

COMPUTATIONAL MICROBIOME ANALYSIS: METHODS AND  
APPLICATIONS

(Spine title: Computational microbiome analysis: methods and applications)  
(Thesis format: Integrated Article)

by

Ruth Wong

Graduate Program in Biochemistry

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

The School of Graduate and Postdoctoral Studies  
The University of Western Ontario  
London, Ontario, Canada

© Ruth Grace Wong 2016

THE UNIVERSITY OF WESTERN ONTARIO  
School of Graduate and Postdoctoral Studies

**CERTIFICATE OF EXAMINATION**

Supervisor:

.....  
Dr. Gregory B. Gloor

Supervisory Committee:

.....  
Dr. Lindi M. Wahl

.....  
Dr. David R. Edgell

Examiners:

.....  
Dr. ExaminerA

.....  
Dr. ExaminerB

.....  
Dr. ExaminerC

The thesis by

**Ruth Grace Wong**

entitled:

**Computational microbiome analysis: methods and applications**

is accepted in partial fulfillment of the  
requirements for the degree of  
Masters of Science

.....  
Date

.....  
Chair of the Thesis Examination Board

# Abstract

With the advent of next generation sequencing, scientists can obtain a more comprehensive snapshot of the bacterial composition of the microbiome, what genes they have, and what proteins they produce. We are in a phase of developing the experiments and accompanying statistical techniques to elucidate the exact mechanisms by which the human microbiome affects health and disease. In this thesis we explore alternatives to the standard weighted and unweighted UniFrac metric for measuring the difference between microbiome samples, to elucidate different trends and outliers. We also apply next generation sequencing and computational analysis techniques to gut microbiome data to examine relationship of the microbiota to atherosclerosis and non alcoholic fatty liver disease.

**Keywords:** Human microbiome, next generation sequencing, bioinformatics, atherosclerosis, non alcoholic fatty liver disease

# Contents

|  |            |
|--|------------|
| <b>Certificate of Examination</b>  | <b>ii</b>  |
| <b>Abstract</b>  | <b>iii</b> |
| <b>List of Figures</b>   | <b>vi</b>  |
| <b>List of Tables</b>  | <b>x</b>   |
| <b>List of Appendices</b>  | <b>xi</b>  |
| <b>1 Introduction</b>  | <b>1</b>   |
| 1.1 The human microbiome . . . . .   | 1          |
| 1.2 Exploring the human microbiome . . . . .   | 2          |
| 1.3 Illumina next generation sequencing platform . . . . .   | 2          |
| 1.4 Gene tag abundance . . . . .   | 3          |
| 1.4.1 16S rRNA gene sequencing experiment . . . . .  | 3          |
| 1.4.2 Operational Taxonomic Units . . . . .  | 4          |
| 1.4.3 General protocol and rationale . . . . .   | 4          |
| 1.4.4 Data analysis . . . . .  | 7          |
| 1.5 The metagenomic experiment . . . . .   | 9          |
| 1.5.1 Sequencing . . . . .   | 10         |
| 1.5.2 Imputation . . . . .   | 13         |
| 1.5.3 Data analysis . . . . .  | 13         |
| 1.6 Points of failure . . . . .  | 14         |
| 1.6.1 Collection methods differ . . . . .  | 14         |
| 1.6.2 Microbiome data is highly variable between individuals . . . . .                                       | 14         |
| 1.6.3 Microbiome data involves the comparison of many features . . . . .                                     | 15         |
| 1.6.4 Microbiome data is compositional . . . . .   | 15         |
| 1.6.5 Microbiome data is sparse . . . . .  | 16         |
| 1.7 The gut microbiome in atherosclerosis-susceptible and atherosclerosis-resistant patients . . . . .       | 17         |
| 1.8 The gut microbiome in patients with non-alcoholic steatohepatitis compared to healthy controls . . . . . | 17         |
| <b>2 Expanding the UniFrac toolbox</b>   | <b>19</b>  |
| 2.0.1 Data . . . . .   | 20         |

|          |  |           |
|----------|--|-----------|
| 2.0.2    | Unweighted UniFrac . . . . .   | 21        |
| 2.0.3    | Weighted UniFrac . . . . .   | 22        |
| 2.0.4    | Analytical techniques . . . . .  | 23        |
| 2.0.5    | Data preparation . . . . .   | 25        |
| 2.0.6    | Unweighted Unifrac is highly sensitive to rarefaction variants . . . . .               | 27        |
| 2.0.7    | Why does Unweighted Unifrac have discrepancies when analyzing rarefied data? . . . . . | 28        |
| 2.0.8    | Information UniFrac . . . . .  | 29        |
| 2.0.9    | Tongue and buccal mucosa comparison . . . . .  | 29        |
| 2.0.10   | Breast milk Data . . . . .   | 30        |
| <b>3</b> | <b>The human microbiome and atherosclerosis</b>  | <b>37</b> |
| 3.1      | Introduction . . . . .   | 37        |
| 3.1.1    | Atherosclerosis risk . . . . .   | 37        |
| 3.1.2    | Metabolic potential of gut microbiota . . . . .  | 42        |
| 3.1.3    | Metabolic potential of mouth microbiota . . . . .                                      | 43        |
| 3.1.4    | Bacteria and atherosclerotic plaques . . . . .   | 44        |
| 3.1.5    | Project proposal . . . . .   | 44        |
| 3.2      | Methods . . . . .  | 44        |
| 3.3      | Results . . . . .  | 44        |
| 3.4      | Discussion . . . . .   | 44        |
| <b>4</b> | <b>The human microbiome and non-alcoholic fatty liver disease</b>                      | <b>45</b> |
| 4.1      | Introduction . . . . .   | 45        |
| 4.2      | Methods . . . . .  | 47        |
| 4.2.1    | 16S rRNA gene tag experiment . . . . .   | 47        |
| 4.2.2    | Metagenomic experiment . . . . .   | 47        |
| 4.2.3    | MetaPhlAn . . . . .  | 48        |
| 4.3      | Results . . . . .  | 48        |
| 4.3.1    | 16S rRNA gene tag experiment . . . . .   | 48        |
| 4.3.2    | Metagenomic experiment . . . . .   | 48        |
| 4.4      | Discussion . . . . .   | 48        |
| <b>5</b> | <b>Theorems</b>  | <b>56</b> |
| 5.1      | Basic Theorems . . . . .   | 56        |
|          | <b>Bibliography</b>  | <b>57</b> |
|          | <b>A Proofs of Theorems</b>  | <b>65</b> |
|          | <b>Curriculum Vitae</b>  | <b>66</b> |

# List of Figures

|     |   |    |
|-----|---|----|
| 1.1 | <b>16S rRNA gene tag experiment workflow.</b> This shows the workflow from sample collection to data generation. The end result is a count table of reads per operational taxonomic unit per sample. . . . .  | 6  |
| 1.2 | <b>Unweighted UniFrac.</b> When two samples do not share any branches of the phylogenetic tree, the unweighted UniFrac distance is maximized at 1. When two samples share half of their branch lengths on the phylogenetic tree, the unweighted UniFrac distance is 0.5. If the two samples contain exactly the same taxa, the unweighted UniFrac distance is minimized at 0, since the samples share all branches. . . . .   | 8  |
| 1.3 | <b>Metagenomic experiment workflow.</b> This shows the workflow from sample collection to data generation. The end result is a table of number of sequencing reads per functionally annotated gene per sample. . . . .  | 12 |
| 1.4 | <b>Example of stripcharts for subsystem 2 and 3 functional categorizations.</b> Dots on the left side are subsystem 4 annotations found to be more abundant in the healthy condition while dots on the right side are subsystem 4 annotations found to be more abundant in the bacterial vaginosis condition. Colored dots were found to be significantly differentially abundant. Figure taken from [63]. .  | 13 |
| 2.1 | <b>Unweighted UniFrac.</b> When two samples do not share any branches of the phylogenetic tree, the unweighted UniFrac distance is maximized at 1. When two samples share half of their branch lengths on the phylogenetic tree, the unweighted UniFrac distance is 0.5. If the two samples contain exactly the same taxa, the unweighted UniFrac distance is minimized at 0, since the samples share all branches. . . . .   | 22 |
| 2.2 | <b>Sample migration in different rarefactions, plotted on principal coordinates, measured with unweighted UniFrac.</b> Red samples have moved from the left cluster to the right cluster between rarefactions. Blue samples have moved from the right cluster to the left. Samples are taken from the tongue dorsum body site from the Human Microbiome Project database. If the experiment were run once, one might mistakenly assume that there are two clusters of data, however, the inconsistent sample membership of the two groups between rarefactions proves the clustering irreproducible. Note that the variance explained by the first and second coordinate is merely 16.1% and 9.8% respectively, indicating that the data is rather spherical, even though the points on the plot appear to show two separated clusters. . . . . | 32 |

|     |  |    |
|-----|--|----|
| 2.3 | <b>Maximum relative deviation of rarefactions versus median deviation for traditional and non-traditional microbiome dissimilarity metrics.</b> Sixty samples from the tongue dorsum were taken from the Human Microbiome Project [110], and rarefied 100 times. The maximum relative deviation was plotted against the median relative deviation of the rarefied data, and ellipses were drawn at the 95% confidence interval, around the cloud of points for each metric. A higher maximum and median deviation indicates lower reproducibility of results between rarefaction instances. Both the maximum relative deviation of rarefied data and the median relative deviation of rarefied data are greater in unweighted UniFrac than in weighted UniFrac, Bray Curtis distance, centered ratio UniFrac, and information UniFrac. . . . . | 33 |
| 2.4 | <b>Phylogenetic tree with long isolated branches.</b> Variation in different rarefactions of data in unweighted UniFrac analysis is exacerbated by the presence of long isolated branches in the phylogenetic tree. . . . .  | 34 |
| 2.5 | <b>Unifrac weights.</b> Each UniFrac weighting is plotted with the corresponding proportional abundance. The black line is unweighted UniFrac, the red line is weighted UniFrac, and the blue line is information UniFrac. . . . .   | 34 |
| 2.6 | <b>Analysis of tongue and buccal mucosa data using different UniFrac weightings.</b> A principal coordinate analysis of a 16S rRNA experiment done on samples from the tongue and buccal mucosa, selected from the Human Microbiome Project [110]. All weightings show separation between the samples by body site. Note that the variance explained by the first and second principal coordinate axis is higher than in the tongue-tongue data set from Figure 2, which had 16.1% and 9.8% variance explained, respectively. . . . .  | 35 |
| 2.7 | <b>Analysis of breast milk data using different UniFrac weightings.</b> A principal coordinate analysis of a 16S rRNA experiment done on samples from a 16S rRNA experiment on breast milk. The circled sample is infected with 97% <i>Pseudomonas</i> , compared to 15-20% in the other samples. . . . .  | 36 |
| 3.1 | <b>Carotid Ultrasound showing intima-media thickness,</b> picture borrowed from Harley Street Cardiologists 2014 London Cardiovascular Clinic. The intima is the innermost layer of the artery bordering the lumen, and the media is the layer just outside that. . . . .  | 38 |
| 3.2 | <b>Carotid Ultrasound showing plaque area measurement,</b> picture borrowed from Spence [101] . . . . .  | 39 |
| 3.3 | <b>Risk factors, predicted carotid plaque area, and actual plaque area for two patients.</b> Two patients' predicted carotid plaque area based on their risk factors, and their actual plaque area. One patient, a young non-smoker, has a plaque much larger than predicted, and the other patient, an old smoker with a high LDL/HDL ratio has a plaque much smaller than predicted. Both patients are in the top tenth percentile for unexplained atherosclerosis progression or regression. Figure borrowed from Spence [100]. . . . .   | 40 |
| 3.4 | <b>Predicted risk vs. total plaque area.</b> The two arrows represent the residual measures of the two patients in the previous figure. Figure borrowed from Spence [100]. . . . .   | 41 |

|     |   |    |
|-----|---|----|
| 3.5 | <b>Kaplan-Meier estimates of major adverse cardiovascular events, according to the quartile of TMAO level.</b> “Data are shown for 4007 participants in the clinical-outcomes study. The P-value is for all comparisons” Figure borrowed from Tang [103]. Each line of this graph represents data from patients in one of four quartiles for TMAO levels. The risk of myocardial infarction, stroke, or death is more than double for patients who are in the top 25% for TMAO levels, compared to patients in the bottom quartile. . . . .   | 43 |
| 4.1 | <b>Venn diagram of genus found to be differentially abundant by different studies between NASH/NAFLD and healthy controls.</b> Boursier et al 2015 is not included as they reported a p-value of less than 0.05 for the Bacteroides genus only, which was not reported in any of the other studies. Only 3 out of the 16 genus claimed to be differentially abundant were the same in two studies: Escherichia was found in the Zhu [120] and Jiang [46] studies, and Lactobacillus and Oscillibacter were found in the Jiang [46] and Raman [84] studies. . . . .  | 46 |
| 4.2 | <b>Principal Components Analysis of 16S rRNA gene tag sequencing data with different UniFrac weightings.</b> Each point represents one sample, and the distances between the samples have been calculated using different UniFrac metrics, taking into account phylogenetic as well as abundance information. There is no obvious separation between groups by any of the UniFrac weightings. Furthermore the variance explained by each principal component axis is not notably high, indicating a rather uniform data set. . . . .  | 50 |
| 4.3 | <b>Difference within vs. difference between groups.</b> Each point represents one OTU, and the differential abundance of that OTU within groups is plotted against the differential abundance between groups. None of the OTUs are more different between groups than within groups. The healthy samples used for these comparisons are the 10 healthy samples used for the metagenomic study. The extreme NASH samples used for these comparisons are the subset of the NASH patients selected for the metagenomic study. . . . .  | 51 |
| 4.4 | <b>Correlation in effect sizes of different group experiments.</b> Each point represents one OTU, and the effect size of that OTU in one comparison (for example, comparing the gut microbiome of healthy patients with patients who have simple steatosis) is plotted against the effect size of that OTU in another comparison. The healthy samples used for these comparisons are the 10 healthy samples used for the metagenomic study. The extreme NASH samples used for these comparisons are the subset of the NASH patients selected for the metagenomic study. The y intercepts of the regression lines are all between 0.005 and 0.025, close to zero. The median difference in the absolute effect sizes is -0.02076 for Healthy vs. NASH - Healthy vs. SS, 0.017070 for Healthy vs. extreme NASH - Healthy vs. SS, and 0.04256 for Healthy vs. extreme NASH - Healthy vs. NASH. . . . . | 52 |
| 4.5 | <b>Taxa barplot dendrogram derived from MetaPhlAn.</b> The metagenomic reads were input into MetaPhlAn to generate a count table. The taxa in the count table were filtered such that only taxa with at least 1% abundance in any sample was kept. . . . .  | 53 |



|     |   |    |
|-----|---|----|
| 4.6 | <b>Biplot derived from MetaPhlAn.</b> This biplot was generated from the count table inferred by MetaPhlAn, with taxa filtered such that only taxa with at least 1% abundance in any sample was kept. Note that the variance explained by the first and the second coordinate is not particularly high, indicating that there is not a clear unidirectional separation between groups. Samples from healthy controls are colored black while samples from patients with NASH are colored red. . . . . | 54 |
| 4.7 | <b>Difference within groups vs. difference between groups per taxa, derived from MetaPhlAn.</b> This plot was generated from the count table inferred by MetaPhlAn, with taxa filtered such that only taxa with at least 1% abundance in any sample was kept. No taxa are more differential between groups than within groups. . . . .  | 55 |

# List of Tables

|     |  |    |
|-----|--|----|
| 2.1 | <b>Original abundance of taxa and rarefied abundance of taxa.</b> This data was simulated to demonstrate how rarefaction can change the distances reported by the unweighted UniFrac metric. . . . . | 28 |
|-----|--|----|

# List of Appendices

65Appendix.a.A

# Chapter 1

## Introduction

This thesis focuses on the human microbiome, its relation to human diseases, and techniques used in the data analysis and exploration of it. During the course of my thesis, I conducted one study about non-alcoholic fatty liver disease, one study about atherosclerosis, and written a conference paper about alternate weightings of a common microbiome analysis technique (UniFrac). Each of these topics is represented as a chapter of my thesis.

### 1.1 The human microbiome

Approximately half of the cells that make up the human body are bacterial [93]. Trillions of these bacteria live in the gut [41], and have a massive metabolic potential. For example, the gut microbiome has been shown to produce changes in hormone levels [65], short chain fatty acid levels [108], and ethanol levels [53], to name a few. The human gut microbiome can even digest polysaccharides otherwise unusable by humans [32].

This massive metabolic potential produces measurable symptomatic effects. Transplanting gut bacteria from obese mice to lean mice have been shown to convert lean mice to absorb more calories from the same food [107]. The microbiome can also affect behavior: Completely germ free mice exhibit more anxiety-like behaviors than specific pathogen free mice [72].

The human microbiome opens up a host of possibilities for reducing the effects of disease and improving quality of life. However, until recently, a deep understanding of the human microbiome has been beyond the reach of available technology. For example, *Escherichia coli* is a common model gut bacteria because it is easy to culture, however in reality only makes up about 1% of the average human gut microbiome [4].

With the advent of next generation sequencing, scientists can obtain a more comprehensive snapshot of the bacterial composition of the microbiome, what genes they have, and what proteins they produce [24]. We are in a phase of developing the experiments and accompanying statistical techniques to elucidate the exact mechanisms by which the human microbiome affects health and disease. Armed with a deeper understanding of how the microbiome works, we may be able to develop probiotic techniques to improve quality of life.

## 1.2 Exploring the human microbiome

The advent of next generation sequencing has prompted the development of a number of different experiments that can be run on biological samples of the human microbiome. Samples can be collected by swabbing the target body site or collecting excretions such as saliva or stool. Products such as DNA or RNA may be extracted as appropriate for the analysis.

Usually a study involves an experimental group and a control group. These can be patients with disease and healthy controls [63], people who are susceptible and resistant to a condition [104], or patients before and after a medical intervention [40]. The questions that scientists in this field generally want to answer are: Is the human microbiome driving or associated with the difference between the two groups? If so, what is the mechanism of action? There are also exploratory studies which try to determine what the core microbiome for a body site in a single condition is by examining what people who fit the condition have in common.

The questions that the data can answer directly are: Is there a statistically significant difference in the microbiome between the control and the experimental groups, in terms of the types of microbes present or the microbial genes present? Do separated groups exist in the data? Are the proportional abundances of certain taxa or genes correlated with each other, or with patient metadata? These questions can be answered by metagenomic experiments and statistical analysis, leading to clues about the larger questions of the mechanism of action.

The two metagenomic experiments that can be done with microbiome next generation sequencing data used in this thesis are gene tag abundance and deep metagenomic sequencing [85]. The tag used for gene tag abundance here is the 16S rRNA gene [37]. The process and resulting data of each experiment is described in the next section, followed by a piece about data analysis and points of failure.

## 1.3 Illumina next generation sequencing platform

Illumina is a next generation sequencing platform. The Illumina MiSeq machines yields up to 25 million reads of paired end 300 nucleotide sequences, and the Illumina HiSeq machines yield up to 4 billion reads of paired end 125 nucleotide sequences, as stated on the official Illumina website (<http://www.illumina.com/systems.html>). The sequencing works as follows:

1. DNA is amplified or fragmented to smaller pieces
2. Adaptors are ligated to the ends of the DNA
3. The DNA is denatured into single strands
4. The DNA washed on a flow cell covered in primers, such that complementary DNA sticks
5. The DNA on the flow cell is replicated to form clusters of identical sequences
6. The DNA is made single stranded again
7. Primers, nucleotides, DNA polymerase, and fluorescently labelled nucleotide terminators are added

8. A camera can detect the fluorescently labelled nucleotide terminators for each added base on each cluster of identical sequences, allowing the DNA to be sequenced.

The Illumina technology has been used for years [8], and standard protocols exist for library preparation, with kits available commercially.

## 1.4 Gene tag abundance

Gene tag abundance experiments provide an estimate of the proportion of different types of bacteria in the sample. This can be used to answer questions such as:

*What bacterial taxa make up the microbial community?* Scientists often want to characterize microbiomes for certain conditions. For example, the core gut microbiome was described by one group to have three enterotypes [4], however, when another group studied a diverse population including non-Western people, the enterotypes did not hold [119]. The vaginal microbiome is known to be *Lactobacillus* dominated, except in bacterial vaginosis, where the microbiome is much more diverse [44]. The idea is that characterizing the core microbiome can lead to insight on core functions and how they can be altered when the core microbiome is disrupted.

*Are there any differentially abundant taxa between conditions?* Some theories of disease progression include the involvement of bacteria as pathogens. Others involve bacteria as probiotics, preventing disease progression. Salient examples include atopic dermatitis where flare-ups are associated with an increase in the proportion of *Staphylococcus aureus* on the skin [51], and RePOOPulate, a probiotic therapy where 33 microbes cultured from a healthy donor were used to successfully treat symptoms of *C. difficile* [80].

Historically, Koch's postulates have been used to determine if a microbe is a disease-causing pathogen: First, the microbe must be present in all cases of the disease. Second, the microbe must not be present and non-pathogenic in other diseases. Third, if the microbe is isolated in pure culture, it can be used to induce the disease [49]. One group has created a modified set of postulates that takes DNA sequencing into account [33] which can be applied to differentially abundant taxa detected by gene tag sequencing. However, Koch's postulates do not account for when the same bacteria can have a very different expression profile in health and disease, such as *Lactobacillus iners* in bacterial vaginosis [63].

*Do samples from different conditions cluster together?* Sometimes when the data is plotted, there appears to be separation between groups, even if specific taxa are not differentially abundant. One example of this is a study on discordant gut microbiomes between twins in Malawi where one twin has kwashiorkor and the other is healthy [96]. In this case the microbiomes diverge the most during treatment with ready-to-use therapeutic food.

### 1.4.1 16S rRNA gene sequencing experiment

The gene tag chosen throughout this thesis is the gene for the 16S subunit of ribosomal RNA. The 16S rRNA gene is present in all known bacteria and has regions of variability interspersed with regions of high conservation. This allows primers to be made to match the conserved regions, such that the variable regions can be amplified, sequenced, and used to infer taxa.

Entire databases exist specifically to match the 16S rRNA gene with taxonomy, such as SILVA [83], the Ribosomal Database Project [19], and Greengenes [23].

Specifically, we have been using the 16S rRNA primers from the Earth Microbiome Project protocol [35], which amplify the V4 variable region of the 16S rRNA gene. This region was identified by PrimerProspector to be nearly universal to archaea and bacteria [113].

### 1.4.2 Operational Taxonomic Units

Unlike more distinct species, such as mammalian species, bacterial species are not well defined. Bacterial genomes are highly variable, and regions used to identify bacteria vary in a continuum rather than clusters of similar sequences.

Historically bacteria that have 97% identity in a variable region are considered to be the same taxa. The 97% cutoff was arbitrarily chosen to best map sequence data to bacterial classifications. This threshold maximizes the grouping of bacteria classified as the same species while minimizing the grouping of bacteria classified as different species. Before sequencing bacterial classification was often done by appearance or by metabolic products, so there are outliers where bacteria classified in the same species are actually genetically very different, or bacteria classified in different genus are genetically very similar.

However, it is difficult to determine how a batch of sequences should be partitioned into groups of 97% identity. One way is to perform a clustering algorithm that optimally partitions the groups and then later assign taxonomic identity by matching the sequences with public databases. Another way is to start off with seed sequences from known bacteria and perform the clustering such that the 97% identity groups are centered on the seed sequences. In any case, the resulting taxonomic groupings are known as Operational Taxonomic Units (OTUs), and are used consistently within the same experiment. While OTUs can be annotated with standard taxonomic names such that results can be compared between experiments, technically the taxonomic groupings used by different experiments are not the same.

### 1.4.3 General protocol and rationale

The 16S rRNA gene sequencing experiment uses next generation sequencing to estimate the proportional abundance of different bacterial taxa. Samples are extracted and prepared for sequencing, and then the sequenced reads are collated into counts per assumed taxa per sample. The resulting table undergoes statistical analysis.

#### Pre-sequencing processing

There are several very general steps to the pre-sequencing process:

1. Take a biological sample and extract the DNA The sample can be collected swabbing the target body site or by collecting samples in some other way. DNA extraction is usually done with common commercial kits.
2. Run a PCR amplification As discussed previously, the gene tag experiments in this thesis amplify the V4 region of the 16S rRNA gene, following the Earth Microbiome Project

protocol [15]. The set of primers that we use are barcoded, so that we can sequence all the samples in the same sequencing run and differentiate them afterwards.

3. Run sequencing We use 150 nucleotide paired-end sequencing on the Illumina MiSeq platform. The 150 nucleotide paired ends allow us to overlap paired sequences in the middle to reconstitute the full sequence of the variable region.

### **Post-sequencing processing**

Here are the steps for going from raw sequenced reads to a table of counts per taxa per sample.

1. Demultiplex the raw sequence The barcodes are used to separate the sequences according to what sample they came from.
2. Assemble the paired ends of sequenced DNA The paired sequences are overlapped in the middle, resulting in the full variable region amplified by the primers.
3. Group the reads into operational taxonomic units (OTUs) We used UCLUST to cluster the reads into groups of 97% identity [28].
4. Annotate the OTUs with bacterial taxonomy Annotation was done by matching our OTUs to the SILVA database [83].

Generate a phylogenetic tree This can be done using the center-most sequence of each cluster that forms each OTU, and putting the sequences in a multiple sequence alignment, using software such as MUSCLE [27].

Alternatively, an Individual Sequence Unit (ISU) based approach rather than an OTU based approach can be taken, where the individual sequences are preserved even after grouping into OTUs, so that different strains within the same OTU can be analyzed separately [13].



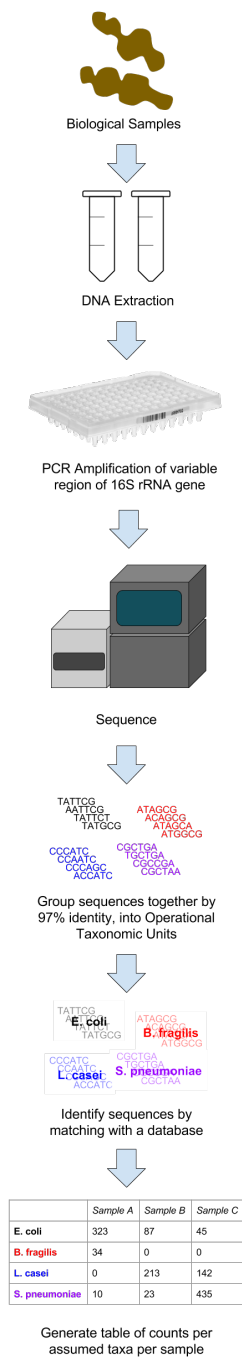


Figure 1.1: **16S rRNA gene tag experiment workflow.** This shows the workflow from sample collection to data generation. The end result is a count table of reads per operational taxonomic unit per sample.

### 1.4.4 Data analysis

There are two goals in gene tag data analysis. First, is there any structure in the data (separation, clustering, correlations, differentials, etc.)? Second, what drives the structure in the data?

Separation or clustering can be examined by determining the distance between each sample, and using these distances to plot the samples as points on a graph. The following sections will go over the most commonly used distance metric in microbiome research, called UniFrac, as well as the Principal Components Analysis multidimensional scaling method for plotting the points on a graph. Afterwards the data can be visually or mathematically inspected for separation or clustering.

The technique used for determining if taxa are differentially abundant between groups is the same technique used for determining if gene annotations are differentially abundant between groups in the metagenomic experiment, and has its own section, titled “Compositional data analysis”.

#### UniFrac

Principal Component Analysis is necessary for multivariate statistics, and It is well known that the Principal Component Analysis cannot be performed on proportions, such as the OTU abundances derived from gene tag sequencing. Instead, a Euclidean distance is required [3].

In 2005, Lozupone et al introduced the UniFrac distance metric, a measure to calculate the difference between microbiomes that incorporated phylogenetic distance [61]. The goal of UniFrac was to enable objective comparison between microbiome samples from different conditions. In 2007, Lozupone added a proportional weighting to the original unweighted method [60]. Since then, papers reporting these metrics have garnered over a thousand citations, and enabled research about everything from how kwashiorkor causes malnutrition [96] to how people can have similar microbiomes to their pet dogs [97]. Except for Generalized UniFrac, used to make hybrid unweighted and weighted UniFrac comparisons [17], few advances in the metric have occurred since 2007.

#### Unweighted UniFrac

Unweighted UniFrac uses an inferred evolutionary distance to measure similarity between samples. It requires a reference phylogenetic tree containing all the taxa present in the samples to be examined. The calculation is performed by dividing the branch lengths shared between the two samples by the branch lengths covered by either sample. A distance of 0 means that the samples have an identical set of taxa detected, and a distance of 1 means that the two samples share no taxa in common.

The qualitative rather than quantitative nature of unweighted UniFrac makes the metric very sensitive to sequencing depth. A greater sequencing depth generally results in the detection of a greater number of taxa. To account for this problem, ecologists use a technique called rarefaction to normalize the sequencing depth across samples by random sampling without replacement [16]. However, in unweighted UniFrac samples move relative to the other samples in different rarefaction instances, to the point where they can switch from being a member of one cluster of data to another, as demonstrated in the chapter Expanding the UniFrac Toolbox.

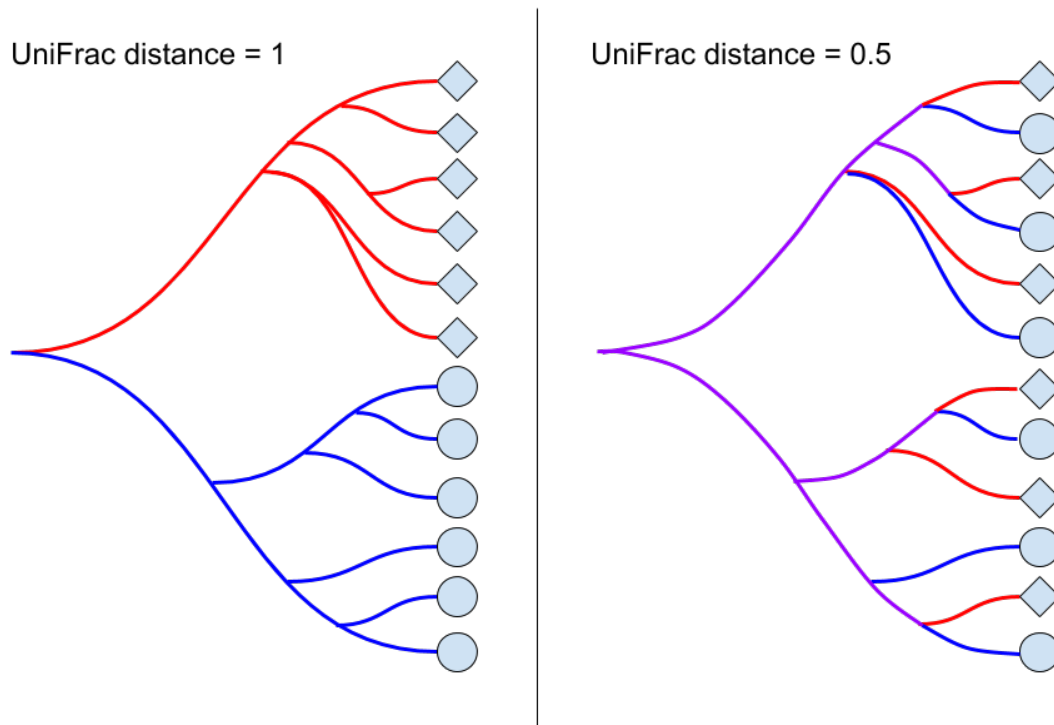


Figure 1.2: **Unweighted UniFrac.** When two samples do not share any branches of the phylogenetic tree, the unweighted UniFrac distance is maximized at 1. When two samples share half of their branch lengths on the phylogenetic tree, the unweighted UniFrac distance is 0.5. If the two samples contain exactly the same taxa, the unweighted UniFrac distance is minimized at 0, since the samples share all branches.

### Weighted UniFrac

Weighted UniFrac is an implementation of the Kantorovich-Rubinstein distance in mathematics, also known as the earth mover's distance [29]. Rather than looking only at the presence or absence of taxa, each branch length of the phylogenetic tree is weighted by the difference in proportional abundance of the taxa between the two samples. This technique reduces the problem of low abundance taxa being represented as a 0 or by a low count depending on sampling depth. In unweighted UniFrac, such taxa would flip from absent to present, and could skew the measurement: this would be especially problematic if the taxa are on a long branch. In weighted UniFrac, low abundance taxa have a much lower weight and so will have a lower impact on the total distance reported by the metric.

UniFrac is constituted as either a presence/absence (unweighted UniFrac) [61], a linear proportion in the form of weighted UniFrac [60], or some combination of the two in the form of Generalized UniFrac [17]. However, the data are not linear, because the sum of the total number

of reads is constrained by the sequencing machinery [34]. Alternative weightings and non-linear transformations of data need to be explored.

### Principal Components Analysis

Once the distances between each pair of samples has been calculated, they can be visualized on a plot, with each sample represented as one point. For visualization, the data should be placed so distances are preserved as much as possible, so that clustering and separation of samples can be clearly seen. This is done using the Principal Coordinate Analysis method of multidimensional scaling [25], shortened as PCoA.

To plot all of the samples as points in space such that the distances between each pair of samples are preserved, multiple dimensions are required. In this data specifically, the number of dimensions required is equal to one less than the number of samples. PCoA rescales all the dimensions as components, so that the first component captures the largest variation, or spread of the data, the second component captures the largest variation remaining in the data after the first component, and so on. This way, even if only the first two components are used to plot all the samples as points on a two dimensional graph, the data is spread out to enable visualization of separation or clustering.

After multidimensional scaling the data can be analyzed in several ways. The data can be examined for clustering by k-means analysis [105]. The points can also be measured for separation by looking only at their position on the first principal component axis, especially if the first axis covers the majority of the variation in the data set. With each sample associated with a number on the first principal component axis, one can examine the effect size of two different groups by taking the mean positions and dividing by the standard deviation.

## 1.5 The metagenomic experiment

Deep metagenomic sequencing provides an estimate of the proportion that each type of gene comprises out of the total genes present in the genetic material of the sample. This can be used to answer questions such as:

*What is the metabolic potential of the microbial community?* The metabolic potential is made up of all the protein functions that are coded by the genetic material present in the sample. Biologically speaking, these protein functions represent the enzymatic reactions that the microbiome could produce if all the genes were expressed. For example, the human gut microbiome has more genes related to methanogenesis, compared to the average sequenced microbe [36].

*Are any genes, functional categories of genes, or metabolic pathways made up of genes differentially abundant between groups?* In 2006, Turnbaugh et al published a paper showing that an obesity associated gut microbiome in mice had an increased capacity for energy harvest [107], sparking more research into the gut microbiome and obesity related ailments such as diabetes [56] and non-alcoholic fatty liver disease [120]. The ability to check if genes, functional categories of genes, or pathways are differentially abundant between groups allows scientists to find clues about the mechanisms by which the microbiome affects certain diseases.

All of this information can be determined by either imputation or actual sequencing, discussed in the next sections.

### 1.5.1 Sequencing

The goal of metagenomic analysis is to examine the metabolic potential of the microbiota in the microbiome. This is done by identifying genes, sorting them by the known function of the protein for which they code (such as the catalyzation of a certain reaction), and checking if any functions are differentially present between conditions. Further analysis can also include checking for pathway enrichment, and assembling the sequenced reads into genomes. The general protocol for metagenomic analysis is as follows:

1. Take a biological sample and perform DNA extraction The sample can be collected by swabbing the target body site or collecting excretions.
2. Prepare the DNA for sequencing Fragment the DNA, and filter for the desired size. These steps are all part of the standard Illumina library prep protocol for the HiSeq. There are two options for fragment size, either 50 or 100 nucleotides in length, and we chose the longer one for easier assembly and mapping.
3. Sequence the DNA. We performed single end sequencing on the Illumina HiSeq platform, with our samples barcoded so that they could be pooled into the same sequencing run.
4. Create an annotated library of reference sequences The annotated library contains annotations about what kind of protein each sequence codes for. The first step to creating the annotated library is to gather a database of sequences. The database of sequences can be created before the sequencing is complete by gathering all the genomes of all the bacterial strains predicted to be present in the sample, or it can be created after sequencing by assembling the sequenced reads into parts of genomes. The second step is to annotate the sequences with predicted protein functions. Some publically available genomes already have protein annotations. For genomes or partial genomes without annotations, the placement of genes can be predicted by looking for open reading frames, and these predicted genes can be aligned with databases such as SEED [75] or KEGG [48] to match them with functional annotations, using the BLAST algorithm [2].
5. Map the sequenced reads to the library. Mapping is the process of annotating the sequenced reads by aligning them with sequence that has already been annotated. We used Bowtie2 [55] to map our sequenced reads to the annotated library created in the previous step. Bowtie2 aligns similar sequences together.
6. Determine how many mapped reads match each functional annotation. Once the sequenced reads have been mapped to the annotated reference sequence, the number of reads sequenced for each annotation can be counted up. The end result is a table of counts per gene annotation per sample.

Issues with sequencing and the analysis of sequencing data arise from sampling and the fat nature of the data. The sequences that are read by the sequencer are only a small fraction of the DNA from the sample. Additionally, primers used for sequencing may be biased for certain sequences more than others. Lastly, the data is very fat, which is to say that there are magnitudes more variables (in the form of functional annotations of genes) than there are

samples. This makes it difficult to have enough power to detect small differences in the data, a concept expanded upon in the Points of Failure section below.

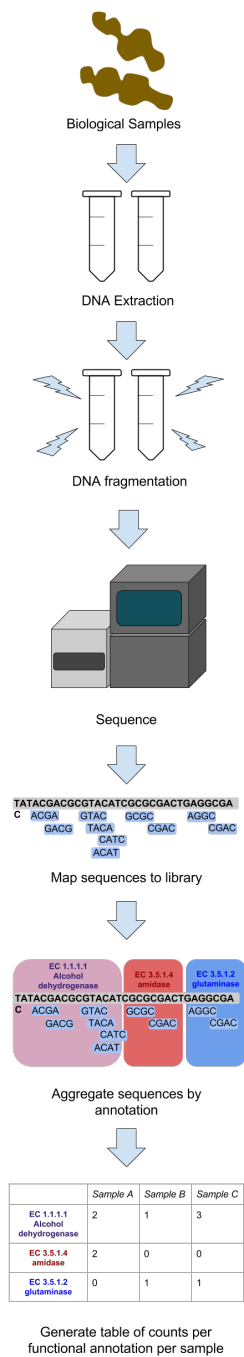


Figure 1.3: **Metagenomic experiment workflow.** This shows the workflow from sample collection to data generation. The end result is a table of number of sequencing reads per functionally annotated gene per sample.

## 1.5.2 Imputation

Deep metagenomic sequencing can be imputed using a tool called PiCrust from a gene tag experiment [54]. PiCrust uses the Greengenes database [23] to identify the bacterial taxa in the sample, and pulls their genomes from the Integrated Microbial Genomes database [66]. With the genomes, the program tries to predict what would be seen if the samples underwent deep metagenomic sequencing. For taxa without a fully sequenced genome, PiCrust infers the genetic content based on ancestors in the phylogenetic tree. PiCrust produces metagenome predictions with Spearman  $r = 0.7$  [54], compared to a full metagenomic sequencing experiment.

Imputation is useful for identifying potential correlations that should be explored and validated further, but should not be used to make conclusions. The issues with imputation include all the issues with sequencing, plus the added variation in its imperfect correlation.

## 1.5.3 Data analysis

Data analysis can be performed by seeing if functions are differentially abundant between samples in different groups (described in the “Compositional data analysis” section), examining functional categorizations, and checking for pathway enrichment.

### Functional categorization

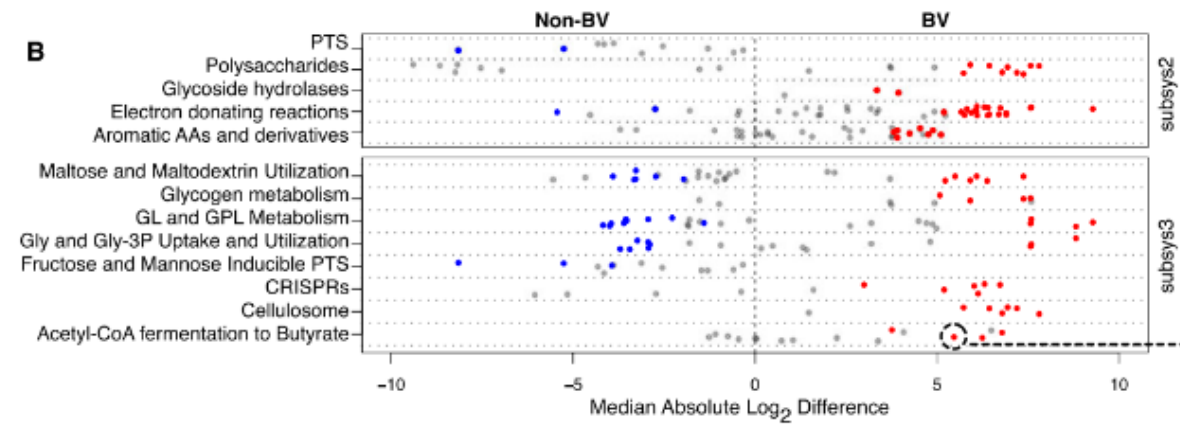


Figure 1.4: **Example of stripcharts for subsystem 2 and 3 functional categorizations.** Dots on the left side are subsystem 4 annotations found to be more abundant in the healthy condition while dots on the right side are subsystem 4 annotations found to be more abundant in the bacterial vaginosis condition. Colored dots were found to be significantly differentially abundant. Figure taken from [63].

We use the SEED annotation, which has four different levels of categorization. Subsystem 4 is the most atomic categorization level and describes the specific function of the protein group, for example, “Isovaleryl-CoA dehydrogenase (EC 1.3.99.10)”. Subsystem 3, 2, and 1 are increasing more general levels of categorizations, from enzyme families to large categorizations such as genes related to carbohydrate metabolism.



Even if the subsystem 4 functional categories are not significantly different between groups, they each have an effect size with a direction. Stripcharts can be used to plot the effect sizes of the subsystem 4 categories for a larger category. For example, by plotting the effect sizes of all the subsystem 4 categorizations under Carbohydrate Metabolism, one can visually see if there are any obvious directional trends for carbohydrate metabolism functions being more present in the experimental group compared to the control.

### **Pathway enrichment**

Biological pathways can be thought of as made up of a series of chemical reactions, each catalyzed by a protein enzyme, which is encoded by a gene. KEGG (Kyoto Encyclopaedia of Genes and Genomes) is a manually curated annotation database that matches genes to pathways [48]. This database allows researchers to see if there is differential abundance of pathways encoded by functionally annotated genes, even when the genes may not be differentially abundant by themselves.

## **1.6 Points of failure**

The Huttenhower lab has organized the Microbiome Quality Control project (MBQC) at <http://www.mbqc.org/>. Preliminary results show that despite being given the same samples, different participating labs can come up with vastly different results. This lack of reproducibility is caused by a lack of consensus on the correct way to analyze microbiome data. The following sections explore different aspects of microbiome data that contribute to this.

### **1.6.1 Collection methods differ**

These experiments are very sensitive to batch effects because microbiome composition can be very variable within groups such that the effect size of a difference between groups can be small. Wherever possible, all samples should be processed in the same batch. Analysis should also be done to check if samples extracted on different dates or sequenced with different primers separate into clusters, to make sure that there is no systematic bias in the data.

### **1.6.2 Microbiome data is highly variable between individuals**

One highly studied body site is the gut, and the gut microbiome can be affected very strongly by diet [109]. This among other factors lead to a highly diverse gut microbiome between subjects for reasons unrelated to the disease being studied, creating a lot of noise, potentially obscuring real effects or even creating the appearance of false effects.

Generally experiments of this nature typically have low sample sizes due to budget constraints, sample collection difficulties, patient compliance, and other issues. To increase cost effectiveness and reduce batch effects, we run all the samples in an experiment on the same sequencing run, by means of a primer design [37].

There are several models for computationally analyzing the variance within conditions in order to determine if operational taxonomic units are significantly differentially abundant, most of which were originally designed for RNA-seq experiments on single organisms [76]. Currently

the most popular tools for analyzing differential abundance are EdgeR [87], DESeq2 [58], and MetagenomeSeq [78]. EdgeR was cited by 1,130 papers in 2015 according to Google Scholar. DESeq2 and MetagenomeSeq are part of the QIIME pipeline, which was cited by 1,620 papers in 2015.

EdgeR and DESeq2 use the negative binomial distribution. The negative binomial distribution allows the variance of data to be estimated given the mean, through a function. The function is determined by collecting the mean and variance for all the counts for each OTU in each experimental condition, and fitting the variances according to the negative binomial distribution. This vastly underestimates the variance at low counts, which represent the sampling of low abundance OTUs, and can be very different between replicates. Underestimating the variance at low counts produces spurious low p-values for low count OTUs [30].

MetagenomeSeq uses the Zero-Inflated Gaussian (ZIG) model, which is a binomial distribution of counts (that may include zero counts), plus a function to predict how many extra zeros there will be. This doesn't work well when the total number of reads are not well matched, because then there will be much more zeros in the data set with less reads, due to having a lower sequencing depth, and a consistent total read count is required between samples according to page 2 of the supplementary material in the first metagenomeSeq paper [78].

For my differential abundance analysis, I've used ALDEx2, which samples from the Dirichlet distribution to model variation in the data [31]. After a number of samples, the mean value and mean variance are used to determine if OTUs are differentially abundant between groups, an approach that is believed to result in greater sensitivity and equivalent specificity compared to the DESeq2 approach [31].

### **1.6.3 Microbiome data involves the comparison of many features**

Oftentimes, the number of taxa or gene functions comparisons is a magnitude larger than the sample size. This is known in statistics as having more variables than observations, or having fat data. The higher the ratio of variables to observations are, the less likely the principal components analysis is to be reliable [74].

Researchers should include multiple test corrections to ensure that the results they are reporting are true, at the expense of having p-values less than 0.05. Unfortunately many studies have been published in high impact journals without multiple test corrections, including a famous paper linking the gut microbiome to autism published in Cell [43].

### **1.6.4 Microbiome data is compositional**

There are several core truths about microbiome data that should be considered when making an analysis strategy.

First, the total number of reads per sample is irrelevant to the biological implications of the data, as it is limited by how the samples were processed and the sequencing platform. Based on spurious correlations discovered in organ size research, it is known that given compositional data (such as bone lengths as a proportion of height, or OTU abundances that add up to the total number of counts per sample), analysis with the assumption that the variables (bone lengths or OTU counts) are independent lead to spurious positive correlations [79]. The variables thought to be independent are related by the sum they are divided by. Additionally, the constrained sum

causes the abundance of different taxa to appear to be negatively correlated with each other when analyzed by conventional statistics. When one taxa increases in abundance, the counts detected in other taxa decrease in abundance, even if the taxa are not decreasing in abundance biologically.

Second, removing an entire variable (an OTU in gene tag sequencing, or a functional annotation in deep metagenomic sequencing) from the analysis should not change correlations between OTUs. Removing variables occur routinely in microbiome research, such as when rare OTUs are discarded. Without a data transformation, removing variables will change the correlation between variables [1].

To ensure that these conditions are met, data should be analyzed in a compositional way. Several types of log ratio data transformations are recommended to allow the data to be analyzed by standard Euclidean methods [1]. The type that makes the most sense for microbiome data is the centered log ratio transform. The centered log ratio transform is performed by dividing each proportional abundance by the geometric mean of all the proportional abundances, and taking the logarithm. The geometric mean acts as a low level baseline abundance in microbiome data. Taking the logarithm of the ratio allows for a consistent measurement whether the large number is in the numerator or denominator of the ratio.

The centered log ratio transform prevents the total number of reads from affecting the measurement, so long as the geometric mean is a stable baseline, a condition met in a typical microbiome data set [CITATION NEEDED]. The centered log ratio transform also allows for coherent subcompositional data analysis as remaining values are not affected when entire variables are removed.

Compositional techniques such as those espoused in the Analysis of Composition of Microbiomes (ANCOM) framework [64] and the ANOVA-Like Differential Expression 2 (ALDEx2) software [31] should be used to prevent spurious correlations and promote consistent data analysis. However, these techniques are not yet mainstream in the field.

### 1.6.5 Microbiome data is sparse

One of the fundamental challenges in analyzing differential abundance is accounting for zeroes. Unlike a presence/absence test, a zero does not necessarily mean that the expression is not there. The expression could be present in an amount smaller than the resolution of the test. This is a problem because when statistical methods are used to examine significantly different expression, the comparison of zero values to non-zero values are likely to come out as significant whether or not the expression is differential. Additionally, the log transformations used in compositional data analysis cannot be performed on zeros.

Two methods have been suggested in the literature to account for zeros. The first is simply to add a small arbitrary value to each zero, as suggested in the original literature about the statistical analysis of compositional data [1]. This is used in ALDEx2, and the arbitrary value is chosen to be 0.5, representing complete uncertainty in whether or not a zero count in one sample (where the OTU or gene has non zero counts in other samples) would be a 0 or a 1 in a technical replicate [30].

The second method is to take a Bayesian approach where the likelihood that a zero could be changed to a positive count if the sample were resequenced is estimated, based on . This is implemented by the `cmultRepl` command in the `zCompositions` package in R [77]. Based on the

shape of the rest of the data for the same sample, the average value of the count detected if a zero were resequenced is determined, and the zeros are all replaced by this fraction.

The microbiome field is quite new, and has been undergoing many exciting developments. Gold standards must be set to ensure that studies are replicable, and that published research represents the biological reality.

## **1.7 The gut microbiome in atherosclerosis-susceptible and atherosclerosis-resistant patients**

In 2010, over half a million deaths in the United States were due to cardiovascular disease [70]. Atherosclerosis, a chronic disease in which fatty plaques build up in arteries leading to blood clot formation and blockage of the blood stream is a strong contributor to cardiovascular disease. Modifiable risk factors for atherosclerosis include obesity, smoking, physical inactivity, stress, and more. By imaging the carotid artery for atherosclerotic plaques, Spence [100] found that, while the progression of most patients' atherosclerosis is predicted by their risk factors, some patients present with few factors, and yet their atherosclerosis progresses. Conversely, other patients have many risk factors and their atherosclerosis regresses. Patients on the extreme ends of this spectrum, who exhibit unexplained progressive atherosclerosis or unexplained regressive atherosclerosis will be the focus of this chapter.

Currently the main therapies for atherosclerosis and other cardiovascular disease include a diet low in cholesterol, the cessation of smoking, and the administration of statins to inhibit cholesterol production in the liver. However, in some patients, such as those in this study, these interventions may not be effective. A characterization of the gut microbiome and their effect on cardiovascular disease is necessary to explore atherosclerosis risk factors which are beyond the patient's explicit control. With this knowledge, the mortality and morbidity of victims of cardiovascular disease may be improved.

Products of gut metabolism have been shown to affect atherosclerotic progression. For example, high levels of trimethylamine N-oxide (TMAO) as measured in blood plasma and urine have been associated with higher atherosclerosis risk in humans [103]. In humans, TMAO has been shown to be produced from L-carnitine, present in red meat [50], and dosing human patients with antibiotics decreased their TMAO levels [103].

More research is necessary to elucidate the role of the gut microbiome in atherosclerosis progression. One chapter of this thesis is dedicated to comparing the gut microbiota in atherosclerosis-resistant and atherosclerosis-susceptible patients through 16S rRNA gene tag sequencing.

## **1.8 The gut microbiome in patients with non-alcoholic steatohepatitis compared to healthy controls**

Non alcoholic fatty liver disease (NAFLD) has been on the rise along with obesity, affecting a fifth to a third of the North American population [81]. Most people with NAFLD remain asymptomatic, however, in up to a third of patients NAFLD can progress to non-alcoholic

steatohepatitis (NASH), causing inflammation and scarring in the liver, and decreasing the 5 year survival rate to 67% [82]. If we can shed some light on the process by which people progress from NAFLD to NASH, we might be able to find treatments to prevent NASH.

Several genetic [99] [86], epigenetic [71], hormonal [118], and metabolite [84] factors are known to affect the risk progression to NASH. The relationship between the gut microbiome and non alcoholic fatty liver disease is less clear.



A 2001 paper performed C-D-xylose-lactulose breath tests and measured tumor necrosis factor alpha levels to determine presence of bacterial overgrowth, and found increased bacterial overgrowth in 22 patients with NASH compared to 23 healthy controls [115]. Some papers claim a link between ethanol-producing gut bacteria and NAFLD [120] [46], however, no multiple test correction was performed in these studies. Five published studies claiming to have found differentially abundant bacteria in the gut microbiome between healthy controls and patients with non alcoholic fatty liver disease have nearly non-overlapping results [120] [116] [84] [46] [11].

These five studies do not form a consistent story about the gut microbiome and NAFLD. In one chapter of this thesis we report the results of our own analysis, which we have attempted to run rigorously, such that our results are replicable. Additionally, we are running a deeply sequenced metagenomic study, which hasn't been done in the past.

## **Chapter 2**

### **Expanding the UniFrac toolbox**

# Expanding the UniFrac toolbox

Ruth G Wong<sup>1</sup> , Jia R Wu<sup>1</sup> , Gregory B Gloor<sup>1</sup> \*

**1 Department of Biochemistry, University of Western Ontario, London, Ontario, Canada**

 These authors contributed equally to this work.

\* ggloor@uwo.ca

## Abstract

Microbiome analysis is frequently performed using the UniFrac distance metric to separate groups. Here we demonstrate that unweighted UniFrac is highly sensitive to rarefaction instance and to sequencing depth in uniform data sets with no clear structure or separation between groups. We show that this arises because of subcompositional effects. We introduce information UniFrac and centered ratio UniFrac, two new weightings that are not sensitive to rarefaction and allow greater separation of outliers than classic unweighted and weighted UniFrac. With this expansion of the UniFrac toolbox, we hope to empower researchers to extract more varied information from their data.

## Introduction

In 2005, Lozupone et al introduced the UniFrac distance metric, a measure to calculate the difference between microbiomes that incorporated phylogenetic distance [61]. The goal of UniFrac was to enable objective comparison between microbiome samples from different conditions. In 2007, Lozupone added a proportional weighting to the original unweighted method [60]. Since then, papers reporting these metrics have garnered over a thousand citations, and enabled research about everything from how kwashiorkor causes malnutrition [96] to how people can have similar microbiomes to their pet dogs [97]. Except for generalized UniFrac, used to make hybrid unweighted and weighted UniFrac comparisons [17], few advances in the metric have occurred since 2007. In this paper we examine a data set where unweighted UniFrac gives misleading results, and discuss some alternative weightings for UniFrac.

### 2.0.1 Data

UniFrac requires two pieces of data: A phylogenetic tree and a table of counts per inferred taxa per sample. This data is derived from a gene tag experiment, the most common of which is the 16S rRNA gene tag experiment for microbiome research [106], which has highly conserved regions interspersed with variable regions. A gene tag experiment is run by amplifying the

gene tag through PCR, and then sequencing the resulting amplicon. The variable region in the amplicon can be used to make inferences about the taxonomy of the detected bacteria.

### Operational Taxonomic Units

Unlike more distinct species, such as mammalian species, bacterial species are not well defined. Bacterial genomes are highly variable, and regions used to identify bacteria vary in a continuum rather than clusters of similar sequences.

Historically bacteria that have 97% identity in a variable region are considered to be the same taxa. The 97% cutoff was arbitrarily chosen to best map sequence data to bacterial classifications. This threshold maximizes the grouping of bacteria classified as the same species while minimizing the grouping of bacteria classified as different species. Before sequencing bacterial classification was often done by appearance or by metabolic products, so there are outliers where bacteria classified in the same species are actually genetically very different, or bacteria classified in different genus are genetically very similar.

However, it is difficult to determine how a batch of sequences should be partitioned into groups of 97% identity. One way is to perform a clustering algorithm (using software such as UCLUST [28]) that optimally partitions the groups and then later assign taxonomic identity by matching the sequences with public databases, such as SILVA [83], the Ribosomal Database Project [19], and Greengenes [23]. Another way is to start off with seed sequences from known bacteria and perform the clustering such that the 97% identity groups are centered on the seed sequences. In any case, the resulting taxonomic groupings are known as Operational Taxonomic Units (OTUs), and are used consistently within the same experiment. While OTUs can be annotated with standard taxonomic names such that results can be compared between experiments, technically the taxonomic groupings used by different experiments are not the same.

Grouping of amplicon sequences into OTUs allows for the data to be summarized into a table of counts per OTU per sample.

### Phylogenetic tree

To generate the phylogenetic tree [18], a representative sequence is taken from each of the cluster of sequences that belong to each OTU. If seed sequences were used to generate the OTU, then these can be used for the phylogenetic tree, otherwise a sequence in the center of the cluster should be selected.

The phylogenetic tree is created through a multiple sequence alignment with the representative OTU sequences, using software such as MUSCLE [27]. Each leaf of the tree represents one of the OTUs, and each of the branches of the tree has a length. Additionally, the tree needs to be rooted for the UniFrac calculation to be performed. This can be done by rooting the tree by its midpoint, which can be performed by software such as the phangorn R package [90].

## 2.0.2 Unweighted UniFrac

Unweighted UniFrac [61] uses an inferred evolutionary distance to measure similarity between samples. It requires a reference phylogenetic tree containing all the taxa present in the samples to be examined, plus information about which taxa were detected in each sample. The calculation



is performed by dividing the branch lengths that are not shared between the two samples by the branch lengths covered by either sample. A distance of 0 means that the samples are identical, and a distance of 1 means that the two samples share no taxa in common.

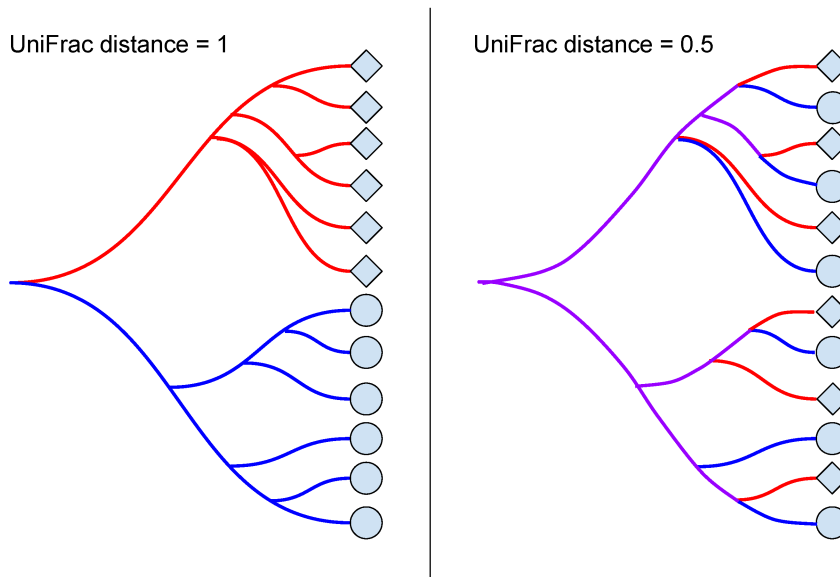


Figure 2.1: **Unweighted UniFrac.** When two samples do not share any branches of the phylogenetic tree, the unweighted UniFrac distance is maximized at 1. When two samples share half of their branch lengths on the phylogenetic tree, the unweighted UniFrac distance is 0.5. If the two samples contain exactly the same taxa, the unweighted UniFrac distance is minimized at 0, since the samples share all branches.

As UniFrac is a presence/absence test, it is sensitive to sequencing depth, and assumes that the data has been rarefied to a common sequencing depth [62], and rarefaction prior to unweighted UniFrac has become a standard part of the microbiome analysis workflow, with built in rarefaction functions in QIIME [14] and mothur [91].

### 2.0.3 Weighted UniFrac

Weighted UniFrac [60] is an implementation of the Kantorovich–Rubinstein distance in mathematics, also known as the earth mover’s distance [29]. Rather than looking only at the presence or absence of taxa, each branch length of the phylogenetic tree is weighted by the difference in proportional abundance of the taxa between the two samples.

This technique reduces the problem of low abundance taxa being represented as a 0 or by a low count depending on sampling depth. In unweighted UniFrac, such taxa would flip from absent to present, and could skew the measurement: this would be especially problematic if the taxa are on a long branch. In weighted UniFrac, low abundance taxa have a much lower weight and so will have a lower impact on the total distance reported by the metric.

UniFrac is constituted as either a presence/absence (unweighted UniFrac) [61], a linear proportion (weighted UniFrac) [60], or some combination of the two (generalized UniFrac) [17]. However, the data are not linear, because the sum of the total number of reads is constrained by the sequencing machinery [34] [30] [31] [59]. Alternative weightings and non-linear transformations of data need to be explored. Furthermore, unweighted UniFrac is known to be unreliable, but it is not generally known or understood how this can impact results.

## Materials and Methods

### 2.0.4 Analytical techniques

#### Rarefaction

Rarefaction normalizes the samples OTU counts to a standard sequencing depth [95]. This resulting table can be thought of as a random point estimate of the dataset, as the output is a sub-sample of the original table. This standardization process is recommended by the authors of UniFrac [16] in order to account for the sensitivity of UniFrac to sequencing depth.

Rarefactions can be performed using the QIIME software [14] or using the vegan package in R [73].

#### Unweighted UniFrac

Unweighted UniFrac is calculated based on the presence or absence of counts for each branch in the phylogenetic tree, when comparing two samples. A branch belongs to a sample when at least one of the OTUs in the leaves below it have a non-zero abundance. The formula for unweighted UniFrac is as follows, where  $b$  is the set of branch lengths in the phylogenetic tree, and  $A$  and  $B$  represent the two samples being compared:

$$Unweighted_{AB} = \frac{\sum b_A \Delta b_B}{\sum b_A \cup b_B}$$

The sum of the branch lengths that belong to one sample but not the other is divided by the sum of the branch lengths that belong to one or both samples.

#### Weighted UniFrac

Weighted UniFrac [60] also incorporates each branch length of the phylogenetic tree, and weights them according to proportional abundance of the two samples. The formula for weighed UniFrac is as follows, where  $A$  and  $B$  are the two samples,  $b$  is the set of branch lengths, and  $\frac{A_i}{A_T}$  and  $\frac{B_i}{B_T}$  are the proportional abundances associated with branch length  $b_i$ :

$$Weighted_{AB} = \sum_i^n b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|$$

### Information UniFrac

Information UniFrac is calculated by weighing each branch length by the difference in the uncertainty of the taxa abundance between the two samples. Uncertainty is calculated as follows, where  $p$  is the proportional abundance [94]:

$$uncertainty = -p \times \log_2(p) \quad (2.1)$$

If a sample is 50% taxa A and 50% taxa B, then the proportional abundances have maximum uncertainty about what taxa is likely to be seen in a given sequence read. If a sample is 80% taxa A and 20% taxa B, then there is less uncertainty, because a given sequence read is more likely to be taxa A. When the amount of uncertainty that a taxa has in one sample corresponds with the amount of uncertainty the same taxa has in a different sample, the abundance of that taxa is mutually informative between samples. Weighting UniFrac by uncertainty combines the the concept of uncertainty with phylogenetic relationships to identify taxa that are differentially informative between groups.

The formula for Information UniFrac is as follows:

$$Information_{AB} = \sum_i^n b_i \times \left| \frac{A_i}{A_T} \log \left( \frac{A_i}{A_T} \right) - \frac{B_i}{B_T} \log \left( \frac{B_i}{B_T} \right) \right|$$

### Centered Ratio UniFrac

In complex microbiome communities, there are very many bacterial taxa with a low level of counts. Taking the geometric mean of the proportional abundances of taxa in a microbiome sample represents an unbiased baseline [1]. Experiments generally do not have power to detect differences at abundances below the mean [30]. Centering the proportional abundances around the geometric mean thus allows one to examine the data in context, muting differences that are close to baseline and accentuating outliers. The formula for centered ratio UniFrac is as follows, where  $gm$  is the geometric mean:

$$CenteredRatio_{AB} = \sum_i^n b_i \times \left| \frac{\frac{A_i}{A_T}}{gm(A_i)} - \frac{\frac{B_i}{B_T}}{gm(B_i)} \right|$$

Note that the geometric mean is calculated by combining all children in the subtree of  $b_i$  into  $\frac{A_i}{A_T}$  for sample A or  $\frac{B_i}{B_T}$  for sample B, and including the rest of the single taxa proportional abundances separately. The one combined proportional abundance and the remaining single taxa proportional abundances are input into the geometric mean formula, as set  $a$ :

$$gm(a) = \left( \prod_i^n a_i \right)^{1/n}$$

One challenge when it comes to the analysis of read count data is the presence of zero counts. Whether a low-abundance taxa appears in the data as a zero or a low positive count is up to chance, and assuming that a zero count represents the absence of a taxa can be very misleading

[30]. A Bayesian approach can be used to estimate the likelihood that a zero could be changed to a positive count if the sample were resequenced, implemented by the `cmultRepl` command in the `zCompositions` package in R [77].

The use of this weighting for UniFrac produces measurements that violate the triangle inequality, such that Euclidean statistics are technically invalid. Thus this, like the Bray-Curtis metric, is a dissimilarity, not a distance.

For this paper, we calculate UniFrac metrics using a custom R script, which includes unweighted UniFrac, weighted UniFrac, information UniFrac, and centered ratio UniFrac: <https://github.com/ruthgrace/exponentUnifrac/blob/master/UniFrac.r>

### Bray-Curtis dissimilarity metric

The Bray Curtis dissimilarity metric [6] quantifies how dissimilar two sites are based on counts. A index of 0 means that two samples are identical, while a index of 1 means samples do not share any species. It is computed as a proportion through the formula:

$$C_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

where  $C_{ij}$  = dissimilarity index bound by [0,1]

$S_i$  = Specimen counts at site i

$S_j$  = Specimen counts at site j

## 2.0.5 Data preparation

The data used comes in the form of a table of counts per operational taxonomic unit per sample, plus a phylogenetic tree. All of our data are derived from 16S rRNA gene tag sequencing experiments.

### Tongue dorsum data set

The tongue dorsum data set is a collection of 60 microbiome samples taken from the tongues of healthy participants. There were 324789 reads across 554 OTUs, and a minimum and maximum of 659 and 17176 reads per sample.

Samples from this experiment were sourced from the Human Microbiome Project [110] Qiime Community profiling v35 otu tables (<http://hmpdacc.org/HMQCP/>).

Rarefaction was conducted through Qiime version 1.8.0-20140103 to 659 reads (the lowest number of reads for a sample), and generation of the ellipse figures was done in R version 3.2.3 (2015-12-10) "Wooden Christmas-Tree" x86\_64-apple-darwin13.4.0 (64 bit).

A principal coordinate analysis is drawn from each distance matrix per metric, and for the first principal coordinate of each metric, Vres is computed per each first principal coordinate as defined by the formula:

$$V_{res} = \frac{|V_1 - V_i|}{range(V_1, V_i)}$$

where  $V_{res}$  = Set of computed PC1s,

$V_1$  = Reference PC1 (the first),

$V_i$  = Each subsequent PC1,

### Tongue dorsum and buccal mucosa data set

The tongue dorsum and buccal mucosa data set is a collection of 30 microbiome samples taken from the tongues of healthy participants, plus 30 microbiome samples taken from the buccal mucosa (cheek) of a different set of healthy participants. There were 399188 reads across 12701 OTUs, and a minimum and maximum of 5028 and 9861 reads per sample. Note that if the OTUs that are less than 1% abundant in all samples are filtered out, only 179 OTUs remain.

To create this data set, thirty random samples were selected from the tongue site of the Human Microbiome Project [110] and thirty random samples from the buccal mucosa site. Samples were filtered so that only samples with 5000 to 10,000 reads were included.

Read counts from the HMP data set were rarefied to the smallest total read count per sample using the vegan R package [73] before the unweighted UniFrac distance was calculated. Weighted, information, and centered log UniFrac were calculated on the data set without rarefaction. The resulting distances were plotted for principal coordinate analysis.

The script used to run this analysis can be referenced at [https://github.com/ruthgrace/exponentUnifrac/blob/master/tongue\\_cheek\\_script.r](https://github.com/ruthgrace/exponentUnifrac/blob/master/tongue_cheek_script.r).

### Breast milk data set

The breast milk data set is a collection of 58 microbiome samples taken from lactating Caucasian Canadian women. The breast milk data set used here has also been published in the Microbiome Journal [111]. There were a total of 5318357 reads across 115 OTUs, and a minimum and maximum of 3072 and 2791000 reads per sample. Note that 2790735 reads came from a sample that was taken from a patient with an infection, and the next largest number of reads per sample was 282485 (ten times less).

The count table was analyzed using our custom UniFrac script. Data was rarefied to the sample with the smallest number of read counts (3072) before the unweighted UniFrac distance matrix was calculated. Non-rarefied data was used for weighted, information, and centered ratio UniFrac. Data was plotted using a principal components or coordinate analysis as appropriate.

The script used to run this analysis can be referenced at [https://github.com/ruthgrace/exponentUnifrac/blob/master/breastmilk\\_script.r](https://github.com/ruthgrace/exponentUnifrac/blob/master/breastmilk_script.r).

## Results

### 2.0.6 Unweighted Unifrac is highly sensitive to rarefaction variants

A commentary by Lozupone et al. 2011 [62] addressed the sensitivity of Unweighted UniFrac to sampling. They utilized mean UniFrac values to compute a confidence ellipse. However, we observed that this approach under-represented the true variability of unweighted UniFrac as a distance metric by highlighting how individual samples vary. In the absence of true differences and in the presence of uneven sampling, unweighted UniFrac can be sensitive to rarefaction variants. We show this by analyzing two rarefactions of the same body site with the rationale that if there is no true difference in the data, separation of these samples should not be observed.

Sixty tongue dorsum subsamples were drawn from the Human Microbiome Project data without replacement. Rare OTUs with less than 100 total counts across all the samples were removed. The minimum sample count for the subset of 60 we analyzed was 659, therefore we rarefied (subsampling) to the minimum of 659 to normalize the samples. For Fig. 2.2, two independent rarefactions of the data were conducted in order to observe the effect of rarefaction variants on the metrics. The unweighted UniFrac distance was then computed for each rarefaction, and Procrustes adjustment was applied in order to overlay the second rarefaction onto the first. A PCA of rarefaction 1 was plotted, and any samples that changed between rarefactions one and two were visualized with red and blue on the plot. If the sample moved from one cluster to another between the rarefactions, it was indicated with either a blue or a red arrow.

In both rarefactions on Fig. 2.2, samples separated distinctly into two clusters on principal coordinate 1. Principal coordinate 1 explains the most variation in the data, and is thus useful to visualize if any associated metadata is behind the sample separation. However, the separation was not explainable by any metadata associated with the HMP experiment, and is thus an undesirable result. When plotting the rarefactions against each other, various samples are observed moving between the various clusters. This example demonstrates that samples with little difference can appear to be different through the unweighted UniFrac distance metric.

For the ellipse plot in Fig. 2.3, 60 tongue dorsum subsamples were randomly drawn without replacement. Rare OTUs with less than 100 total counts across all samples were removed. A hundred separate rarefactions were conducted on the data to a minimum sampling depth of 378. For each individual rarefied OTU table, a distance matrix was computed using the unweighted UniFrac, weighted UniFrac, Bray-Curtis Dissimilarity, information UniFrac, and centered ratio UniFrac as a weighting method. By generating 100 separate datasets for each metric, it is possible to assess the magnitude of difference each metric has by analyzing what is essentially the same data. In other words, what does the effect of random sampling (rarefaction) have on the output of each metric? Each distance matrix generated per metric was adjusted with a Procrustes adjustment to overlay the subsequent rarefactions onto the first.

Given the wide use of unweighted UniFrac in the literature with small principal component 1 and 2 effects, we suggest caution in their interpretation. For example, see the use of unweighted UniFrac in these papers about the human microbiome published in Cell[43] and Nature [98].

The maximum value of Vres for each rarefaction is plotted against the median value per rarefaction in Fig. 2.3. This plotting serves to highlight the maximum potential change for an analysis given that there is no difference in the data. Unweighted UniFrac shows by far the

highest maximum potential change between rarefactions, compared to weighted, information, and centered ratio UniFrac, as well as Bray-Curtis.

## 2.0.7 Why does Unweighted Unifrac have discrepancies when analyzing rarefied data?

One point to note is that rarefaction carries the assumption that microbiota within samples are homogeneous and randomly distributed. However, this assumption is only valid if proper sampling protocols are observed [39]. A combination of unevenly sampled OTUs and distantly related OTUs will contribute to UniFrac's variability when OTUs are ultimately rarefied. Distance matrices between samples will be affected when rare OTUs are left out during the rarefaction processes. It becomes intuitive to see how similar samples may grow dissimilar from each other through unweighted UniFrac on rarefied samples as the number of unshared branches increases as OTUs are removed.

Table 2.1: **Original abundance of taxa and rarefied abundance of taxa.** This data was simulated to demonstrate how rarefaction can change the distances reported by the unweighted UniFrac metric.

| OTU.ID    | A   | B   | A<br>R1 | B<br>R1 | A<br>R2 | B<br>R2 |
|-----------|-----|-----|---------|---------|---------|---------|
| OTU.16340 | 52  | 1   | 8       | 1       | 12      | 1       |
| OTU.17317 | 17  | 4   | 3       | 4       | 5       | 4       |
| OTU.20    | 70  | 18  | 14      | 18      | 20      | 18      |
| OTU.37867 | 59  | 10  | 9       | 10      | 11      | 10      |
| OTU.37990 | 7   | 59  | 0       | 59      | 1       | 59      |
| OTU.38187 | 646 | 115 | 132     | 115     | 122     | 115     |
| OTU.38446 | 6   | 8   | 0       | 8       | 1       | 8       |
| OTU.45429 | 218 | 6   | 55      | 6       | 49      | 6       |

242

*Distance<sub>A:B</sub> for Rarefaction 1*

$$\begin{aligned}
 Distance_{A:B} &= \frac{\sum UnsharedBranches}{\sum TotalBranches} \\
 &= \frac{(0.2889 + 0.1706)}{1.12} \\
 &= \frac{0.5281}{1.12} \\
 &= 0.4175
 \end{aligned}$$

*Distance<sub>A:B</sub>forRarefaction2*

$$\begin{aligned} \text{Distance}_{A:B} &= \frac{\sum \text{UnsharedBranches}}{\sum \text{TotalBranches}} \\ &= \frac{0}{1.12} \\ &= 0 \end{aligned}$$

With rare OTUs and long branch lengths in the phylogenetic tree (Fig. 2.4), the Unweighted UniFrac distance metric on rarefied data is highly variable, declaring the samples A and B identical (distance of 0) with 1 rarefaction, and different with another (distance of 0.4175), as demonstrated in Table 2.1 and the calculations above.

While an improvement on unweighted UniFrac, weighted UniFrac can overweight differences between large proportional abundances and underweight differences between small proportional abundances. If one bacterial taxa increased in proportion from 5/1000 to 10/1000 and another taxa increased in proportion from 95/1000 to 100/1000, they would have the same weight in weighted UniFrac. However, the first taxa has doubled in proportion between samples, and this is much more biologically significant than the change in proportional abundance in the second taxa. Additionally, it does not account for how the counts add up to a constrained sum determined by the sequencing machine model. Because the sum is constrained, as an example, an increase in growth of one taxa can make the data look like there is a decrease in abundance in other taxa, even if in reality the population of the other taxa stayed the same.

Here we explore some alternatives to unweighted and weighted UniFrac, and discuss their merits and shortfalls.

## 2.0.8 Information UniFrac

The difference in information content between low proportional abundances (which make up the bulk of microbiome data) is generally higher than the difference between the proportional abundances themselves, potentially allowing scientists to differentiate groups with subtle differences.

Near the 0, 0 point in Fig. 2.5, the proportional abundances are low. Here there is higher differentiation between weights of different pairs of low proportional abundances for information UniFrac, as shown by the higher slope of the curved graph. The centered ratio UniFrac (not depicted) depends on the geometric mean of the taxonomic abundances, and would have a different slope in the weight graph depending on how evenly the abundances were distributed.

## 2.0.9 Tongue and buccal mucosa comparison

We next explore two other datasets, one with a defined difference between groups (tongue dorsum compared to buccal mucosa), and one with an outlier that is only apparent when analyzed by certain dissimilarity metrics.

Fig. 2.6 shows a principal coordinate analysis plot with four different metrics: unweighted UniFrac, weighted UniFrac, information UniFrac, and centered ratio UniFrac. We observe



that the difference in the microbiome between the human tongue and buccal mucosa are well defined by all metrics (Fig. 2.6), since all of the weightings show separation between the samples according to body site. We conclude from (Fig. 2.3) that weighted UniFrac, information UniFrac, and centered ratio UniFrac do not tend to show spurious separation in uniform data sets to the degree that unweighted UniFrac does, while reliably separating samples in data with a defined difference between groups.

## 2.0.10 Breast milk Data

Fig. 2.7 is a principal coordinate analysis of a 16S rRNA gene sequencing experiment done on microbiome samples from breast milk [111]. Breast milk samples were collected and the V4 region of the 16S rRNA gene was sequenced. One of the patients who provided a sample had an active infection, producing a sample that consisted of 97% *Pasteurella*. We noted that this sample was not distinct in unweighted and weighted UniFrac because the distance from the *Pasteurella* branches of the phylogenetic tree to the root of the tree (rooted by midpoint) were not particularly short or long, measuring at just over the 3rd quartile of all root-to-leaf distances. In addition, the *Pasteurella* leaves shared a clade with many other taxa.

The reason the infected sample in the breast milk study is so distinct from the rest of the samples in Information UniFrac and Centered Ratio UniFrac is because of the weighting. The infected sample was 97% *Pasteurella*, while the other samples generally had 15-20% each of *Staphylococcus* and *Pseudomonas*, and little or no *Pasteurella*. Unweighted UniFrac does not differentiate between high and low abundance. Weighted UniFrac does, placing the infected sample in the bottom right corner of that plot. Information UniFrac weights everything in the infected sample close to zero, as taxa are present in either very high or very low abundance, while weighting *Staphylococcus* and *Pseudomonas* in the other samples highly (around 0.4) due to their 15-20% abundance. Centered ratio UniFrac recognizes that the infected sample has a taxonomic abundance very far from the geometric mean abundance. For these reasons information and centered ratio UniFrac are more adept at picking up outliers with uneven distributions, even if the taxa are shared by other samples.

## Discussion

As shown in the tongue and buccal mucosa data set, unweighted UniFrac is perfectly sufficient for data sets with a notable difference. However, in data sets with no difference or a very small difference between groups such the uniform tongue dorsum data set, unweighted UniFrac is the least reliable and we found that it may produce wildly different results depending on rarefaction and sequencing depth. This can result in spurious groups, or inclusion of samples in the wrong groups.

We found weighted UniFrac, information UniFrac, centered ratio UniFrac, and Bray-Curtis methods to be more reliable choices. We suggest that investigators use several methods as they can detect outliers in different circumstances. When an outlier is detected by any metric, an investigation is warranted, as with our example in the breast milk data set.

In summary, with the addition of information UniFrac and centered ratio UniFrac, biologists have more tools at their disposal to prevent spurious interpretations, detect outliers, and

ultimately understand their data better.

315

## **Acknowledgments**

316

Thanks to Camilla Urbaniak for providing the data from her breast milk study [111].  
unifrac

317

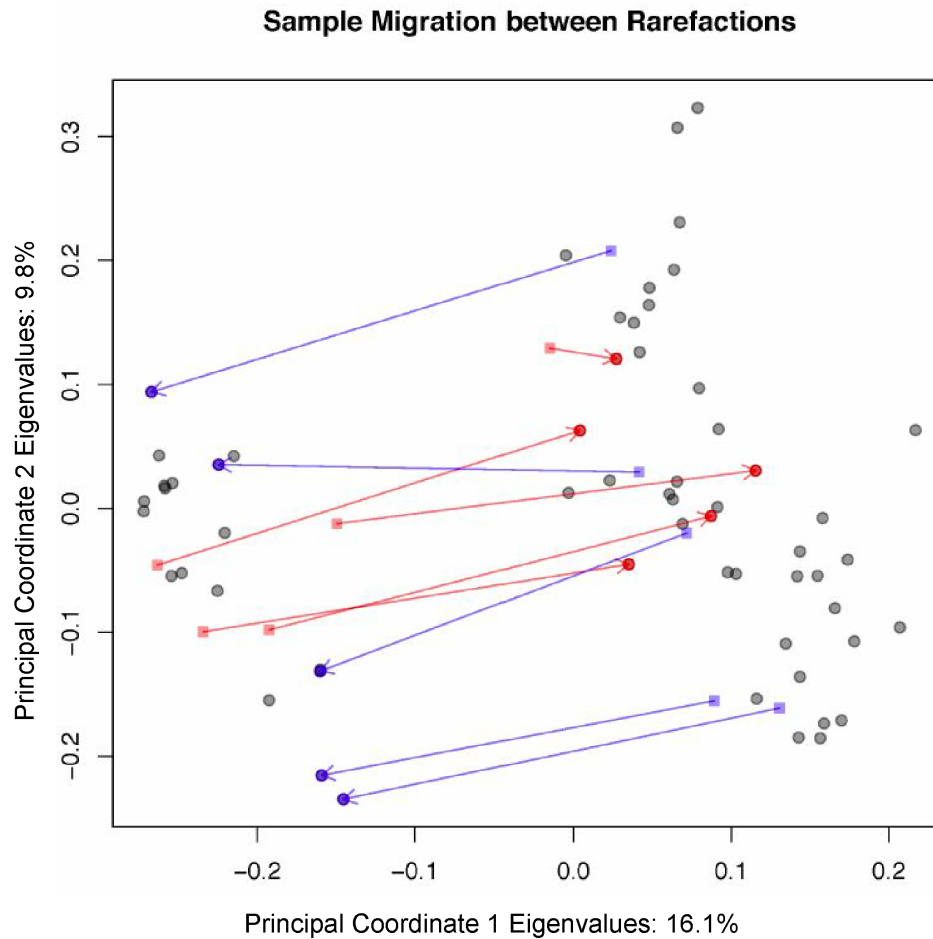


Figure 2.2: **Sample migration in different rarefactions, plotted on principal coordinates, measured with unweighted UniFrac.** Red samples have moved from the left cluster to the right cluster between rarefactions. Blue samples have moved from the right cluster to the left. Samples are taken from the tongue dorsum body site from the Human Microbiome Project database. If the experiment were run once, one might mistakenly assume that there are two clusters of data, however, the inconsistent sample membership of the two groups between rarefactions proves the clustering irreproducible. Note that the variance explained by the first and second coordinate is merely 16.1% and 9.8% respectively, indicating that the data is rather spherical, even though the points on the plot appear to show two separated clusters.

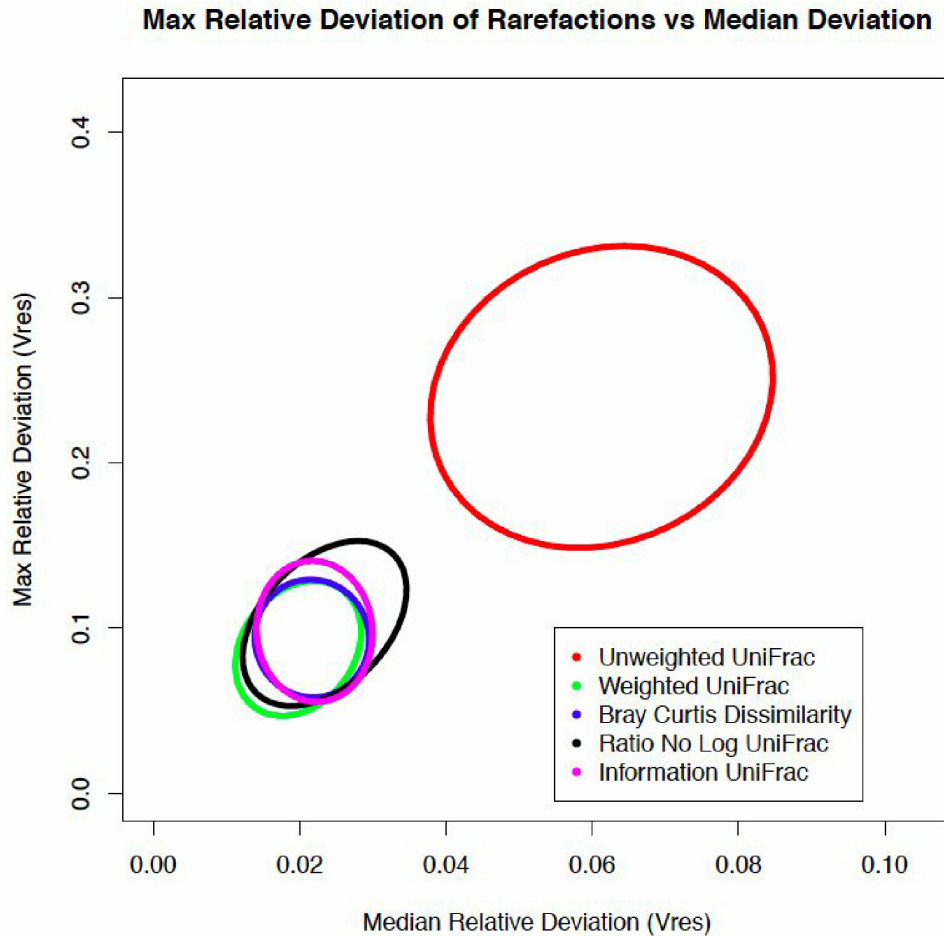


Figure 2.3: **Maximum relative deviation of rarefactions versus median deviation for traditional and non-traditional microbiome dissimilarity metrics.** Sixty samples from the tongue dorsum were taken from the Human Microbiome Project [110], and rarefied 100 times. The maximum relative deviation was plotted against the median relative deviation of the rarefied data, and ellipses were drawn at the 95% confidence interval, around the cloud of points for each metric. A higher maximum and median deviation indicates lower reproducibility of results between rarefaction instances. Both the maximum relative deviation of rarefied data and the median relative deviation of rarefied data are greater in unweighted UniFrac than in weighted UniFrac, Bray Curtis distance, centered ratio UniFrac, and information UniFrac.

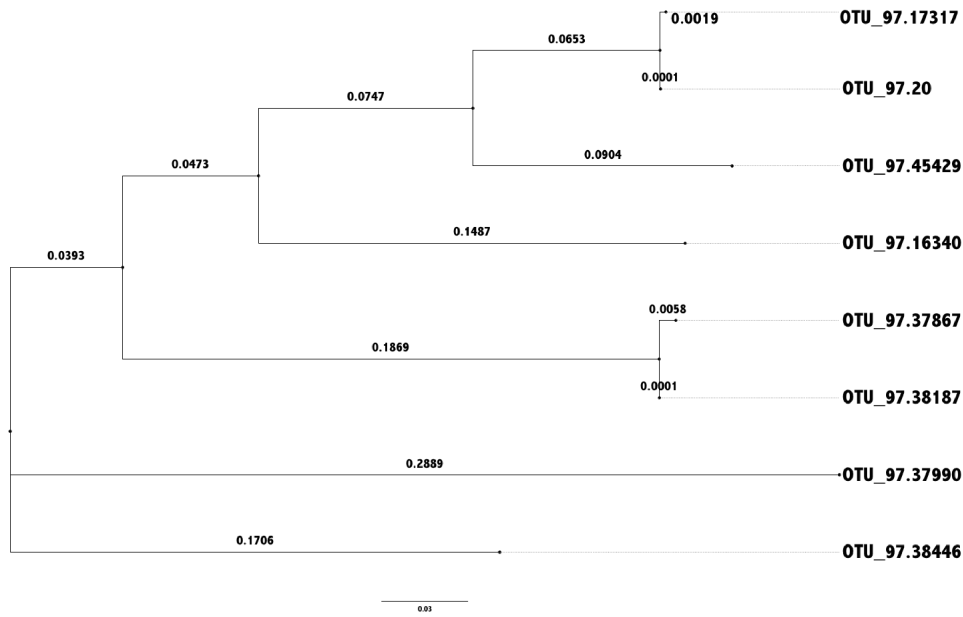


Figure 2.4: **Phylogenetic tree with long isolated branches.** Variation in different rarefactions of data in unweighted UniFrac analysis is exacerbated by the presence of long isolated branches in the phylogenetic tree.

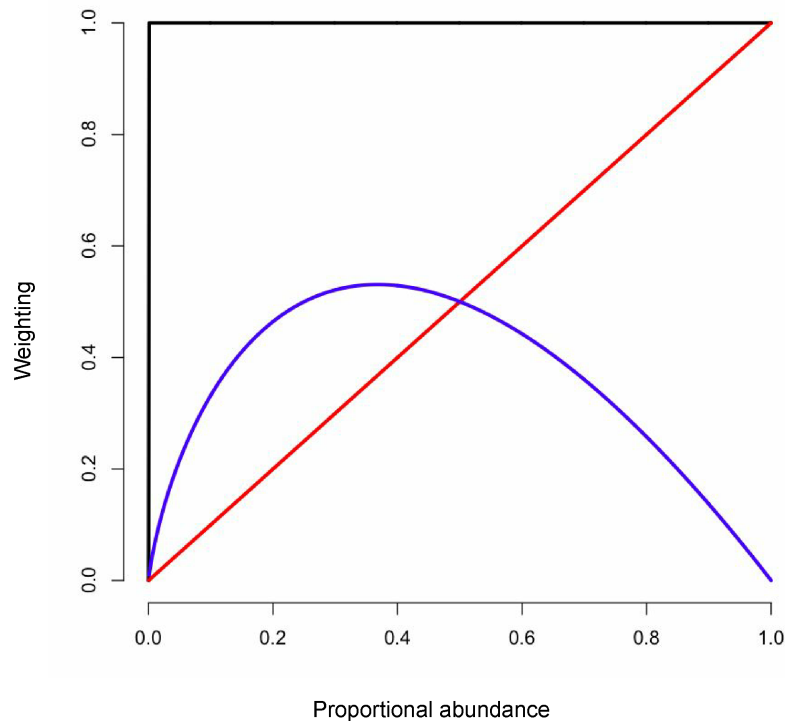
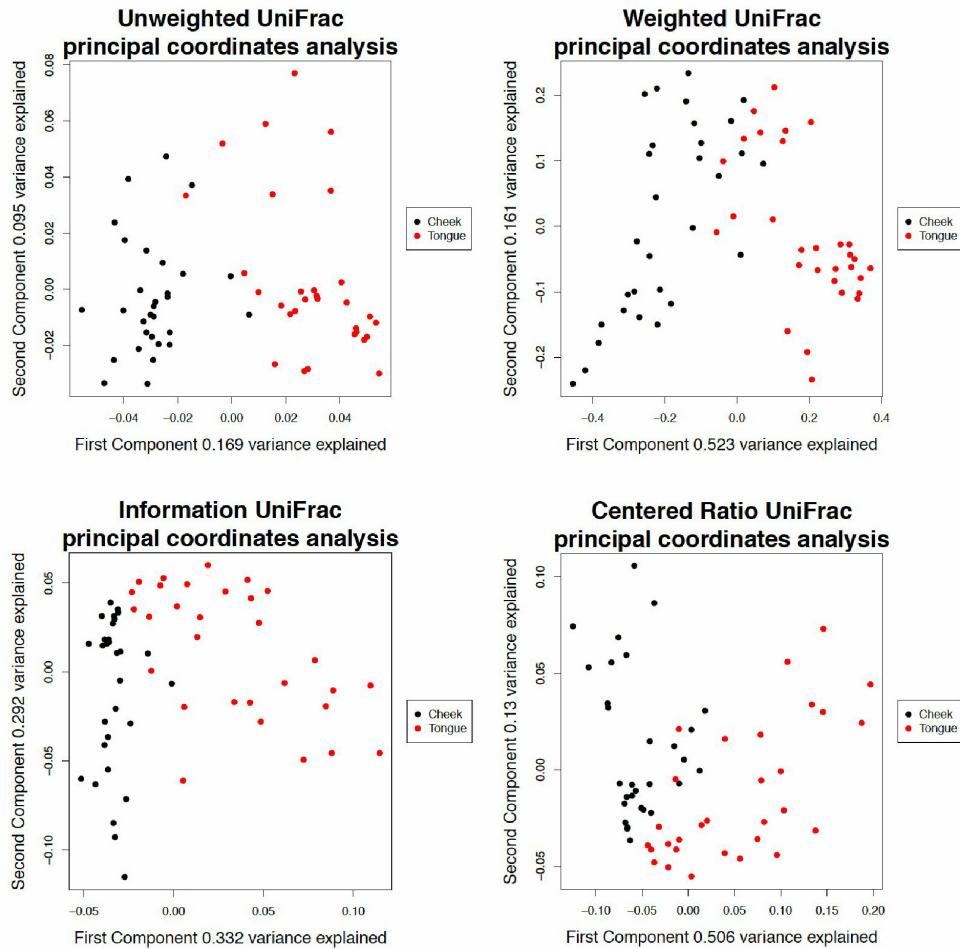
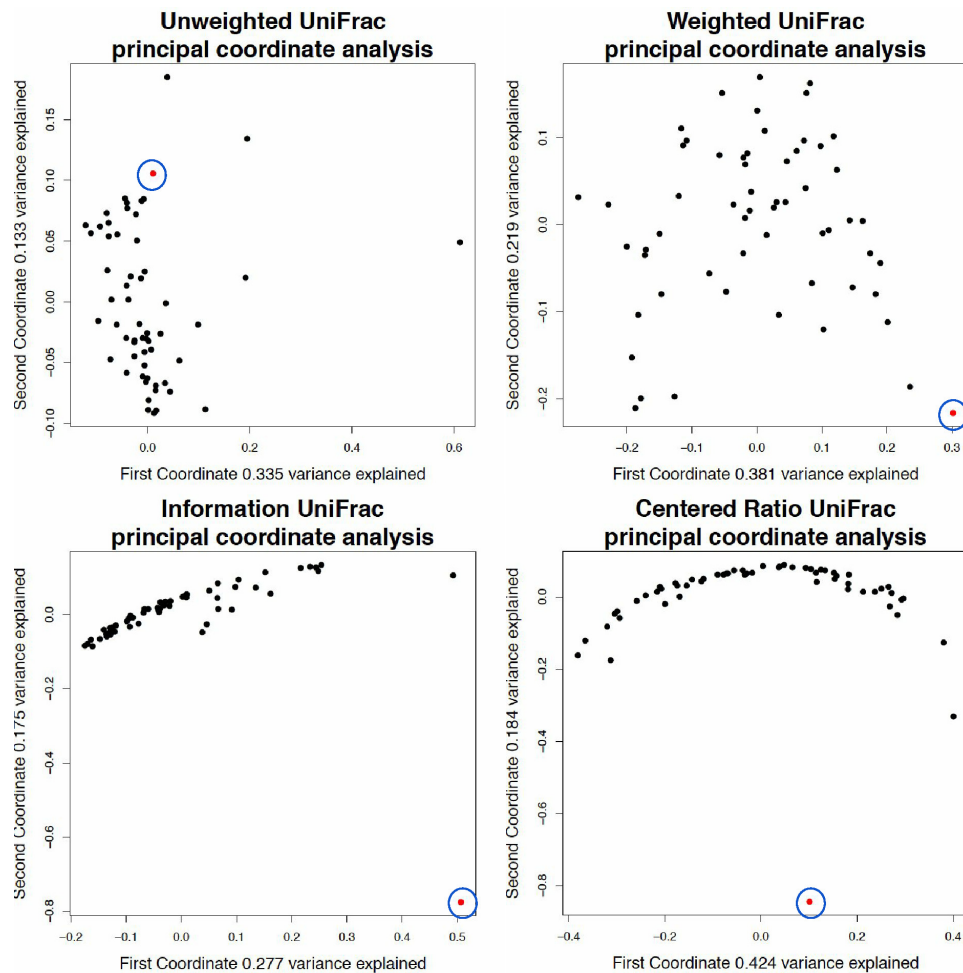


Figure 2.5: **UniFrac weights.** Each UniFrac weighting is plotted with the corresponding proportional abundance. The black line is unweighted UniFrac, the red line is weighted UniFrac, and the blue line is information UniFrac.



**Figure 2.6: Analysis of tongue and buccal mucosa data using different UniFrac weightings.** A principal coordinate analysis of a 16S rRNA experiment done on samples from the tongue and buccal mucosa, selected from the Human Microbiome Project [110]. All weightings show separation between the samples by body site. Note that the variance explained by the first and second principal coordinate axis is higher than in the tongue-tongue data set from Figure 2, which had 16.1% and 9.8% variance explained, respectively.



**Figure 2.7: Analysis of breast milk data using different UniFrac weightings.** A principal coordinate analysis of a 16S rRNA experiment done on samples from a 16S rRNA experiment on breast milk. The circled sample is infected with 97% *Pseudomonas*, compared to 15-20% in the other samples.

# Chapter 3

## The human microbiome and atherosclerosis

### 3.1 Introduction

In 2010, over half a million deaths in the United States were due to cardiovascular disease [70]. Atherosclerosis, a chronic disease in which fatty plaques build up in arteries leading to blood clot formation and blockage of the blood stream is a strong contributor to cardiovascular disease. Modifiable risk factors for atherosclerosis include obesity, smoking, physical inactivity, stress, and more. By imaging the carotid artery for atherosclerotic plaques, Spence [100] found that, while the progression of most patients' atherosclerosis is predicted by their risk factors, some patients present with few factors, and yet their atherosclerosis progresses. Conversely, other patients have many risk factors and their atherosclerosis regresses. Patients on the extreme ends of this spectrum, who exhibit unexplained progressive atherosclerosis or unexplained regressive atherosclerosis will be the focus of this chapter.

Currently the main therapies for atherosclerosis and other cardiovascular disease include a diet low in cholesterol, the cessation of smoking, and the administration of statins to inhibit cholesterol production in the liver. However, in some patients, such as those in this study, these interventions may not be effective. A characterization of the gut and oral microbiome and their effect on cardiovascular disease is necessary to explore atherosclerosis risk factors which are beyond the patient's explicit control. With this knowledge, the mortality and morbidity of victims of cardiovascular disease may be improved.

#### 3.1.1 Atherosclerosis risk

Prior to the advent of measuring atherosclerosis progression directly by carotid ultrasound, pioneered by Dr. J. David Spence, clinicians generally used the measure of intima-media thickness. Tunica intima and tunica media are the two innermost layers of the carotid artery. A greater thickness is due to medial hypertrophy resulting from high blood pressure, and high intima media thickness is correlated with stroke but only weakly predictive of heart attacks [101].

[TO DO: EXPLAIN R SQUARED]





Figure 3.1: **Carotid Ultrasound showing intima-media thickness**, picture borrowed from Harley Street Cardiologists 2014 London Cardiovascular Clinic. The intima is the innermost layer of the artery bordering the lumen, and the media is the layer just outside that.

Dr. Spence’s carotid ultrasound technique measures the plaque area. Carotid plaque size appears to have some concordance with both the risk of stroke and myocardial infarction (more so for the latter than the former), while having a much improved correlation with a patient’s risk factors [101]. The quartile of carotid plaque size a patient is in correlates with the risk of a stroke or myocardial infarction. Traditional risk factors are able to predict carotid plaque area and volume with  $R^2 = 0.52$  (compare with  $R^2 = 0.15$  for intima-media thickness), using the regression model developed by Dr. J. D. Spence [100]. The regression model took the following risk factors into account: Age, sex, diabetic status, total cholesterol, triglycerides, HDL cholesterol, LDL cholesterol, systolic pressure, diastolic pressure, lipid or blood pressure medication, and smoking.

Most patients’ actual carotid plaque sizes were relatively close to their predicted carotid plaque sizes, however, some patients presented with plaques much smaller than expected, termed “unexplained atherosclerotic regression” and other patients presented with plaques much larger than expected (“unexplained atherosclerotic progression”). The experiments described in this thesis compare the gut and oral microbiota of the patients with extreme (top 10



Figure 3.2: **Carotid Ultrasound showing plaque area measurement**, picture borrowed from Spence [101]

|                      | Baseline    |           |
|----------------------|-------------|-----------|
|                      | Unexplained | Protected |
| Age                  | 46          | 82        |
| Sex                  | Male        | Male      |
| Diabetic             | No          | No        |
| Total cholesterol    | 5.02        | 4.04      |
| Triglycerides        | 1.54        | 2.0       |
| HDL cholesterol      | 1.03        | 0.72      |
| LDL cholesterol      | 3.29        | 2.40      |
| Systolic pressure    | 170         | 167       |
| Diastolic pressure   | 105         | 85        |
| Smoking (pack-years) | 30          | 195       |
| Plaque area          | 704         | 173       |
| Residual score       | 3.68        | -2.31     |

Figure 3.3: **Risk factors, predicted carotid plaque area, and actual plaque area for two patients.** Two patients' predicted carotid plaque area based on their risk factors, and their actual plaque area. One patient, a young non-smoker, has a plaque much larger than predicted, and the other patient, an old smoker with a high LDL/HDL ratio has a plaque much smaller than predicted. Both patients are in the top tenth percentile for unexplained atherosclerosis progression or regression. Figure borrowed from Spence [100].

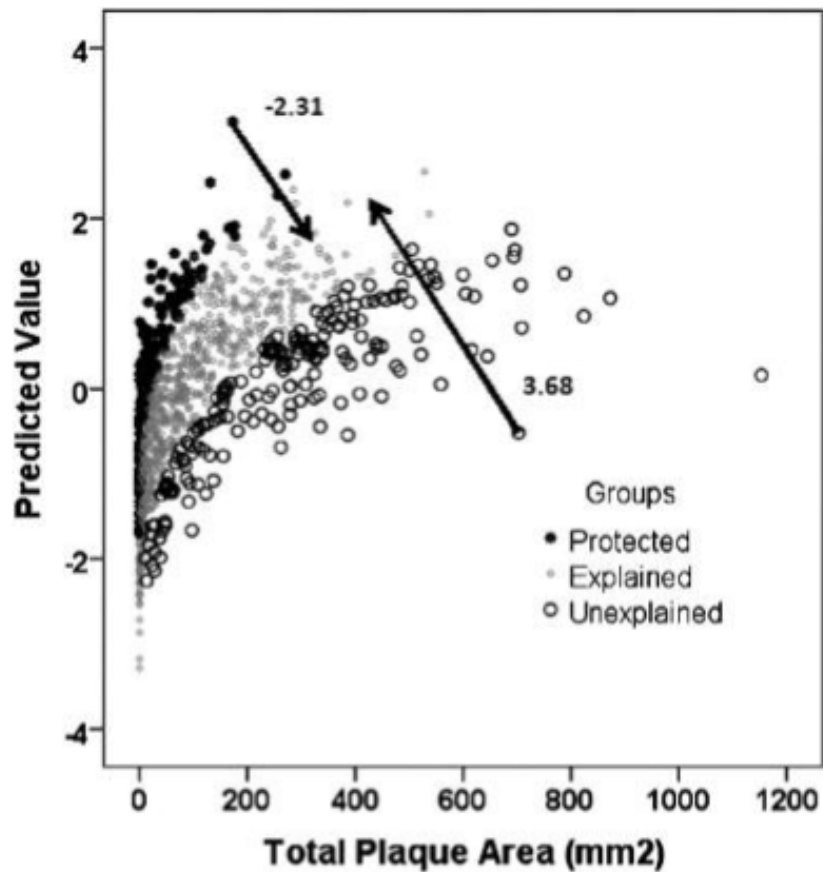


Figure 3.4: **Predicted risk vs. total plaque area.** The two arrows represent the residual measures of the two patients in the previous figure. Figure borrowed from Spence [100].

### 3.1.2 Metabolic potential of gut microbiota

A large contributor to a person's metabolism is the bacteria in their gut. Studies in both mouse and human show that the microorganisms living in each individual [41] can produce hormones [102] and vitamins [12] and even affect brain chemistry [20]. The following sections will focus on evidence that the gut microbiota contributes to atherosclerosis risk.

#### Mouse

In mice, it has been found that genetically obese mice [107] and mice who are obese due to their diet [108] have a lower ratio of members of the phyla Bacteroidetes to members of the phyla Firmicutes, compared to lean mice. Additionally, if gut bacteria from the obese mice are transplanted into lean mice, the lean mice gain weight, despite eating the same diet [107]. Furthermore, germ-free mice appear to be protected from the effects of diet-induced obesity [5], and benefit from the increased life span associated with caloric restriction [38].

#### Human

Human obese patients also have a lower Bacteroidetes to Firmicutes ratio [107] in most studies. Certain strains of Firmicutes decreased in proportion as patients dieted, relative to the proportion of the other strains present, increasing the Bacteroidetes : Firmicutes ratio [26]. Many metabolites are produced by gut bacteria, including hippurate, phenylacetylglycine, dimethylamine [117], and TMAO [103], the latter of which will be examined in depth in the next section.

#### TMAO: an example of gut bacteria affecting atherosclerosis risk

[TO DO: ADD TMAO MECHANISM OF ACTION]

Products of gut metabolism have been shown to affect atherosclerotic progression. For example, high levels of trimethylamine N-oxide (TMAO) as measured in blood plasma and urine have been associated with higher atherosclerosis risk in humans [103]. The same association has been reported in mouse models where trimethylamine (TMA) is formed from free choline [112] and phosphatidylcholine [114] by bacteria in the mouse gut. TMA that enters the bloodstream from the intestinal tract is later converted to TMAO in the liver. In humans, TMAO has been shown to be produced from L-carnitine, present in red meat [50], and dosing human patients with antibiotics decreased their TMAO levels [103].

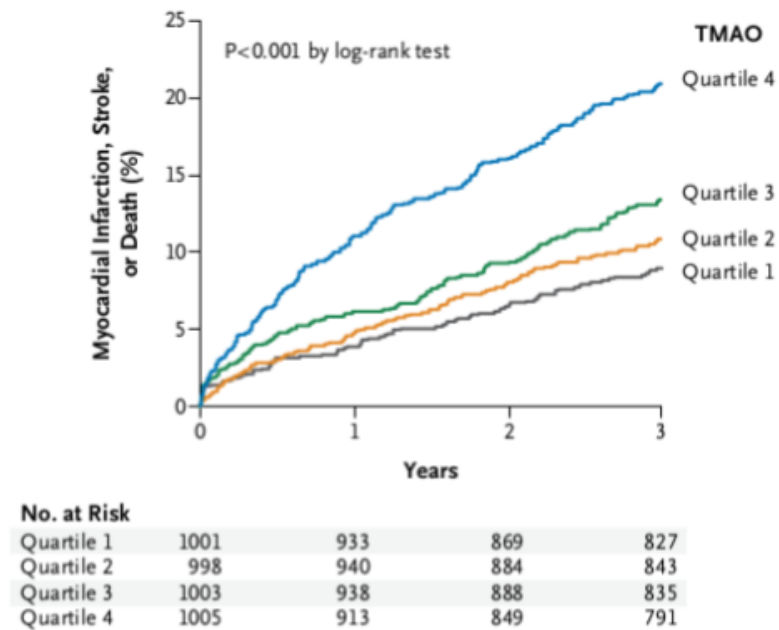


Figure 3.5: **Kaplan-Meier estimates of major adverse cardiovascular events, according to the quartile of TMAO level.** “Data are shown for 4007 participants in the clinical-outcomes study. The P-value is for all comparisons” Figure borrowed from Tang [103]. Each line of this graph represents data from patients in one of four quartiles for TMAO levels. The risk of myocardial infarction, stroke, or death is more than double for patients who are in the top 25% for TMAO levels, compared to patients in the bottom quartile.

### 3.1.3 Metabolic potential of mouth microbiota

There are few studies done on the metabolic contribution of oral microbiota, compared to that of the intestinal microbiota. However, it is known that the mouth can serve as a microbial reservoir for the gut [22].

Many unexpected systemic effects are instigated by, or correlated with, bacteria in the mouth. Plaque containing microbiota may arrive at the bloodstream through ulceration in the gums as a result of gum disease [88] or dental work. Patients have been found to be more at risk for heart attacks and strokes in the 4 weeks immediately following dental treatment [69].

Oral microbiota are certainly associated with gum health [67], and cardiovascular disease has been correlated with gum infection [7]. Patients with poor gum health are more likely to develop diabetes and diabetic complications [10], rheumatoid arthritis [89], and even Alzheimer’s [47]. Pregnant patients with periodontal disease are more likely to experience undesirable pregnancy outcomes [45], such as low birth weight, preterm birth, and pre-eclampsia.

Further proof that oral bacteria may contribute to atherosclerotic plaques is that the combined abundance of *Streptococcus* and *Veillonella* appear to be correlated between atherosclerotic plaques and saliva sample taken from the same patient [52].

[TO DO: TALK ABOUT THE ORAL MICROBES IN PLAQUE]

### **3.1.4 Bacteria and atherosclerotic plaques**

A 2011 paper by Koren et al. characterized the microbiota present in the atherosclerotic plaques of 15 patients. They found that *Pseudomonas luteola*, three types of *Staphylococcus*, three types of *Propionibacterineae*, and one type of *Burkholderia* were highly abundant in plaques but not in the oral or gut microbiota. The amount of DNA present in their samples which coded for the 16S ribosomal RNA subunit (indicative of the quantity of bacteria present in the plaque) correlated very strongly with the number of leukocytes in the plaques [52]. The inflammation response that comes along with increased bacteria in the atherosclerotic plaques plays an important role in plaque growth. Vascular cell adhesion molecule 1 (VCAM-1) binds to these leukocytes, and mice who do not express VCAM-1 exhibit slowed atherosclerotic plaque formation, compared to control mice with the same diet and lipoprotein profiles [21]. Bacteria-related inflammation status may be a way to explain some of the unexplained atherosclerotic progression or regression.

### **3.1.5 Project proposal**

The main objective of the research I am proposing is to determine the differences in the microbiota of the intestinal tract and oral cavity, between the extreme unexplained progressive atherosclerosis patients and the extreme unexplained regressive atherosclerosis patients. The project will have three different dimensions of analysis. First, a cursory examination of the differential abundance of bacteria between the groups will be done. Then, a metagenomic analysis will follow, to determine the metabolic potential of the microbiome. Lastly, the metabolomics data will connect the differential transcriptomic data to differential metabolite production. All together, these analyses may provide a clearer picture of the processes that produce unexplained progressive and unexplained regressive atherosclerosis.

## **3.2 Methods**

## **3.3 Results**

## **3.4 Discussion**

## Chapter 4

# The human microbiome and non-alcoholic fatty liver disease

### 4.1 Introduction

Non alcoholic fatty liver disease (NAFLD) has been on the rise along with obesity, affecting a fifth to a third of the North American population [81]. Most people with NAFLD remain asymptomatic, however, in up to a third of patients NAFLD can progress to non-alcoholic steatohepatitis (NASH), causing inflammation and scarring in the liver, and decreasing the 5 year survival rate to 67% [82]. If we can shed some light on the process by which people progress from NAFLD to NASH, we might be able to find treatments to prevent NASH.

There are several known genetic factors that increase the risk of progression to NASH. The I148M variant of the Patatin-Like Phospholipase Domain Containing 3 gene (PNPLA3) correlates with a 3.2 fold increased risk of NASH from NAFLD when present homozygously, compared to to patients without the variant [99]. Additionally, mice with a toll-like receptor 4 knockout had lower lipid and injury accumulation markers when fed a methionine/choline-deficient diet which would normally induce steatohepatitis in wild type mice [86].

On the epigenetic level, genes are differentially methylated in advanced NAFLD compared to mild NAFLD. 11% of genes are differentially hypomethylated in advanced NAFLD (compared to 3% hypermethylated), leading to increased expression [71]. On a hormonal level, estrogen has been shown to inhibit fibrosis in rats [118]. On a metabolite level, Raman et al. found differences in the number of volatile organic compounds detected in patients with NAFLD compared to obese patients without NAFLD [84]. Reactive oxygen species have also been implicated in NASH due to their involvement in the mechanism of steatohepatitis-inducing drugs [9].

A 2001 paper performed C-D-xylose-lactulose breath tests and measured tumor necrosis factor alpha levels to determine presence of bacterial overgrowth, and found increased bacterial overgrowth in 22 patients with NASH compared to 23 healthy controls [115]. Some papers claim a link between ethanol-producing gut bacteria and NAFLD [120] [46], however, no multiple test correction was performed in these studies.

Several papers have already been published in the literature on the topic of NAFLD and the gut microbiome:



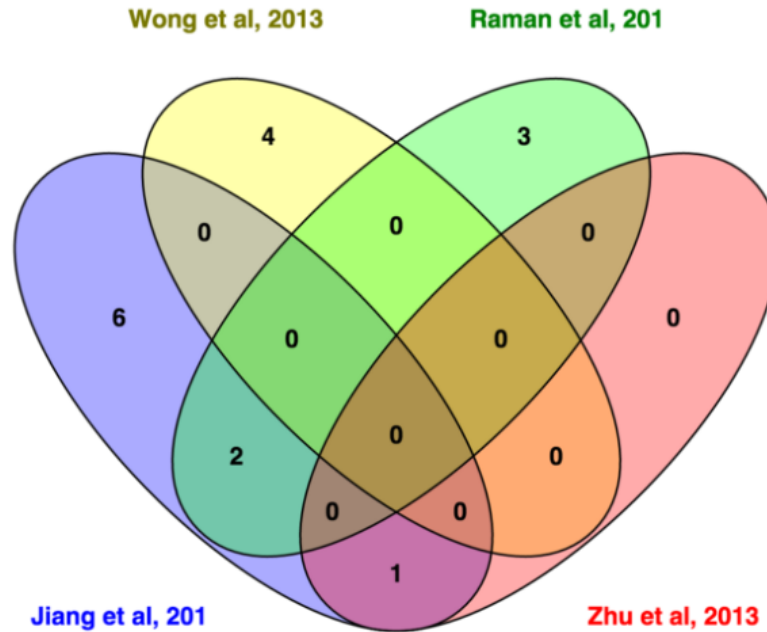


Figure 4.1: **Venn diagram of genus found to be differentially abundant by different studies between NASH/NAFLD and healthy controls.** Boursier et al 2015 is not included as they reported a p-value of less than 0.05 for the *Bacteroides* genus only, which was not reported in any of the other studies. Only 3 out of the 16 genus claimed to be differentially abundant were the same in two studies: *Escherichia* was found in the Zhu [120] and Jiang [46] studies, and *Lactobacillus* and *Oscillibacter* were found in the Jiang [46] and Raman [84] studies.

- Jiang et al, 2015 [46] compared 53 NAFLD patients with 32 healthy controls
- Zhu et al, 2013 [120] compared 16 non-obese controls, 25 obese patients, and 22 NASH patients
- Raman et al, 2013 [84] compared 30 NAFLD patients with 30 healthy controls
- Wong et al, 2013 [116] compared 16 NASH patients with 22 healthy controls
- Boursier et al, 2015 [11] compared 30 patients with F0 or F1 fibrosis to 27 patients with F2 or greater fibrosis, 35 of which had NASH

Of these, only Raman et al [84] reported using a multiple test correction.

These five studies do not form a consistent story about the gut microbiome and NAFLD. We hope to run our own analysis rigorously, such that our results are replicable. Additionally, we are running a deeply sequenced metagenomic study, which hasn't been done in the past.

## 4.2 Methods

In total, 92 samples were collected: 41 from patients with NASH, 18 from patients with SS, and 33 from healthy controls.

[TO DO: include information about sample collection and exclusion criteria]

DNA extraction was performed with the E.Z.N.A.® Stool DNA Kit, and the protocol was followed with the addition of lysozyme with an extra 30 minute incubation at 37 degrees Celcius, between steps 2 and 3.

### 4.2.1 16S rRNA gene tag experiment

DNA was amplified by PCR according to the Earth Microbiome protocol [15], with the addition of barcodes so that all the samples could be sequenced in the same sequencing run [37]. The DNA was sequenced on the Illumina MiSeq platform with paired end 150 nucleotide reads, producing 34955148 reads in total.

Reads were overlapped with Pandaseq [68], clustered into Operational Taxonomic Units using UCLUST [28], and annotated with the SILVA database [83], producing a table of counts per operational taxonomic unit per sample. 16809756 reads (48%) were successfully overlapped and annotated with 232 OTUs. Differential abundance was analyzed using ALDEx2 [31].

### 4.2.2 Metagenomic experiment

A deep metagenomic sequencing experiment was performed using samples from 10 healthy controls and 10 of the patients with NASH. Samples from healthy patients were selected to exclude confounding factors, such as having a different country of birth, which could affect diet. Samples from NASH patients were selected for the strongest NASH phenotype.

The DNA was sequenced on the Illumina HiSeq platform, with single end 100 nucleotide reads. Samples were barcoded and sequenced on the same sequencing run. After sequencing, the reads were quality filtered and demultiplexed to separate the reads for each sample, yielding 1914714572 reads in total.

To annotate the reads, we used a two pronged strategy:

First, we created a reference library using the inferred taxa from the 16S rRNA gene tag experiment. For each genus the OTUs were annotated with, we randomly picked 10 strain genomes from the NCBI bacterial genome database. For genus where there were less than 10 fully sequenced representatives, we selected all genomes available. The library was made with 1134 genomes from 104 bacterial genus. The library was then clustered at 99% identity for each genus using CD-HIT [57] to decrease the number of sequences in the library from 3495887 to 2256844. Annotation was performed with the SEED database [75], and sequenced reads were mapped onto this library. Out of 1914714572 reads total, 585382507 (30.6%) were annotated by this method, over 5836 unique SEED hierarchy annotations. The code for the reference library creation and annotation is on GitHub.

Second, we assembled the reads per sample de novo using Trinity [42], producing 8847816 sequences, and removed sequences that matched our reference library with 90% identity as determined by BLAST [2], leaving 5876423 sequences. [FILL IN THIS] of these assembled sequences were successfully annotated with the SEED database [75], and sequenced reads were

mapped onto this. [FILL IN THIS] additional reads were annotated by this method, over [FILL IN THIS] unique SEED hierarchy annotations. The code for the custom assembly pipeline is on GitHub. The data from both prongs was amalgamated into a single table of counts per annotation per sample.

Differential abundance was analyzed using ALDEx2 [31].

### 4.2.3 MetaPhlAn

MetaPhlAn (Metagenomic Phylogenetic Analysis) [92] is a piece of software that allows one to infer the taxa present based on the metagenomic sequencing experiment. We used this to generate a count table per taxa per sample, and will compare it to our experimental results from the 16S rRNA gene tag sequencing experiment.

## 4.3 Results

### 4.3.1 16S rRNA gene tag experiment

[consider adding nice bar graphs]

No obvious structure or separation is evident from the principal components analysis in Fig. 4.3.1. Furthermore the variance explained by each principal component axis is not notably high, indicating a rather uniform data set.

A differential expression analysis performed with ALDEx2 yielded no significantly differentially abundant OTUs (Fig. 4.3.1). However, the effect size of each OTU in each comparison is correlated, and the regression line indicates that the effect sizes are higher in the Healthy vs. extreme NASH compared to the Healthy vs. SS or Healthy vs. NASH comparison.

### 4.3.2 Metagenomic experiment

#### MetaPhlAn

[TO DO: find proper 16S taxa assignments and compare effect sizes with MetaPhlAn results]

## 4.4 Discussion

Given the inconsistency in the five papers that have been published about NAFLD and the gut microbiome, we have performed our analysis with care in an effort to find true effects. We found that there was no significant difference between groups by sample clustering (Fig. 4.3.1) or at the level of the individual OTUs (Fig. 4.3.1).

There are several factors that would make such a study underpowered. First, the gut microbiome is highly diverse between individuals. This is compounded by the fact that the samples were taken from a diverse Toronto population, including people who immigrated from other countries who likely have different diets. Additionally, the nature of microbimoe data is that there are very many more variables (in the form of OTUs or annotated gene functions) than samples.

From Fig. 4.3.1, the correlation shows that even though there is not enough power to detect a significant difference, the difference from the healthy baseline are moving in the same direction through simple steatosis to nonalcoholic steatohepatitis to extreme NASH.

We hypothesize that there is a characterizable difference in the gut microbiome between patients prone to NASH and healthy controls. Further study with a higher sample size, a more homogenous population, and a greater phenotypic difference between groups may provide the statistical power required to detect the nature of this difference.

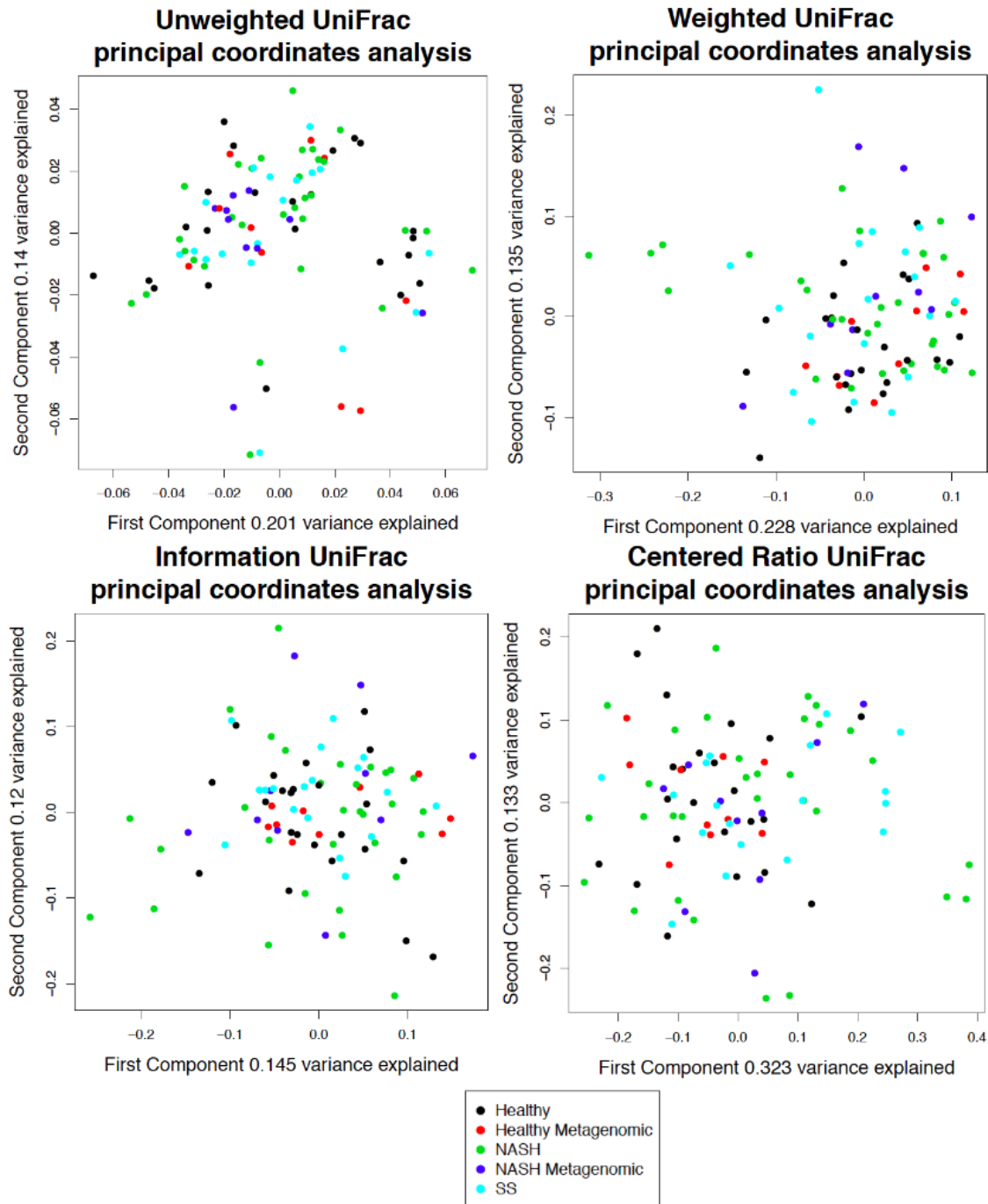


Figure 4.2: **Principal Components Analysis of 16S rRNA gene tag sequencing data with different UniFrac weightings.** Each point represents one sample, and the distances between the samples have been calculated using different UniFrac metrics, taking into account phylogenetic as well as abundance information. There is no obvious separation between groups by any of the UniFrac weightings. Furthermore the variance explained by each principal component axis is not notably high, indicating a rather uniform data set.

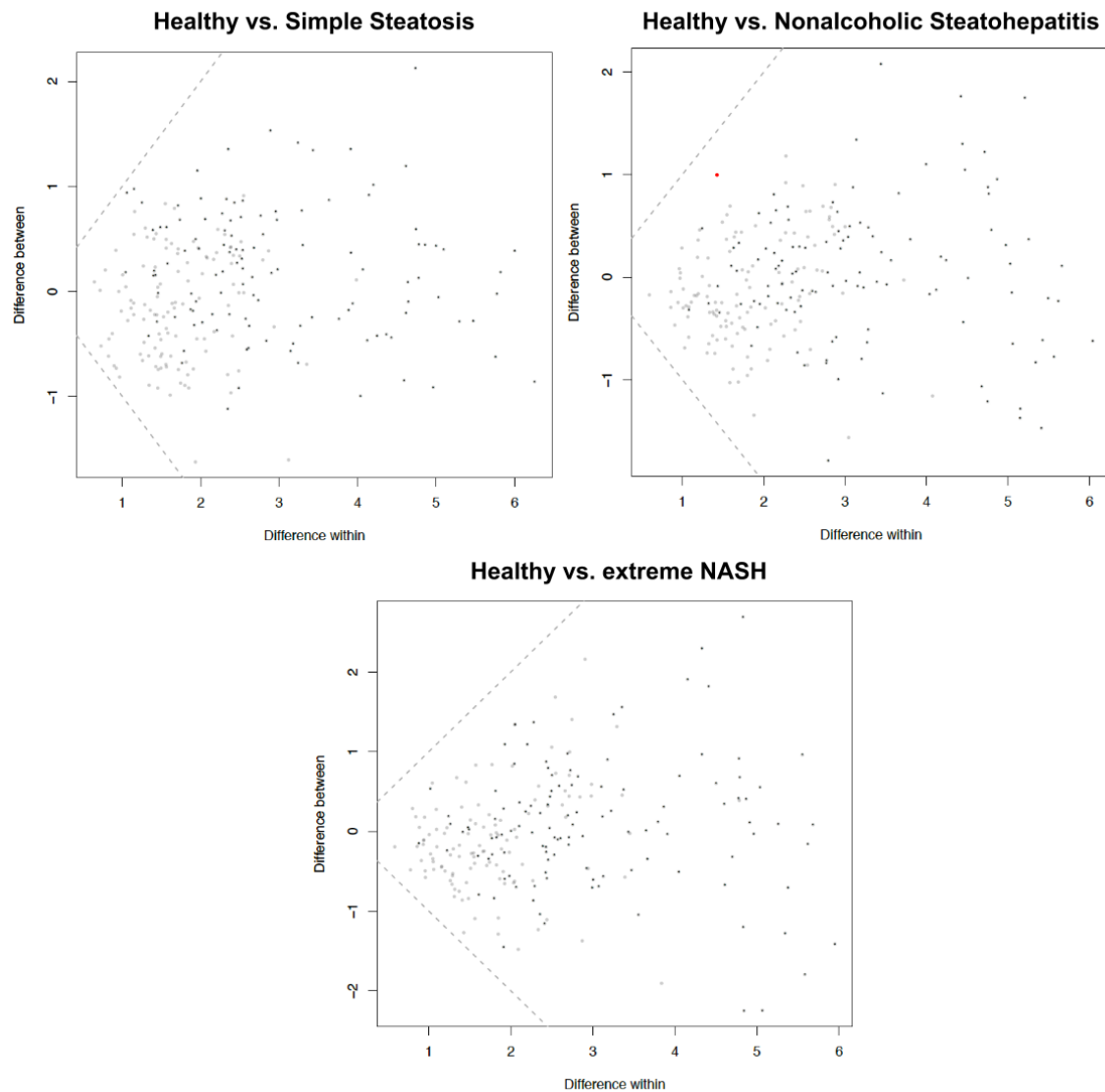


Figure 4.3: **Difference within vs. difference between groups.** Each point represents one OTU, and the differential abundance of that OTU within groups is plotted against the differential abundance between groups. None of the OTUs are more different between groups than within groups. The healthy samples used for these comparisons are the 10 healthy samples used for the metagenomic study. The extreme NASH samples used for these comparisons are the subset of the NASH patients selected for the metegenomic study.

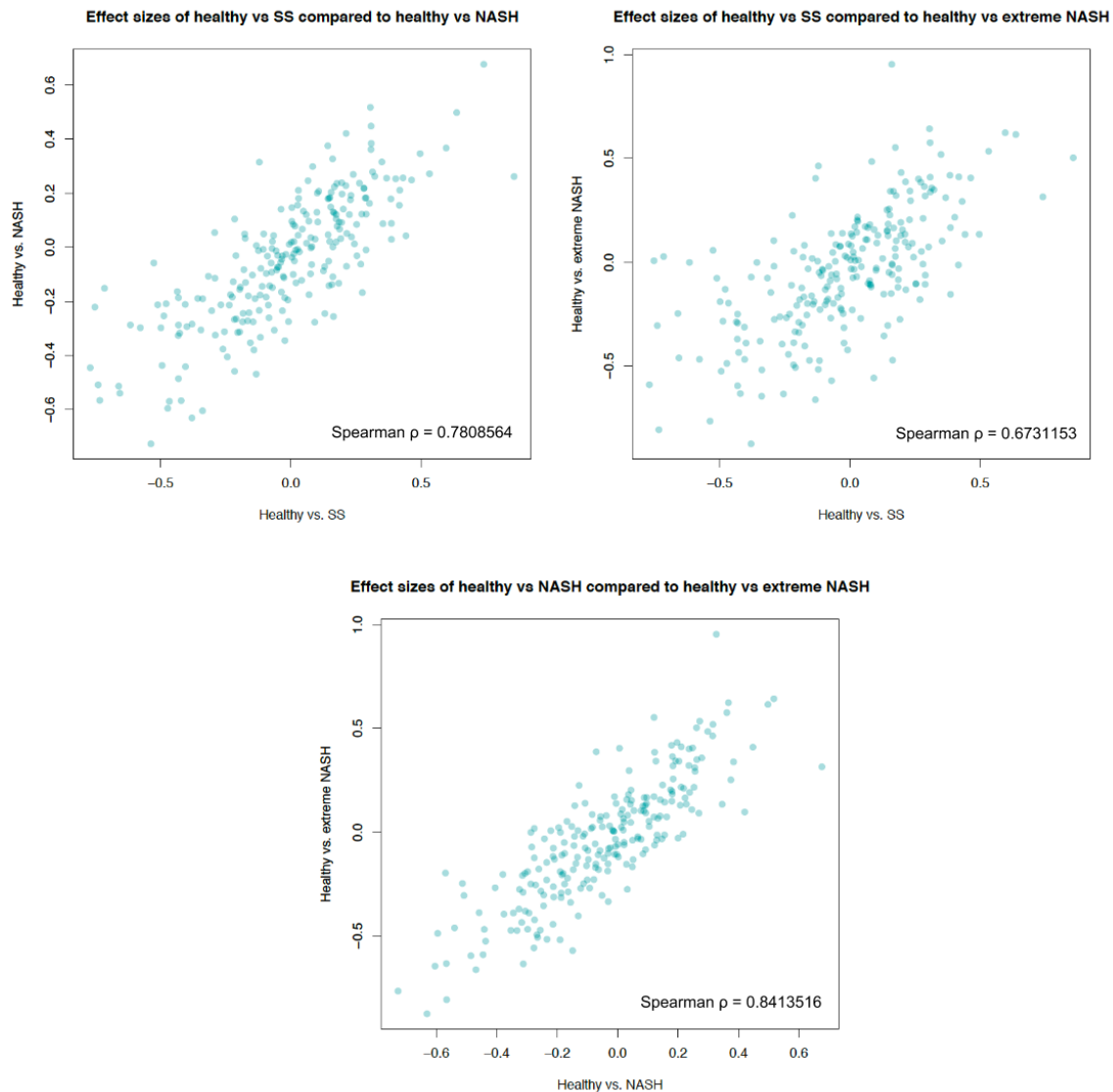


Figure 4.4: **Correlation in effect sizes of different group experiments.** Each point represents one OTU, and the effect size of that OTU in one comparison (for example, comparing the gut microbiome of healthy patients with patients who have simple steatosis) is plotted against the effect size of that OTU in another comparison. The healthy samples used for these comparisons are the 10 healthy samples used for the metagenomic study. The extreme NASH samples used for these comparisons are the subset of the NASH patients selected for the metagenomic study. The y intercepts of the regression lines are all between 0.005 and 0.025, close to zero. The median difference in the absolute effect sizes is -0.02076 for Healthy vs. NASH - Healthy vs. SS, 0.017070 for Healthy vs. extreme NASH - Healthy vs. SS, and 0.04256 for Healthy vs. extreme NASH - Healthy vs. NASH.

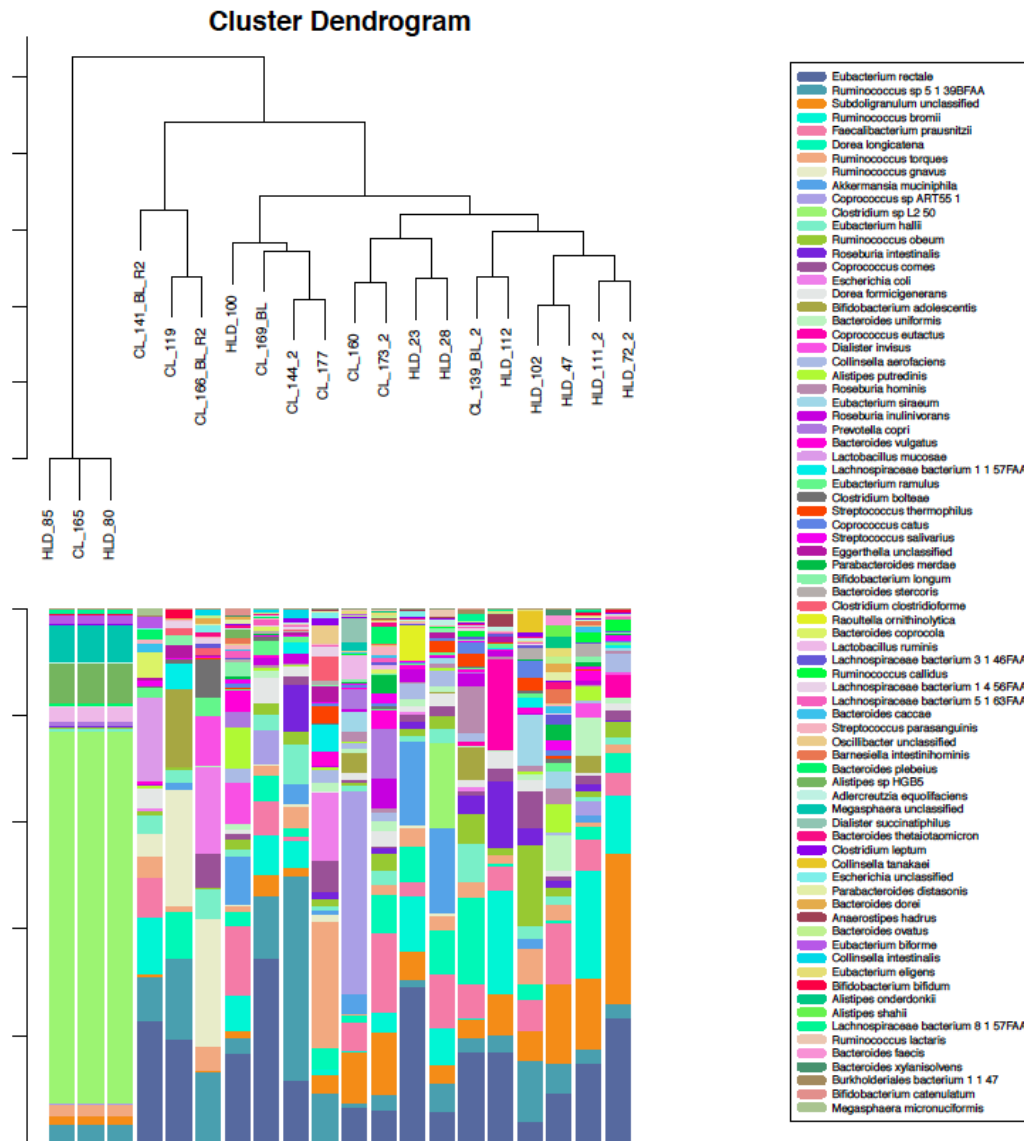


Figure 4.5: **Taxa barplot dendrogram derived from MetaPhlAn.** The metagenomic reads were input into MetaPhlAn to generate a count table. The taxa in the count table were filtered such that only taxa with at least 1% abundance in any sample was kept.



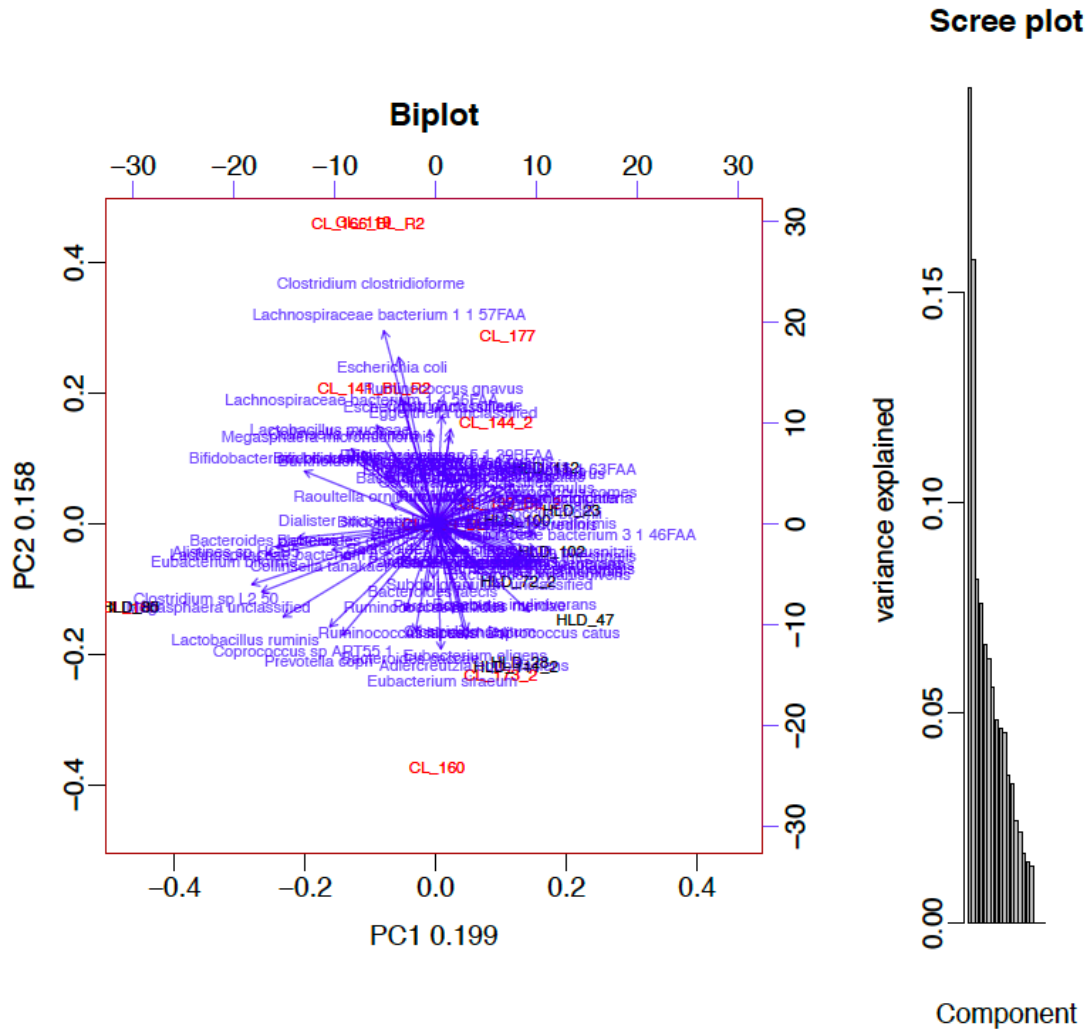


Figure 4.6: **Biplot derived from MetaPhlAn.** This biplot was generated from the count table inferred by MetaPhlAn, with taxa filtered such that only taxa with at least 1% abundance in any sample was kept. Note that the variance explained by the first and the second coordinate is not particularly high, indicating that there is not a clear unidirectional separation between groups. Samples from healthy controls are colored black while samples from patients with NASH are colored red.

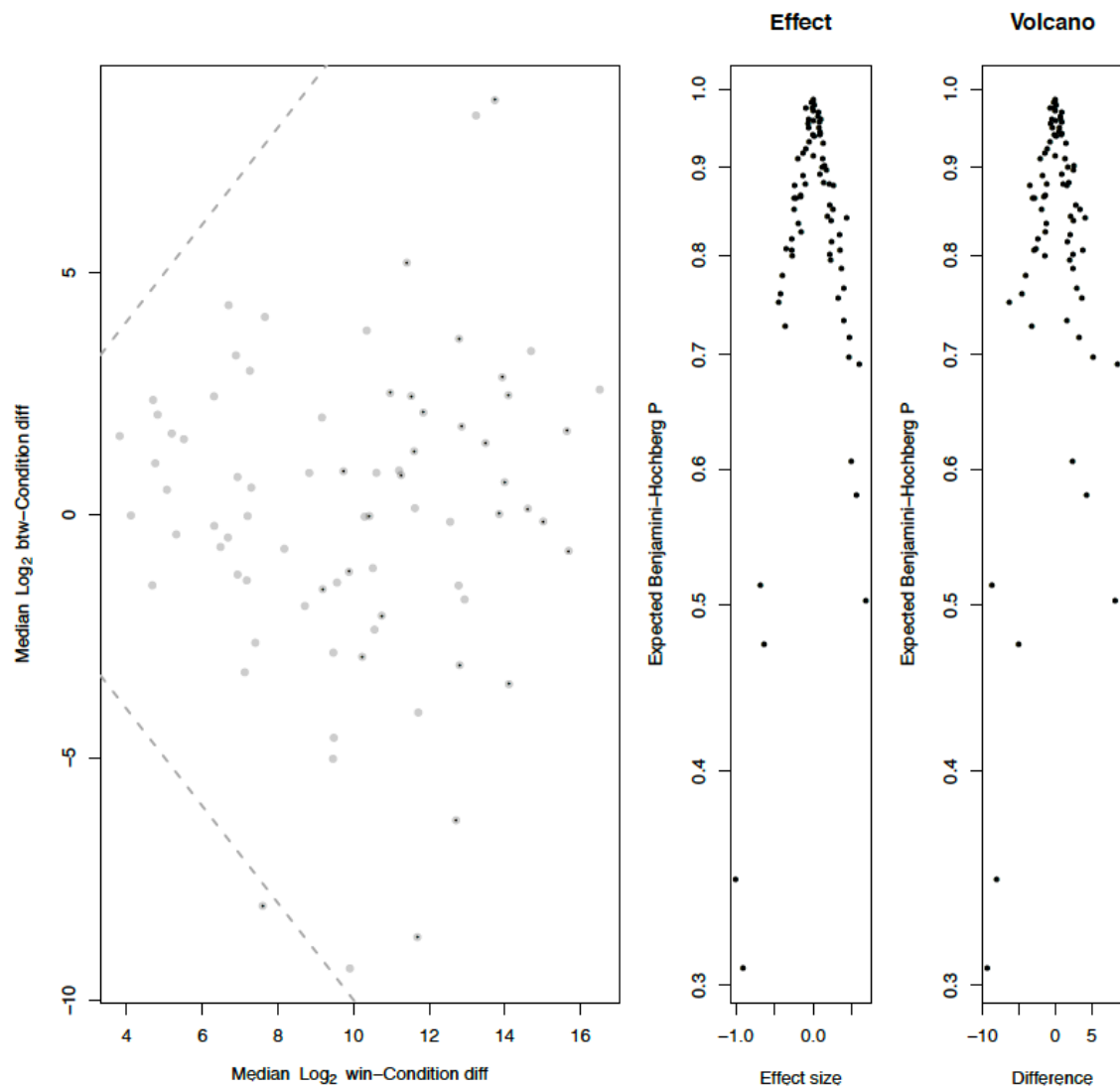


Figure 4.7: **Difference within groups vs. difference between groups per taxa, derived from MetaPhlAn.** This plot was generated from the count table inferred by MetaPhlAn, with taxa filtered such that only taxa with at least 1% abundance in any sample was kept. No taxa are more differential between groups than within groups.

# Chapter 5

## Theorems

### 5.1 Basic Theorems

**Theorem 5.1.1**  $e^{i\pi} = -1$

# Bibliography

- [1] John Aitchison. “The statistical analysis of compositional data”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1982), pp. 139–177.
- [2] Stephen F Altschul et al. “Basic local alignment search tool”. In: *Journal of molecular biology* 215.3 (1990), pp. 403–410.
- [3] Marti J Anderson and Trevor J Willis. “Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology”. In: *Ecology* 84.2 (2003), pp. 511–525.
- [4] Manimozhiyan Arumugam et al. “Enterotypes of the human gut microbiome”. In: *nature* 473.7346 (2011), pp. 174–180.
- [5] Fredrik Bäckhed et al. “Mechanisms underlying the resistance to diet-induced obesity in germ-free mice”. In: *Proceedings of the National Academy of Sciences* 104.3 (2007), pp. 979–984.
- [6] Edward W Beals. “Bray-Curtis ordination: an effective strategy for analysis of multi-variate ecological data”. In: *Advances in Ecological Research* 14.1 (1984), p. 55.
- [7] James D Beck and Steven Offenbacher. “Systemic effects of periodontitis: epidemiology of periodontal disease and cardiovascular disease”. In: *Journal of periodontology* 76.11-s (2005), pp. 2089–2100.
- [8] David R Bentley et al. “Accurate whole human genome sequencing using reversible terminator chemistry”. In: *nature* 456.7218 (2008), pp. 53–59.
- [9] Alain Berson et al. “Steatohepatitis-inducing drugs cause mitochondrial dysfunction and lipid peroxidation in rat hepatocytes”. In: *Gastroenterology* 114.4 (1998), pp. 764–774.
- [10] Wenche S Borgnakke et al. “Effect of periodontal disease on diabetes: systematic review of epidemiologic observational evidence”. In: *Journal of periodontology* 84.4-s (2013), S135–S152.
- [11] Jérôme Boursier et al. “The severity of nonalcoholic fatty liver disease is associated with gut dysbiosis and shift in the metabolic function of the gut microbiota”. In: *Hepatology* (2016).
- [12] Paul R Burkholder and Ilda McVeigh. “Synthesis of vitamins by intestinal bacteria”. In: *Proceedings of the National Academy of Sciences of the United States of America* 28.7 (1942), p. 285.
- [13] Benjamin J Callahan et al. “DADA2: High resolution sample inference from amplicon data”. In: *bioRxiv* (2015), p. 024034.

- [14] J Gregory Caporaso et al. “QIIME allows analysis of high-throughput community sequencing data”. In: *Nature methods* 7.5 (2010), pp. 335–336.
- [15] J Gregory Caporaso et al. “Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms”. In: *The ISME journal* 6.8 (2012), pp. 1621–1624.
- [16] Daniel Aguirre de Cárcer et al. “Evaluation of subsampling-based normalization strategies for tagged high-throughput sequencing data sets from gut microbiomes”. In: *Applied and environmental microbiology* 77.24 (2011), pp. 8795–8798.
- [17] Jun Chen et al. “Associating microbiome composition with environmental covariates using generalized UniFrac distances”. In: *Bioinformatics* 28.16 (2012), pp. 2106–2113.
- [18] Francesca D Ciccarelli et al. “Toward automatic reconstruction of a highly resolved tree of life”. In: *science* 311.5765 (2006), pp. 1283–1287.
- [19] James R Cole et al. “The Ribosomal Database Project: improved alignments and new tools for rRNA analysis”. In: *Nucleic acids research* 37.suppl 1 (2009), pp. D141–D145.
- [20] Stephen M Collins, Zain Kassam, and Premysl Bercik. “The adoptive transfer of behavioral phenotype via the intestinal microbiota: experimental evidence and clinical implications”. In: *Current opinion in microbiology* 16.3 (2013), pp. 240–245.
- [21] Myron I Cybulsky et al. “A major role for VCAM-1, but not ICAM-1, in early atherosclerosis”. In: *The Journal of clinical investigation* 107.10 (2001), pp. 1255–1262.
- [22] Fabio Dal Bello and Christian Hertel. “Oral cavity as natural reservoir for intestinal lactobacilli”. In: *Systematic and applied microbiology* 29.1 (2006), pp. 69–76.
- [23] Todd Z DeSantis et al. “Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB”. In: *Applied and environmental microbiology* 72.7 (2006), pp. 5069–5072.
- [24] Julia M Di Bella et al. “High throughput sequencing methods and analysis for microbiome research”. In: *Journal of microbiological methods* 95.3 (2013), pp. 401–414.
- [25] SL Dollhopf, SA Hashsham, and JM Tiedje. “Interpreting 16S rDNA T-RFLP data: application of self-organizing maps and principal component analysis to describe community dynamics and convergence”. In: *Microbial Ecology* 42.4 (2001), pp. 495–505.
- [26] Sylvia H Duncan et al. “Human colonic microbiota associated with diet, obesity and weight loss”. In: *International journal of obesity* 32.11 (2008), pp. 1720–1724.
- [27] Robert C Edgar. “MUSCLE: multiple sequence alignment with high accuracy and high throughput”. In: *Nucleic acids research* 32.5 (2004), pp. 1792–1797.
- [28] Robert C Edgar. “Search and clustering orders of magnitude faster than BLAST”. In: *Bioinformatics* 26.19 (2010), pp. 2460–2461.
- [29] Steven N Evans and Frederick A Matsen. “The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3 (2012), pp. 569–592.
- [30] Andrew D Fernandes et al. “ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq”. In: *PLoS One* 8.7 (2013), e67019.

- [31] Andrew D Fernandes et al. “Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis”. In: *Microbiome* 2.1 (2014), p. 1.
- [32] Harry J Flint et al. “Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis”. In: *Nature Reviews Microbiology* 6.2 (2008), pp. 121–131.
- [33] DN Fredericks and David A Relman. “Sequence-based identification of microbial pathogens: a reconsideration of Koch’s postulates.” In: *Clinical microbiology reviews* 9.1 (1996), pp. 18–33.
- [34] Jonathan Friedman and Eric J Alm. “Inferring correlation networks from genomic survey data”. In: *PLoS Comput Biol* 8.9 (2012), e1002687.
- [35] Jack A Gilbert, Janet K Jansson, and Rob Knight. “The Earth Microbiome project: successes and aspirations”. In: *BMC biology* 12.1 (2014), p. 69.
- [36] Steven R Gill et al. “Metagenomic analysis of the human distal gut microbiome”. In: *science* 312.5778 (2006), pp. 1355–1359.
- [37] Gregory B Gloor et al. “Microbiome profiling by illumina sequencing of combinatorial sequence-tagged PCR products”. In: *PloS one* 5.10 (2010), e15406.
- [38] Helmut A Gordon, Edith Bruckner-kardoss, and Bernard S Wostmann. “Aging in germ-free mice: life tables and lesions observed at natural death”. In: *Journal of gerontology* 21.3 (1966), pp. 380–387.
- [39] Monika A Gorzelak et al. “Methods for Improving Human Gut Microbiome Data by Reducing Variability through Sample Processing and Storage of Stool”. In: *PloS one* 10.8 (2015), e0134802.
- [40] J Graessler et al. “Metagenomic sequencing of the human gut microbiome before and after bariatric surgery in obese patients with type 2 diabetes: correlation with inflammatory and metabolic parameters”. In: *The pharmacogenomics journal* 13.6 (2013), pp. 514–522.
- [41] Francisco Guarner and Juan-R Malagelada. “Gut flora in health and disease”. In: *The Lancet* 361.9356 (2003), pp. 512–519.
- [42] Brian J Haas et al. “De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis”. In: *Nature protocols* 8.8 (2013), pp. 1494–1512.
- [43] Elaine Y Hsiao et al. “Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders”. In: *Cell* 155.7 (2013), pp. 1451–1463.
- [44] Ruben Hummelen et al. “Deep sequencing of the vaginal microbiota of women with HIV”. In: *PloS one* 5.8 (2010), e12078.
- [45] Mark Ide and Panos N Papapanou. “Epidemiology of association between maternal periodontal disease and adverse pregnancy outcomes—systematic review”. In: *Journal of clinical periodontology* 40.s14 (2013).

- [46] Weiwei Jiang et al. “Dysbiosis gut microbiota associated with inflammation and impaired mucosal immune function in intestine of humans with non-alcoholic fatty liver disease”. In: *Scientific reports* 5 (2015).
- [47] Angela R Kamer et al. “Alzheimer’s disease and peripheral infections: the possible contribution from periodontal infections, model and hypothesis”. In: *Journal of Alzheimer’s Disease* 13.4 (2008), pp. 437–449.
- [48] Minoru Kanehisa and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30.
- [49] R Koch. “Über bakteriologische Forschung Verhandlung des X Internationalen Medicinischen Congresses, Berlin, 1890, 1, 35. August Hirschwald, Berlin”. In: *German.) Xth International Congress of Medicine, Berlin*. 1891.
- [50] Robert A Koeth et al. “Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis”. In: *Nature medicine* 19.5 (2013), pp. 576–585.
- [51] Heidi H Kong et al. “Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis”. In: *Genome research* 22.5 (2012), pp. 850–859.
- [52] Omry Koren et al. “Human oral, gut, and plaque microbiota in patients with atherosclerosis”. In: *Proceedings of the National Academy of Sciences* 108.Supplement 1 (2011), pp. 4592–4598.
- [53] HA Krebs and JR Perkins. “The physiological role of liver alcohol dehydrogenase”. In: *Biochemical Journal* 118.4 (1970), pp. 635–644.
- [54] Morgan GI Langille et al. “Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences”. In: *Nature biotechnology* 31.9 (2013), pp. 814–821.
- [55] Ben Langmead and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature methods* 9.4 (2012), pp. 357–359.
- [56] Nadja Larsen et al. “Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults”. In: *PloS one* 5.2 (2010), e9085.
- [57] Weizhong Li and Adam Godzik. “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences”. In: *Bioinformatics* 22.13 (2006), pp. 1658–1659.
- [58] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome biology* 15.12 (2014), pp. 1–21.
- [59] David Lovell et al. “Proportionality: a valid alternative to correlation for relative data”. In: *PLoS Comput Biol* 11.3 (2015), e1004075.
- [60] Catherine A Lozupone et al. “Quantitative and qualitative  $\beta$  diversity measures lead to different insights into factors that structure microbial communities”. In: *Applied and environmental microbiology* 73.5 (2007), pp. 1576–1585.

- [61] Catherine Lozupone and Rob Knight. “UniFrac: a new phylogenetic method for comparing microbial communities”. In: *Applied and environmental microbiology* 71.12 (2005), pp. 8228–8235.
- [62] Catherine Lozupone et al. “UniFrac: an effective distance metric for microbial community comparison”. In: *The ISME journal* 5.2 (2011), p. 169.
- [63] Jean M Macklaim et al. “Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis”. In: *Microbiome* 1.1 (2013), p. 1.
- [64] Siddhartha Mandal et al. “Analysis of composition of microbiomes: a novel method for studying microbial composition”. In: *Microbial ecology in health and disease* 26 (2015).
- [65] Janet GM Markle et al. “Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity”. In: *Science* 339.6123 (2013), pp. 1084–1088.
- [66] Victor M Markowitz et al. “IMG: the integrated microbial genomes database and comparative analysis system”. In: *Nucleic acids research* 40.D1 (2012), pp. D115–D122.
- [67] PD Marsh. “Microbial ecology of dental plaque and its significance in health and disease”. In: *Advances in dental research* 8.2 (1994), pp. 263–271.
- [68] Andre P Masella et al. “PANDAsseq: paired-end assembler for illumina sequences”. In: *BMC bioinformatics* 13.1 (2012), p. 31.
- [69] Caroline Minassian et al. “Invasive dental treatment and risk for vascular events: a self-controlled case series”. In: *Annals of internal medicine* 153.8 (2010), pp. 499–506.
- [70] Sherry L Murphy, Jiaquan Xu, and Kenneth D Kochanek. “Deaths: preliminary data for 2010.” In: *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System* 60.4 (2012), pp. 1–52.
- [71] Susan K Murphy et al. “Relationship between methylome and transcriptome in patients with nonalcoholic fatty liver disease”. In: *Gastroenterology* 145.5 (2013), pp. 1076–1087.
- [72] KM Neufeld et al. “Reduced anxiety-like behavior and central neurochemical change in germ-free mice”. In: *Neurogastroenterology & Motility* 23.3 (2011), 255–e119.
- [73] Jari Oksanen et al. “The vegan package”. In: *Community ecology package* (2007), pp. 631–637.
- [74] Jason W Osborne and Anna B Costello. “Sample size and subject to item ratio in principal components analysis”. In: *Practical assessment, research & evaluation* 9.11 (2004), p. 8.
- [75] Ross Overbeek et al. “The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes”. In: *Nucleic acids research* 33.17 (2005), pp. 5691–5702.



- [76] Lior Pachter. “Models for transcript quantification from RNA-Seq”. In: *arXiv preprint arXiv:1104.3889* (2011).
- [77] Javier Palarea-Albaladejo and Josep Antoni Martin-Fernández. “zCompositions—R package for multivariate imputation of left-censored data under a compositional approach”. In: *Chemometrics and Intelligent Laboratory Systems* 143 (2015), pp. 85–96.
- [78] Joseph Nathaniel Paulson. “metagenomeSeq: Statistical analysis for sparse high-throughput sequencing”. In: *Bioconductor package* 1 (2014).
- [79] Karl Pearson. “Mathematical Contributions to the Theory of Evolution.—On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs”. In: *Proceedings of the royal society of london* 60.359-367 (1896), pp. 489–498.
- [80] Elaine O Petrof et al. “Stool substitute transplant therapy for the eradication of *Clostridium difficile* infection: ‘RePOOPulating’ the gut”. In: *Microbiome* 1.1 (2013), p. 1.
- [81] David Preiss and Naveed Sattar. “Non-alcoholic fatty liver disease: an overview of prevalence, diagnosis, pathogenesis and treatment considerations”. In: *Clinical science* 115.5 (2008), pp. 141–150.
- [82] Albert Propst et al. “Prognosis and life expectancy in chronic liver disease”. In: *Digestive diseases and sciences* 40.8 (1995), pp. 1805–1815.
- [83] Christian Quast et al. “The SILVA ribosomal RNA gene database project: improved data processing and web-based tools”. In: *Nucleic acids research* 41.D1 (2013), pp. D590–D596.
- [84] Maitreyi Raman et al. “Fecal microbiome and volatile organic compound metabolome in obese humans with nonalcoholic fatty liver disease”. In: *Clinical Gastroenterology and Hepatology* 11.7 (2013), pp. 868–875.
- [85] Christian S Riesenfeld, Patrick D Schloss, and Jo Handelsman. “Metagenomics: genomic analysis of microbial communities”. In: *Annu. Rev. Genet.* 38 (2004), pp. 525–552.
- [86] Chantal A Rivera et al. “Toll-like receptor-4 signaling and Kupffer cells play pivotal roles in the pathogenesis of non-alcoholic steatohepatitis”. In: *Journal of hepatology* 47.4 (2007), pp. 571–579.
- [87] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (2010), pp. 139–140.
- [88] Frank A Scannapieco. “Systemic effects of periodontal diseases”. In: *Dental Clinics of North America* 49.3 (2005), pp. 533–550.
- [89] Jose U Scher et al. “Periodontal disease and the oral microbiota in new-onset rheumatoid arthritis”. In: *Arthritis & Rheumatism* 64.10 (2012), pp. 3083–3094.
- [90] Klaus Peter Schliep. “phangorn: Phylogenetic analysis in R”. In: *Bioinformatics* 27.4 (2011), pp. 592–593.

- [91] Patrick D Schloss et al. “Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities”. In: *Applied and environmental microbiology* 75.23 (2009), pp. 7537–7541.
- [92] Nicola Segata et al. “Metagenomic microbial community profiling using unique clade-specific marker genes”. In: *Nature methods* 9.8 (2012), pp. 811–814.
- [93] Ron Sender, Shai Fuchs, and Ron Milo. “Revised estimates for the number of human and bacteria cells in the body”. In: *bioRxiv* (2016), p. 036103.
- [94] Claude Elwood Shannon. “A mathematical theory of communication”. In: *ACM SIG-MOBILE Mobile Computing and Communications Review* 5.1 (2001), pp. 3–55.
- [95] Daniel Simberloff. “Use of rarefaction and related methods in ecology”. In: *Biological data in water pollution assessment: quantitative and statistical analyses*. ASTM International, 1978.
- [96] Michelle I Smith et al. “Gut microbiomes of Malawian twin pairs discordant for kwashiorkor”. In: *Science* 339.6119 (2013), pp. 548–554.
- [97] Se Jin Song et al. “Cohabiting family members share microbiota with one another and with their dogs”. In: *Elife* 2 (2013), e00458.
- [98] Erica D Sonnenburg et al. “Diet-induced extinctions in the gut microbiota compound over generations”. In: *Nature* 529.7585 (2016), pp. 212–215.
- [99] Silvia Sookoian and Carlos J Pirola. “Meta-analysis of the influence of I148M variant of patatin-like phospholipase domain containing 3 gene (PNPLA3) on the susceptibility and histological severity of nonalcoholic fatty liver disease”. In: *Hepatology* 53.6 (2011), pp. 1883–1894.
- [100] J David Spence. “Genetics of atherosclerosis: The power of plaque burden and progression”. In: *Atherosclerosis* 223.1 (2012), pp. 98–101.
- [101] J David Spence. “Technology insight: ultrasound measurement of carotid plaque—patient management, genetic research, and therapy evaluation”. In: *Nature Clinical Practice Neurology* 2.11 (2006), pp. 611–619.
- [102] Vanessa Sperandio et al. “Bacteria–host communication: the language of hormones”. In: *Proceedings of the National Academy of Sciences* 100.15 (2003), pp. 8951–8956.
- [103] WH Wilson Tang et al. “Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk”. In: *New England Journal of Medicine* 368.17 (2013), pp. 1575–1584.
- [104] Casey M Theriot et al. “Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to *Clostridium difficile* infection”. In: *Nature communications* 5 (2014).
- [105] Robert Tibshirani and Guenther Walther. “Cluster validation by prediction strength”. In: *Journal of Computational and Graphical Statistics* 14.3 (2005), pp. 511–528.
- [106] Susannah G Tringe and Philip Hugenholtz. “A renaissance for the pioneering 16S rRNA gene”. In: *Current opinion in microbiology* 11.5 (2008), pp. 442–446.

- [107] Peter J Turnbaugh et al. “An obesity-associated gut microbiome with increased capacity for energy harvest”. In: *nature* 444.7122 (2006), pp. 1027–131.
- [108] Peter J Turnbaugh et al. “Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome”. In: *Cell host & microbe* 3.4 (2008), pp. 213–223.
- [109] Peter J Turnbaugh et al. “The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice”. In: *Science translational medicine* 1.6 (2009), 6ra14–6ra14.
- [110] Peter J Turnbaugh et al. “The human microbiome project: exploring the microbial part of ourselves in a changing world”. In: *Nature* 449.7164 (2007), p. 804.
- [111] Camilla Urbaniak et al. “Human milk microbiota profiles in relation to birthing method, gestation and infant gender”. In: *Microbiome* 4.1 (2016), pp. 1–9.
- [112] M Al-Waiz et al. “The exogenous origin of trimethylamine in the mouse”. In: *Metabolism* 41.2 (1992), pp. 135–136.
- [113] William A Walters et al. “PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers”. In: *Bioinformatics* 27.8 (2011), pp. 1159–1161.
- [114] Zeneng Wang et al. “Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease”. In: *Nature* 472.7341 (2011), pp. 57–63.
- [115] AJ Wigg et al. “The role of small intestinal bacterial overgrowth, intestinal permeability, endotoxaemia, and tumour necrosis factor  $\alpha$  in the pathogenesis of non-alcoholic steatohepatitis”. In: *Gut* 48.2 (2001), pp. 206–211.
- [116] Vincent Wai-Sun Wong et al. “Molecular characterization of the fecal microbiota in patients with nonalcoholic steatohepatitis—a longitudinal study”. In: *PLoS One* 8.4 (2013), e62885.
- [117] Ivan KS Yap et al. “Metabonomic and microbiological analysis of the dynamic effect of vancomycin-induced gut microbiota modification in the mouse”. In: *Journal of proteome research* 7.9 (2008), pp. 3718–3728.
- [118] Mitugi Yasuda et al. “Suppressive effects of estradiol on dimethylnitrosamine-induced fibrosis of the liver in rats”. In: *Hepatology* 29.3 (1999), pp. 719–727.
- [119] Tanya Yatsunenko et al. “Human gut microbiome viewed across age and geography”. In: *Nature* 486.7402 (2012), pp. 222–227.
- [120] Lixin Zhu et al. “Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: a connection between endogenous alcohol and NASH”. In: *Hepatology* 57.2 (2013), pp. 601–609.

# Appendix A

## Proofs of Theorems

### Proof of Theorem 5.1.1

$$e^{i\pi} = \cos(\pi) + i \sin(\pi) \tag{A.1}$$

$$= -1 \tag{A.2}$$

■

# Curriculum Vitae

**Name:** Ruth Wong

**Post-Secondary Education and Degrees:** The University of Western Ontario  
London, ON  
2010-2014 B.M.Sc.

University of Western Ontario  
London, ON  
2014-2016 M.Sc.

**Honours and Awards:** Western Gold Medal  
2014

Leland Ritcey Prize  
2011

**Related Work Experience:** Summer Intern, Persistent Disk Team  
Google Inc., New York office  
Summer 2015

Google Summer of Code Participant  
Bader Lab, University of Toronto  
Summer 2014

## **Publications:**

Wong, Ruth G., Jia R. Wu, Gregory B. Gloor. "Expanding the UniFrac toolbox." Full length paper accepted for oral presentation at the Great Lakes Bioinformatics and the Canadian Computational Biology Conference 2016.