_1. This explains that with careful choice of the tuning parameters λ_1 and λ_2, some column group coefficients are set to zero and the within-group sparsity is achieved by the plain ℓ_1 penalty.

4.4.2. Tuning parameter selection

There are two tuning parameters λ_1 and λ_2 in the penalized likelihood estimation that need to be tuned with data by v-fold cross validation or a BIC criterion. To facilitate computation, we reparameterize λ_1 and λ_2 as $\lambda_1 = c\lambda\sqrt{q}$ and $\lambda_2 = (1-c)\lambda$. The multiplier $\sqrt{q}$ in the group penalty is used so that the group ℓ_1 penalty and overall ℓ_1 penalty are on a similar scale. Here we use λ to control the overall sparsity level and use $c \in [0,\ 1]$ to control the proportion of group ℓ_1 in the composite sparse group penalty. When $c = 0$, the penalty is reduced to lasso; when $c = 1$, it is reduced to group lasso. We consider the tuning parameter c from the set $\{0, 0.05, 0.1, 0.2, 0.4\}$. For each c, to search for the best tuning parameter value, we run the algorithm from $\lambda_{\max}$ so that it produces the sparsest model with the intercepts $\boldsymbol{\beta}_0$ only. The value $\lambda_{\max}$ can be roughly determined by using the starting value $\boldsymbol{\beta}^{(0)}$ with components $\boldsymbol{\beta}_j^{(0)} = \mathbf{0}$ $(j \neq 0)$ and $\boldsymbol{\beta}_0^{(0)}$ the MLE of (4.7) without covariates, and choosing the smallest value of λ so that the iteration converges in the first iteration, that is, $\boldsymbol{\beta}^{(0)}$ is a stationary point. We then decrease λ value and use the estimate of $\boldsymbol{\beta}$ from the last λ as a warm start. The grid of λ can be chosen to be equally spaced on

a log scale, *e.g.*, $\lambda_j = 0.96^j \lambda_{\max}$ $(j = 1, \cdots, m)$, where m is set so that $\lambda_{\min} = 0.2\lambda_{\max}$ or alternatively we could terminate the loop until the model receives more than the maximum number of nonzero coefficients allowed.

4.5. Simulation studies

4.5.1. Simulation strategies

We simulate n microbiome samples, p nutrients and q OTUs to mimic the real data set that we analyze in Section 4.6. The nutrient intake vector is simulated using a multivariate normal distribution with mean $\mathbf{0}$ and a covariance matrix $\Sigma_{i,j} = \rho^{|i-j|}$. We simulate p_r relevant nutrients with each nutrient being associated with q_r OTUs. For each nutrient, the association coefficients β_{ij} for the q_r OTUs are equally spaced over the interval $[0.6f, 0.9f]$ with alternative signs, where f controls the association strength. We consider two growth models to relate the OTU abundances to the covariates. In the exponential growth model, the proportion of the jth OTU of ith sample is determined as:

$$\phi_{ij} = \frac{\exp\left(\beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_{ik}\right)}{\sum_{j=1}^{q} \exp\left(\beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_{ik}\right)}. \tag{4.12}$$

The intercepts $\boldsymbol{\beta}_0$, which determines the base abundances of the OTUs, are taken from a uniform distribution over $(-2.3, 2.3)$ so that the base OTU abundances can differ up to 100 folds. The exponential growth model is a common model for bacteria growth in response to environmental stimuli. We also consider a linear growth model, in which the proportion of the jth OTU of ith sample is determined as:

$$\phi_{ij} = \frac{\beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_{ik}}{\sum_{j=1}^{q} (\beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_{ik})}.$$

The intercepts $\boldsymbol{\beta}_0$ are now drawn from a uniform distribution over $(0.02, 2)$ so that the base OTU abundances can also differ up to 100 folds. To deal with possible negative $\sum_{k=0}^{p} \beta_{jk} x_{ik}$, we add a small constant to make it positive.

We then generate the count data using DM model of parametrization II (4.4) with a common dispersion θ. The number of individuals (sequence reads) for ith sample m_i is generated from a uniform distribution over $(m, 2m)$. Note that the data are not generated exactly according to our model assumptions, which are based on parametrization I (4.2) and link (4.6). This can further demonstrate the robustness of our proposed model.

4.5.2. Evaluating the LRT for detecting environmental effect on microbiome composition

We compare our LRT with the pseudo-F statistic based permutation test for association between one covariate and the microbiome. The count data are normalized to proportions before performing the permutation test. The parameter values used in this simulation are: $n = 100,\ p = p_r = 1,\ q = 40,\ m = 500,\ \theta = 0.025$. We vary the number of relevant OTUs (q_r) and the association strength (f), and compare the powers under both exponential and linear growth models. The power is calculated based on $1,000$ replications at type I error 0.05. The results are shown in Table 4. Both methods control the type I error around 0.05 $(f = 0)$. In the exponential growth model, LRT is almost always more powerful than the permutation test, especially in situations where there are few relevant species and relatively strong covariate effect ($q_r = 2;\ f = 0.6, 0.8$). The LRT performs much better for models with more relevant OTUs and relatively weak covariate effect ($q_r = 8, 16;\ f = 0.2$). In summary, LRT has overall better power than the permutation test under almost all the models we considered, even when the model assumption is violated.

4.5.3. Evaluation of the penalized likelihood approach for selecting covariates affecting the microbiome composition

To evaluate the variable selection performance of the proposed sparse penalized likelihood approach with group ℓ_1 penalty, we first simulate the count data using the exponential growth model with $n = 100,\ p = 100,\ p_r = 4,\ q = 40,\ q_r = 4,\ m = 1000,\ \theta = 0.025$, and $\rho = 0.4$, totaling $4,000$ variables. We compare the results to the corresponding penalized estimation of the DM model using only ℓ_1 penalty function and two other sparse group ℓ_1

estimation based on multinomial or Dirichlet regression. In sparse multinomial regression, we use the multinomial model for count data and the link function is given by (4.12). We set $\beta_{10} = 0$ to make the coefficients identifiable. In sparse Dirichlet regression, instead of modeling the counts directly, we model the proportions using Dirichlet distribution and the link function is the same as that of the DM regression. Since the count data contain zeros, we add 0.5 to the cells with 0 counts. We also include results from LRT based univariate testing procedure for group selection controlling the false discovery rate (FDR) at 0.05.

We measure the selection performance using

$$\text{recall} = \frac{TP}{TP + FN}, \quad \text{precision} = \frac{TP}{TP + FP}, \quad \text{F}_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},$$

where TP, FN and FP are true positives, false negatives and false positives respectively, and F_1 is an overall measure, which weights the precision and recall equally. The averages of these measures are reported based on 100 replications.

To select the best tuning parameter values, we simulate an independent test data set of $n/2$ samples. We then run the penalized procedure over the training data set and re-estimate the selected coefficients using an unpenalized procedure ("nlm" function in R). The log-likelihood of the test data set is calculated based on the re-estimated coefficients and the tuning parameter is selected to maximize the log likelihood over the test data set. We choose the tuning parameter c from the set $\{0, 0.05, 0.1, 0.2, 0.4\}$. Figure 13 shows that a small c is sufficient to identify the groups efficiently, while further increase of c only improves the group selection marginally. On the other hand, within-group selection exhibits a unimode pattern indicating slight grouping could lead to better identification of within-group elements. In the following simulations, we tune both c and λ to achieve the maximum likelihood values in the test data sets.

Table 5 shows the simulation results. The sparse group ℓ_1 penalized DM regression has a much higher precision rate in group selection than the corresponding ℓ_1 penalized procedure,

while both achieve similar recall rates, demonstrating the gain from including the group ℓ_1 penalty in the regularization. Interestingly, the sparse group penalized DM regression also performs better in within-group selection as shown by a higher recall rate and F_1, indicating better group selection could also facilitate better overall variable selection. Compared to models based on the sparse Dirichlet regression and multinomial regression, DM model performs better in variable selection especially for within-group selection, suggesting the DM model is more appropriate than multinomial or Dirichlet models when the counts exhibit overdispersion. The Dirichlet model performs slightly better than the multinomial model. At 5% FDR, the LRT based univariate testing procedure selects far more variables than these penalized procedures, yielding a higher recall rate but a much worse precision rate.

4.5.4. Effects of overdispersion and model misspecification

We further investigate the effect of overdispersion and simulate the count data with different degrees of overdispersion and present the results in Figure 14. We observe that larger overdispersion makes the selection more difficult for all three models, as shown by smaller F_1 values. When the data have small overdispersion ($\theta = 0.005$), the selection performances of the three models are similar. On the other hand, when the data have a large overdispersion ($\theta = 0.1$), the DM based procedure performs much better than the other two in terms of both group selection and within-group selection. Therefore, modeling overdispersion can lead to gain in power of identifying relevant variables if the data are overdispersed and remains as powerful when the data do not exhibit overdispersion.

To assess the sensitivity to model misspecification, we simulate the counts using the linear growth model instead and compare the results with the exponential growth model (see Figure 14). Interestingly, both the Dirichlet and DM model are very robust to model misspecification and their selection performances do not decrease significantly. On the other hand, the multinomial model suffers a large performance loss with the F_1 measure for group selection decreasing from 0.79 to 0.56. We also study the effect of the total counts for each

sample (data not shown). Even increasing the total count by 10 folds, the DM model is still better than the proportion based Dirichlet model. Therefore, even we have much deeper sequencing of the microbiome that results in larger counts for each sample, using the DM model can still lead to improved performance over the model that considers only the proportions.

4.5.5. Effects of the number of the covariates and the relevant OTUs

We next study the effect of the number of relevant OTUs in each group on the performance of different models and present the results in Figure 15. When each relevant group contains only one relevant OTU, the grouping is not very helpful, so the sparse group regularized DM model and ℓ_1 regularized DM model do not differ much in selecting the relevant groups. When the relevant group contains 8 relevant OTUs, variable grouping becomes much more important and the sparse group regularized DM model performs much better than the ℓ_1 penalized DM. The group penalized multinomial and Dirichlet regression models, on the other hand, select groups as well as the DM regression model, since the grouping effect is much stronger.

Figure 15 also shows the results when we increase the dimension of covariates to 400 ($16,000$ variables in total). Increase of the dimension does not deteriorate the variable selection performance, demonstrating the efficiency of our method in handling high-dimensional data.

4.6. Application to real data analysis

Diet strongly affects the human health, partly by modulating gut microbial community composition. Wu *et al.* (2011a) studied the habitual diet effect on the human gut microbiome, where a cross-sectional 98 healthy volunteers were enrolled in the study. Diet information was collected using food frequency questionnaire (FFQ) and converted to nutrient intake values of 214 micronutrients. Nutrient intake was further normalized using the residual method to adjust for caloric intake and was standardized to have mean 0 and standard deviation 1. Since some nutrient measurements were almost identical and we used

only one representative for these highly correlated nutrients (correlation $\rho > 0.9$), resulting in 118 representative nutrients. Stool samples were collected and DNA samples were analyzed by the 454/Roche pyrosequencing of 16S rDNA gene segments of the V1-V2 region. The pyrosequences were denoised prior to taxonomic assignment yielding an average of $9,265 \pm 3,864(SD)$ reads per sample. The denoised sequences were then analyzed by the QIIME pipeline (Caporaso *et al.*, 2010b) with the default parameter settings. The OTU table contained 3068 OTUs (excluding the singletons) and these OTUs can be further combined into 127 genera (genus-level OTUs). We studied 30 relatively common genera that appeared in at least 25 subjects. Finally, we had the count matrix $\mathbf{Y}_{98\times 30}$ and covariate matrix $X_{98\times 118}$. Our goal is to identify the micronutrients that are associated with the gut microbiomes and the specific genera that the selected nutrients affect.

We applied the sparse group ℓ_1 penalized DM regression to this data set. We used the BIC to select the tuning parameters. The final DM model selected 11 nutrients and 13 associated genera. We refit the DM regression model using the selected variables and obtained the maximum likelihood estimates of the coefficients. We compared the fitted counts (total count $\times$ fitted proportion) against the observed counts in Figure 16 (top panel). The model fits the data quite well with $r^2 = 0.79$. Table 6 shows the MLEs of the regression coefficients for the selected nutrients and genera. Except for Methionine (second column), the coefficients are not too small. Since the nutrient measurements are standardized, the exponentiation of a given coefficient can be interpreted as the amount of change in proportion for a genus when a given nutrient changes by one unit while other nutrients remain constant. The marginal p-value based on the LRT for each of the selected nutrients is also shown in this table. Except for Vitamin E and Eriodictyol, these selected nutrients all show a significant marginal association with the gut microbiome.

To further assess the relevance of the nutrients selected, we used the bootstrap to analyze the stability of the selected nutrients (Bach, 2008). Specifically, we took 100 bootstrap samples and for each sample we ran our algorithm to select the nutrients. Since some nutrients are

highly correlated, we expect that highly correlated nutrients (if the correlation is greater than 0.75) can be selected in different bootstrap samples, we define the bootstrap selection probability of a given nutrient as the number of times that this nutrient or its correlated nutrients were selected. Table 6 shows the bootstrap probabilities of the nutrients that were selected by the sparse DM regression, indicating quite stable selection of most of the selected microbiome-associated nutrients. Vitamin E had the least stable selection over the 100 bootstrap samples.

The identified nutrient-genus associations are visualized in a bipartite graph shown in Figure 17, where the genera and nutrients are depicted with circles and hexagons, respectively. These results further confirmed the findings of Wu *et al.* (2011a), where they found the human gut microbiome can be clustered into two enterotypes characterized by Prevotella and Bacteroides respectively, and the Prevotella enterotype is associated with high carbohydrate diet while the Bacteroides enterotype is associated with high protein/fat/choline diet. Figure 17 shows that two carbohydrates, Maltose and Sucrose, are positively associated with Prevotella and negatively associated with Bacteroides while animal proteins are positively associated with Bacteroides, Parabacteroides and Alistipes, the three genera mostly enriched in the Bacteroides enterotype. Choline is positively associated with Bacteroides and negatively associated with Prevotella. Polyunsaturated fat is strongly associated with Alistipes, Odoribacter, Barnesiella and Parasutterella, indicating the large effect of fat on the human microbiome.

The DM model also identified several other associations that are worth further investigation. For example, we found that Naringenin (flavanone) was positively associated with Faecalibacterium, an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn's disease patients (Sokol *et al.*, 2008). If the association is validated, diet with high Naringenin (e.g. Orange, Grapefruit) can be beneficial for patients with Crohn's disease.

As a comparison, we also run the sparse group ℓ_1 penalized multinomial or Dirichlet regres-

sion models and the identified nutrient-genus associations showed significant overlap with those from the DM regression model. However, the interpretability of the DM regression model was the best. To further demonstrate the advantage of the DM model, we simulated OTU counts for each individual based on the fitted models and the observed total OTU counts. The bottom plot of Figure 16 shows that the simulated counts produced by the fitted sparse DM model resemble the observed counts better than those from the sparse multinomial model, where the simulated counts are apparently over-smoothed. This indicates the importance of considering the overdispersion in modeling the gut microbiome data. We also performed LRT based univariate testing procedure. At FDR=0.05, the LRT identified 13 nutrients, 8 of which are also identified or highly correlated with the nutrients identified by the sparse group ℓ_1 penalized DM model.

4.7. Discussion

We have proposed a sparse group ℓ_1 penalized estimation for the DM regression in order to select covariates associated with the microbiome composition. The sparse group ℓ_1 penalty encourages both group-level and within-group sparsity, with which we can select the relevant OTUs associated with the selected covariates. We have performed extensive simulations to evaluate our proposed penalized estimation procedure for both group and within-group selections. We demonstrated the procedure with a real data set on associating nutrient intakes with gut microbiome composition and confirmed the major findings in Wu *et al.* (2011a).

In our penalized likelihood estimation of the DM model, we use a combination of group ℓ_1 and individual ℓ_1 penalties, which result in a convex and separable (in groups of parameters) penalty function. This property facilitates the application of the general coordinate gradient descent method of Tseng and Yun (2009) to implement an efficient optimization algorithm. In each iteration, we have a closed form solution for a block update. For a given set of the sparsity tuning parameters, our algorithm is fully automatic and does not require the specification of an algorithmic tuning parameter to ensure convergence. For

example, it took about 3 minutes on a standard laptop (Core i5, 2G memory) to finish the analysis of the real data set using an R implementation of the algorithm (available on *http://statgene.med.upenn.edu/*). Beside the sparse l_1 group penalty, other group penalty functions such as the sup-norm penalty in Zhang *et al.* (2008) and the composite absolute penalties in Zhao *et al.* (2009) can be also be used in the setup of the Dirichlet multinomial regression. However, efficient implementation of the optimization problems with these penalty functions is challenging.

In microbiome data analysis literature, one commonly used approach is to normalize the counts into proportions and perform statistical analysis using the proportions. However, by converting into the proportions, the variation associated with multinomial sampling process is lost. In 16S rRNA sequencing, the sequencing depths (total counts) for samples can vary up to 10-fold. Obviously, the accuracy of the proportion estimates under sequencing depth of 500 reads is very different from that of $10,000$ reads. As shown in our simulations, modeling counts directly can result in gain of power in selecting relevant variables even when the number of sequence reads is very large. Another problem associated with proportions is the existence of numerous zeros in the OTU count data. Many proportion based approaches require taking logarithms of the proportions, which is problematic for the zero proportions. To circumvent this problem, either a pseudo count (e.g. 0.5) is added to these zero counts before converting into proportions or an arbitrary small proportion is substituted for these zero proportions. The effects of creating pseudo counts have not been evaluated thoroughly when the data contain excessive zeros.

Besides overdispersion, the OTU count data can also exhibit zero-inflation (Barry and Welsh, 2002) where the count data contain more zeros than expected from the DM model. How to model the microbiome count data that allows overdispersion, zero-inflation and possibly the phylogenetic correlations among the OTUs is an important future research topic. Multilevel zero-inflated DM regression model for overdispersed count data with extra zeros (Moghimbeigi *et al.*, 2008; Lee *et al.*, 2006) can potentially provide a solution

to this problem. Another problem associated with the DM model is its inflexibility in modeling the covariance structure among the OTU counts. The multinomial model for counts compounded by a logistic normal model (Aitchison, 1982) for proportions provides a possible solution. This needs to be investigated further.

Table 4: **Comparison of the power of pseudo-F statistic based permutation test (Perm) and the DM model based likelihood ratio test (LRT) in detecting the covariate effect.** The power is calculated based on 1,000 replications.

Spe No. (q_r)	Method	Exponential growth, Covariate Effect (f) 0	0.2	0.4	0.6	0.8	Linear growth, Covariate Effect (f) 0	0.2	0.4	0.6	0.8
2	Perm	0.04	0.19	0.45	0.59	0.69	0.05	0.10	0.42	**0.75**	**0.96**
	LRT	0.06	0.19	**0.63**	**0.86**	**0.96**	0.06	**0.20**	**0.50**	0.71	0.90
4	Perm	0.05	0.31	0.69	0.85	0.91	0.05	0.18	0.71	**0.98**	1.00
	LRT	0.05	**0.35**	**0.89**	**0.99**	**1.00**	0.06	**0.36**	**0.80**	0.97	1.00
8	Perm	0.04	0.54	0.92	0.98	1.00	0.05	0.39	0.96	1.00	1.00
	LRT	0.06	**0.67**	**0.99**	**1.00**	1.00	0.05	**0.70**	**0.99**	1.00	1.00
16	Perm	0.06	0.89	1.00	1.00	1.00	0.05	0.66	1.00	1.00	1.00
	LRT	0.05	**0.97**	1.00	1.00	1.00	0.05	**0.95**	1.00	1.00	1.00

Table 5: **Comparison of sparse group ℓ_1 and ℓ_1 penalized procedures for variable selection under Dirichlet-multinomial (DM), Dirichlet (D) and multinomial (M) regression models.** The selection performance, both group selection and within-group selection, is evaluated using recall rate (R), precision rate (P) and F_1 (F), all averaged over 100 runs (standard deviation in parenthesis). Selection based on univariate likelihood ratio test (LRT) at FDR=0.05 is also indicated.

	Sparse group ℓ_1 penalization						ℓ_1 penalization					
	Within-group			Group			Within-group			Group		
Model	R	P	F	R	P	F	R	P	F	R	P	F
	Exponential growth, $p=100, q_r=4, \theta=0.025$											
DM	0.59	0.70	0.59	0.86	0.92	0.87	0.42	0.76	0.48	0.88	0.68	0.70
	(0.23)	(0.23)	(0.18)	(0.23)	(0.16)	(0.18)	(0.21)	(0.23)	(0.18)	(0.22)	(0.29)	(0.22)
D	0.48	0.73	0.52	0.83	0.89	0.82	0.36	0.82	0.45	0.82	0.77	0.72
	(0.23)	(0.23)	(0.20)	(0.26)	(0.18)	(0.21)	(0.20)	(0.21)	(0.19)	(0.26)	(0.27)	(0.23)
M	0.46	0.72	0.50	0.82	0.85	0.79	0.36	0.76	0.44	0.84	0.70	0.69
	(0.23)	(0.26)	(0.21)	(0.27)	(0.24)	(0.25)	(0.19)	(0.24)	(0.18)	(0.26)	(0.28)	(0.24)
LRT	-	-	-	0.96	0.54	0.66	-	-	-	0.96	0.54	0.66
	-	-	-	(0.11)	(0.21)	(0.16)	-	-	-	(0.11)	(0.21)	(0.16)

Table 6: **Estimated regression coefficients from the sparse group ℓ_1 penalized DM regression for the diet-gut microbiome data.** The exponentiation of a given coefficient can be interpreted as the amount of change in proportion for a genus when a given nutrient changes by one unit while other nutrients remain constant. The columns 1-11 represent the selected nutrients: Polyunsaturated fat, Methionine, Sucrose, Animal Protein, Vitamin E- Food Fortification, Maltose, Added Germ from wheats, Choline-Phosphatidylcholine, Taurine, Naringenin-flavanone and Eriodictyol-flavonone. The rows 1-13 represent the selected genus-level OTUs(genera): Bacteroides, Barnesiella, Odoribacter, Parabacteroides, Prevotella, Alistipes, Coprococcus, Faecalibacterium, Oscillibacter, Ruminococcus, Subdoligranulum, Phascolarctobacterium and Parasutterella. Marginal p-value based on the LRT and the bootstrap selection probability of each of the selected nutrients are also shown.

Row: genus; Column: nutrient										
-	-0.03	-0.08	0.09	-0.08	-0.10	-0.02	0.02	0.10	-	-0.03
-0.32	-	-0.33	-	-	-	0.22	-	-	-	-
-0.38	-	-	-	-	-	-	-	-	-0.29	-
-	-0.01	-0.08	0.13	-0.07	-	-	-	0.02	-0.23	-
-	-	0.23	-	-	0.36	0.63	-0.72	-	-	-
-0.19	-0.04	-	0.16	-	-	-	-	-	-	0.05
-	-	-	-	-	-	-	-	-	-	0.16
-	-	-	-	-	-0.08	-	-	-	0.07	-
-	-0.02	-	-	-	-	-	-	-0.10	-	-
-	-	0.19	-	-	-	-	-	-	-	-
-	0.02	-	-	-	-	-	-	-0.12	-0.12	0.14
-	-	-	-	-0.35	-	-	-	-	-	-
-0.26	-	-0.29	-	-	-	-	-	-	-	-
Marginal p-value										
4.5×10^{-3}	2.2×10^{-4}	8.4×10^{-4}	3.6×10^{-4}	1.1×10^{-1}	6.0×10^{-3}	9.5×10^{-6}	2.7×10^{-3}	5.9×10^{-3}	5.8×10^{-2}	5.2×10^{-3}
Bootstrap selection probability										
0.50	0.93	0.72	0.94	0.35	0.67	0.43	0.58	0.92	0.61	0.60

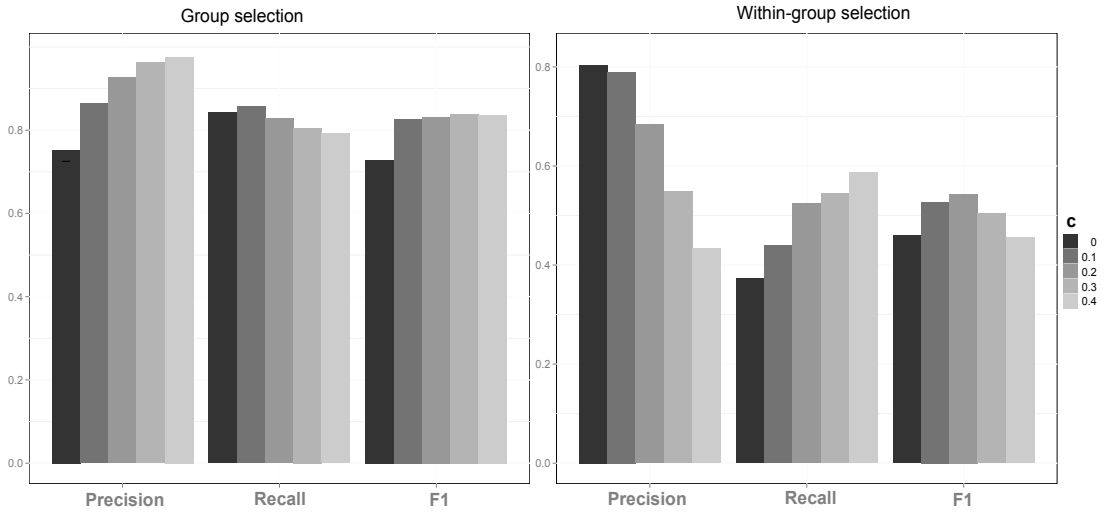

Figure 13: **Effects of the tuning parameter c on variable selection.** The tuning parameter c is varied from 0 to 0.4. Under each value of c, the best λ value, which maximizes the likelihood of the test data set, is selected to generate the sparse model. Group (left) and within-group (right) selection performance are then evaluated using measures of recall, precision and F_1 based on 100 replications. Simulation setting: $n=100,\ p=100,\ p_r=4,\ q=40,\ q_r=4,\ m=500,\ \theta=0.025,\ \rho=0.4$.

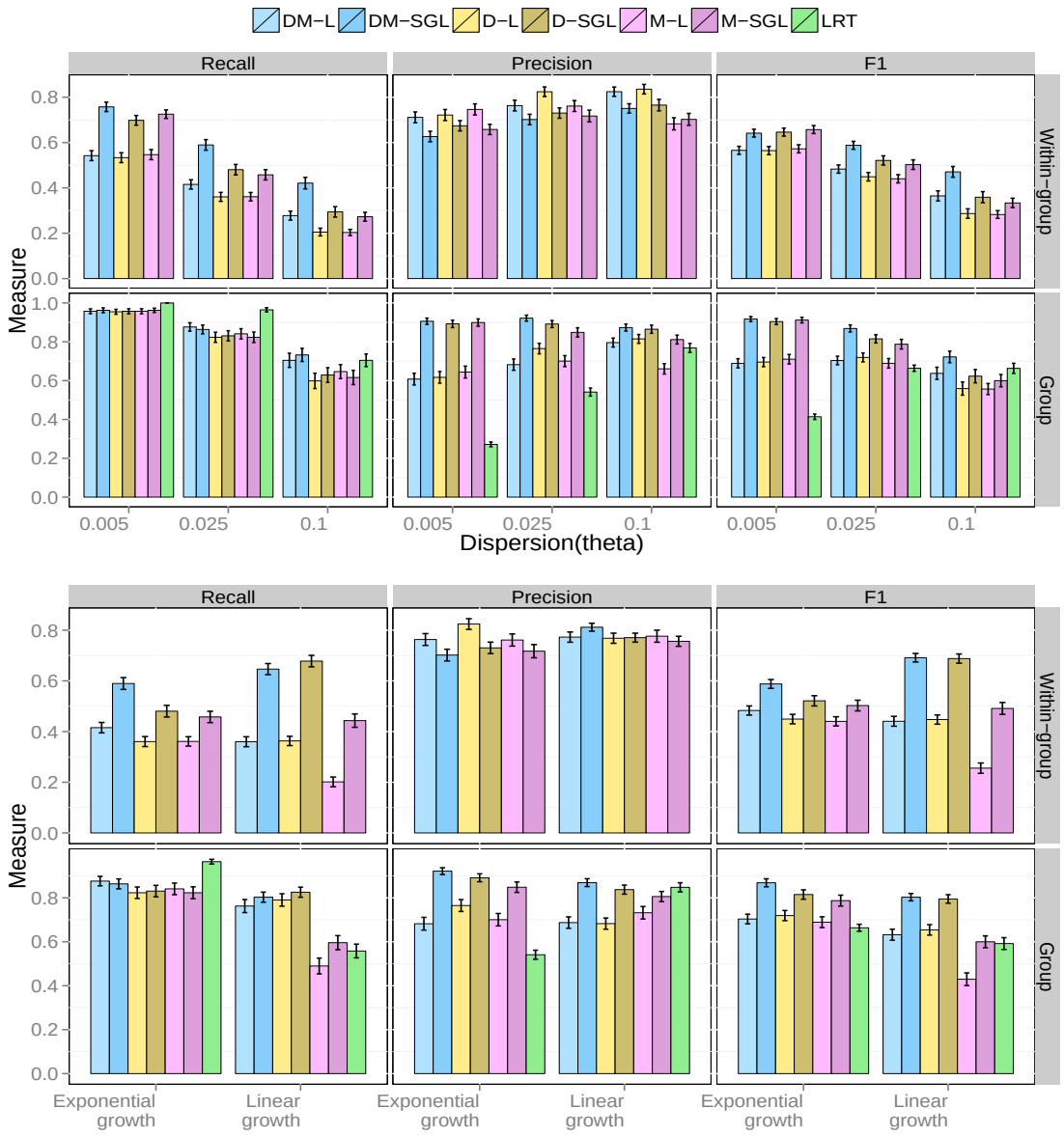

Figure 14: **Effects of overdispersion (top panel) and model-misspecification (bottom panel) on the performance of several models and methods.** DM-SGL: sparse group ℓ_1 penalized Dirichlet-multinomial model; DM-L: ℓ_1 penalized Dirichlet-multinomial model; M-SGL: sparse group ℓ_1 penalized multinomial model; M-L: ℓ_1 penalized multinomial model; D-SGL: sparse group ℓ_1 penalized Dirichlet model; D-L: ℓ_1 penalized Dirichlet model. For each bar, mean±standard error is presented based on 100 replications.

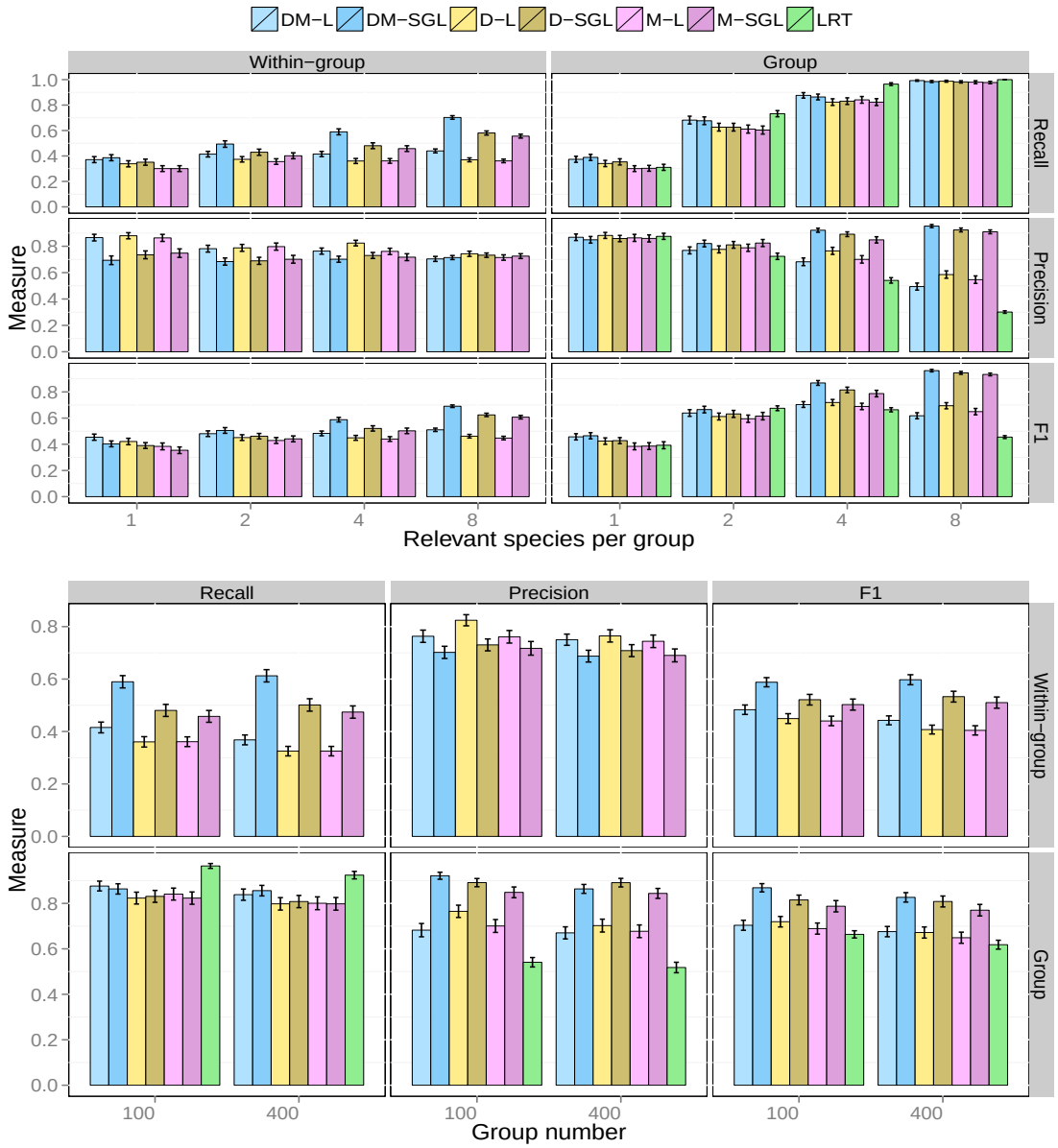

Figure 15: **Effects of the number of relevant OTUs (top panel) and the number of covariates (bottom panel) on the performances of several models and methods.** DM-SGL: sparse group ℓ_1 penalized Dirichlet-multinomial model; DM-L: ℓ_1 penalized Dirichlet-multinomial model; M-SGL: sparse group ℓ_1 penalized multinomial model; M-L: ℓ_1 penalized multinomial model; D-SGL: sparse group ℓ_1 penalized Dirichlet model; D-L: ℓ_1 penalized Dirichlet model. For each bar, mean±standard error is presented based on 100 replications.

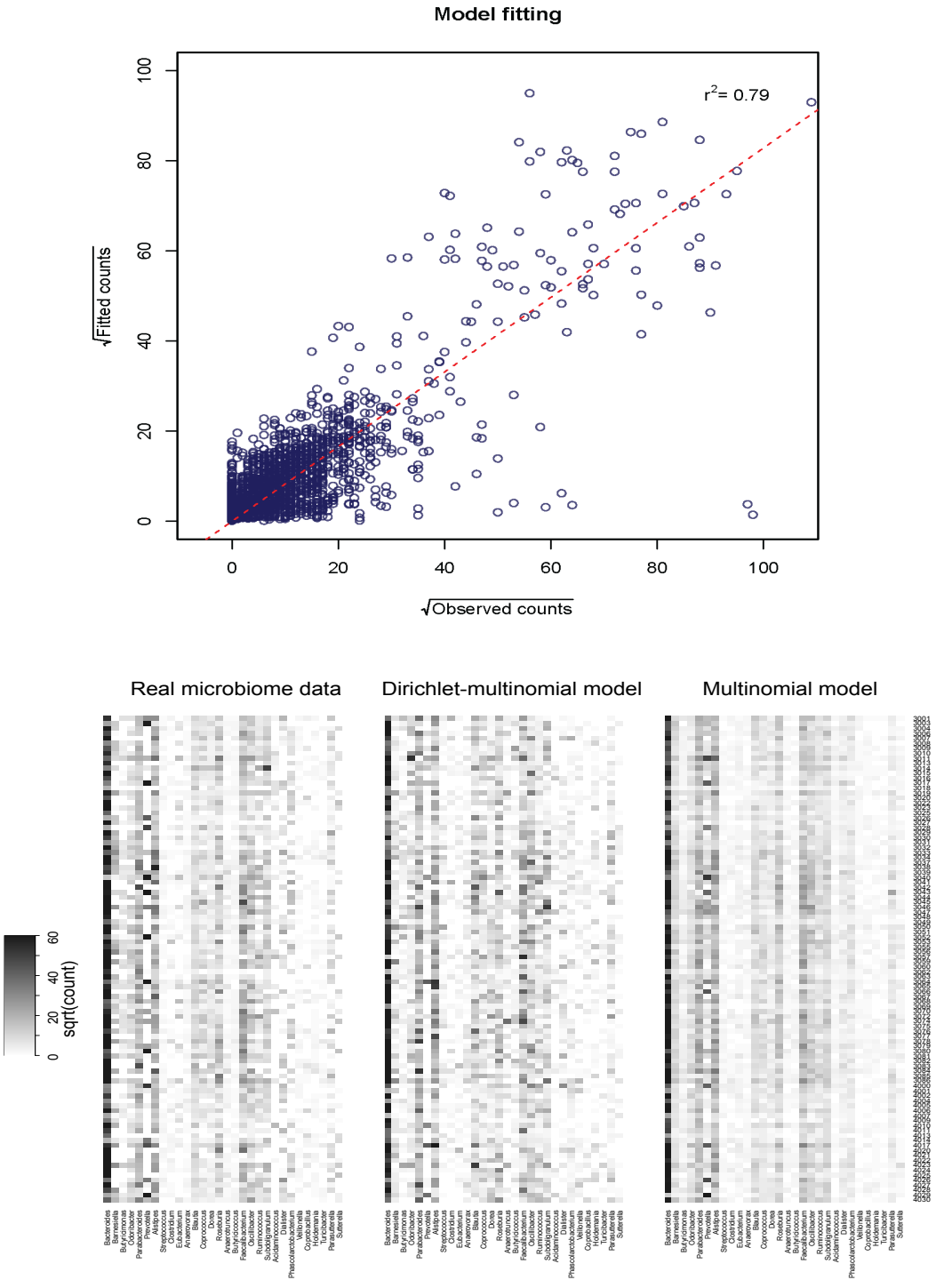

Figure 16: **Model fit using the variables selected by the sparse group l_1 penalized DM regression model.** Top plot: square root of the fitted counts versus square root of the observed counts based on the DM model with the selected nutrients; bottom plots: Observed counts and simulated counts produced by the fitted sparse DM model and multinomial model.

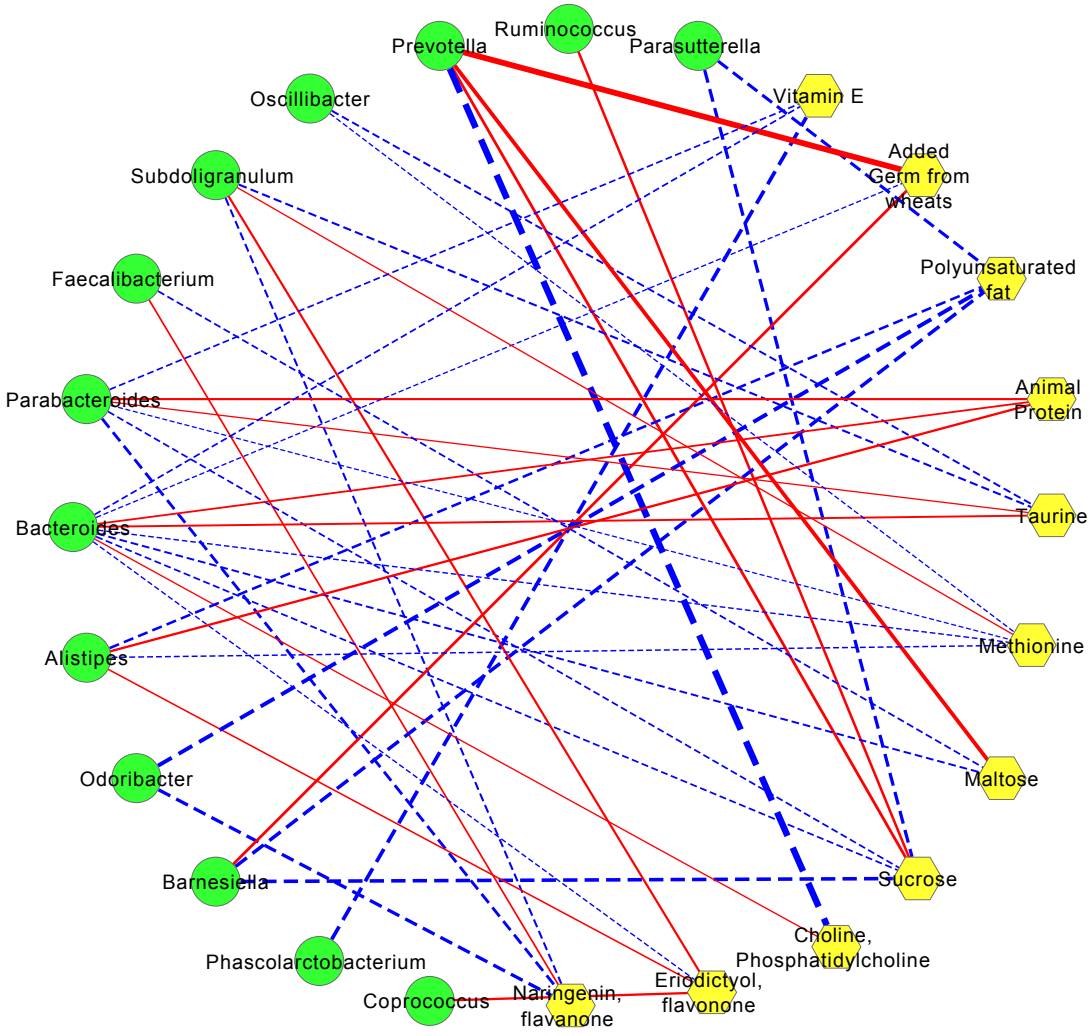

Figure 17: **Nutrient-genus association in the human gut identified by the sparse group ℓ_1 penalized DM regression model.** We use a bipartite graph to visualize the selected nutrients and their associated genera based on sparse group ℓ_1 penalized DM regression. Circle: genus; hexagon: nutrient; solid line: positive correlation; dashed line: negative correlation. The thickness of the line represent the association strength.

CHAPTER 5 : Future work

This dissertation presents three statistical methods developed specifically for 16S metagenomic data analysis in the context of associating the microbiome composition with environmental covariates. These methods have taken into account some important aspects of the 16S data such as the high-dimensionality of the OTU counts (GUniFrac, ssCCA, SDMR), the phylogenetic constraint among the OTUs (GUniFrac, ssCCA), the overdispersion of the OTU counts (SDMR) and the skewness of the OTU abundance distribution (GUniFrac). For future research, I have identified the following problems of interest. Some have immediate solutions based on previously proposed methods from other fields while others need more careful consideration.

Kernel methods for testing the significance of microbiome composition on a disease outcome. One important goal of human microbiome studies is to test the significance of microbiome composition on a biological/disease outcome after adjusting for other covariates (*e.g* gut microbiome on inflammatory bowel disease adjusting for nutrient intakes and genotypes). The kernel based semi-parametric regression methods, which was initially developed for testing the significance of the gene pathway effect on an outcome (Liu *et al.*, 2007, 2008) and was later extended to genome-wide association studies (Wu *et al.*, 2011b), can be potentially applied to this problem. The kernel method can allow for flexible modeling of nonlinear OTU effects and OTU interactions. Incorporation of the phylogenetic information is possible by designing a phylogenetic tree based kernel. Adjustment of covariate effects is very natural in this framework and the corresponding score test is computationally efficient.

Empirical Bayes method for OTU differential abundance analysis. Identification of OTUs that show differential abundance between two conditions such as smoking vs nonsmoking is a very important problem in microbiome studies. Currently, methods for identifying OTUs with differential abundance under two conditions are very simple and fail to utilize the unique property of the composition data (Rodriguez-Brito *et al.*, 2006; White *et al.*, 2009;

Parks and Beiko, 2010; Wagner *et al.*, 2011). They treat the OTUs as separate fixed effects, which may have reduced efficiency when compared to empirical Bayes method. Empirical Bayes method (Efron *et al.*, 2001) is the most successful method for gene differential expression analysis due to its ability of pooling information across genes. It can also be applied to differential abundance analysis by pooling information across OTUs. Under the empirical Bayes framework, we can model the taxa counts using a beta-binomial model and the prior distribution of the mean of the taxa proportion can be taken to be another beta distribution. Inference can then be based on the posterior odds.

*Compositing distance measures for microbiome data analysis.*There are numerous distance measures (>20) for comparing microbiomes (Kuczynski *et al.*, 2010b; Swenson, 2011). They can be quantitative or qualitative, phylogenetic or non-phylogenetic. Each distance is only capable of revealing a certain aspect of the microbiome difference and no distance measure is optimal across all conditions. On the other hand, many distances are highly correlated. Selection of representative distances and compositing these distances in a distance-based statistical framework is expected to increase statistical power.

Sparse clustering for microbiome data. Cluster analysis has been recently applied to microbiome data analysis. For example, three robust clusters (enterotypes) have been discovered for the human gut microbiomes (Arumugam *et al.*, 2011) using a distance-based clustering method. Holmes *et al.* (2012) proposed a model-based clustering method using Dirichlet multinomial mixtures. These methods use the abundances of all OTUs to produce the clusters. However, many OTUs are not informative. Including them in the analysis increases the noise level and leads to failure of establishing clear clusters. Sparse clustering (Pan and Shen, 2007; Witten and Tibshirani, 2010), where clustering and feature selection are integrated, may provide a useful alternative. This method will select OTUs responsible for the clustering pattern, hence providing more mechanistic insights into the formation of clusters.

Genotype, microbiome and disease relationship studies. Genome-wide association studies

have identified certain disease susceptibility loci for Crohn's disease (Barrett *et al.*, 2008). Meanwhile, the gut microbiome composition is also associated with the disease (Manichanh *et al.*, 2006). It is interesting to know the relationship between genotype (G), microbiome (M) and disease (D). It could be causal ($G \to M \to D$), independent ($G \to D; M \to D$), reactive ($G \to D \to M$) or interactive ($G \times M \to D$). Building a likelihood model to distinguish these possible relationships is crucial for understanding the etiology of the microbiome-associated genetic diseases.

Statistical and computational analysis of shotgun metagenomic data. Shotgun metagenomic approach has become increasingly popular for microbiome studies due to its ability to reveal both taxonomic and functional content of the microbiome (Tringe *et al.*, 2005; Gill *et al.*, 2006; von Mering *et al.*, 2007; Grice *et al.*, 2009; Iverson *et al.*, 2012). The shotgun metagenomic data are much more complex than 16S data. Each sequence is randomly sampled from any location within a random microbial genome from an unknown mixture of microbial genomes of different sizes, abundances and phylogenetic divergence. Statistical modeling of the sampling process can produce more accurate characterization of the microbiomes. Many problems from shotgun metagenomics have not yet been satisfactorily solved. De Novo assembly for metagenomes, analysis of taxonomic composition of metagenomes, gene/pathway level analysis, studies of other communities (viruses/phages, microeukaryotes etc) are all interesting and challenging problems.

The problems listed above only scratch the surface of the exciting field of microbiome data analysis. Statistical and computational analysis of microbiome data is still in its infancy. In the next a few years, it is expected that more computational and statistical tools will emerge to make sense of the metagenomic data. Hopefully, this dissertation will motivate more research in this field.

APPENDIX

A.1. Theorem

Theorem 1. *Let* $\mathbf{b}, \mathbf{x} \in \mathbb{R}^n$,$\lambda_1, \lambda_2, c$ *are non-negative constants and* $\mathbf{x}^0$ *is the minimizer of the following function*

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{x} + \mathbf{b}^T\mathbf{x} + c + \lambda_1 \|\mathbf{x}\|_2 + \lambda_2 \|\mathbf{x}\|_1 , \tag{A.1}$$

then $\mathbf{x}^0_S = \mathbf{0}$ *and* $\mathbf{x}^0_{\bar{S}} = \arg\min_{\mathbf{x}_{\bar{S}}} \left\{ \frac{1}{2}\mathbf{x}^T_{\bar{S}}\mathbf{x}_{\bar{S}} + (\mathbf{b}_{\bar{S}} - \lambda_2 sgn(\mathbf{b}_{\bar{S}}))^T \mathbf{x}_{\bar{S}} + c + \lambda_1 \|\mathbf{x}_{\bar{S}}\|_2 \right\}$, *where* $S = \{i \in \{1, \cdots, n\} |\ |b_i| < \lambda_2\}$ *and* $\bar{S} = \{1, \cdots, n\} \backslash S$ *and sgn(.) is the sign function.*

Proof. We prove $\mathbf{x}^0_S = \mathbf{0}$ by contradiction. If $x^0_i \neq 0$ $(i \in S)$, then we can construct a new $\mathbf{x}^1$ with $x^1_i = 0$ and other components being the same as $\mathbf{x}^0$. Clearly, $\frac{1}{2}\mathbf{x}^{1T}\mathbf{x}^1 + \mathbf{b}^T\mathbf{x}^1 + c + \lambda_2 \left\|\mathbf{x}^1\right\|_1 < \frac{1}{2}\mathbf{x}^{0T}\mathbf{x}^0 + \mathbf{b}^T\mathbf{x}^0 + c + \lambda_2 \left\|\mathbf{x}^0\right\|_1$ and $\lambda_1 \left\|\mathbf{x}^1\right\|_2 < \lambda_1 \left\|\mathbf{x}^0\right\|_2$. The former is due to the fact that $\frac{1}{2}(x^0_i)^2 + b_i x^0_i + \lambda_2 |x^0_i| > 0$ for $|b_i| < \lambda_2$. Hence $\mathbf{x}^0$ is not the minimizer of $f(\mathbf{x})$, which is contradictory. Therefore, $\mathbf{x}^0_S = \mathbf{0}$.

To prove the second part, we note that x^0_i must be either 0 or have an opposite sign of b_i for $i \in \{1, \cdots, n\}$. So the minimization of $\mathbf{f}(\mathbf{x})$ is equivalent to minimizing

$$\begin{aligned} f^*(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{x} + (\mathbf{b} - \lambda_2 \mathrm{sgn}(\mathbf{b}))^T\mathbf{x} + c + \lambda_1 \|\mathbf{x}\|_2 , \\ \text{subject to } \mathrm{sgn}(x_i) = -\mathrm{sgn}(b_i) \ \text{ or } \ x_i = 0. \end{aligned}$$

Since $\mathbf{x}^0_S = \mathbf{0}$, we can restrict the minimization over only $\mathbf{x}_{\bar{S}}$

$$\begin{aligned} f^*(\mathbf{x}_{\bar{S}}) = \frac{1}{2}\mathbf{x}^T_{\bar{S}}\mathbf{x}_{\bar{S}} + (\mathbf{b}_{\bar{S}} - \lambda_2 \mathrm{sgn}(\mathbf{b}_{\bar{S}}))^T\mathbf{x}_{\bar{S}} + c + \lambda_1 \|\mathbf{x}_{\bar{S}}\|_2 , \\ \text{subject to } \mathrm{sgn}(x_i) = -\mathrm{sgn}(b_i) \ \text{ or } \ x_i = 0 \ (i \in \bar{S}). \end{aligned} \tag{A.2}$$

Since $\mathbf{x}^0_{\bar{S}}$ is the minimizer of $f^*(\mathbf{x}_{\bar{S}})$ without the constraint, the sign of $\mathbf{x}^0_{\bar{S}}$ should be the

opposite of the sign of $(\mathbf{b}_{\bar{S}} - \lambda_2 \text{sgn}(\mathbf{b}_{\bar{S}}))$. Because $|b_i| \geq \lambda_2$ for $i \in \bar{S}$, the sign of $(\mathbf{b}_{\bar{S}} - \lambda_2 \text{sgn}(\mathbf{b}_{\bar{S}}))$ is the same as $\mathbf{b}_{\bar{S}}$. So the sign of $\mathbf{x}^0_{\bar{S}}$ is the opposite of that of $\mathbf{b}_{\bar{S}}$. Therefore, $\mathbf{x}^0_{\bar{S}}$ satisfies the constraint. □

Using simple variable substitution, we have the following Corollary.

Corollary 1. *Let* $\mathbf{b}, \boldsymbol{\beta}, \mathbf{d} \in \mathbb{R}^n$,$\lambda_1, \lambda_2, c$ *are non-negative constants and* $\mathbf{d}^0$ *is the minimizer of the following function*

$$f(\mathbf{d}) = \frac{1}{2}\mathbf{d}^T\mathbf{d} + \mathbf{b}^T\mathbf{d} + c + \lambda_1 \|\boldsymbol{\beta} + \mathbf{d}\|_2 + \lambda_2 \|\boldsymbol{\beta} + \mathbf{d}\|_1 , \tag{A.3}$$

then $\mathbf{d}^0_S = -\boldsymbol{\beta}_S$ *and*

$$\mathbf{d}^0_{\bar{S}} = \arg\min_{\mathbf{d}_{\bar{S}}} \left\{ \frac{1}{2}\mathbf{d}^T_{\bar{S}}\mathbf{d}_{\bar{S}} + (\mathbf{b}_{\bar{S}} - \lambda_2 sgn(\mathbf{b}_{\bar{S}} - \boldsymbol{\beta}_{\bar{S}}))^T \mathbf{d}_{\bar{S}} + c + \lambda_1 \|\mathbf{d}_{\bar{S}} + \boldsymbol{\beta}_{\bar{S}}\|_2 \right\},$$

where $S = \{i \in \{1, \cdots, n\} |\ |b_i - \beta_i| < \lambda_2\}$, $\bar{S} = \{1, \cdots, n\} \backslash S$ *and sgn(.) is the sign function.*

A.2. Supplementary Tables and Figures

Table A1: Differential OTUs between smokers and nonsmokers in the oropharyngeal microbiome.

OTU ID	Lineage	Proportion	Test	Raw P value	Smoker
1490	Firmicutes;Veillonella	6.1E-02	Wilcoxon	5.7E-04	+
411	Firmicutes;Veillonella	2.2E-03	Fisher	1.2E-03	+
2434	Bacteroidetes;Prevotella	3.6E-02	Wilcoxon	1.4E-03	+
3538	Actinobacteria;Atopobium	1.0E-02	Wilcoxon	1.9E-03	+
1280	Firmicutes;Lachnospiraceae	5.9E-04	Fisher	2.4E-03	-
4363	Firmicutes;Lachnospiraceae	3.9E-04	Fisher	3.1E-03	+
2831	Bacteroidetes;Prevotella	2.2E-02	Wilcoxon	3.7E-03	+
2893	Bacteroidetes;Prevotella	2.1E-03	Wilcoxon	3.8E-03	+
2300	Bacteroidetes;Prevotella	6.1E-03	Wilcoxon	4.1E-03	-
4703	Firmicutes;Megasphaera	1.2E-02	Wilcoxon	4.5E-03	+
3227	Proteobacteria;Neisseria	5.0E-02	Wilcoxon	9.0E-03	-
4912	Firmicutes;Dialister	3.8E-04	Fisher	9.1E-03	+
4357	Spirochaetes;Treponema	1.7E-03	Wilcoxon	9.9E-03	-
3954	Fusobacteria;Fusobacterium	5.1E-02	Wilcoxon	1.1E-02	-
4440	Firmicutes;Streptococcus	1.3E-03	Fisher	1.2E-02	+
1766	Firmicutes;Lachnospiraceae I.S.	1.0E-04	Fisher	1.8E-02	+
913	Spirochaetes;Treponema	1.3E-04	Fisher	1.8E-02	+
5603	Actinobacteria;Eggerthella	8.0E-05	Fisher	1.8E-02	+
4813	Bacteroidetes;Prevotellaceae	7.3E-03	Wilcoxon	1.9E-02	-
4871	Proteobacteria;Haemophilus	2.0E-02	Wilcoxon	2.7E-02	-
2913	Firmicutes;Veillonella	9.0E-05	Fisher	2.9E-02	-
171	Bacteroidetes;Prevotella	6.3E-04	Fisher	3.5E-02	+
1633	Actinobacteria;Actinomyces	1.8E-03	Wilcoxon	3.7E-02	+
1365	Bacteroidetes;Prevotella	1.5E-03	Wilcoxon	4.1E-02	-
4847	Firmicutes;Erysipelotrichaceae I.S.	4.0E-05	Fisher	4.2E-02	+
1796	Firmicutes;Streptococcus	1.5E-04	Fisher	4.2E-02	+
5402	Bacteroidetes;Prevotella	1.0E-04	Fisher	4.2E-02	+
3529	Firmicutes;Lachnospiraceae	4.0E-05	Fisher	4.2E-02	+
4816	Fusobacteria;Fusobacterium	2.2E-04	Fisher	4.2E-02	+
4365	Actinobacteria;Atopobium	6.0E-05	Fisher	4.2E-02	+
2228	Firmicutes;Lachnospiraceae	7.4E-04	Fisher	4.3E-02	-
4036	Firmicutes;Moryella	1.3E-02	Wilcoxon	4.5E-02	+

When the frequency of OTU occurrence is less than 1/3, Fisher's exact test is used; otherwise, Wilcoxon rank sum test is used. "+" indicates increase in smokers relative to nonsmokers.

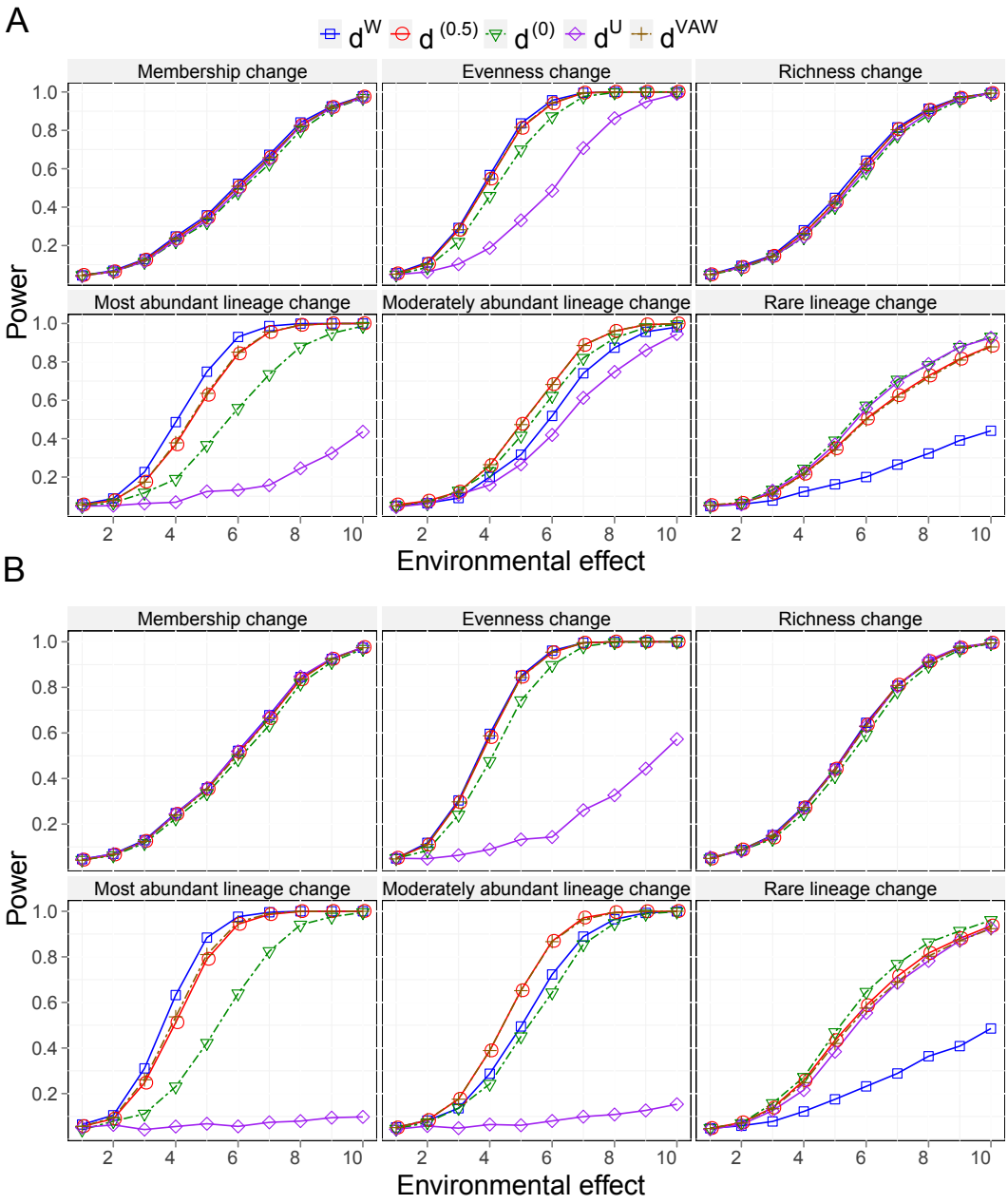

Figure A1: **Power comparison of different UniFrac variants for detecting environmental effects using 2D circle based simulation and different bin sizes for OTU formation.** Ten samples from each of the two environmental conditions are generated using 2D circle based simulation. A bin size of 0.01 (A) or 0.03 (B) is used in OTU formation. UniFrac distance matrices are constructed based on the simulated OTU abundances and NJ tree. PERMANOVA is used for testing hypotheses. d^W, $d^{(0.5)}$, $d^{(0)}$, d^U and d^{VAW} are compared and indicated by different colors. The specific community difference caused by different environmental conditions is indicated in the panel title. The power curves are created by varying the degree of environmental effect. The initial point of the power curve is the power when there is no environmental effect.

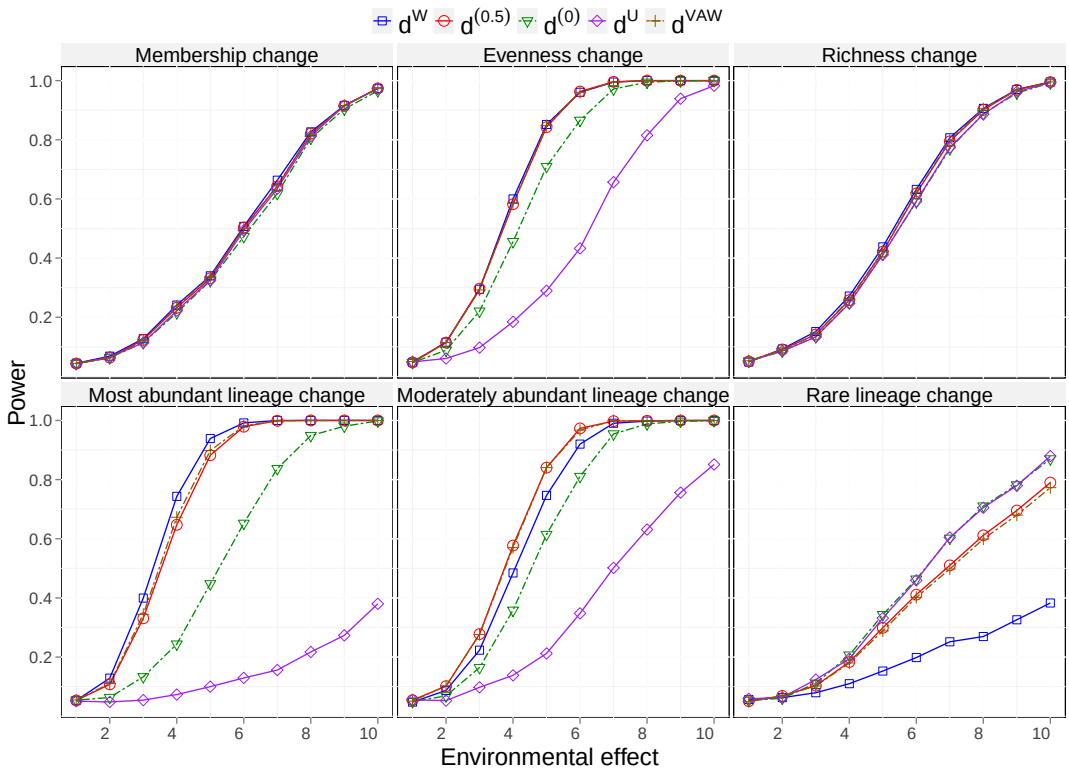

Figure A2: **Power comparison of different UniFrac variants for detecting environmental effects using 2D circle based simulation and UPGMA tree.** Ten samples from each of the two environmental conditions are generated using 2D circle based simulation. A bin size of 0.015 is used in OTU formation. UniFrac distance matrices are constructed based on the simulated OTU abundances and UPGMA tree. PERMANOVA is used for testing hypotheses. Four representative UniFrac variants d^W, $d^{(0.5)}$, $d^{(0)}$, d^U and d^{VAW} are compared and indicated by different colors. The specific community difference caused by different environmental conditions is indicated in the panel title. The power curves are created by varying the degree of environmental effects. The initial point of the power curve is the power when there is no environmental effect.

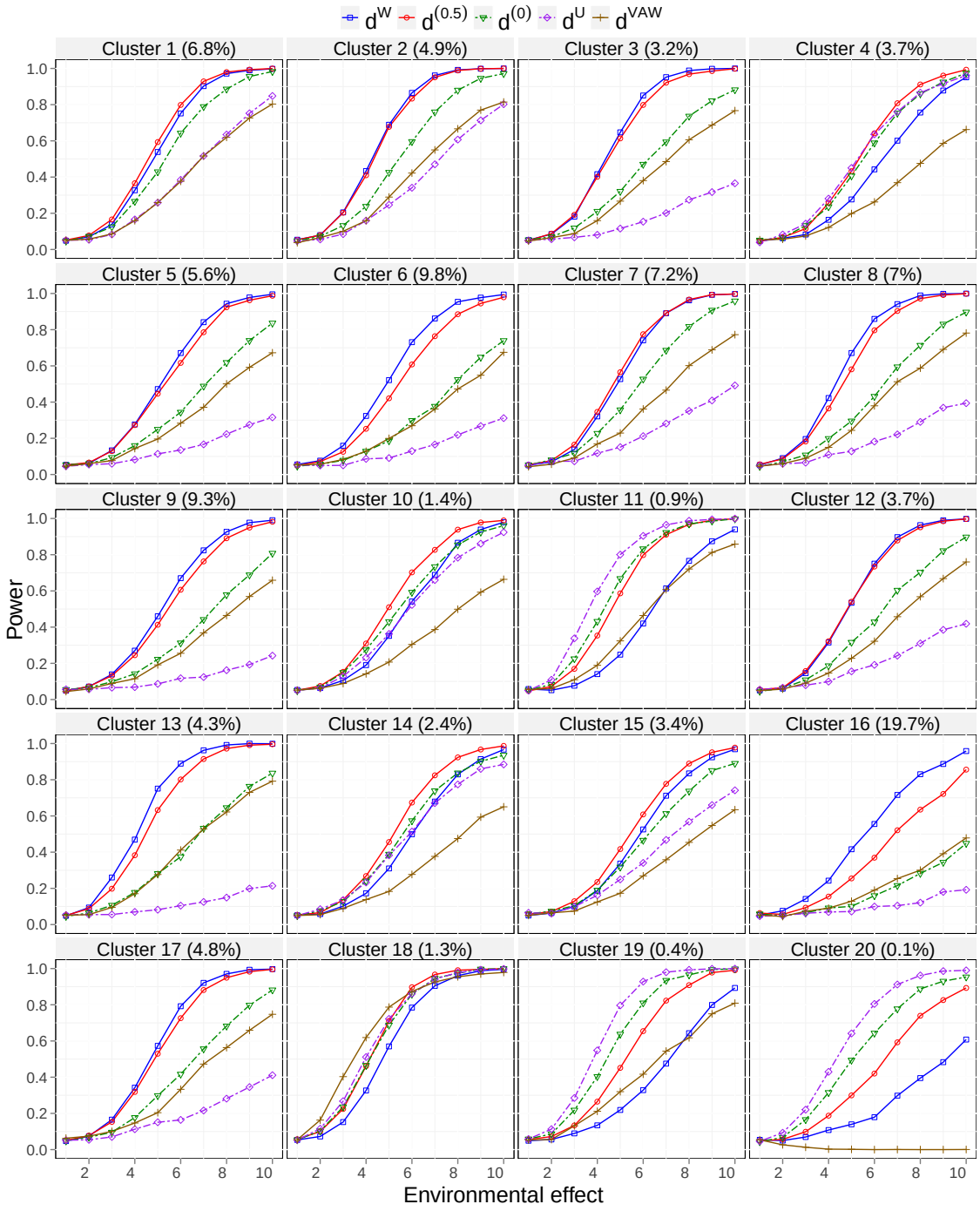

Figure A3: **Power comparison of different UniFrac variants for detecting environmental effect using tree based simulation (all lineages)**. Ten samples from each of the two environmental conditions are generated using tree based simulation. UniFrac distance matrices are constructed based on the simulated OTU abundances and the phylogenetic tree. PERMANOVA is used for testing hypotheses. The figure shows all the 20 lineages that are affected by the environment. The lineage abundance is given in parentheses in the panel title.

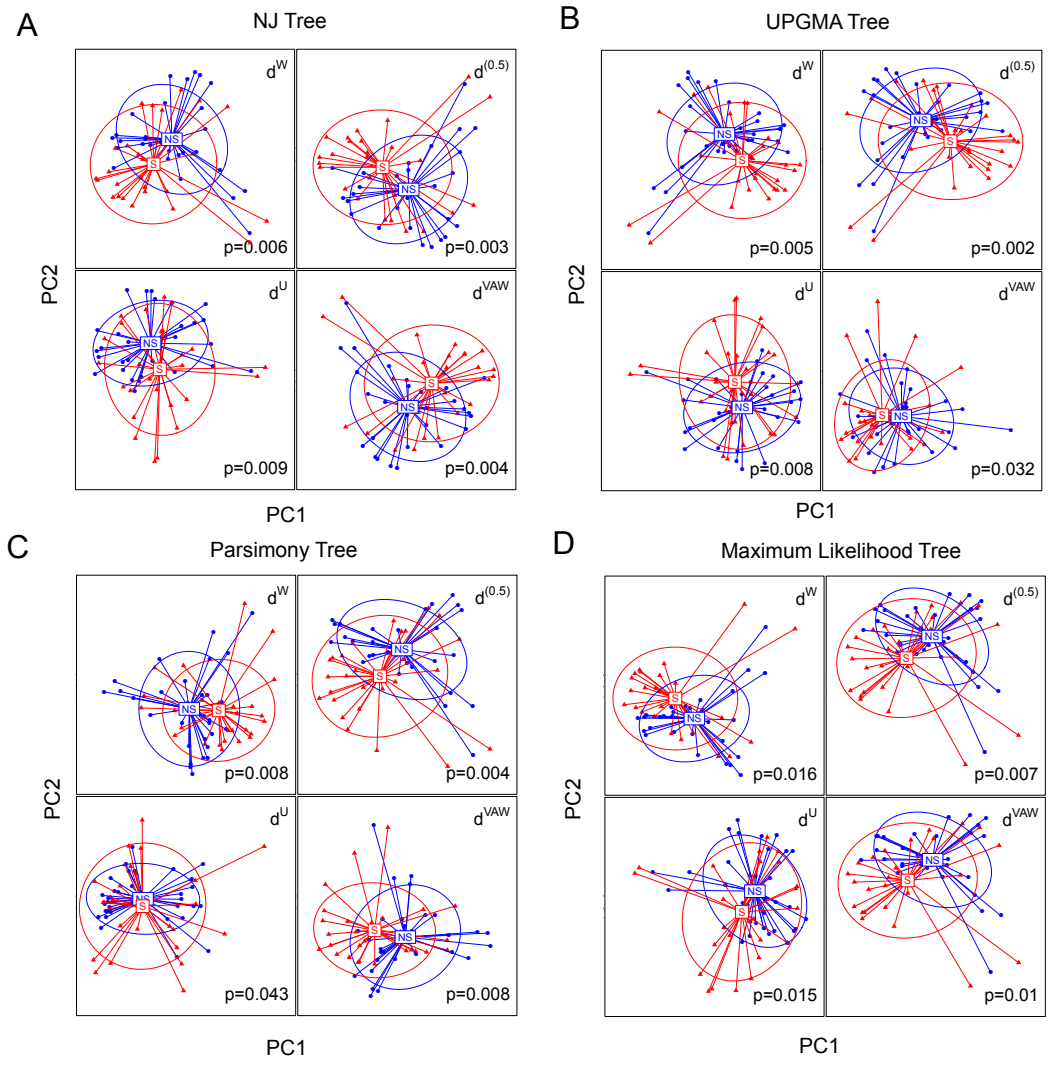

Figure A4: **Comparison of UniFrac variants for separating the oropharyngeal microbial communities of smokers from nonsmokers using various tree construction methods**. Various UniFrac distance matrices are constructed based on the OTU abundances and the phylogenetic tree constructed by NJ (A), UPGMA (B), Parsimony (C) or Maximum likelihood method (D). Samples from smokers (28) and nonsmokers (32) are indicated by "S" and "NS" respectively. Four representative UniFrac variants d^W, $d^{(0.5)}$, d^{VAW} and d^U are compared. Principle coordinate analysis is performed to embed the samples into 2D plane using the first two principle coordinates.The ellipse center indicates groups means, its main axis corresponds to the first two principle components from principle component analysis, and the height and width are variances on that direction. The p values from PERMANOVA for testing difference are also indicated as a measure of separation.

BIBLIOGRAPHY

Aitchison, J. (1982). The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B*, **44**(2), 139–177.

Andersson, A. *et al.* (2008). Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS One*, **3**(7), e2836.

Angiuoli, S. *et al.* (2011). Clovr: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC bioinformatics*, **12**(1), 356.

Arumugam, M. *et al.* (2011). Enterotypes of the human gut microbiome. *Nature*, **473**(7346), 174–180.

Bach, F. (2008). Bolasso: Model consistent lasso estimation through the bootstrap. In *ICML '08: Proceedings of the 25th international conference on machine learning*, New York, NY, USA.

Barrett, J. *et al.* (2008). Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nat. genet.*, **40**(8), 955–962.

Barry, S. and Welsh, A. (2002). Generalized additive modelling and zero inflated count data. *Ecol. Model.*, **157**(2/3), 179–188.

Benson, A. *et al.* (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl. Acad. Sci. U. S. A.*, **107**(44), 18933–18938.

Biagi, E. *et al.* (2010). Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS One*, **5**(5), e10667.

Caporaso, J. *et al.* (2010a). Pynast: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, **26**(2), 266–267.

Caporaso, J. *et al.* (2010b). Qiime allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**(5), 335–336.

Castellarin, M., Warren, R., Freeman, J., Dreolini, L., Krzywinski, M., Strauss, J., Barnes, R., Watson, P., Allen-Vercoe, E., Moore, R., *et al.* (2012). Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res.*, **22**(2), 299–306.

Chang, Q. *et al.* (2011). Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*, **12**, 118.

Charlson, E. *et al.* (2010). Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. *PLoS One*, **5**(12), e15216.

Charlson, E. *et al.* (2011). Topographical continuity of bacterial populations in the healthy human respiratory tract. *Am. J. Respir. Crit. Care. Med.*, **184**(8), 957–963.

Cho, I. and Blaser, M. (2012). The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.*, **13**(4), 260–270.

Chung, F. (1997). *Spectral graph theory*. American Mathematical Society.

Clemente, J. *et al.* (2012). The impact of the gut microbiota on human health: An integrative view. *Cell*, **148**(6), 1258–1270.

Cole, J. *et al.* (2009). The ribosomal database project: improved alignments and new tools for rrna analysis. *Nucleic Acids Res.*, **37**, D141–D145.

Collison, M. *et al.* (2012). Data mining the human gut microbiota for therapeutic targets (in press). *Briefing Bioinform.*

Costello, E. *et al.* (2009). Bacterial community variation in human body habitats across space and time. *Science*, **326**(5960), 1694–1697.

Dauxois, J. and Nkiet, G. (1997). Canonical analysis of two euclidean subspaces and its applications. *Linear algebra appl.*, **264**, 355–388.

De Filippo, C. *et al.* (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from europe and rural africa. *Proc. Natl. Acad. Sci. U. S. A.*, **107**(33), 14691–14696.

DeSantis, T. *et al.* (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Appl. Environ. Microbiol.*, **72**(7), 5069–5072.

Dinsdale, E. *et al.* (2008). Functional metagenomic profiling of nine biomes. *Nature*, **452**(7187), 629–632.

Dominguez-Bello, M. *et al.* (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U. S. A.*, **107**(26), 11971.

Dudoit, S. *et al.* (2001). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Edgar, R. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, **26**(19), 2460–2461.

Efron, B. *et al.* (2001). Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**(456), 1151–1160.

Ehrlich, S. (2011). Metahit: The european union project on metagenomics of the human intestinal tract. *Metagenomics of the Human Body*, pages 307–316.

Evans, S. and Matsen, F. (2012). The phylogenetic kantorovich–rubinstein metric for environmental sequence samples(in press). *J. Roy. Statist. Soc. Ser. B*.

Felsenstein, J. (2003). *Inferring Phylogenies*. Sinauer Associates.

Friedman, J. *et al.* (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.*, **1**(2), 302–332.

Friedman, J. *et al.* (2010). A note on the group lasso and a sparse group lasso. *Arxiv preprint arXiv10010736*, pages 1–8.

Fukuyama, J. *et al.* (2012). Comparisons of distance methods for combining covariates and abundances in microbiome studies. *Pac. Symp. Biocomput.*, pages 213–224.

Gill, S. *et al.* (2006). Metagenomic analysis of the human distal gut microbiome. *Science*, **312**(5778), 1355–1359.

Goll, J. *et al.* (2010). Metarep: Jcvi metagenomics reportsan open source tool for high-performance comparative metagenomics. *Bioinformatics*, **26**(20), 2631–2632.

Grice, E. *et al.* (2009). Topographical and temporal diversity of the human skin microbiome. *Science*, **324**(5931), 1190–1192.

Haas, B. a. (2011). Chimeric 16s rrna sequence formation and detection in sanger and 454-pyrosequenced pcr amplicons. *Genome res.*, **21**(3), 494–504.

Hamady, M. *et al.* (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. methods*, **5**(3), 235–237.

Hildebrandt, M. *et al.* (2009). High-fat diet determines the composition of the murine gut microbiome independently of obesity. *Gastroenterology*, **137**(5), 1716–1724.

Holmes, E. *et al.* (2011). Understanding the role of gut microbiome-host metabolic signal disruption in health and disease. *Trends microbiol.*, **19**(7), 349–59.

Holmes, I. *et al.* (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS one*, **7**(2), e30126.

Huson, D., Auch, A., Qi, J., and Schuster, S. (2007). Megan analysis of metagenomic data. *Genome res.*, **17**(3), 377–386.

Iverson, V. *et al.* (2012). Untangling genomes from metagenomes: Revealing an uncultured class of marine euryarchaeota. *Science*, **335**(6068), 587–590.

James, G. *et al.* (2010). Sparse regulatory networks. *Ann. Appl. Stat.*, **4**(2), 663–686.

Kinross, J. *et al.* (2011). Gut microbiome-host interactions in health and disease. *Genome Med.*, **3**(3), 14.

Knights, D. *et al.* (2011). Human-associated microbial signatures: Examining their predictive value. *Cell Host Microbe*, **10**(4), 292–296.

Kostic, A. *et al.* (2012). Genomic analysis identifies association of fusobacterium with colorectal carcinoma. *Genome Res.*, **22**(2), 292–298.

Kuczynski, J. *et al.* (2010a). Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol.*, **11**, 210–218.

Kuczynski, J. *et al.* (2010b). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. methods*, **7**(10), 813–819.

Kuczynski, J., Lauber, C., Walters, W., Parfrey, L., Clemente, J., Gevers, D., and Knight, R. (2011). Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.*, **13**(1), 47–58.

Lee, A. *et al.* (2006). Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Stat. Methods Med. Res.*, **15**(1), 47–61.

Ley, R. *et al.* (2005). Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. U. S. A.*, **102**(31), 11070.

Ley, R. *et al.* (2006). Microbial ecology: human gut microbes associated with obesity. *Nature*, **444**(7122), 1022–1023.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**(9), 1175–1182.

Li, C. and Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl. Stat.*, **4**(3), 1498–1516.

Littman, D. and Honda, K. (2012). The microbiome in infectious disease and inflammation. *Annu. Rev. Immunol.*, **30**(1).

Liu, D. *et al.* (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, **63**(4), 1079–1088.

Liu, D. *et al.* (2008). 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.

Lozupone, C. *et al.* (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, **73**(5), 1576–1585.

Lozupone, C. *et al.* (2010). Unifrac: an effective distance metric for microbial community comparison. *ISME J.*, **5**(2), 169–172.

Lozupone, C. and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**(12), 8228–8835.

Manichanh, C. *et al.* (2006). Reduced diversity of faecal microbiota in crohn's disease revealed by a metagenomic approach. *Gut*, **55**(2), 205–211.

Markowitz, V. *et al.* (2008). Img/m: a data management and analysis system for metagenomes. *Nucleic acids res.*, **36**(suppl 1), D534–D538.

Matsen, F. *et al.* (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, **11**(1), 538.

McArdle, B. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, **82**(1), 290–297.

Meier, L. *et al.* (2008). The group lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B*, **70**(1), 53–71.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Stat.*, **34**, 1436–1462.

Meyer, F. *et al.* (2008). The metagenomics rast server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, **9**(1), 386.

Minot, S. *et al.* (2011). The human gut virome: Inter-individual variation and dynamic response to diet. *Genome res.*, **21**(10), 1616–1625.

Moghimbeigi, A. *et al.* (2008). Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *J. of Appl. Stat.*, **35**(10), 1193–1202.

Mosimann, J. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, **49**(1/2), 65–82.

Muegge, B. *et al.* (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, **332**(6032), 970–974.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn Res.*, **8**, 1145–1164.

Parkhomenko, E. *et al.* (2009). Sparse canonical correlation analysis with application to genomic data integration. *Stat. Appl. Genet. Mol. Biol.*, **8**(1), 1.

Parks, D. and Beiko, R. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, **26**(6), 715–721.

Peng, J. *et al.* (2009). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.*, **4**(1), 53–77.

Peterson, J. *et al.* (2009). The nih human microbiome project. *Genome res.*, **19**(12), 2317–2323.

Pflughoeft, K. and Versalovic, J. (2011). Human microbiome in health and disease. *Annu. Rev. Pathol.*, **7**, 99–122.

Plottel, C. and Blaser, M. (2011). Microbiome and malignancy. *Cell Host Microbe*, **10**(4), 324–335.

Price, M. *et al.* (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.*, **26**(7), 1641–1650.

Pruesse, E. *et al.* (2007). Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic acids res.*, **35**(21), 7188–7196.

Purdom, E. (2011). Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree. *Ann. Appl. Stat.*, **5**(4), 2326–2358.

Qin, J. *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**(7285), 59–65.

Quince, C. *et al.* (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. methods*, **6**(9), 639–641.

Rodriguez-Brito, B. *et al.* (2006). An application of statistics to comparative metagenomics. *BMC bioinformatics*, **7**(1), 162.

Schloss, P. (2008). Evaluating different approaches that test whether microbial communities have the same structure. *ISME J.*, **2**(3), 265–275.

Schloss, P. *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**(23), 7537.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome biol.*, **12**(6), R60.

Seshadri, R. *et al.* (2007). Camera: a community resource for metagenomics. *PLoS biol.*, **5**(3), e75.

Sokol, H. *et al.* (2008). Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci. U. S. A.*, **105**(43), 16731–16736.

Spencer, M. *et al.* (2011). Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. *Gastroenterology*, **140**(3), 976–986.

Spor, A. *et al.* (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.*, **9**(4), 279–290.

Sun, Y. *et al.* (2012). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief. Bioinform.*, **13**(1), 107–121.

Swenson, N. (2011). Phylogenetic beta diversity metrics, trait evolution and inferring the functional beta diversity of communities. *PloS One*, **6**(6), e21264.

Sze, M. *et al.* (2012). The lung tissue microbiome in chronic obstructive pulmonary disease (in press). *Am. J. Respir. Crit. Care. Med.*

Tibshirani, R. *et al.* (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Stat. Sci.*, **18**(1), 104–117.

Tringe, S. *et al.* (2005). Comparative metagenomics of microbial communities. *Science*, **308**(5721), 554–557.

Tseng, P. and Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program.*, **117**(1-2), 387–423.

Turnbaugh, P. *et al.* (2009). A core gut microbiome in obese and lean twins. *Nature*, **457**(7228), 480–484.

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

Virgin, H. and Todd, J. (2011). Metagenomics and Personalized Medicine. *Cell*, **147**(1), 44–56.

von Mering, C. *et al.* (2007). Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, **315**(5815), 1126–1130.

Waaijenborg, S. *et al.* (2008). Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Stat. Appl. Genet. Mol. Biol.*, **7**(1), 3.

Wagner, B. *et al.* (2011). Application of two-part statistics for comparison of sequence variant counts. *PloS one*, **6**(5), e20296.

Wang, Q. *et al.* (2007). Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**(16), 5261.

White, J. *et al.* (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS computat. biol.*, **5**(4), e1000352.

Whitman, W. *et al.* (1998). Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. U. S. A.*, **95**(12), 6578.

Witten, D. and Tibshirani, R. (2010). A framework for feature selection in clustering. *J. Am. Stat. Assoc.*, **105**(490), 713–726.

Witten, D. M. *et al.* (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**(3), 515–534.

Wooley, J. and Ye, Y. (2010). Metagenomics: facts and artifacts, and computational challenges. *J. Comput. Sci. Technol.*, **25**(1), 71–81.

Wu, G. *et al.* (2010). Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiol.*, **10**(1), 206.

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

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

Zhang, H. *et al.* (2008). Variable selection for multicategory svm via sup-norm regularization. *Electron. J. Stat.*, **2**, 149–167.

Zhao, P. *et al.* (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Stat.*, **37**(6A), 3468–3497.