MICROBIOME ANALYSIS: METHODS AND APPLICATIONS

(Spine title: 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

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. Patrick O'Donoghue

.....
Dr. Chris J. Brandl

.....
Dr. Jeremy Burton/Gregory Thorn

The thesis by

Ruth Grace Wong

entitled:

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 DNA sequencing, scientists can obtain a more comprehensive snapshot of the composition of the microbiome, what genes are present, and what proteins are produced. The scientific community is in a phase of developing the experiments and accompanying statistical techniques to investigate the mechanisms by which the human microbiome affects health and disease. In this thesis I explore alternatives to the standard weighted and unweighted UniFrac difference metric that measure the difference between microbiome samples. I show that alternative weightings provide novel insight and allow the extraction of trends and outliers that are not visible with traditional methods. I also apply next generation DNA sequencing and computational analysis techniques to gut microbiome data from a nonalcoholic fatty liver disease cohort to examine the potential role of the microbiota in this condition.

Keywords: Human microbiome, next generation sequencing, bioinformatics, nonalcoholic fatty liver disease

Contents

Certificate of Examination ii

Abstract iii

List of Figures vii

List of Tables ix

List of Appendices x

1 Introduction 1
- 1.1 The human microbiome . . . 1
- 1.2 Exploring the human microbiome . . . 2
- 1.3 Illumina next generation sequencing . . . 2
- 1.4 Gene tag abundance . . . 3
  - 1.4.1 16S rRNA gene sequencing experiment . . . 5
  - 1.4.2 Operational Taxonomic Units . . . 5
  - 1.4.3 General protocol and rationale . . . 6
  - 1.4.4 Data analysis . . . 9
- 1.5 The metagenomic experiment . . . 14
  - 1.5.1 Sequencing . . . 14
  - 1.5.2 Imputation . . . 18
  - 1.5.3 Data aggregation, categorization, and amalgamation . . . 18
- 1.6 Points of failure . . . 20
  - 1.6.1 Collection methods differ . . . 20
  - 1.6.2 Microbiome data is highly variable between individuals . . . 20
  - 1.6.3 Microbiome data involves the comparison of many features . . . 21
  - 1.6.4 Microbiome data is compositional . . . 22
  - 1.6.5 Microbiome data is sparse . . . 24
- 1.7 The gut microbiome in patients with nonalcoholic steatohepatitis compared to healthy controls . . . 25

2 Expanding the UniFrac toolbox 28
- 2.0.1 Data . . . 30
- 2.0.2 Compositional Data Analysis . . . 30
- 2.0.3 Unweighted UniFrac . . . 31

2.0.4 Weighted UniFrac . . . 32
2.0.5 Analytical techniques . . . 33
2.0.6 Data preparation . . . 35
2.0.7 Unweighted Unifrac is highly sensitive to rarefaction instance . . . 37
2.0.8 The cause of rarefaction variation by Unweighted Unifrac . . . 39
2.0.9 Information UniFrac . . . 41
2.0.10 Tongue and buccal mucosa comparison . . . 42
2.0.11 Breast milk Data . . . 44
2.0.12 Monoculture data . . . 46

3 The human microbiome and nonalcoholic fatty liver disease 55
3.1 Introduction . . . 55
3.1.1 NASH progression risk . . . 55
3.1.2 Data . . . 56
3.1.3 Literature . . . 57
3.2 Methods . . . 59
3.2.1 16S rRNA gene tag experiment . . . 59
3.2.2 MetaPhlAn . . . 60
3.2.3 Metagenomic experiment . . . 60
3.2.4 Compositional data analysis . . . 61
3.3 Results . . . 63
3.3.1 Data sets . . . 63
3.3.2 16S rRNA gene tag experiment . . . 63
3.3.3 Metagenomic experiment . . . 72
3.4 Discussion . . . 85

4 Discussion 88
4.1 Lack of reproducibility . . . 88
4.2 Recommendations . . . 89
4.3 Summary . . . 90

Bibliography 91

Appendix A Workflows 100
A.1 Non-alcoholic fatty liver disease metagenomic workflow . . . 100
A.1.1 Filter OTUs . . . 100
A.1.2 Reference library annotation strategy . . . 100
A.1.3 De novo assembly annotation strategy . . . 101
A.1.4 Map sequenced reads to reference library . . . 102

Appendix B NAFLD study data collection 103
B.0.1 Study Participants . . . 103
B.0.2 Study Visits . . . 103
B.0.3 Clinical Data, Environmental Questionnaire, and Anthropometric Measurements . . . 104

B.0.4 Nutrition and Activity Assessment 104
B.0.5 Biochemistry 105
B.0.6 Serum Metabolites 105
B.0.7 Liver Histology 106
B.0.8 Stool Sample Collection and Analysis 107
B.0.9 Stool Homogenization 107
B.0.10 DNA Extraction 107

Curriculum Vitae **109**

List of Figures

1.1 16S rRNA gene tag experiment workflow. . . . 8
1.2 Unweighted UniFrac. . . . 11
1.3 Sample migration in different rarefactions, plotted on principal components, measured with unweighted UniFrac. . . . 12
1.4 Unweighted vs. weighted UniFrac weights. . . . 13
1.5 Metagenomic experiment workflow. . . . 17
1.6 Example stripcharts for subsystem 2 and 3 functional categorizations. . . . 19
1.7 Counting vs. sequencing. . . . 23
1.8 Venn diagram of genera found to be differentially abundant by different studies between NASH/NAFLD and healthy controls. . . . 26

2.1 Unweighted UniFrac. . . . 32
2.2 Sample migration in different rarefactions, plotted on principal coordinates, measured with unweighted UniFrac. . . . 38
2.3 Maxiumum relative deviation of rarefactions versus median deviation for traditional and non-traditional microbiome dissimilarity metrics. . . . 39
2.4 Phylogenetic tree with long isolated branches. . . . 40
2.5 UniFrac weights. . . . 42
2.6 Analysis of tongue and buccal mucosa data using different UniFrac weightings. 43
2.7 Analysis of breast milk data using different UniFrac weightings. . . . 45
2.8 Analysis of simulated monocultures using different UniFrac weightings. . . . 47
2.9 Principal Coordinate Analysis derived from GUniFrac distance matrices. . . . 49
2.10 Principal Coordinate Analysis derived from GUniFrac distance matrices. . . . 50
2.11 Principal Coordinate Analysis derived from GUniFrac distance matrices. . . . 51
2.12 Principal coordinate analysis derived from tongue dorsum samples using unweighted UniFrac distance matrices with no tree pruning. . . . 52
2.13 Principal coordinate analysis derived from tongue dorsum and buccal mucosa samples using unweighted UniFrac distance matrices with no tree pruning. . . . 53
2.14 Weighted UniFrac is a dissimilarity with tree pruning. . . . 54

3.1 Venn diagram of genera found to be differentially abundant by different studies between NASH/NAFLD and healthy controls. . . . 58
3.2 Bar plot of 16S rRNA gene tag sequencing experiment. . . . 64
3.3 Principal co-ordinate analysis of 16S rRNA gene tag sequencing data with different UniFrac weightings. . . . 65
3.4 16S rRNA gene tag sequencing experiment biplot. . . . 66
3.5 Effect plot showing difference within vs. difference between groups. . . . 67

3.6 Correlation in effect sizes of different group experiments. 68
3.7 Difference within vs. difference between healthy and extreme NASH for metagenomic data. 73
3.8 Difference within vs. difference between healthy and extreme NASH for metagenomic data. 74
3.9 Principal components analysis of metagenomic data. 75
3.10 Principal components analysis of carbohydrate metagenomic data. 77
3.11 Principal components analysis of lipid metagenomic data. 78
3.12 Principal components analysis of lipid metagenomic data. 79
3.13 Taxa barplot dendogram derived from MetaPhlAn. 81
3.14 Biplot derived from MetaPhlAn. 82
3.15 Biplot derived from 16S rRNA gene experiment. 83
3.16 Difference within groups vs. difference between groups per taxa, derived from MetaPhlAn. 84
3.17 NAFLD comparison effect sizes, with non overlapping healthy samples. 87

List of Tables

2.1 Original abundance of taxa and rarefied abundance of taxa. 40

3.1 List of overall study inclusion and exclusion criteria. 61
3.2 List of inclusion and exclusion criteria for metagenomic study. 62
3.3 Top decile of OTUs relatively increased in NASH based on effect size from healthy vs. NASH comparison. 70
3.4 Top decile of OTUs relatively increased in healthy based on effect size from healthy vs. NASH comparison. 71
3.5 Subsystem 4 label key for Figure 3.9. 76

List of Appendices

Appendix A Workflows 100
Appendix B NAFLD study data collection 103

Chapter 1

Introduction

This thesis focuses on the human microbiome, its relation to human diseases, and techniques used in the analysis and exploration of data derived from it. During the course of my thesis, I conducted one study about nonalcoholic fatty liver disease, and investigated 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 [112]. Trillions of these bacteria live in the gut [44], and have a massive metabolic potential. For example, the gut microbiome has been shown to produce changes in hormone levels [74], short chain fatty acid levels [125], and ethanol levels [58], to name a few. The human gut microbiome can also digest polysaccharides otherwise unusable by humans [33].

This massive metabolic potential produces measurable physiological effects. Transplanting gut bacteria from obese mice to lean mice has been shown to allow lean mice to absorb more calories from the same amount of food, thereby becoming more obese [124]. The microbiome can also affect behavior: Completely germ free mice exhibit more anxiety-like behaviors than specific pathogen free mice that contain a complex gut microbiome [86].

Study of 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 this species makes up less than 1% of the average human gut microbiome [4].

With the advent of next generation DNA sequencing, scientists can obtain a more comprehensive snapshot of the bacterial composition of the microbiome, what genes are present, and what proteins are produced [22]. 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 modulate the microbiome to improve quality of life.

1.2 Exploring the human microbiome

The advent of next generation DNA sequencing prompted the development of various experiments on sampling 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 from these samples as appropriate for the analysis.

A comparative study design involves an experimental group and a control group. The study subjects can be patients with disease and healthy controls [70], people who are susceptible and resistant to a condition [121], or patients before and after a medical intervention [43]. The questions that scientists in this field 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 that try to determine the similarities in the microbiomes of specific body sites among patients with similar medical conditions.

Next generation sequencing experiments can deduce: Is there a significant difference in the microbiome between the control and the experimental groups? Is this difference due to different types of microbes present or the microbial genes present? Do separated groups exist in the data? Are the abundances of certain taxa or genes correlated with each other, or with patient metadata? Through metagenomic experiments and statistical analysis we can gather clues about the larger questions of the mechanism of action.

In this thesis, I have performed two experiments that can be done with microbiome next generation DNA sequencing data: gene tag abundance (Fig. 1.1) and metagenomic sequencing (Fig. 1.5) [105]. The tag used for gene tag abundance here is the conserved 16S rRNA gene that is presumed to track taxonomic identity [40].

1.3 Illumina next generation sequencing

The Illumina MiSeq and HiSeq are next generation sequencing platforms. The Illumina MiSeq machines yields up to 25 million paired end reads up to 300 nucleotides long. The Illumina HiSeq machines yield up to 4 billion paired end reads up to 125 nucleotides long, as stated on the official Illumina website (http://www.illumina.com/systems.html). The general sequencing workflow is as follows:

1. DNA is amplified or fragmented to smaller pieces of approximately 1000 nucleotides or less

2. Adaptors are joined to the ends of the DNA

3. The DNA is denatured

4. The DNA is placed on a flow cell covered in oligonucleotides complimentary to the adaptor sequences, such that the DNA fragments are bound to the oligonucleotides

5. The DNA on the flow cell is replicated *in situ* to form clusters of identical sequences

6. The DNA is denatured

7. Primers, nucleotides, DNA polymerase, and fluorescently labelled deoxyribonucleotide triphosphate terminators are added

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

9. Fluorescent terminators are removed, exposing a 3'OH

10. Steps 7-9 are repeated until the desired number of cycles is complete.

The Illumina sequencing technology is the industry standard for metagenomic studies [6], and library preparation kits and protocols are available commercially. Roche 454 pyrosequencing [73] and sequencing on the SOLiD platform [115] [79] are other next generation sequencing options. The pyrosequencing platform has the advantage of longer reads, but yields higher error rates and is more expensive. SOLiD uses shorter read lengths and also has a higher error rate [72] [114].

One unappreciated feature of Illumina and other next generation DNA sequencing technologies is that they deliver data as parts per machine limit, not a count of molecules in the sample [31] [32] [40] [38] [39].

1.4 Gene tag abundance

Historically, Koch's postulates have been used to determine if a microbe is a disease-causing pathogen. Koch's postulates are:

1. The microbe must be present in all cases of the disease.

2. The microbe must not be present but non-pathogenic in other diseases.

3. If the microbe is isolated in pure culture, it can be used to induce the disease [56].

Fredricks et al. have created a modified set of postulates that takes DNA sequencing into account which can be applied to differentially abundant taxa detected by gene tag sequencing [34]:

1. A nucleic acid sequence belonging to a putative pathogen should be present in most cases of an infectious disease. Microbial nucleic acids should be found preferentially in those organs or gross anatomic sites known to be diseased and not in those organs that lack pathology.

2. Fewer, or no, copy numbers of pathogen-associated nucleic acid sequences should occur in hosts or tissues without disease.

3. With resolution of disease (for example, with clinically effective treatment), the copy number of pathogen-associated nucleic acid sequences should decrease or become undetectable. With clinical relapse, the opposite should occur.

4. When sequence detection predates disease, or sequence copy number correlates with severity of disease or pathology, the sequence-disease association is more likely to be a causal relationship.

5. The nature of the microorganism inferred from the available sequence should be consistent with the known biological characteristics of that group of organisms. When phenotypes are predicted by sequence-based phylogenetic relationships, the meaningfulness of the sequence is enhanced.

6. Tissue-sequence correlates should be sought at the cellular level: efforts should be made to demonstrate specific in situ hybridization of microbial sequence to areas of tissue pathology and to visible microorganisms or to areas where microorganisms are presumed to be located.

7. These sequence-based forms of evidence for microbial causation should be reproducible.

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 [70]. Gene tag abundance takes us beyond Koch's postulates to the effect of consortia of microbiota.

Gene tag abundance experiments provide an estimate of the proportion of different bacterial taxa 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. The idea is that characterizing what the core microbiome is can lead to insight on core functions and how they can be altered when the core microbiome is disrupted.

For example, the core gut microbiome was described by one group to have three enterotypes [4]. The enterotype structure would have been very useful for measuring the association of certain enterotypes with conditions, and for observing how gut microbiomes transition across enterotypes. However, when another group studied a diverse population including non-Western people, the enterotypes did not hold [135]. A second group showed that these enterotypes were artifacts induced by the analysis methods that depended on the abundance of the predominant taxa [41]. The gut microbiome is highly diverse between individuals, and the enterotype model does not capture this diversity.

An example of a successfully characterized body site is the vaginal microbiome. The vaginal microbiome is known to be *Lactobacillus* dominated, except in bacterial vaginosis [50] or anaerobic vaginitis [24], where the microbiome is much more diverse. The bacterial composition of vaginal microbiomes in bacterial vaginosis have high variation, however their expression profiles are similar, allowing for functional characterization in the absence of taxonomic characterization [70].

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.

For example, in atopic dermatitis, a flare-up is defined as an acute exacerbation of disease despite standard treatment. Flare-ups are associated with a significant increase in the proportion of *Staphylococcus aureus* on the skin [57]. However, the exact mechanisms by which *Staphylococcus aureus* causes atopic dermatitis are unknown.

Bacteria have also been used for therapy in the treatment of *Clostridium difficile*. Stool transplants have been found to be more effective than the vancomycin antibiotic [87]. Work is being done to standardize treatments: In one study, 33 microbes cultured from a healthy donor were used to successfully treat symptoms, with no recurrence throughout the 6 month follow up period [96].

Do samples from different conditions cluster together? Beta diversity distance similarities between microbiomes can be examined for distinct sites or conditions. This is often done with distance or dissimilarity metrics, such as the UniFrac distance and the Bray Curtis dissimilarity.

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 [117]. In this case the microbiomes were seen to diverge the most during treatment with ready-to-use therapeutic food.

1.4.1 16S rRNA gene sequencing experiment

The gene tag chosen for analysis 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 taxonomy. Entire databases exist specifically to match the 16S rRNA gene with taxonomy, such as SILVA [101], the Ribosomal Database Project [18], and Greengenes [21].

Specifically, this work uses the 16S rRNA gene primers from the Earth Microbiome Project protocol [36], 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 [130].

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 examples where bacteria classified in the same species are actually genetically very different, or bacteria classified in different genera are genetically very similar [17].

However, it is difficult to determine how a batch of sequences should be partitioned into groups of 97% identity. Two common ways of doing this are open reference OTU picking [104] and closed reference OTU picking. Open reference OTU picking performs a clustering algorithm that partitions the groups and then later assign taxonomic identity by matching the sequences with public databases [27]. Closed reference OTU picking starts off with seed sequences from known bacteria and performs 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 likely to be the same.

1.4.3 General protocol and rationale

The 16S rRNA gene sequencing experiment (Fig. 1.1) uses next generation DNA 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. The process by which these data are collected and analyzed has many steps.

Pre-sequencing processing
There are several very general steps to the pre-sequencing process (Fig. 1.1):

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 [13]. The set of primers that we use are combinatorial barcoded, so that we can sequence all the samples in the same sequencing run and differentiate them afterwards [40].

3. Perform sequencing
   We use 2x220 nucleotide paired-end sequencing on the Illumina MiSeq platform. The 220 nucleotide paired ends allow us to overlap paired sequences in the middle to reconstitute the full sequence of the V4 variable region, which is 298 base pairs in size.

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

1. 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.

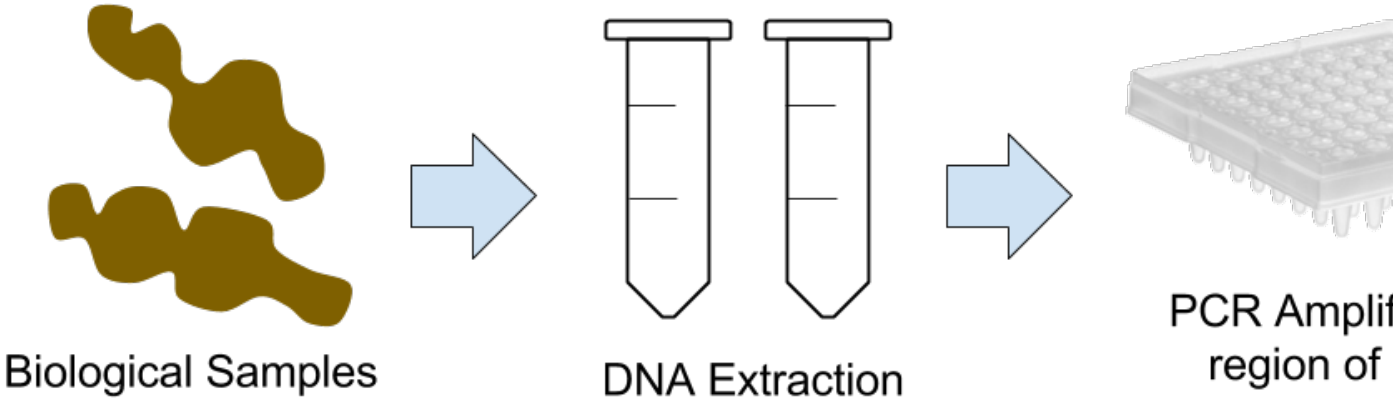
2. Demultiplex the raw sequence
   The barcodes are used to separate the sequences according to what sample they came from.

3. Group the reads into operational taxonomic units (OTUs)
   We used UCLUST [26] to cluster the reads into groups of 97% identity and UCHIME [28] to filter out chimeric sequences.

4. Annotate the OTUs with bacterial taxonomy
   Annotation was done by matching our OTUs to the SILVA database [101].

5. 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 [25]. The multiple sequence alignment can be converted into a phylogenetic tree using software such as FastTree [99].

Alternatively, an Individual Sequence Unit (ISU) based approach rateher 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 [10].



	Sample A	*Sample B*	*Sample C*
E. coli	323	87	45
B. fragilis	34	0	0
L. casei	0	213	142
S. pneumoniae	10	23	435

Generate table of counts per assumed taxa per sample

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 dissimilarity between each sample, and using these dissimilarities to plot the samples as points on a principal co-ordinate graph. The following sections will go over the most commonly used distance metric in microbiome research, called UniFrac, as well as the Principal Co-ordinate 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 *Microbiome data is compositional*.

Principal Component Analysis (PCA) is a data reduction strategy used for multivariate statistics. However, the OTU abundances derived from the 16S rRNA gene sequencing experiment are proportional, so the PCA (which assumes a linear differences) is not applicable. Instead, the data must be transformed in some way into a Euclidean distance [3]. This is the rationale behind the development of the UniFrac distance metric. Using the pairwise UniFrac distances between the samples, a PCA can be performed to analyze the data.

UniFrac

In 2005, Lozupone et al. introduced the UniFrac distance metric, a measure to calculate the difference between microbiomes that incorporated phylogenetic distance [68]. 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 [67]. Since then, papers reporting these metrics have garnered over a thousand citations, and enabled insights about everything from how kwashiorkor causes malnutrition [117] to how people can have a similar microbiome to their pet dog [118]. Except for Generalized UniFrac, used to make hybrid unweighted and weighted UniFrac comparisons [16], few advances in the metric have occurred since 2007.

Unweighted UniFrac

Unweighted UniFrac uses an inferred evolutionary distance to measure similarity between samples (Fig. 1.2). 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, microbial ecologists use a technique called rarefaction to normalize the sequencing depth across samples by random sampling without replacement [14], although this is controversial [80].

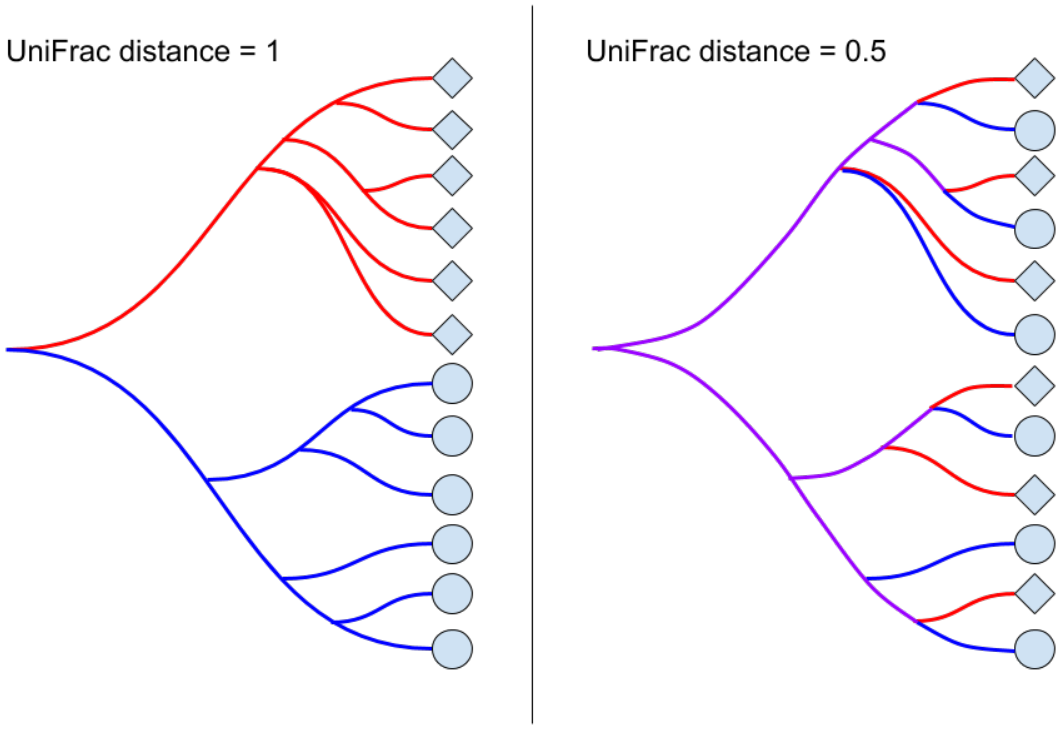


Figure 1.2: **Unweighted UniFrac.** When two samples do not share any branches of the phylogenetic tree, the unweighted UniFrac distance is maximized at 1. When two samples share half of their branch lengths on the phylogenetic tree, the unweighted UniFrac distance is 0.5. If the two samples contain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

In the UniFrac paper, the authors pointed out that UniFac was unstable with rarefaction and recommended that users take the average of multiple UniFrac instances [69]. This is generally disregarded, sometimes even by the original authors [78] [82].

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 (Fig. **??**). Any published finding done with an unweighted UniFrac analysis is suspect, especially if only one rarefaction instance is reported or if most of the variance is not explained by the first and second principal components. This is further explored in the chapter *Expanding the UniFrac Toolbox*.

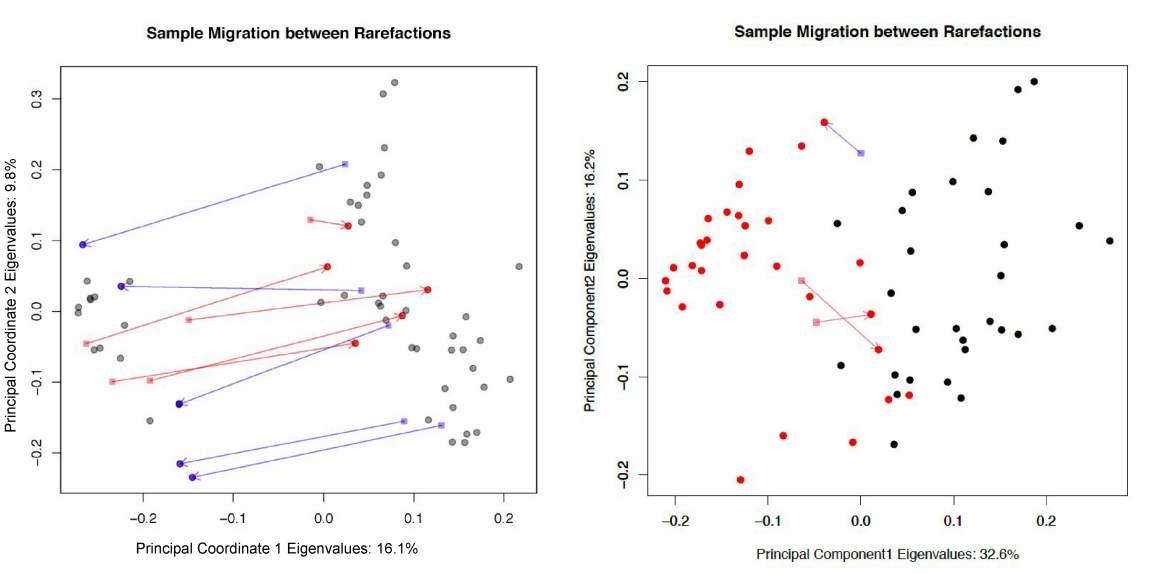


Figure 1.3: **Sample migration in different rarefactions, plotted on principal components, measured with unweighted UniFrac.** The left panel shows the movement across clusters for a set of randomly selected tongue dorsum samples from healthy volunteers in the Human Microbiome Project database. The right plot has samples from tongue dorsum and buccal mucosa, which have real separation. If the tongue only 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 in the tongue data set by the first and second component 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.

Weighted UniFrac

Weighted UniFrac is an implementation of the Kantorovich-Rubinstein distance in mathematics, also known as the earth mover's distance [30]. 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 (Fig. 1.4). 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 reduced impact on the total distance reported by the metric.

UniFrac is constituted as either a presence/absence (unweighted UniFrac) [68], a linear proportion in the form of weighted UniFrac [67], or some combination of the two in the form of Generalized UniFrac [16]. However, the data are not linear, because the sum of the total number of reads is constrained by the sequencing machinery [35]. Alternative weightings and nonlinear transformations of data need to be explored, and this is the focus of Chapter 2.

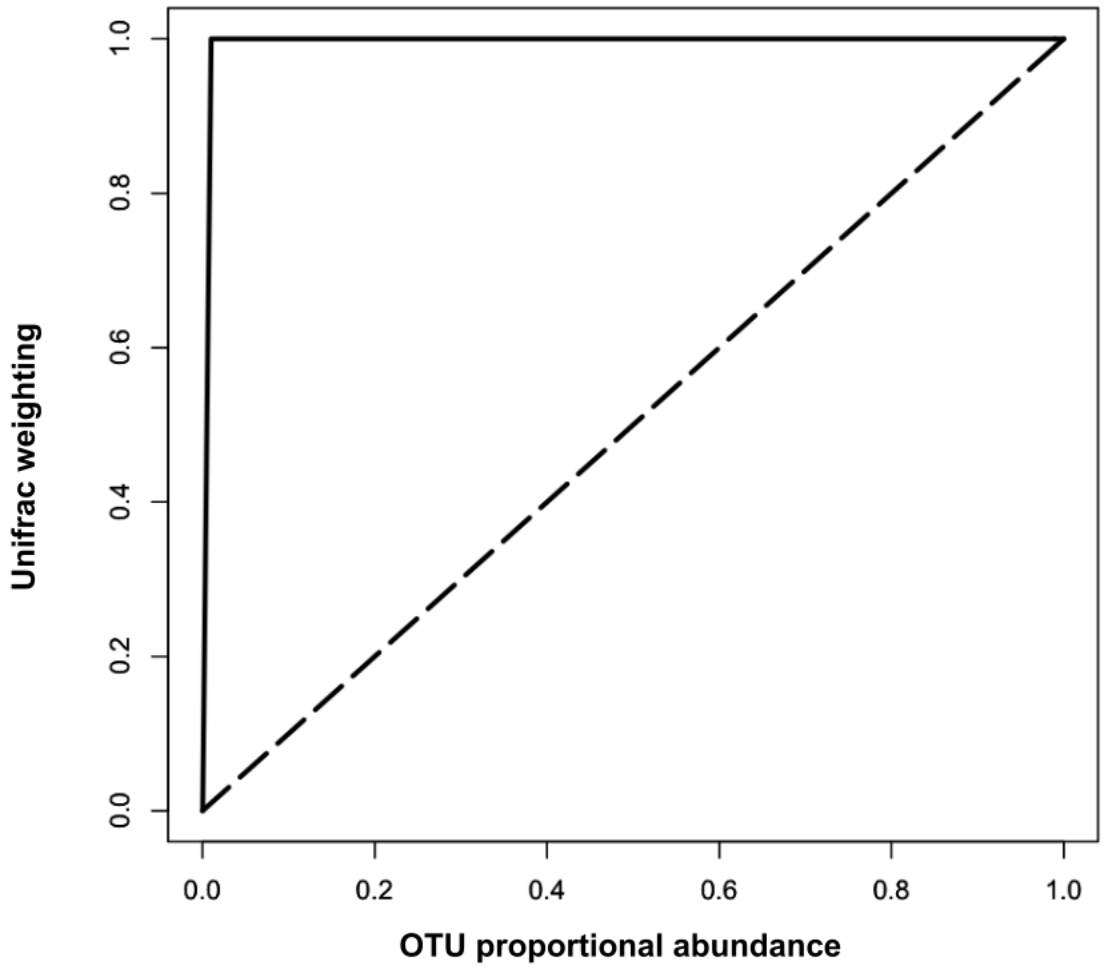


Figure 1.4: **Unweighted vs. weighted UniFrac weights.** The solid line represents the weight that unweighted UniFrac would produce with the given proportional abundance, and the dotted line represents the weight that weighted UniFrac would produce.

Principal Co-ordinate Analysis

Once the dissimilarity 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 Co-ordinate Analysis method of multidimensional scaling [23], shortened as PCoA. PCoA is a singular value decomposition of the distance relationships.

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 distances, or spread of the data, the second component captures the largest distances 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. Ideally the first and the second principal components should explain most of the distance in the data. For example, there is much less variance explained in the uniform tongue dorsum data set, compared to the data set with separation between tongue dorsum and buccal mucosa samples (Fig. 1.3).

After multidimensional scaling the data can be analyzed in several ways. The data can be examined by k-means analysis clustering [122] or by unsupervised clustering.

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.

There are several limitations of principal co-ordinate analysis. First, it is an indirect analysis. If a separation between groups is found, further examination is necessary to determine the source of separation. Second, the PCoA is only as good as the dissimilarity metric used. Lastly, one cannot easily determine the contributions of each OTU to the principal components.

1.5 The metagenomic experiment

Deep metagenomic sequencing provides an estimate of the proportion that each type of gene composes 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 encoded 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 is known to facilitate methanogenesis. However, methanogenesis is a common function in rare taxonomic groups (mostly *Archaea*). Metagenomic sequencing shows that the human gut microbiome has more genes related to methanogenesis than expected - a feature that would have not been as prominent in 16S rRNA gene sequencing analysis [37].

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 [124], sparking more research into the gut microbiome and obesity related ailments such as diabetes [62] and nonalcoholic fatty liver disease [136]. 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 DNA sequencing analysis is to examine the metabolic potential of the microbiota in the microbiome. This is done by identifying genes by DNA sequence, sorting

them by the known function of the protein that they encode (such as the catalysis of a certain reaction), and checking if any functions are differentially abundant between conditions. Further analysis can also include checking for pathway enrichment, and assembling the sequenced reads into genomes. The general protocol for metagenomic analysis (Fig. 1.5) 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.

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. There are two options for read length: either 50 or 100 nucleotides. We chose the longer one for ease of assembly and mapping.

4. Create an annotated library of reference sequences
   The annotated library contains DNA sequence 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. Most 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 [90] or KEGG [55] 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 [61] to map our sequenced reads to the annotated library created in the previous step. Bowtie2 aligns similar sequences together.

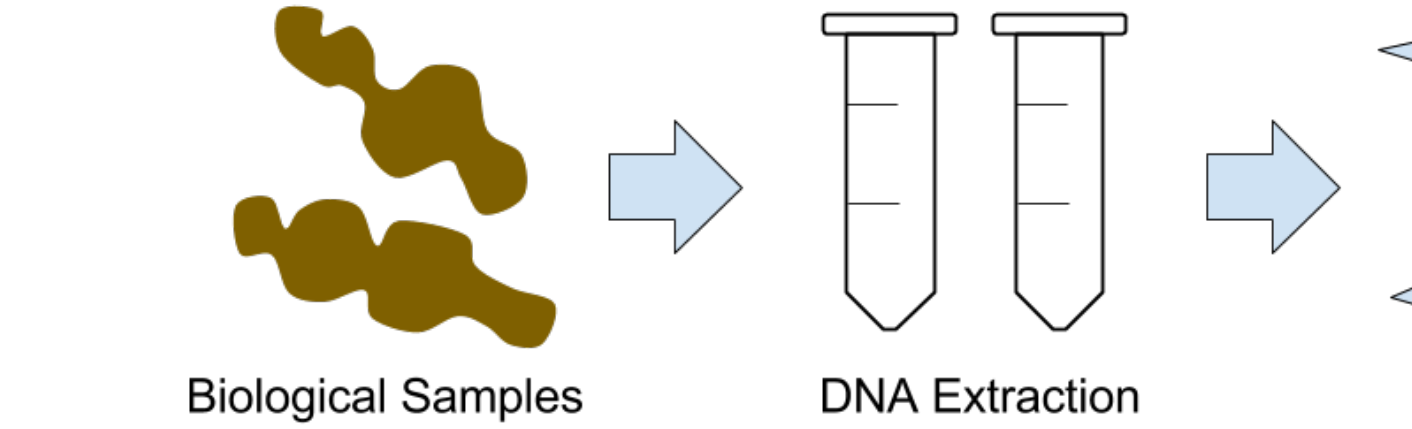
6. Determine how many mapped reads match each functional annotation
   Once the sequenced reads have been mapped to the annotated reference sequence, the number of reads sequenced for each annotation can be counted up. The end result is a table of counts per gene annotation per sample.

Issues with sequencing and the analysis of sequencing data arise from sampling and the "fat" nature of the data.

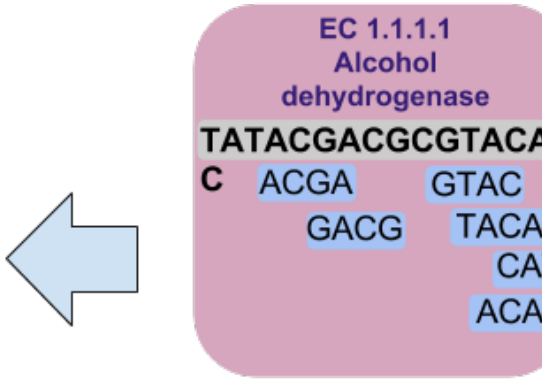
The DNA has been randomly sampled multiple times before the sequencing data is retrieved: The biological sample collected from the patient is only part of the full bacterial community. The amount of DNA extracted is a sample of that sample. Only a fraction of the extracted DNA is sequenced, and finally the DNA fragments that the sequence reads are sampled out of the

input DNA. As a result, on top of the biological variation present in the microbiomes being sampled, there is an additional layer of technical and random variation. High variation or noise in the data can obscure small but biologically significant differences between experimental conditions.

Additionally, primers used for sequencing may be biased for certain sequences more than others [54]. Lastly, the data is very fat, which is to say that there are many 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 *Microbiome data involves the comparison of many features* section.



	Sample A	*Sample B*	*Sample C*
EC 1.1.1.1 Alcohol dehydrogenase	2	1	3
EC 3.5.1.4 amidase	2	0	0
EC 3.5.1.2 glutaminase	0	1	1



Generate table of counts per functional annotation per sample

Aggregate
anr

Figure 1.5: **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

When it is not financially feasible to perform deep metagenomic sequencing, the sequencing results can be imputed using a tool called PICRUSt from a gene tag experiment [60]. PICRUSt uses the Greengenes database [21] to identify the bacterial taxa in the sample, and pulls their genomes from the Integrated Microbial Genomes database [75]. 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 a Spearman correlation coefficient of about 0.7 [60], 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 aggregation, categorization, and amalgamation

Data analysis can be performed to determine if functions are differentially abundant between samples in different groups (described in the *Microbiome is compositional* section), examining functional categorizations, and checking for pathway enrichment. Sequenced genes (open reading frames) can be grouped by common function through annotation by querying the SEED or KEGG functional annotation databases. These functions can be analyzed to see if any are differentially abundant between experimental conditions.

Functional categorization

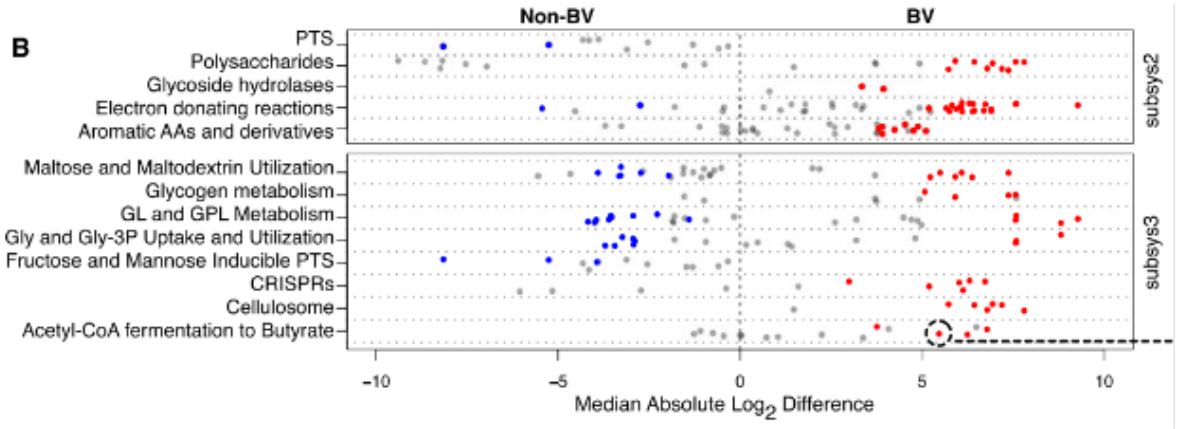


Figure 1.6: **Example stripcharts for subsystem 2 and 3 functional categorizations.** Dots on the left side are SEED subsystem 4 annotations found to be more abundant in the healthy condition while dots on the right side are subsystem 4 annotations found to be more abundant in the bacterial vaginosis condition. Colored dots were found to be significantly differentially abundant. Figure taken from [70].

We typically 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. These levels are simply aggregations of subsystem 4 levels, and one subsystem 4 annotation can be found in one or more higher level groups.

An effect size is measured by taking the difference in means between two groups of data, and dividing by the standard deviation within groups. The effect size is stronger when there is less overlap between the two groups. 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 (Fig. 1.6). 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 relatively abundant 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 [55]. 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. This is an alternate amalgamation of the data.

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 collection, sequencing, and data that contribute to this.

1.6.1 Collection methods differ

The 16S rRNA gene sequencing experiments are very sensitive to batch effects. Microbiome composition is often naturally highly diverse between different individuals. The high amount of variation means that the effect size of a difference between groups can be small [32]. Next generation sequencing is also a sensitive technology, and the data can be confounded by contaminants [108] or batch effects. These artifacts can overpower real biological effects. 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

The gut is often studied but the gut microbiome can be affected very strongly by diet [126]. This among other factors lead to a highly diverse gut microbiome between subjects for reasons unrelated to the disease being studied. This can create a lot of variability, 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 (such as from kit contamination [108]), we run all the samples in an experiment on the same sequencing run, by means of a combinatorial barcode primer design [40].

There are several models for computationally analyzing the variance within conditions in order to determine if operational taxonomic units are significantly differentially abundant, such as LEfSe [110] and Metastats [93] for microbiome analysis. Most of these were originally designed for RNA-seq experiments on single organisms [91]. Currently the most popular tools for analyzing differential abundance are EdgeR [107], DESeq2 [65], MetagenomeSeq [94], and LEfSe [110]. EdgeR was cited by 52 papers that mentioned the microbiome in 2015 according to Google Scholar. DESeq2 and MetagenomeSeq are part of the QIIME pipeline, which was cited by 1,080 microbiome papers in 2015. LEfSe has been cited by 145 microbiome 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 [31].

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 does not work well when the total number of reads are not well matched, because then there will be many 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 [94].

LefSe stands for the Linear discriminant analysis (LDA) Effect Size method [110]. It identifies significantly differentially abundant OTUs, checks for consistency in the differential abundance with the Wilcoxin rank sum test, and uses linear discriminant analysis to estimate an effect size per feature. This method assumes that microbial communities can be split by a linear combination of OTUs (for example the line $ax + by$, where a and b are constants and x and y are OTU abundances). However, bacterial abundances in both count form and proportions do not grow at a linear rate.

EdgeR, DESeq2, and MetagenomeSeq all work by using a statistical model to make a point estimate of the mean and variance of the data. Using the estimated mean and variance, differential abundance is tested statistical significance. However, a point estimate obscures technical variation in the data. That is, if a technical replicate were performed by resequencing the samples, a different set of point estimates would be calculated. In contrast, Bayesian methods

model the distribution of the mean and variance.

For my differential abundance analysis, I've used ALDEx2, which samples from the Dirichlet distribution to model variation in the data [32]. 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 specificity and equivalent sensitivity compared to the point estimate approach [32].

1.6.3 Microbiome data involves the comparison of many features

Oftentimes, the number of taxa or gene functions is more than a magnitude larger than the samples. 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 or pairwise comparison is to be reliable [89].

One way to conceptualize the problem is through combinatorics. Hypothetically, there could exist many bacterial taxa that are equally present in both conditions, but have a low abundance such that a 0 or 1 count is often detected. In this case the chances of observing all 0 counts in one condition and all 1 counts in another condition by simple combinatorics is quite high. This situation is common in microbiome data, where most of the bacteria are abundant in low proportions, sample sizes are often low, and a typical 16S rRNA gene sequencing experiment detects hundreds of OTUs.

Multiple test corrections assume that the experiment has more samples than variables, and result in very high p-values when this condition is not met. However, when using p-value based tests, researchers should include multiple test corrections to ensure that the results they are reporting are not all false positives. Unfortunately many studies have been published in peer reviewed journals without multiple test corrections. For example, four out of five papers in the literature about the gut microbiome and non-alcoholic fatty liver disease did not use a multiple test correction (Fig. 1.8).

1.6.4 Microbiome data is compositional

In both gene tag sequencing and metagenomic sequencing experiments, the data is in the form of a list of counts per feature, with the features composing an aspect of the microbiome for each sample. Therefore, the number of counts observed is arbitrary and not related to number of counts in the original environment. For example, an oral sample (10^7 Colony Forming Units/ml) and a gut sample (10^{10} CFU/ml) will give same number of counts (1×10^5) after DNA sequencing. This is compositional data. There are several core truths about microbiome data and its compositional nature 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. The number of reads is determined mainly by the chosen sequencing platform (i.e. Roche 454, Illumina MiSeq, or Illumina HiSeq). The absolute abundance of reads per sample cannot be used to make biological inferences.

Second, spurious correlations can arise from proportional microbiome data, and should be avoided. In the late 19th century, many studies were being published about how organ sizes (normalized by dividing the size by the individual's height) were correlated. However, it was discovered that when two sets of uncorrelated data are both divided by a third set of uncorrelated

data, the two sets will appear spuriously correlated [95]. This is analogous to microbiome data where raw counts are normalized by dividing by the total number of counts [95].

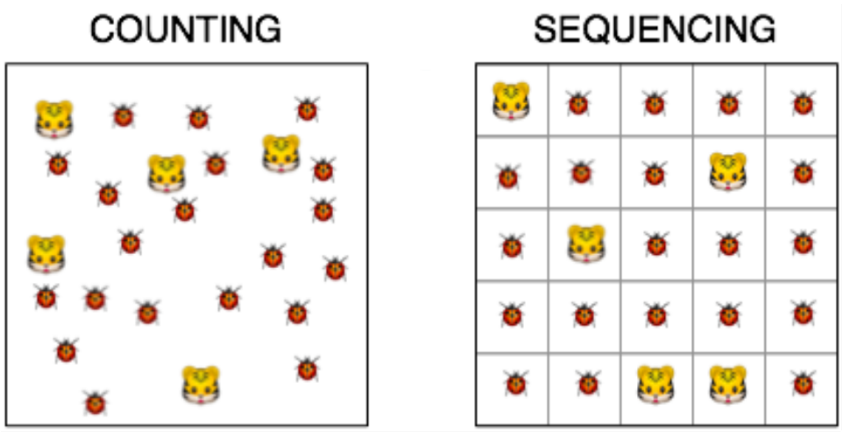


Figure 1.7: **Counting vs. sequencing.** In ecology, where all the animals in a given area are counted, there is no upper limit on the number of animals that can be detected. In sequencing, there is a constraint on the total number of sequences detected by the sequencing platform. If one were to detect one more ladybug, one would have to detect one less lion. Figure courtesy of Dr. Greg Gloor, http://gloorlab.blogspot.ca/2016/04/sequencing-is-not-counting-idea-of.html.

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 proportional abundance, even if the taxa are not decreasing in absolute abundance biologically. This negative correlation bias arises when the data are treated as univariate, when it should be analyzed as multivariate data. All non-parametric tests (Principal Co-ordinates Analysis, correlations, etc.) assume that the abundance measured for each feature is independent. This is true in ecology where the abundance of different species is measured in an area of land where animals can move freely in and out. It is not true in next generation sequencing where the abundance of different species is measured by a sequencing platform with a limited number of measurements, such that detecting extra members of one species means detecting less members of another species (Fig. 1.7).

Third, 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. This is true with counts but not true with proportions [66]. A correlation between two OTUs is suspect if it is dependent on the presence of an additional unrelated OTU. Removing variables occur routinely in microbiome research. For example, rare OTUs are thought to not be very informative, and low counts have high variability, so they are often filtered out. Additionally, primers may be biased against certain taxa, which are underrepresented in the data. Finally, some experiments are performed only on taxa of interest (as is the case with qPCR), and all other OTUs are not considered in the analysis. Without the proper data transformation, removing variables from the full set will change the correlation between variables [1].

To ensure that these conditions are met, data should be analyzed in a compositional way.

In Euclidean space, data points can increase or decrease freely. Compositional data is under a sum constraint, and exist in a non-Euclidean space known as the Aitchison simplex [1]. Many operations are not valid for data on a simplex. These include addition and subtraction, covariance, and correlation. A data transformation can be performed to put the data into Euclidean space, so that it can be analyzed with standard statistical methods that depend on Cartesian co-ordinates and linear relationships.

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. Here x_i is one proportional abundance within a sample, and there are n OTUs in total.

$$clr(x_i) = \frac{x_i}{\sqrt[n]{\prod_{i=1}^{n} x_i}}$$

The geometric mean acts as a baseline abundance in microbiome data. Taking the logarithm of the ratio allows for a symmetric 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 relatively stable baseline. The geometric mean is stable when the total number of reads is constant, or the per feature variation is random. The latter condition is met in a typical microbiome data set. The centered log ratio transform also allows for coherent subcompositional data analysis as the ratios between the remaining values are not affected when entire variables are removed. Note that a logarithm cannot be performed when the data contain one or more zero counts, which is problematic as microbiome data is sparse. This issue is discussed in the next section.

Compositional techniques such as those espoused in the ANOVA-Like Differential Expression 2 (ALDEx2) software [32] and the Analysis of Composition of Microbiomes (ANCOM) framework [71] should be used to promote consistent data analysis. ALDEx2 models the technical variation using the Dirichlet distribution and then performs a log ratio transform while ANCOM uses log ratio analysis to make point estimates of the variance and mean, without any distributional assumptions.

However, these techniques are not yet mainstream in the field, resulting in many conclusions that are not reproducible. One example of this is referenced in the chapter about the gut microbiome and non alcoholic fatty liver disease, where five papers have been published on the same topic with almost completely non overlapping results (Fig. 1.8).

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 OTU is not there. The OTU could be present in an amount smaller than the resolution of the test, or it could be present but missed due to random sampling. This is a problem because when statistical methods are used to examine significantly different OTU abundance, as the comparison of

zero values to non-zero values are likely to come out as significant whether or not the OTU abundance is differential. However, a 0 and a 1 count are easily interchangeable between technical replicates and the difference is not biologically significant [39] [31] [32]. Additionally, the log transformations used in compositional data analysis cannot be performed on zeros. Statisticians often recommend that any sample with at least one zero count be removed during compositional data analysis, but for microbiome data this would often result in the removal of all the samples [1].

One solution to make zeros compatible with a compositional data log transformation is to add a small arbitrary value to each zero [1]. The value used can be 0.5, representing uncertainty as to whether the zero represents an absence of the feature or if the feature is actually present but was missed due to random sampling or sequencing depth. In a Bayesian model, the 0.5 value is a prior. The second method is to estimate the likelihood that a zero was observed because of sampling depth. This is implemented by the cmultRepl command in the zCompositions package in R, with the 'count multiplicative zeros' option [92].

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 patients with nonalcoholic 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 [97]. Most people with NAFLD remain asymptomatic, however, in up to a third of patients NAFLD can progress to nonalcoholic steatohepatitis (NASH), causing inflammation and scarring (fibrosis) in the liver, and decreasing the 5 year survival rate to 67% [100]. It is thus important to shed some light on the process by which people progress from NAFLD to NASH to find interventions that prevent NASH.

There are several known genetic and chemical factors that increase the risk of progression to NASH in animal models and humans.

In mice non alcoholic fatty liver disease is often modelled with a methionine/choline-deficient diet (MCD), which induces steatohepatitis in wild type mice. Mice with a toll-like receptor 4 knockout had lower lipid and injury accumulation markers when fed a MCD diet [106]. In rats liver fibrosis can be induced by drugs. One study found that male rats were more prone to this induced liver fibrosis than female rats. Fibrosis biomarkers were reduced when the male rats were dosed with estradiol, and increased when the male rats were additionally given an estradiol-neutralizing antibody. Female rats who had their ovaries removed similarly lost the protective effect [134]. From this, hormones are also a factor in nonalcoholic fatty liver disease progression.

In humans, the I148M variant of the Patatin-Like Phospholipase Domain Containing 3 gene (PNPLA3) correlates with a 3.2 fold increased risk of progression to NASH from NAFLD when homozygous, compared to to patients without the variant [120]. The heterozygous gene was found to be associated with fatty liver disease in genome wide association studies, but some additional studies have failed to replicate the relationship with NASH [120]. On the

epigenetic level, many genes are differentially methylated in the livers of patients with advanced NAFLD compared to mild NAFLD. Eleven percent of genes are differentially hypomethylated in advanced NAFLD (compared to 3% hypermethylated), leading to increased expression [85]. In advanced NASH specifically, some tissue repair genes were hypomethylated while some metabolism pathways such as 1-carbon metabolism were hypermethylated. However, only 7% of the differentially methylated genes were found to be differentially transcribed [85].

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 [102]. Reactive oxygen species have also been implicated in NASH due to their involvement in the mechanism of steatohepatitis-inducing drugs [7].

The microbiome is thought to have an effect on host digestion and absorption of nutrients [37]. Fermenters produce short chain fatty acids, which make up 10% of the calories in a Western diet [81]. Some groups claim a link between ethanol-producing gut bacteria and NAFLD [136] [52], however the evidence was inconclusive since no multiple test correction was performed.

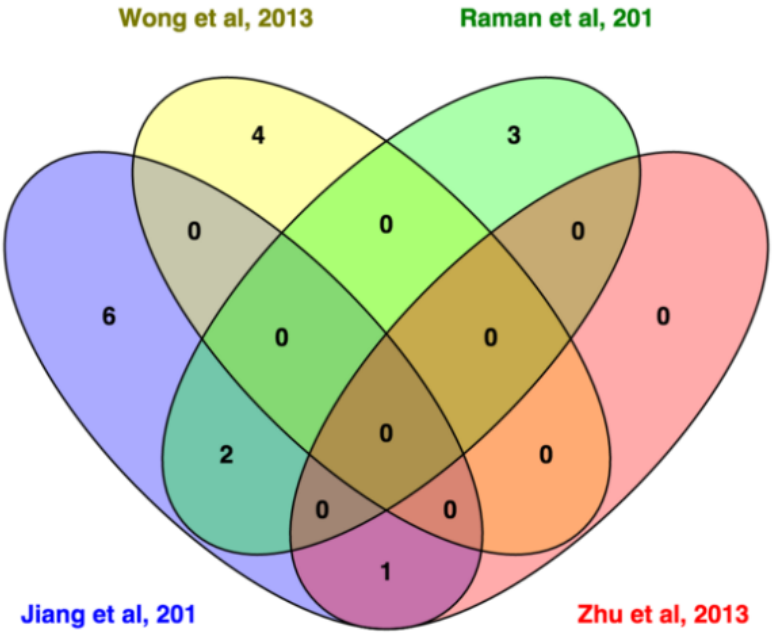


Figure 1.8: **Venn diagram of genera found to be differentially abundant by different studies between NASH/NAFLD and healthy controls.** Only 3 out of the 16 genera claimed to be differentially abundant were found in two studies: members of the *Escherichia* genus were found in the Zhu [136] and Jiang [52] studies, and members of the *Lactobacillus* and *Oscillibacter* genus were found in the Jiang [52] and Raman [102] studies. Of these, only Raman et al [102] reported using a multiple test correction.

Fig. 1.8 shows a Venn diagram illustrating the inconsistency of the literature on the gut microbiome and NAFLD. Of these, only Raman et al [102] reported using a multiple test correction. Many of the studies had healthy controls with a lower BMI, so it is difficult to

separate whether the differences found are related to NAFLD progression or obesity. These five studies do not form a consistent story about the gut microbiome and NAFLD. In one chapter of this thesis, we conduct our own non alcoholic steatohepatosis gut microbiome study, such that our results are replicable. Additionally, we generate the first deeply sequenced metagenomic sample set to examine functional capabilities in this disease.

Our hypothesis for the gene tag sequencing experiment was that we would find significantly differentially abundant taxa. For the metagenomic experiment, we hypothesized that we would find significantly differential groups of gene functions.

We compare the gut microbiome of 29 patients with nonalcoholic steatohepatitis (NASH), 14 patients with simple steatosis (SS), and 24 healthy controls. We found no significantly differentially abundant features between groups. However, the effect sizes of OTUs between extreme and intermediate groups appear to be correlated with a subset of the samples selected with more stringent clinical criteria. We believe that there may be a real difference between patients, if a study were performed with sufficient power.

Chapter 2

Expanding the UniFrac toolbox

Expanding the UniFrac toolbox

Ruth G Wong[1☯], Jia R Wu[1☯], Gregory B Gloor[1*]

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

☯These authors contributed equally to this work.
*** ggloor@uwo.ca**

Abstract

The UniFrac distance metric is often used to separate groups in microbiome analysis, but requires a constant sequencing depth to work properly. 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 ratio UniFrac, two new weightings that are not as 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 microbiome samples that incorporated phylogenetic distance [68]. The goal of the UniFrac distance metric was to enable objective comparison between microbiome samples from different conditions. In 2007, Lozupone added a proportional weighting to the original unweighted method [67]. Since then, papers reporting these metrics have garnered over a thousand citations, and enabled research about everything from how kwashiorkor causes malnutrition [117] to how people can have similar microbiomes to their pet dogs [118]. Except for generalized UniFrac, used to make hybrid unweighted and weighted UniFrac comparisons [16], few advances in the metric have occurred since 2007. In this paper we examine data sets where UniFrac gives misleading results, and present and discuss some alternative weightings for UniFrac.

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 are have 97% identity in a 16S rRNA gene variable region are considered to be the same taxa [17]. The 97% cutoff was arbitrarily chosen to best map sequence data to bacterial classifications. This threshold is thought to maximizes the grouping of bacteria classified as the same species while minimizing the grouping of bacteria classified as different species [11]. 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 [26]) that partitions the groups and then later assign taxonomic identity by matching the seed or central sequences with public databases, such as SILVA [101], the Ribosomal Database Project [18], or Greengenes [21]. Another method is closed reference OTU picking, which starts 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, except with closed reference OTUs, or individual sequence unit methods. Individual sequence unit (ISU) methods which do not use OTUs can be run with software such as DADA2 [10].

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

2.0.1 Data

UniFrac requires two pieces of information: a phylogenetic tree and a table of counts per inferred taxa per sample. These are derived from a gene tag sequencing experiment, such as the commonly used 16S rRNA gene [123]. The sequenced gene contains a variable region, allowing the sequences to be grouped into OTUs as described in the previous section. A count table can then be generated with the number of reads per OTU per sample. The center sequence of each OTU group can be put into a multiple sequence alignment, from which a phylogenetic tree can be inferred.

The phylogenetic tree is created through a multiple sequence alignment with the representative OTU sequences, using software such as MUSCLE [25] and FastTree [99], or using a guide tree, such as through Greengenes [21] or the QIIME software [12]. 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 is often done by rooting the tree at its midpoint.

2.0.2 Compositional Data Analysis

Microbiome data is in the form of a list of counts per feature (OTUs in this case), with the features composing an aspect of the microbiome for each sample. This is compositional data because the total sum of reads for a sample is arbitrary, being determined by the capacity of the

sequencing instrument [31] [32] [40]. There are several core truths about microbiome data and its compositional nature that should be considered when making an analysis strategy.

First, the total number of reads per sample is influenced by sample collection, extraction, sequencing library preparation, and sequencing platform, and is irrelevant to the biological implications of the data. Additionally, the constraint of the count total causes the abundance of different taxa to appear to be negatively correlated with each other when analyzed by conventional statistics [66]. 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. For example, one study compared the microbiome of vaginal swab samples from women with bacterial vaginosis (BV), women without BV, and women with intermediate BV, using qPCR to quantify the taxa [137]. *Prevotella* was found to increase through non-BV to intermediate to BV, while *Lactobacillus iners* stayed relatively the same [137]. If the same samples were put through a gene tag sequencing experiment where the taxa could not be quantified and the total read counts were constrained, one might incorrectly conclude that the abundance of *Lactobacillus iners* was decreasing while *Prevotella* was increasing.

To prevent incorrect conclusions, data should be analyzed in a compositional way. In Euclidean space, data points can increase or decrease freely. Compositional data is under a sum constraint, and exist in a non-Euclidean space known as the Aitchison simplex [1]. A data transformation can be performed to put the data into Euclidean space, so that it can be analyzed with standard statistical methods that depend on Cartesian coordinates and linear relationships. These transformations involve examining the ratios of different OTU abundances to each other, so that the total number of reads do not unduly affect the result [39] [38]. In the example with bacterial vaginosis, using ratios of taxa to each other would elucidate the nature of the biological change in the data.

2.0.3 Unweighted UniFrac

Unweighted UniFrac [68] 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. Figure 2.1 shows example calculations for UniFrac based on the tree overlap. 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.

As UniFrac is a binary test of absence, it is sensitive to sequencing depth, and assumes that the data has been normalized to a common sequencing depth [69]. Thus, rarefaction prior to unweighted UniFrac has become a standard part of the microbiome analysis workflow, with built in rarefaction functions in QIIME [12] and mothur [109].

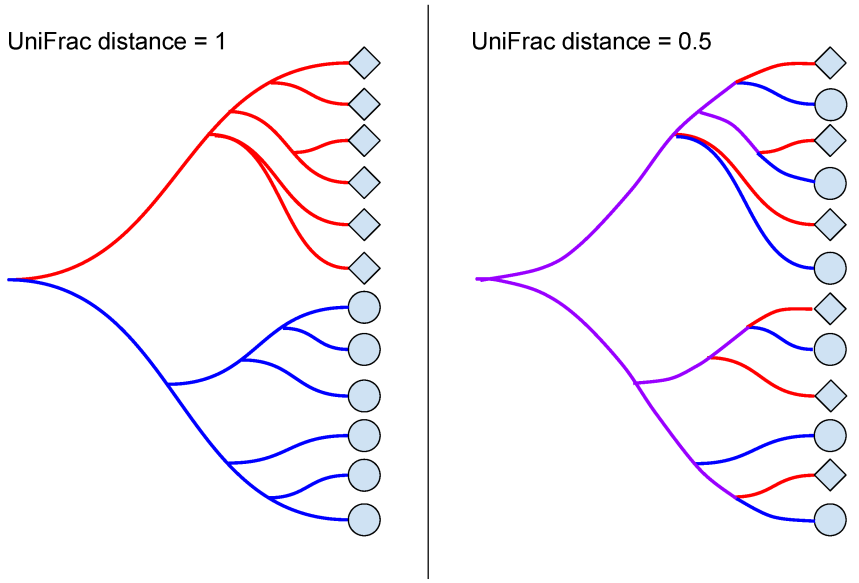


Figure 2.1: **Unweighted UniFrac.** When two samples do not share any branches of the phylogenetic tree, the unweighted UniFrac distance is maximized at 1. When two samples share half of their branch lengths on the phylogenetic tree, the unweighted UniFrac distance is 0.5. If the two samples contain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

2.0.4 Weighted UniFrac

Weighted UniFrac [67] is an implementation of the Kantorovich–Rubinstein distance in mathematics, also known as the earth mover's distance [30]. 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 binary weighting (unweighted UniFrac) [68], a linear proportion (weighted UniFrac) [67], or some combination of the two (generalized UniFrac) [16]. However, it is a misconception that the data are linear because the sum of the total number of reads is constrained by the sequencing machinery [35] [31] [32] [66] as described above.

Microbiome communities can exhibit tremendous variation in their total bacterial count. For example, a stool sample may produce more highly concentrated DNA extract than a skin swab sample, resulting in a different number of input molecules but a similar read count total. Vaginal samples from patients with bacterial vaginosis compared to patients without can have DNA

extract concentrations that differ one magnitude [137]. 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 understood how this can impact results.

Materials and Methods

2.0.5 Analytical techniques

Rarefaction

Rarefaction normalizes the samples OTU counts to a standard sequencing depth by sampling without replacement [116]. This resulting table can be thought of as a random point estimate of the dataset, as the output is a sub-sample without replacement of the original table. This standardization process is recommended by the authors of UniFrac [14] in order to account for the sensitivity of UniFrac to sequencing depth.

Rarefactions can be performed using the QIIME software [12] or using the vegan package in R [88].

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, A and B represent the two samples being compared, $\triangle$ is the symmetric difference between two sets, and $\cup$ is the union between two sets:

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

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.

Note that the implementation of unweighted UniFrac in QIIME (see Fig. 2.2) and also GUniFrac (see Supporting figures 2.9, 2.10, and 2.11) includes a tree pruning procedure, where the tree is pruned to only include OTUs that are present in each pairwise sample comparison. Except for in figure 2.2 and supporting figures 2.9, 2.10, and 2.11, the scripts used in this paper do not prune the tree, in order to be consistent with weighted UniFrac. In weighted UniFrac, pruning the tree makes the measurement a dissimilarity rather than a distance (Supporting Information 2.14).

Weighted UniFrac

Weighted UniFrac [67] 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:

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

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 or information (I) is calculated as follows, where p is the proportional abundance [113]:

$$I = -p \times log_2(p) \tag{2.1}$$

If a sample is composed of 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 about both taxa, because a given sequence read is more likely to be taxa A and less likely to be taxa B. 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:

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

Information UniFrac approaches a minimum of zero (Fig. 2.5) when a sample is composed of a monoculture. It also related to the Aitchison distance in compositional data analysis [29].

Ratio UniFrac

In complex microbiome communities, there may be a large number of bacterial taxa with few counts, such that the data is sparse. Taking the geometric mean of the proportional abundances of taxa in a microbiome sample represents an unbiased baseline of the average abundance of features with geometric growth characteristics - such as bacteria which divide by fission [1]. Experiments generally do not have power to detect differences at abundances below the mean [31]. Centering the proportional abundances around the geometric mean thus allows one to examine the data in this context, muting differences that are close to the baseline abundance and accentuating OTUs that are much more abundant than the mean. The formula for ratio UniFrac is as follows, where gm is the geometric mean:

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

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 that the data is very sparse. Whether a low-abundance taxa or feature 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 [31]. A Bayesian approach can be used to give a posterior estimate of the likelihood for zero count OTUs: this is implemented by the cmultRepl command in the zCompositions package in R [92].

The use of ratio weighting for UniFrac produces measurements that violate the metric triangle inequality, such that Euclidean statistics are technically invalid. Thus this metric, 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 ratio UniFrac [131]:

Bray-Curtis dissimilarity metric

The Bray Curtis dissimilarity metric [5] quantifies how dissimilar two sites are based on counts. A Bray-Curtis index of 0 means that two samples are identical, while a Bray-Curtis 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

2.0.6 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, and the data and scripts can be accessed at `https://github.com/ruthgrace/r_scripts` [132].

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 0.3 million 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 [127] 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, the resultant value (V_{res}) is computed per each first principal coordinate as defined by the formula:

$$V_{res} = \frac{|V_1 - V_i|}{range(V_1, V_i)}$$
$$\text{where } V_{res} = \text{Set of computed PC1s,}$$
$$V_1 = \text{Reference PC1 (the first),}$$
$$V_i = \text{Each subsequent PC1,}$$

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 0.4 million 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 [127] 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 [88] before the unweighted UniFrac distance was calculated. Weighted, information, and ratio UniFrac were calculated on the data set without rarefaction. The resulting distances were plotted for principal coordinate analysis.

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 a recent study [128]. There were a total of 5.3 million reads across 115 OTUs, and a minimum and maximum of 3072 and 2.8 million reads per sample. Note that the 2.8 million 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, which can be accessed at https://github.com/ruthgrace/ruth_unifrac_workshop [131]. 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 ratio UniFrac. Data was plotted using a principal coordinates or component plot as appropriate.

Monoculture data set

The monoculture data set is simulated based on the infected sample from the breast milk data set. Each simulated sample has exactly the same counts per taxa as the infected sample, except that the taxa are shuffled. After taxa shuffling, the data was manipulated into two groups. In one set of 20 samples the taxa with the highest count was swapped with *Pasteurella*, in another set of 20 the taxa with the highest count was swapped with *Staphylococcus*, and in the last set of 20 the taxa with the highest count was swapped with *Pseudomonas*. These three taxa were picked because they were the most highly abundant in the original breast milk data set. This process produced three sets of monocultures, dominated by the three different taxa.

Results

2.0.7 Unweighted Unifrac is highly sensitive to rarefaction instance

A commentary by Lozupone et al. 2011 [69] addressed the sensitivity of Unweighted UniFrac to sampling. Lozupone's group used mean UniFrac values to compute a confidence ellipse between the first and third quartile. 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 instances. 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 (subsampled) to the minimum of 659 to normalize the samples, prior to performing a principal coordinates analysis (PCoA). For Fig. 2.2, two independent rarefactions of the data were conducted in order to observe the effect of rarefaction instance on the metric. The unweighted UniFrac distance was computed for each rarefaction, and Procrustes adjustment was applied in order to overlay the PCoA-derived second rarefaction onto the first. A PCoA 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 side of the first coordinate axis to the other between the rarefaction instances, 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, several samples are observed to be unstable, exhibiting large differences in location. This example demonstrates that samples with little difference can appear to be different through the unweighted UniFrac distance metric and that rarefaction can lead to misleading and non-reproducible results.

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

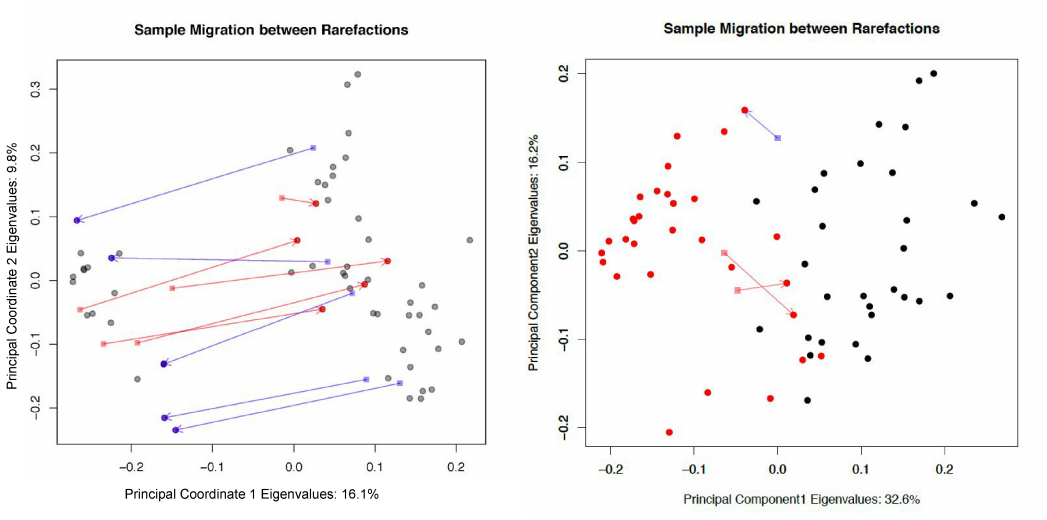


Figure 2.2: **Sample migration in different rarefactions, plotted on principal coordinates, measured with unweighted UniFrac.** The left plot is of the tongue data set while the right plot is the tongue dorsum vs. buccal mucosa data set. On the left panel red samples have moved from the left cluster 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. The tongue dorsum and buccal mucosa data set is included for comparison, with the tongue samples colored black and the buccal mucosa samples colored red. Note that the variance explained in the tongue data set 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 (compare with 32.6% and 16.2% in the tongue dorsum vs. buccal mucosa data set). The variance explained in the first and second coordinate in the 2011 UniFrac commentary [69] was even smaller, at 8.6% and 5.6%.

hundred separate rarefactions were conducted on the data to a minimum sampling depth of 659. For each individual rarefied OTU table, a distance matrix was computed using one of unweighted Unifrac, weighted UniFrac, Bray-Curtis Dissimilarity, information UniFrac, or ratio UniFrac as the weighting method. By generating 100 separate datasets for each metric, it is possible to assess the effect of rarefaction instance on each metric 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.

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 ratio UniFrac, as well as Bray-Curtis.

Given the wide use of unweighted UniFrac in the literature with small principal coordinate 1

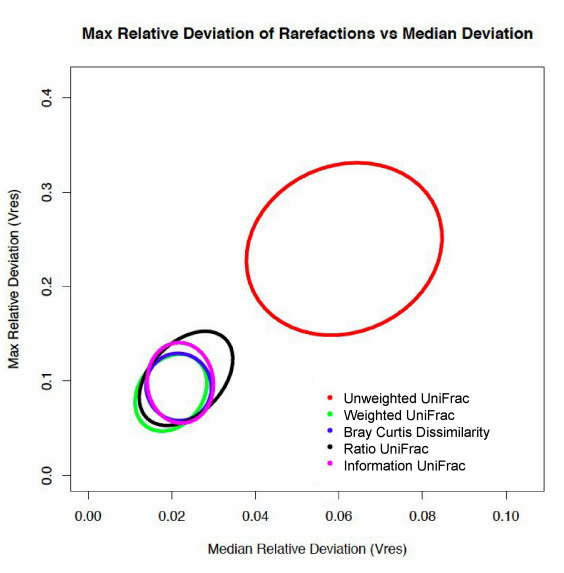


Figure 2.3: **Maxiumum relative deviation of rarefactions versus median deviation for traditional and non-traditional microbiome dissimilarity metrics.** Sixty samples from the tongue dorsum were taken from the Human Microbiome Project [127], and rarefied 100 times. The maximum relative deviation was 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 devation 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 dissimilarity, ratio UniFrac, and information UniFrac.

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[49], where the first and second principal coordinates axis explain 14% and 9.5% of the variation in Figure 2A, as well as in Nature [119], where the first principal coordinate explains 14% of the variation in Figure 1. In both of these examples, less variance is explained by the first principal coordinate than in our uniform tongue data set.

2.0.8 The cause of rarefaction variation by Unweighted Unifrac

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 [42]. A combination of unevenly sampled OTUs and distantly related OTUs will contribute to the variability in unweighted UniFrac 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.

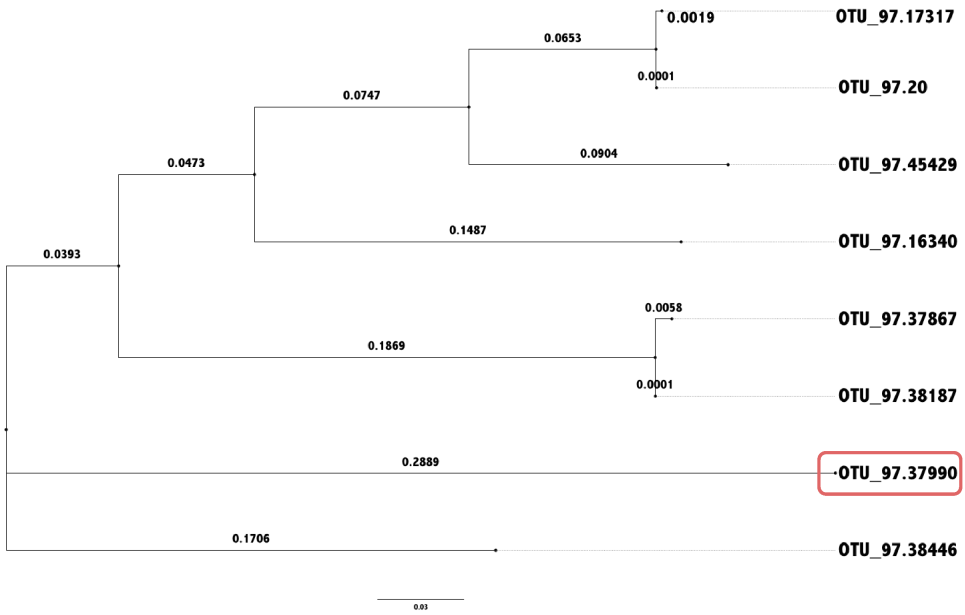


Figure 2.4: **Phylogenetic tree with long isolated branches.** Variation in different rarefactions of data in unweighted UniFrac analysis is exacerbated by the presence of long isolated branches in the phylogenetic tree, such as the circled OTU in this example.

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. Originally, sample A contained 1075 counts and sample B contained 221 counts in total. Both samples were rarefied to 221 counts, twice. The OTU in bold has been rarified to a zero count in sample A for one instance and a non zero count in the other instance. In Rarefaction 1, the unweighted UniFrac distance (unshared over total branches) is 0.4175, while in Rarefaction 2 the distance is 1.12.

OTU.ID	**A**	**B**	**A R1**	**B R1**	**A R2**	**B 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

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 with the bacterial vaginosis sample earlier, 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.

2.0.9 Information UniFrac

The difference in information content between taxa with 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 samples with subtle differences, such as the infected breastmilk sample in Fig. 2.7.

For example, Fig. 2.5 shows the weighting of a taxon in unweighted, weighted, and information UniFrac as a function of the taxon proportional abundance. Near the 0, 0 point the proportional abundances are low and information is 0. However, small increases in abundance result in large changes in contribution to UniFrac weighting, as shown by the slope of the curve. 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 ratio UniFrac (not depicted) depends on the geometric mean of the taxonomic abundances, and each sample would have a different slope in the weight graph depending on how evenly the abundances were distributed.

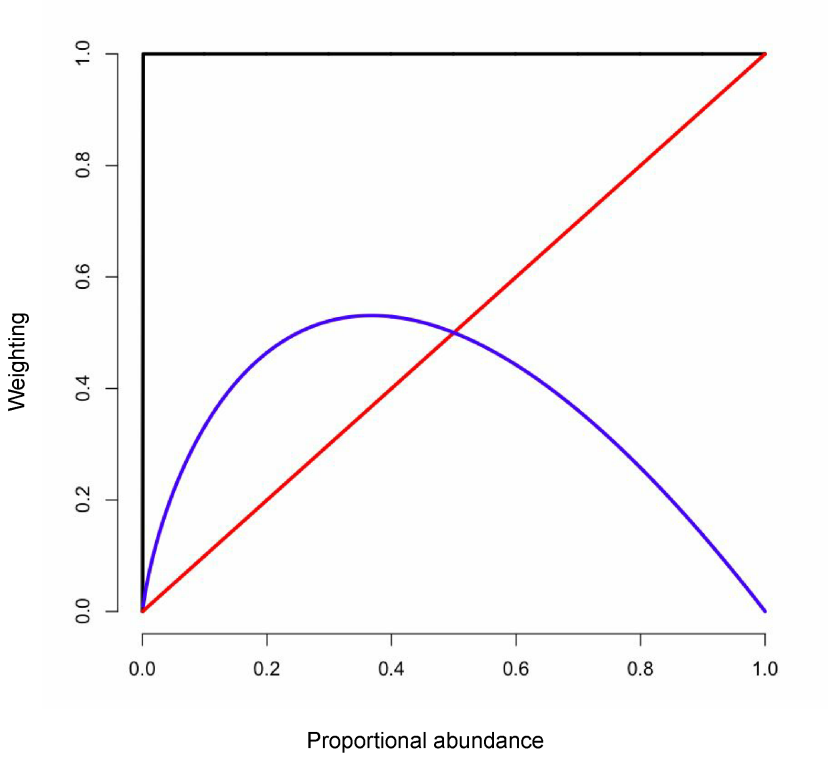


Figure 2.5: **UniFrac weights.** Each UniFrac weighting is plotted with the corresponding proportional abundance. The black line is unweighted UniFrac, the red line is weighted UniFrac,and the blue line is information UniFrac. From 0 to 0.2 on the x-axis information UniFrac has a higher slope, and therefore more discovery power for smaller changes in abundance. As the x-axis approaches 1, changes in abundance add little discovery power to information Unifrac.

2.0.10 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 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 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.

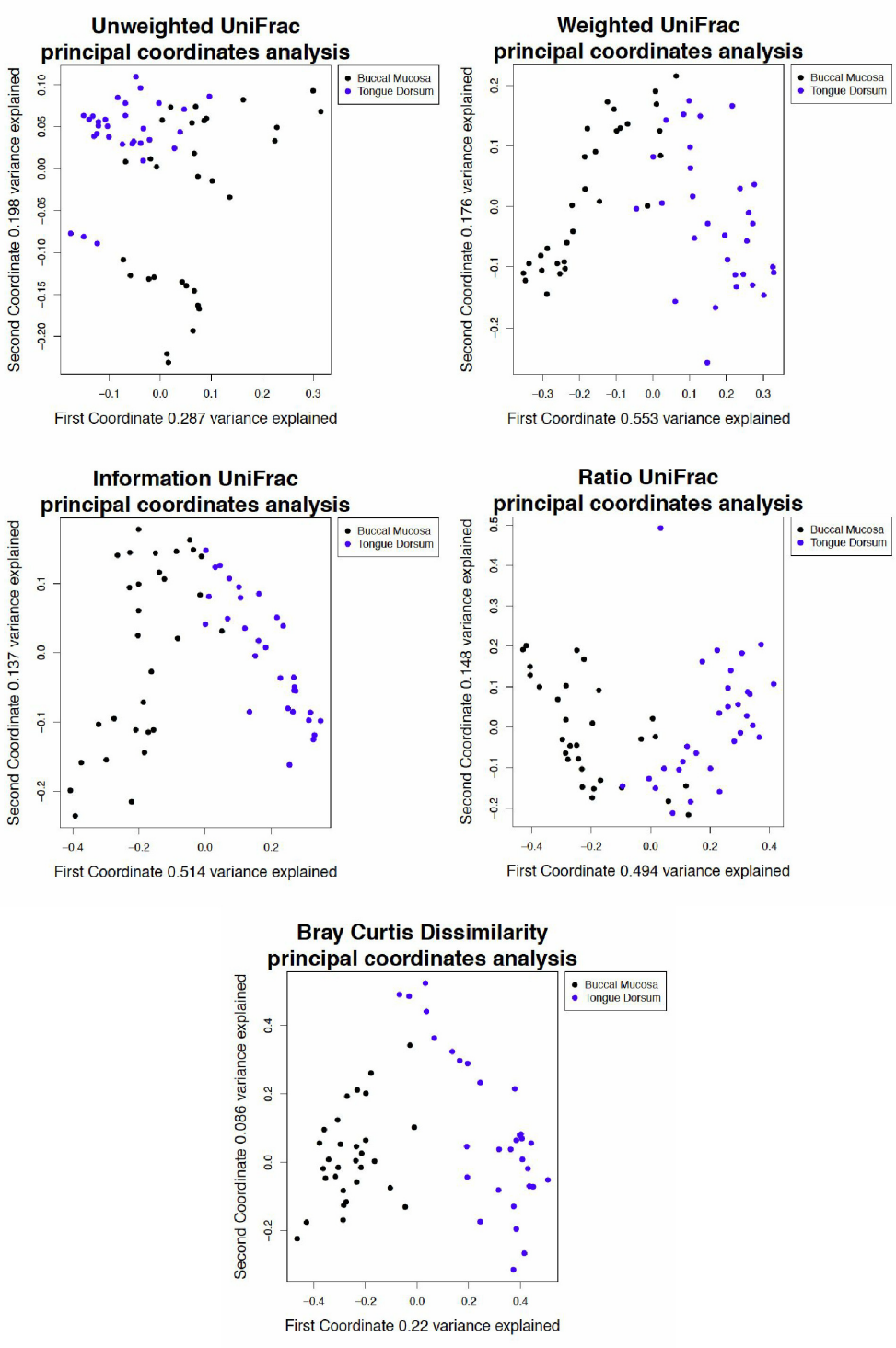


Figure 2.6: **Analysis of tongue and buccal mucosa data using different UniFrac weightings.** A principal coordinate analysis of a 16S rRNA gene tag experiment done on samples from the tongue and buccal mucosa, selected from the Human Microbiome Project [127]. All weightings and the Bray-Curtis dissimilarity 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.

2.0.11 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 [128]. 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 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. Ratio UniFrac recognizes that the infected sample has a taxonomic abundance very far from the geometric mean abundance. For these reasons information and ratio UniFrac are more adept at picking up outliers with uneven distributions, even if the taxa are shared by other samples.

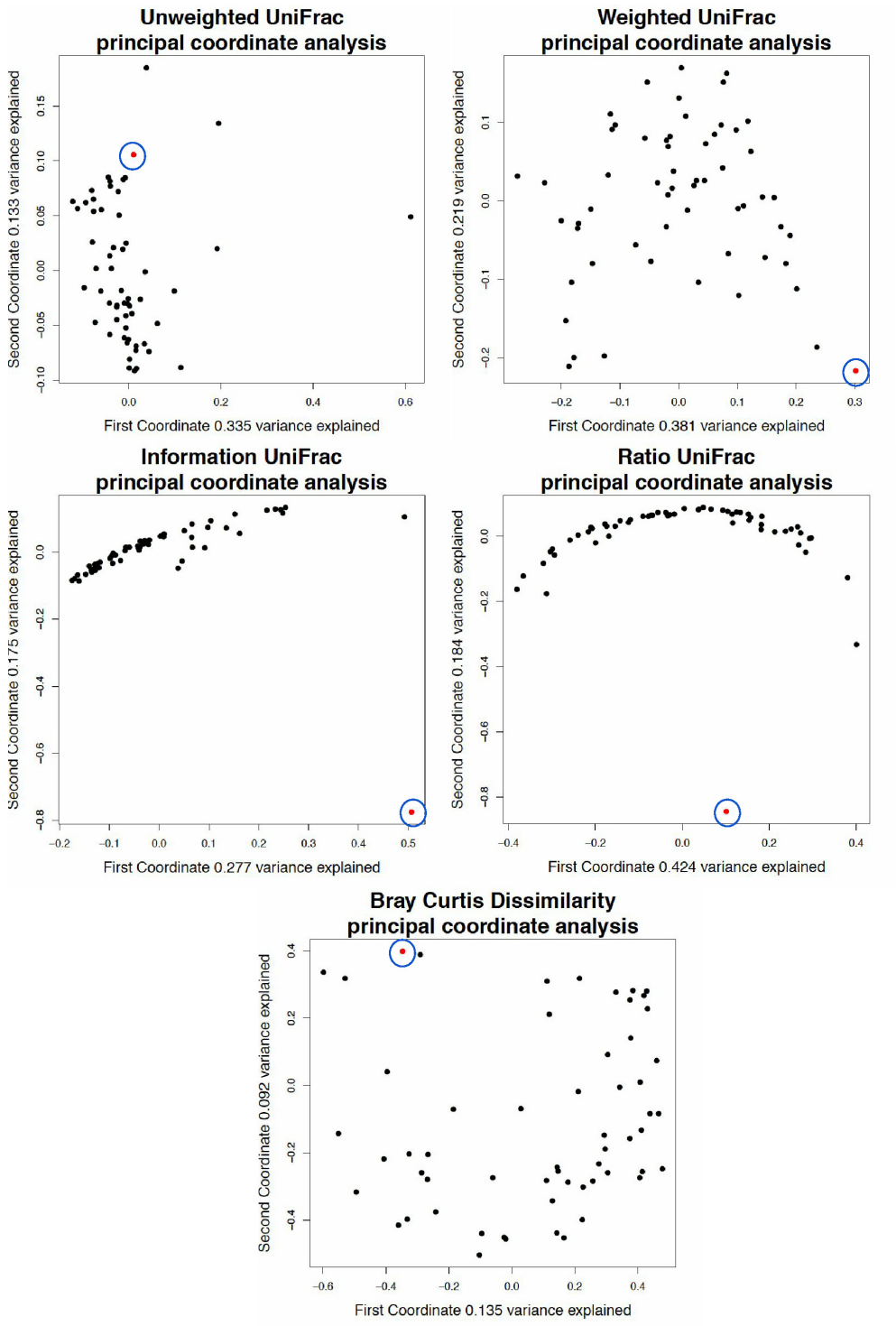


Figure 2.7: **Analysis of breast milk data using different UniFrac weightings.** A principal coordinate analysis of a simulated 16S rRNA gene tag experiment based on the breast milk data. Red samples are dominated at 07% by *Pasteurella*, black samples are dominated by *Staphylococcus*, and cyan samples are dominated by *Pseudomonas*. Note that while information Unifrac appears to separate the samples reasonably well visually, the amount of variance explained by the first two coordinates is much lower than even weighted UniFrac.

2.0.12 Monoculture data

Each sample in the monoculture dataset is 97% dominated by one of three taxa. However, within the remaining 3% there is variation in the counts.

Unweighted UniFrac, being a binary test, detects only the variation in the remaining 3% of counts, without showing the difference in the monocultures. Weighted UniFrac detects only the difference in the identity of the monoculture, and the separation is driven by phylogenetic distance - the pairwise distance from *Pasteurella* to *Staphylococcus* and *Pseudomonas* to *Staphylococcus* is just over 0.9 on the phylogenetic tree while the distance from *Pasteurella* to *Pseudomonas* is 0.45. This is in correspondence with the PCoA plot where the first coordinate (which separates the *Staphylococcus* species from the other two) explains over 90% of the variance in the data set.

Information UniFrac is known to not perform very well for monocultures, due to taxa with very high and low proportional abundances having uncertainty information values close to zero (Fig. 2.5). While the samples separate visually with information UniFrac, the variance explained by the separation is low, and the distance matrix does not separate the three groups by hierarchical clustering. Ratio UniFrac and Bray Curtis both separate the samples by monoculture, and also differentiate the samples by their minor variations, showcasing a more representative perspective of this data set.

If the samples are hierarchically clustered, the three groups separate perfectly with weighted UniFrac, ratio UniFrac, and Bray Curtis dissimilarity, but not with unweighted UniFrac or information UniFrac.

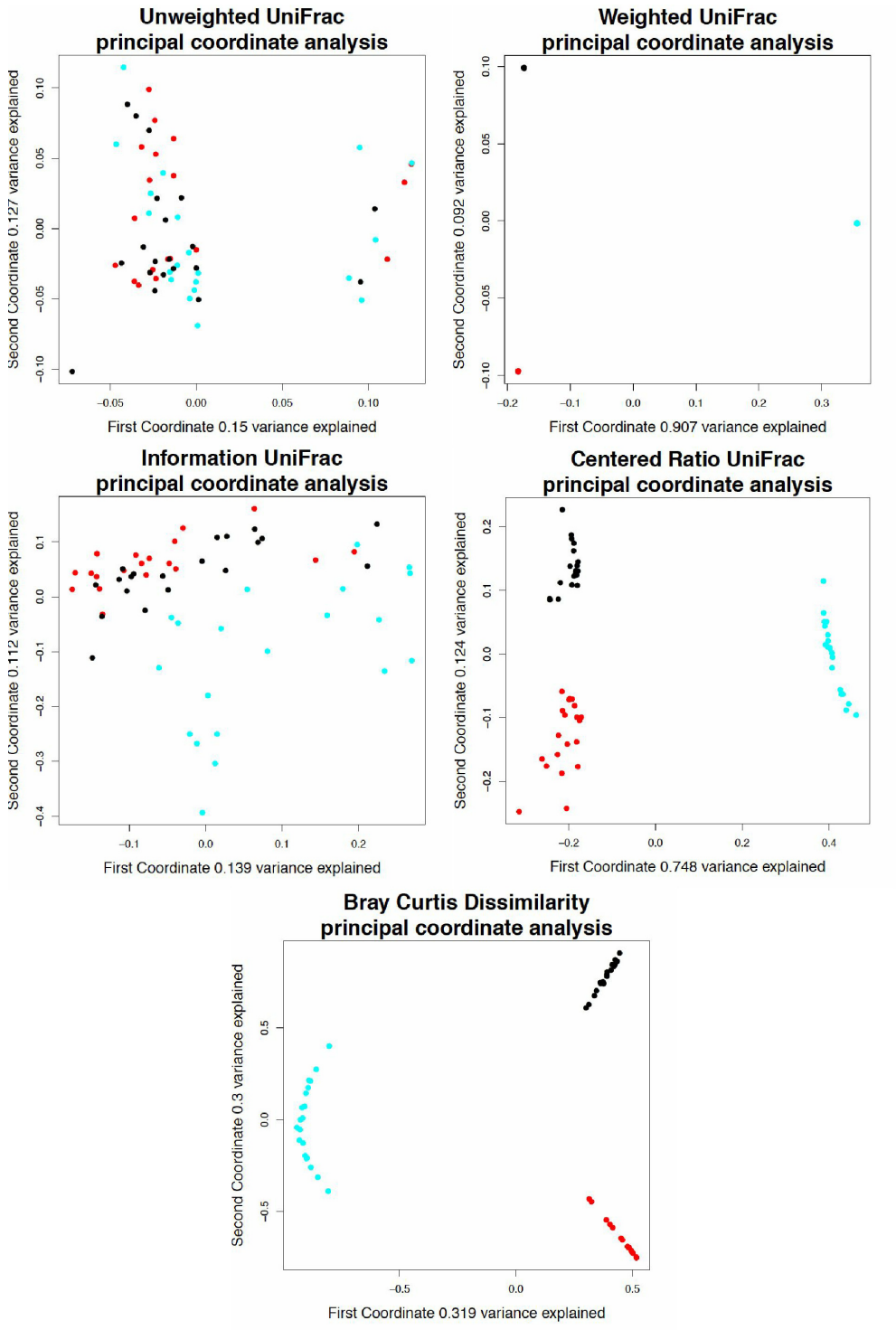


Figure 2.8: **Analysis of simulated monocultures using different UniFrac weightings.** A principal coordinate analysis of a simulated 16S rRNA gene tag experiment based on the breast milk data. Red samples are dominated at 97% by *Pasteurella*, black samples are dominated by *Pseudomonas*, and cyan samples are dominated by *Staphylococcus*. Note that while information Unifrac appears to separate the samples reasonably well visually, the amount of variance explained by the first two coordinates is much lower than even weighted UniFrac.

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, 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 the example in the breast milk data set.

We do not believe that any of these weightings are a perfect model for microbiome data. Each tool is prone to its own set of weaknesses. If the difference in groups is driven by presence/absence then UniFrac is a reasonable choice. If the difference is driven by a linear abundance, then weighted UniFrac is a good choice. Information UniFrac and ratio UniFrac are useful for examining data sets that contain a similar set of taxa between groups. Information and ratio UniFrac are especially useful for examining data sets that have more subtle variations, due to their non linear nature. In any case, inspection should be done to make sure that the tool used accurately represents the data.

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

Supporting Information

Acknowledgments

Thanks to Camilla Urbaniak for providing the data from the breast milk study [128].

unifrac

S1 Fig.

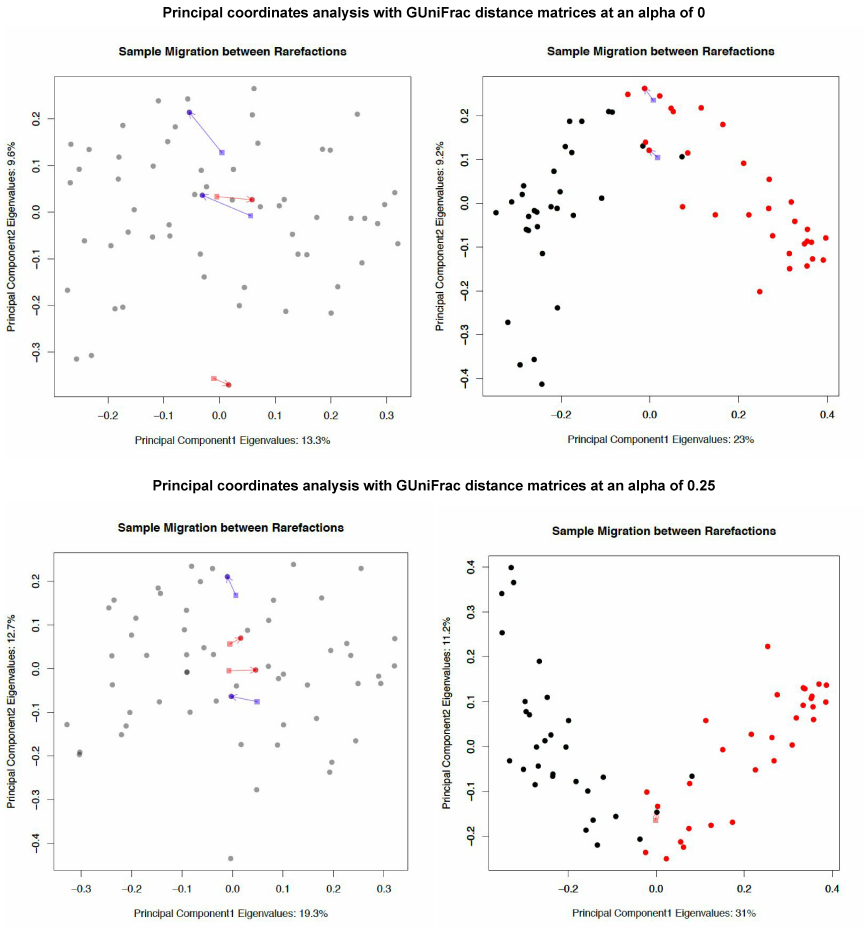


Figure 2.9: **Principal Coordinate Analysis derived from GUniFrac distance matrices.** GUniFrac was run with an alpha of 0 and 0.25. Note that GUniFrac, like QIIME, prunes the tree with every pairwise comparison. That is, the phylogenetic tree used for the distance calculation for each pair of samples can be different. The resulting measurements are a dissimilarity, not a distance. Additionally, QIIME gives slightly different values from GUniFrac, but the source of this (likely an additional normalization) is not known.

S2 Fig.

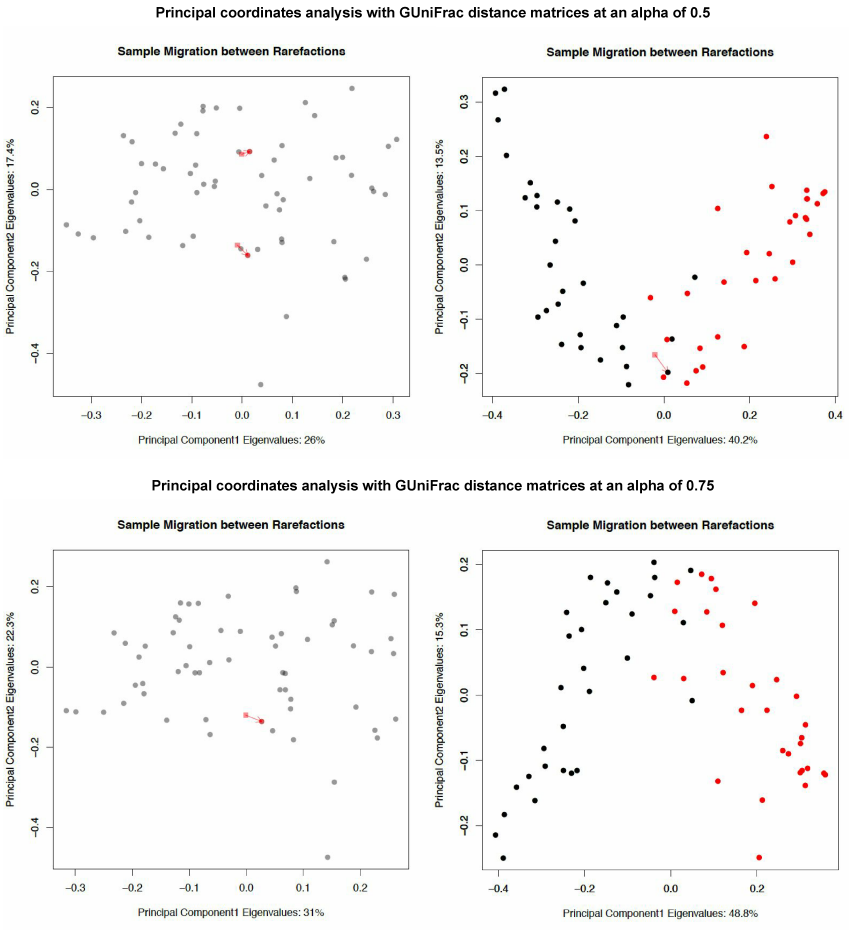


Figure 2.10: **Principal Coordinate Analysis derived from GUniFrac distance matrices.** GUniFrac was run with an alpha of 0.5 and 0.75. Note that GUniFrac, like QIIME, prunes the tree with every pairwise comparison. That is, the phylogenetic tree used for the distance calculation for each pair of samples can be different. The resulting measurements are a dissimilarity, not a distance. Additionally, QIIME gives slightly different values from GUniFrac, but the source of this (likely an additional normalization) is not known.

S3 Fig.

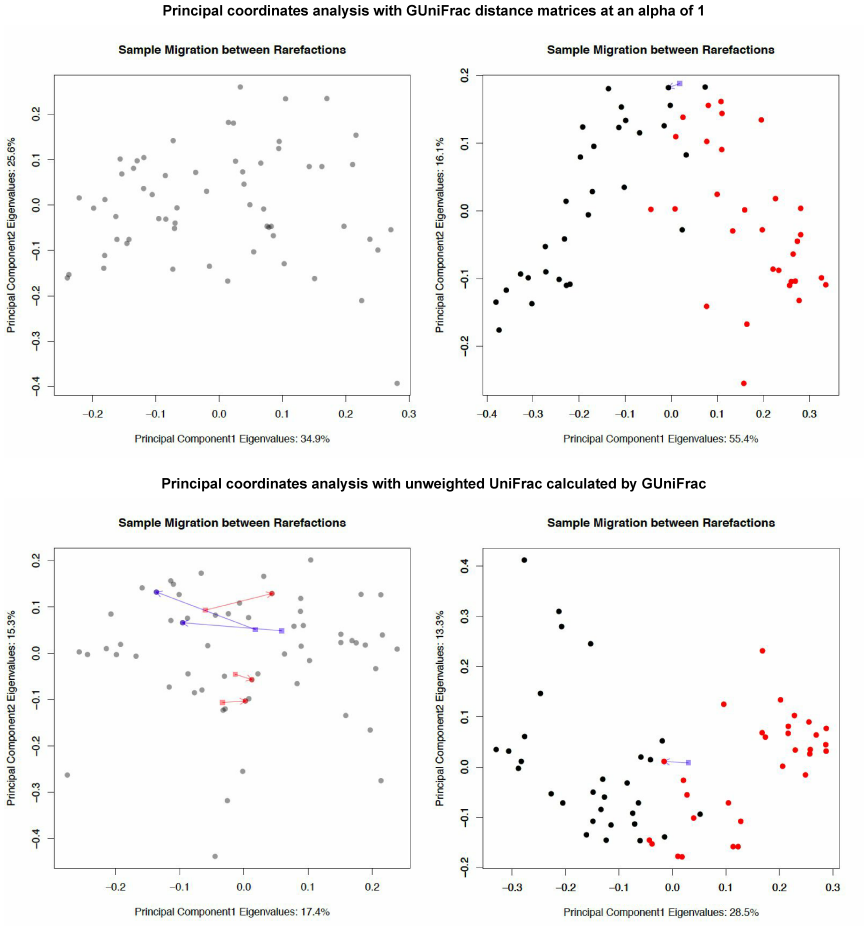


Figure 2.11: **Principal Coordinate Analysis derived from GUniFrac distance matrices.** GUniFrac was run with an alpha of 1 (equivalent to weighted UniFrac), plus unweighted UniFrac for comparison. Note that GUniFrac, like QIIME, prunes the tree with every pairwise comparison. That is, the phylogenetic tree used for the distance calculation for each pair of samples can be different. The resulting measurements are a dissimilarity, not a distance. Additionally, QIIME gives slightly different values from GUniFrac, but the source of this (likely an additional normalization) is not known.

S4 Fig.

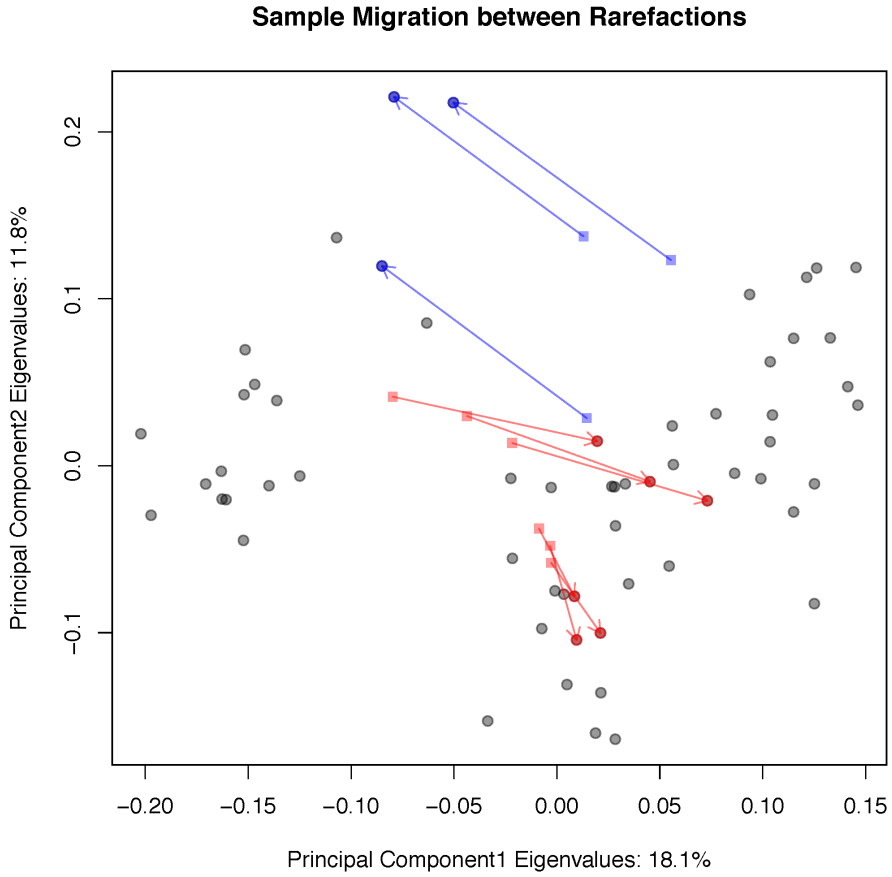


Figure 2.12: **Principal coordinate analysis derived from tongue dorsum samples using unweighted UniFrac distance matrices with no tree pruning.**

S5 Fig.

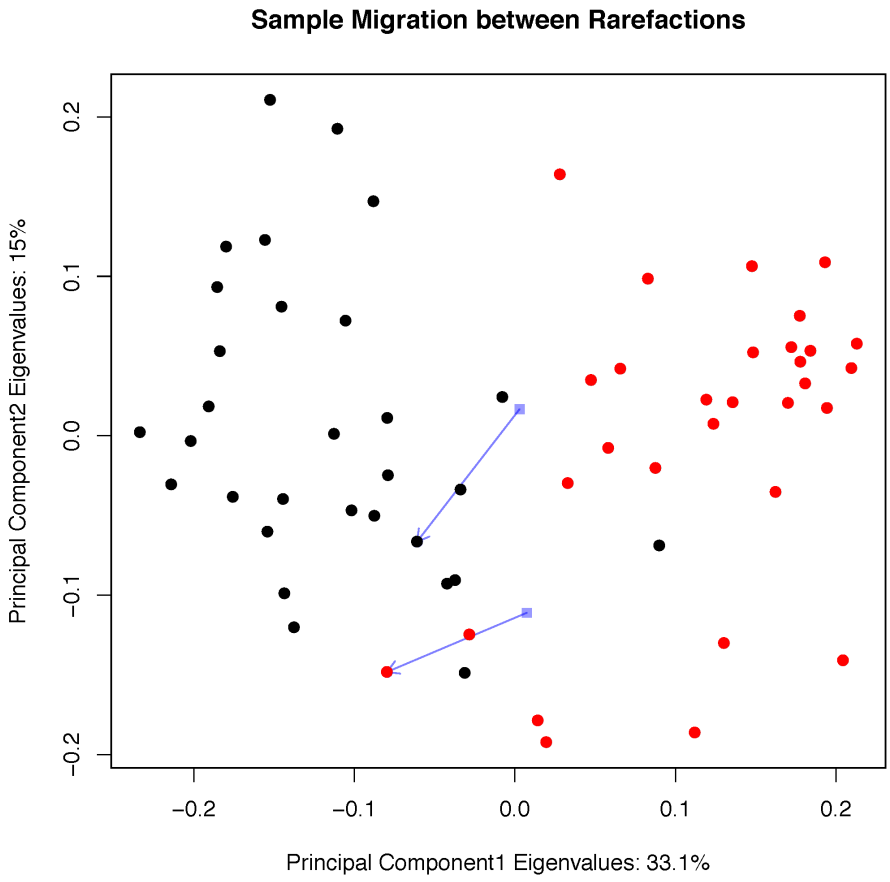
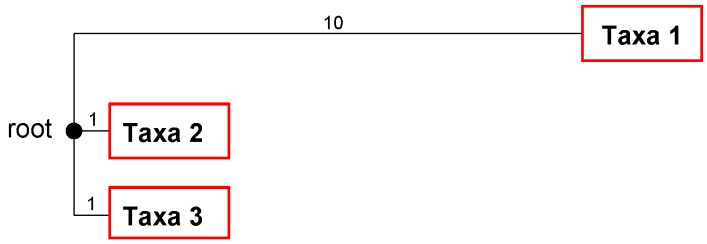


Figure 2.13: **Principal coordinate analysis derived from tongue dorsum and buccal mucosa samples using unweighted UniFrac distance matrices with no tree pruning.**

S6 Fig.



Samples	Taxa 1 counts	Taxa 2 counts	Taxa 3 counts
Sample A	10		
Sample B	10	1	10
Sample C			10

Figure 2.14: **Weighted UniFrac is a dissimilarity with tree pruning.** Here, the weighted UniFrac measurements without tree pruning are: $W_{AB} = \frac{121}{252}$, $W_{BC} = \frac{111}{252}$, and $W_{AC} = \frac{11}{12}$. With tree pruning, the measurements are: $W_{AB} = \frac{121}{252}$, $W_{BC} = \frac{111}{252}$, and $W_{AC} = 1$, which fails the triangle inequality.

Chapter 3

The human microbiome and nonalcoholic fatty liver disease

3.1 Introduction

Non alcoholic fatty liver disease (NAFLD) has been on the rise along with obesity, affecting approximately 25% of the North American population [97]. Most people with NAFLD remain asymptomatic, however, in up to a third of patients NAFLD can progress to nonalcoholic steatohepatitis (NASH), causing inflammation and scarring (fibrosis) in the liver, and decreasing the 5 year survival rate to 67% [100]. It is thus important to shed some light on the process by which people progress from NAFLD to NASH to find interventions that prevent NASH.

3.1.1 NASH progression risk

There are several known genetic and chemical factors that increase the risk of progression to NASH in animal models and humans.

In mice non alcoholic fatty liver disease is often modelled with a methionine/choline-deficient diet (MCD), which induces steatohepatitis in wild type mice. Mice with a toll-like receptor 4 knockout had lower lipid and injury accumulation markers when fed a MCD diet, implying that toll-like receptor 4 has a role in the progression of NAFLD [106].

In rats liver fibrosis can be induced by dosing with dimethylnitrosamine. Yasuda et al. [134] found that male rats were more prone to this induced liver fibrosis than female rats. Fibrosis biomarkers were reduced when the male rats were dosed with estradiol, and increased when the male rats were additionally given an estradiol-neutralizing antibody. Female rats who had their ovaries removed similarly lost the protective effect [134]. From this, it was concluded that hormones are also a factor in nonalcoholic fatty liver disease progression.

In humans, the I148M variant of the Patatin-Like Phospholipase Domain Containing 3 gene (PNPLA3) correlates with a 3.2 fold increased risk of progression to NASH from NAFLD when homozygous, compared to to patients without the variant [120]. The heterozygous gene was found to be associated with fatty liver disease in genome wide association studies, but some additional studies have failed to replicate the relationship with NASH [120].

On the epigenetic level, many genes are differentially methylated in the livers of patients

with advanced NAFLD compared to patients with mild NAFLD. Eleven percent of genes are differentially hypomethylated in advanced NAFLD (compared to 3% hypermethylated), leading to increased expression [85]. In advanced NASH specifically, some tissue repair genes were hypomethylated while some metabolism pathways such as 1-carbon metabolism were hypermethylated. However, only 7% of the differentially methylated genes were found to be differentially transcribed [85].

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 [102]. Reactive oxygen species have also been implicated in NASH due to their involvement in the mechanism of steatohepatitis-inducing drugs [7].

The microbiome is thought to have an effect on host digestion and absorption of nutrients [37]. Anaerobic carbohydrate fermentation produces short chain fatty acids, which make up 10% of the calories in a Western diet [81] Some groups claim a link between ethanol-producing gut bacteria and NAFLD [136] [52], however the evidence was inconclusive since no multiple test correction was performed.

Non alcoholic fatty liver disease is related to obesity, and studies in mice reported that the mouse gut microbiome associated with diet-induced obesity can be transplanted to lean mice and cause them to gain weight while eating the same amount of food [125]. The gut microbiome is an important factor in obesity and potentially obesity related ailments, and its relationship to NAFLD warrants investigation.

3.1.2 Data

Applying next generation DNA sequencing techniques to microbiome research is a relatively new field that has yet to set data analysis standards. There are some considerations that should be made when constructing a data analysis strategy.

Data is multivariate

Experiments of this nature typically have low sample sizes due to budget constraints, sample collection difficulties, patient compliance, and other issues. As a result, the number of taxon or gene function comparisons made are often 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 standard statistical techniques are to be reliable [89].

At a minimum, researchers should include multiple test corrections to ensure that the results they are reporting are more likely to be reproducible, 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 four out of five of the papers in the literature about the gut microbiome and NAFLD (Fig. 3.1).

Data is compositional

In both gene tag sequencing and metagenomic sequencing experiments, the data is in the form of a list of counts per feature, with the features composing an aspect of the microbiome for

each sample. This is compositional data. The total number of reads yielded by the sequencing platform is often platform-dependant and not biologically relevant.

This constrained sum causes the abundance of different taxa to appear to be negatively correlated with each other when analyzed by conventional statistics [39]. 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. In addition, correlation and covariance are invalid when calculated from compositional data. Clustering, and dimension reduction are therefore unreliable [66].

Compositional data should be analyzed in a compositional way. In Euclidean space, data points can increase or decrease freely. Compositional data is under a sum constraint, and exist in a non-Euclidean space known as the Aitchison simplex [1]. Data transformations such as the centered log ratio can be performed to put the data into Euclidean space, so that it can be analyzed with standard statistical methods that depend on Cartesian coordinates and linear relationships.

However, these techniques are not yet mainstream in the field, resulting in a high number of conclusions made that are not reproducible.

3.1.3 Literature

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

Jiang et al, 2015 [52] compared 53 NAFLD patients with 32 healthy controls. The NAFLD patients had a significantly higher BMI ($P < 0.01$). Each sample had an average of 0.6 million reads, from sequencing the V3 region of the 16S rRNA gene on the Illumina sequencing platform. The reads were annotated with the Ribosomal Database Project [18]. Differential abundance was determined using Projection on Latent Structures - Discriminant Analysis (PLS-DA) methods, which is not statistically valid when performed on compositional data. They found a relative increase in members of the *Lentisphaerae* phyla and the *Oscillibacter* and *Flavonifractor* genera in the healthy group, and a relative increase in members of the *Clostridium XI*, *Anaerobacter*-related, *Streptococcus*, and *Lactobacillus* genera in the NAFLD group.

Zhu et al, 2013 [136] compared 16 non-obese controls, 25 obese patients, and 22 NASH patients. All of the patients were pediatric, and the NAFLD group all had a BMI higher than the 85th percentile while the healthy group had BMIs less than the 85th percentile. A 16S rRNA gene tag sequencing experiment was performed and reads were sequenced in a 454 pyrosequencer. This group used MG-RAST [83] and QIIME [12] to analyze their data. Note that in the PCoA plot (Figure 1 in [136]), there is only 11% variance explained by the first co-ordinate axis, and they had to plot the first co-ordinate axis with the 3rd co-ordinate axis (3% variance explained) to show the group separation. By comparing the average absolute read count for each taxa in each group, this group found that members of the *Proteobacteria* phylum, the *Enterobacteriaceae* family, and the *Escherichia* genus had significantly higher average counts in NASH patients compared to obese patients and healthy controls.

Raman et al, 2013 [102] compared 30 NAFLD patients with 30 healthy controls. All the healthy controls had a BMI less than 25 while all the NAFLD patients had a BMI greater than 30. The 16S rRNA gene was amplified and sequenced with 454 pyrosequencing, yielding 2000 reads per sample. Reads were annotated with the Ribosomal Database Project [18]. UniFrac

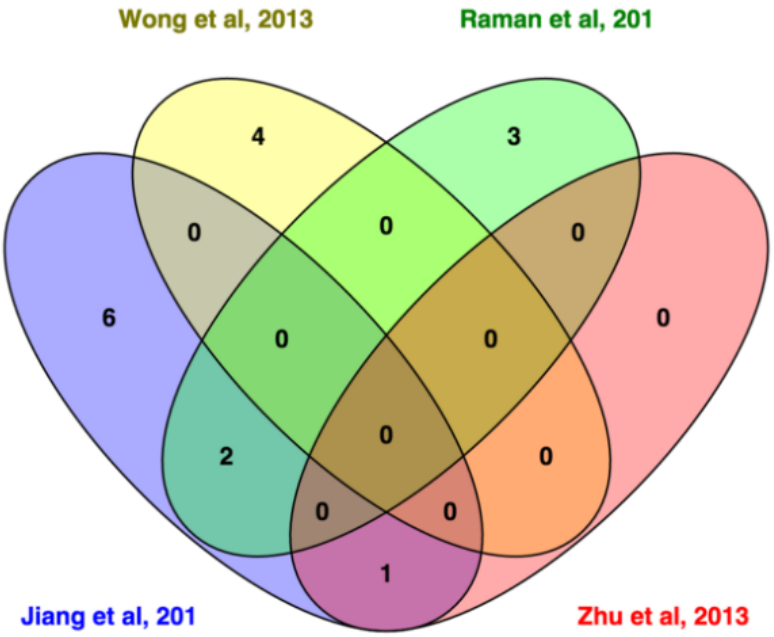


Figure 3.1: **Venn diagram of genera found to be differentially abundant by different studies between NASH/NAFLD and healthy controls.** Only 3 out of the 16 genera claimed to be differentially abundant were found in two studies: members of the *Escherichia* genus were found in the Zhu [136] and Jiang [52] studies, and members of the *Lactobacillus* and *Oscillibacter* genus were found in the Jiang [52] and Raman [102] studies.

analysis was performed with QIIME [12], and differential abundance was tested with Metastats [93]. They found a relative increase in members of the *Lactobacillus*, *Robinsoniella*, *Roseburia*, and *Dorea* genus in NASH patients and a relative increase in members of the *Oscillibacter* in healthy patients.

Wong et al, 2013 [133] compared 16 NASH patients with 22 healthy controls. They amplified the V1-V2 variable region of the 16S rRNA gene with pyrosequencing, yielding 4-11 thousand reads per sample. Reads were clustered with UCLUST [26] and annotated with the Ribosomal Database Project [18]. Members of the the genera *Parabacteroides* and *Allisonella* were found to be relatively increased in NASH patients, while members of the genera *Faecalibacterium* and *Anaerosporobacter* were relatively increased in healthy controls.

Boursier et al, 2015 [8] compared 30 patients with normal liver fibrosis (stage F0 or F1) to 27 patients with stage F2 or greater fibrosis, 35 of which had NASH A gene tag experiment was performed on the V4 region of the 16S rRNA gene, and sequenced on an Illumina platform, yielding an average of 0.2 million reads per sample. Reads were annotated with the Greengenes database [21], and differential abundance was measured by a Mann-Whitney test. A metagenomic imputation was performed with PiCrust [60], annotated with KEGG [55], and analysed with LEfSE [110]. A relative increase in members of the *Bacteroides* phylum and a relative decrease in members of the *Prevotella* phylum was found in NASH, compared to healthy controls. From the metagenomic imputation, the gut microbiome of NASH patients was found to be significantly enriched in functional categories related to carbohydrate,

lipid, amino acid, and secondary metabolism.

Many of the studies had healthy controls with a lower BMI, so it is difficult to separate whether the differences found are related to NAFLD progression or obesity.

Fig. 3.1 shows a Venn diagram illustrating the inconsistency of the literature on the gut microbiome and NAFLD. Of these, only Raman et al. [102] reported using a multiple test correction.

Since these five studies do not form a consistent story about the gut microbiome and NAFLD, we conducted own analysis using a compositional data (CoDa) approach with effect size as the primary measure, such that our results are replicable [47]. Additionally, we generate the first deeply sequenced metagenomic sample set to examine functional capabilities in this disease.

3.2 Methods

In total, 67 samples were collected: 29 from patients with nonalcoholic steatohepatitis (NASH), 14 from patients with simple steatosis (SS), and 24 from healthy controls. The median BMIs were 26.70, 27.34, and 32.06, and the median ages were 36, 49, and 46.5 for healthy, SS, and NASH respectively. A full description of the patient intake, metadata collection, and sample harvesting procedures are provided in Appendix B, written by Hannah Da Silva from Allard research group in Toronto.

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 Celsius, between steps 2 and 3.

3.2.1 16S rRNA gene tag experiment

DNA was amplified by PCR using the Earth Microbiome V4 primer set [13], with the addition of combinatorial in-line barcodes so that all the samples could be sequenced in the same sequencing run [40]. The DNA was sequenced on the Illumina MiSeq platform with paired end 220 nucleotide reads, producing 25 million reads in total.

Reads were overlapped with Pandaseq [76], clustered into Operational Taxonomic Units (OTUs) using UCLUST [26], and annotated with the SILVA database [101] using mothur [109], producing a table of counts per OTU per sample. Twelve million (48%) of the reads were successfully overlapped and annotated into 232 OTUs. Differential abundance was analyzed using ALDEx2 [32].

A generalized workflow for processing 16S rRNA gene sequencing reads is available at https://github.com/ggloor/miseq_bin. The workflow for the 16S rRNA gene tag experiment analysis from the count table stage is on GitHub: https://github.com/ruthgrace/nafld_metaphlan_pca.

3.2.2 MetaPhlAn

MetaPhlAn (Metagenomic Phylogenetic Analysis) [111] 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.

The MetaPhlAn tutorial was followed, using an additional `marker_counts` option in the `merge_metaphlan_tables.py` step to produce a count table instead of a relative abundance table. The workflow for the MetaPhlAn analysis from the count table stage is on GitHub: https://github.com/ruthgrace/nafld_metaphlan_pca.

3.2.3 Metagenomic experiment

Study inclusion criteria
BMI > 40 kg/m2 or BMI > 35-40 kg/m2 with severe weight loss responsive comorbidities, i.e. DM2, hypertension, hyperlipidemia, sleep apnea and/or gastroesophageal reflux disease or physical problems interfering with lifestyle and who have been assessed by the multidisciplinary bariatric team as suitable candidates for laparoscopic RYGB
Male and female
Age 18 years or older
Alcohol consumption < 20g/d
If known to have hyperlipidemia or DM2, need to be stable drug regimen for at least 3 months prior to study entry
Study exclusion criteria
Liver disease of other etiology
Advanced liver disease (need for liver transplantation in one year or complications such as variceal bleeding, ascites or jaundice)
Abnormal coagulation or other reasons contraindicating a liver biopsy
Medications known to precipitate steatohepatitis 6 months prior to entry
Regular intake of non-steroidal anti-inflammatory drugs; prebiotics, probiotics or antibiotics, ursodeoxycholic acid or any experimental drug in the 3 months prior to study entry
Type 1 diabetes
Chronic gastrointestinal diseases
Previous gastrointestinal surgery modifying the anatomy (prior to bariatric surgery)
Smoking
Pregnancy or breastfeeding
Patients not tolerating Optifast, which is a standard weight loss diet given to all patients pre-bariatric surgery

Table 3.1: **List of overall study inclusion and exclusion criteria.** This table lists the inclusion and exclusion criteria for the 16S rRNA gene tag experiment.

Study inclusion criteria
NASH severity
Study exclusion criteria
Took antibiotics at any point
Started Optifast diet early
Sample not frozen immediately after collection
Blood glucose over 7.8 mmol/l

Table 3.2: **List of inclusion and exclusion criteria for metagenomic study.** Patients were selected for the metagenomic study out of the patients selected for the 16S rRNA gene tag sequencing study with the following criteria. Ten healthy and ten patients with NASH were selected in total.

A metagenomic sequencing experiment was performed using total bacterial DNA from 10 healthy controls and 10 of the patients with NASH. Samples from healthy patients were selected to exclude confounding factors (Table 3.2). Samples from NASH patients were selected for the most extreme NASH phenotype, and had higher effect sizes in the 16S rRNA gene tag experiment than the full NASH group.

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 into reads per individual sample, yielding nearly 2 billion reads in total.

We used a two pronged strategy to annotate the reads:

First, we created a reference library using the inferred taxa from the 16S rRNA gene tag experiment. For each genus observed we randomly picked 10 strain genomes from the NCBI bacterial genome database. For genera with less than 10 fully sequenced representatives, we selected all available genomes. The library was made with 1134 genomes from 104 bacterial genera. The open reading frame (ORF) library was then clustered at 99% identity for each genus using CD-HIT [64] to decrease the number of ORFs in the library from 3.5 million to 2.25 million. Annotation was performed by querying the SEED database [90] with the BLAST command line tool [2], and sequenced reads were mapped onto this ORF library. Out of approximately 2 billion reads total, 58.5 million (30.6%) were mapped by this method, over 5836 unique SEED hierarchy annotations. This is comparable to the proportion of reads commonly annotated in transcriptomic analysis, where in one data set, 15.5% of sequenced reads could be matched to a known transcript [15]. The primary limitation of this method is a lack of annotated bacterial sequences. The code for the reference library creation and annotation is on GitHub.

Second, we assembled the reads per sample de novo using Trinity [45], producing 8,847,816 sequences, and removed sequences that matched our reference library with 90% identity as determined by BLAST [2], leaving just under 6 million sequences. Just over three and a half million of these assembled sequences were successfully annotated with the SEED database [90], and sequenced reads were mapped onto this. Approximately 1.32 billion additional reads were annotated by this method, over 7026 unique SEED hierarchy annotations. The code for the

custom assembly pipeline is on GitHub, and described in Appendix A. The data from both prongs was amalgamated into a single table of counts per annotation per sample.

Differential abundance was analyzed using ALDEx2 [32]. A full description of the workflow for this process is included in Appendix A.

3.2.4 Compositional data analysis

The count tables from the 16S rRNA gene sequencing experiment and the metagenomic experiment are compositional data. We use the Centered Log Ratio transform [] to put the compositional data in Euclidean space so that it can be analyzed with conventional statistics, for example, principal co-ordinate analysis (Figures 3.4, 3.9, 3.10, 3.11, and 3.12).

3.3 Results

3.3.1 Data sets

There were originally three groups of patients within these samples: healthy controls, patients with simple steatosis, and patients who had progressed to non alcoholic steatohepatosis. For the metagenomic experiment, 10 of the healthy controls and 10 of the NASH group were selected, based on clinical criteria (Table 3.2).

From this, three comparisons can be made: healthy compared to simple steatosis (SS), healthy compared to non alcoholic steatohepatosis (NASH), and healthy compared to the extreme NASH patients chosen for the metagenomic study. The healthy controls used in the healthy vs. extreme NASH comparison had already been chosen for the metagenomic study. Of the remaining healthy patients, half were used as controls for the healthy vs. SS comparison, and half were used for the healthy vs. NASH comparison. This way, in correlating the results of the three different comparisons, there will not be spurious correlations from using the same control samples. The three data sets will be denoted in figures as 'Healthy vs. SS', 'Healthy vs. NASH', and 'Healthy vs. extreme NASH', respectively.

3.3.2 16S rRNA gene tag experiment

The four most abundant genera detected by 16S rRNA gene sequencing (excluding unclassified bacteria) were: *Bacteroides*, *Faecalibacterium*, *Blautia*, and *Pseudobutyrivibrio* (Fig. 3.2).

No obvious structure or separation is evident from the principal co-ordinates analysis in Fig. 3.3 or the principal component analysis in Fig. 3.4. Furthermore the variance explained by each principal component or co-ordinate axis is not notably high, indicating a rather uniform data set. Additionally, no OTUs are significantly differentially abundant between groups (Fig. 3.5)

The Toronto patient population is very diverse, including patients coming from different cultural backgrounds who consume different diets. With this kind of population and diversity, it can be difficult to have sufficient power to detect significant differences. We explore effect sizes, since these are known to be more robust and predict the p-values. Effect size is a standardized difference between means

When comparing all the healthy samples with all the NASH samples, the genera with the highest effect sizes are *Adlercreutzia*, *Odoribacter*, and *Escherichia-Shigella*. However, when only the select 10 healthy samples and the 10 extreme NASH samples used in the metagenomic study are compared, the genera with the highest effect sizes are *Ruminococcus*, *Adlercreutzia*, and *Alistipes*. This corresponds with the qPCR experiment, where *Bacteriodetes*, *Prevotella*, and *Ruminococcus* were tested, and only *Ruminococcus* was found to be differentially abundant.

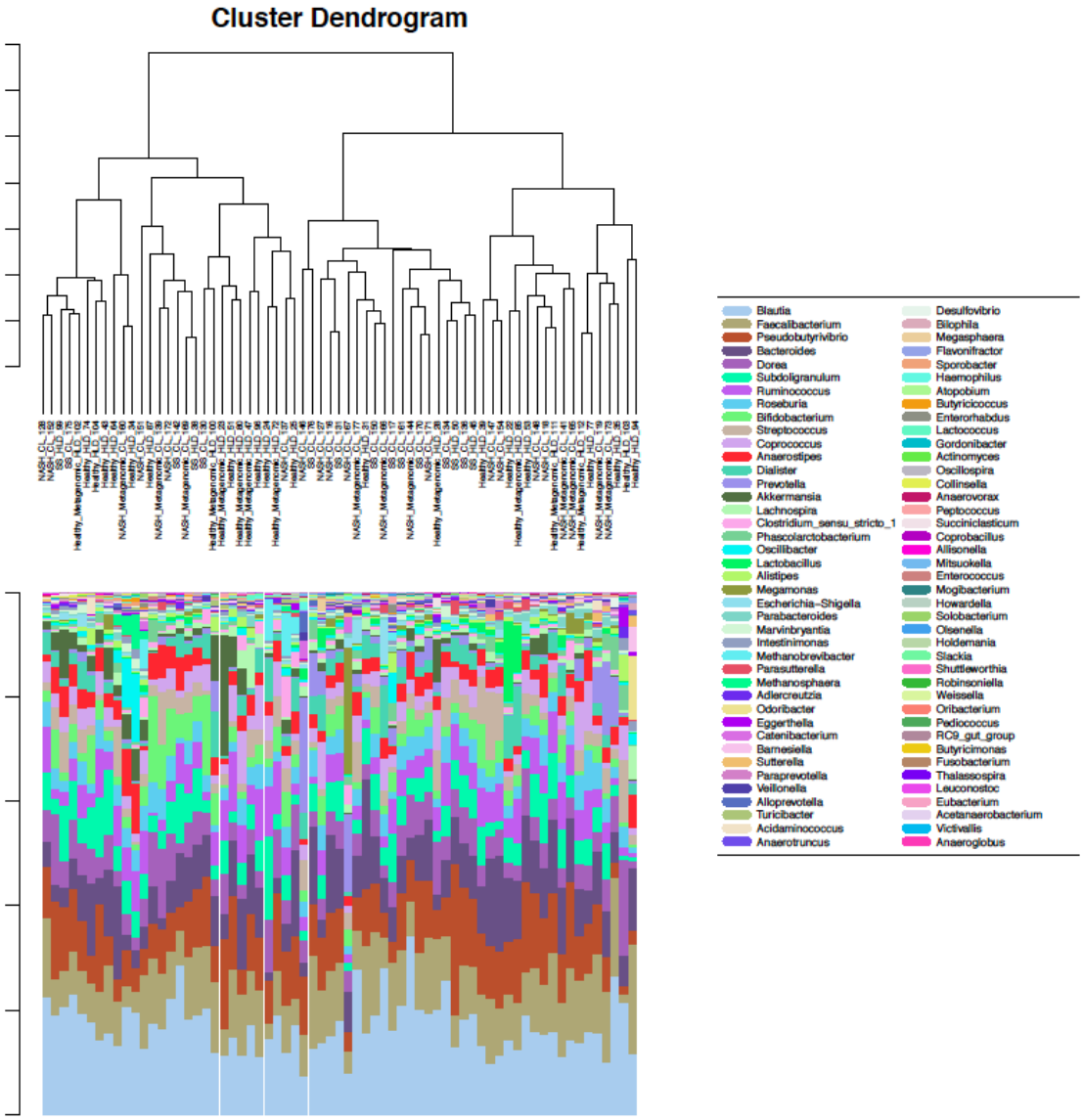


Figure 3.2: **Bar plot of 16S rRNA gene tag sequencing experiment.** Each column of this bar plot represents one sample, and each color represents one bacterial genus. Genus names are listed in the legend in order of decreasing total abundance across all samples. Samples do not cluster according to their condition (healthy, simple steatosis, or nonalcoholic steatohepatitis). Note that OTUs that mapped to unclassified or *Incertae Sedis* were removed, and these made up just over a third of the total abundance. This operation is valid in compositional data analysis, but not in non-compositional approaches such as UniFrac.

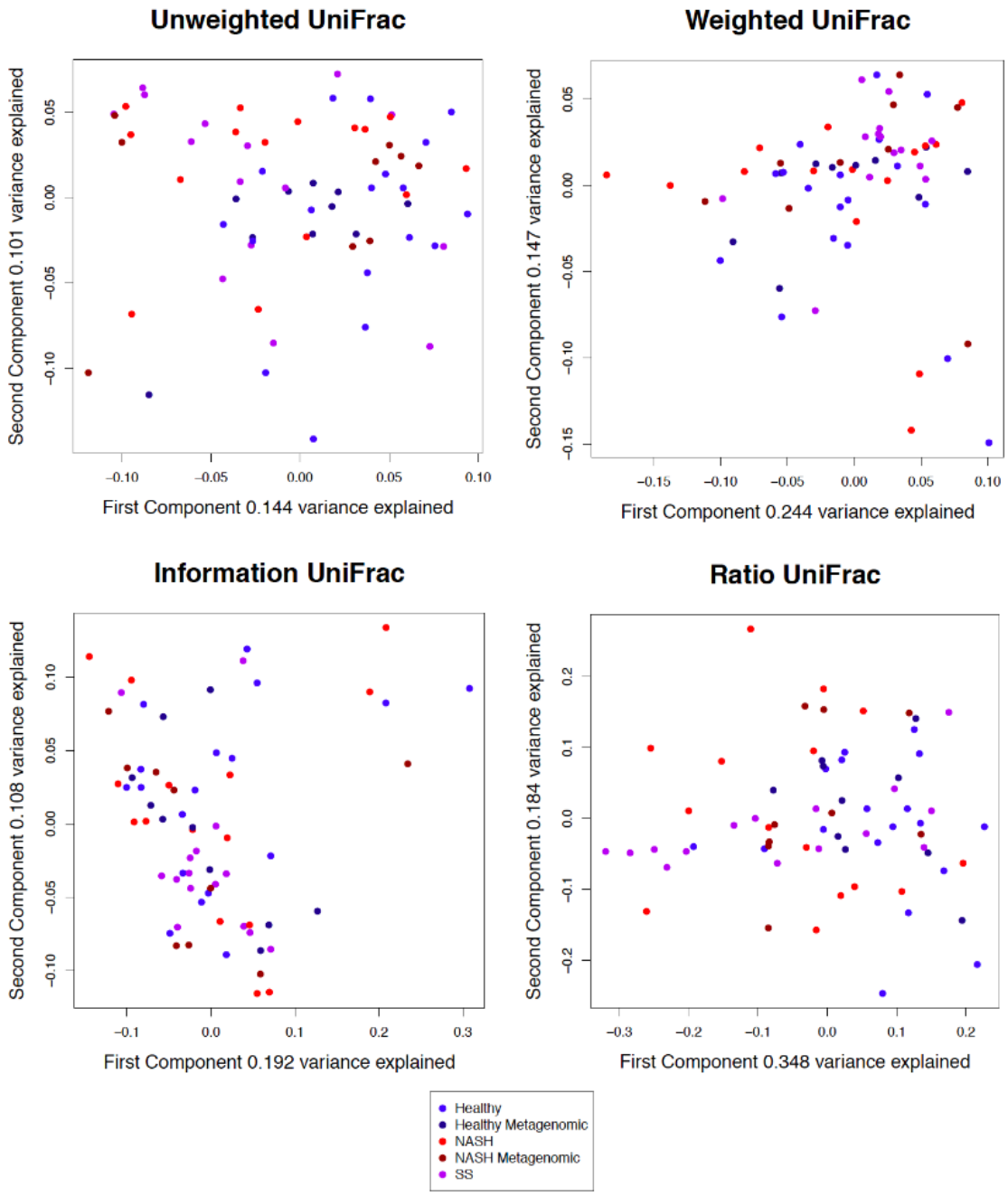


Figure 3.3: **Principal co-ordinate analysis of 16S rRNA gene tag sequencing data with different UniFrac methods.** Each point represents one sample, and the distances between the samples have been calculated using different UniFrac metrics, taking into account phylogenetic as well as abundance information. There is no obvious separation between groups by any of the UniFrac weightings. Furthermore the variance explained by each principal co-ordinate axis is not notably high, indicating a rather uniform data set.

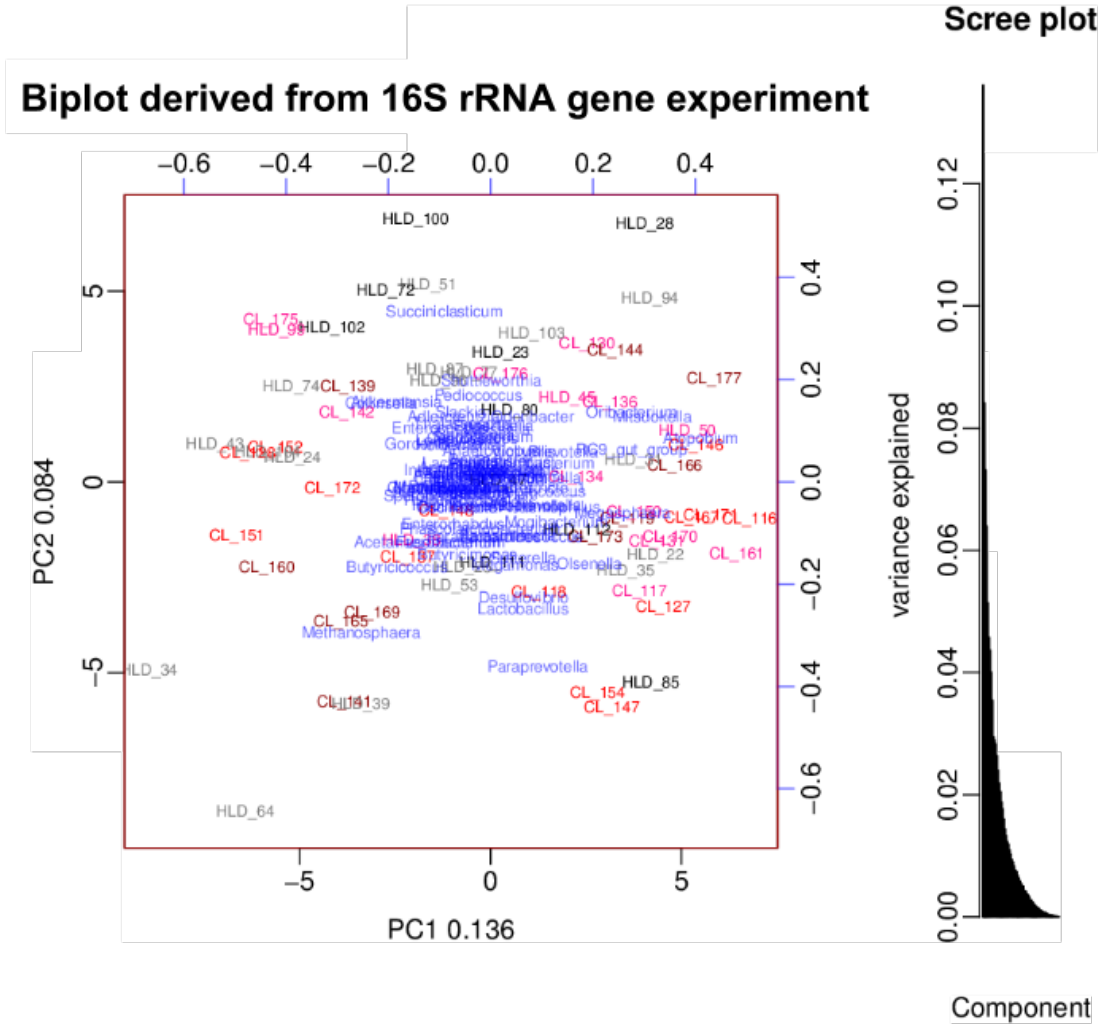


Figure 3.4: **16S rRNA gene tag sequencing experiment biplot.** Compositional data analysis is done by transforming the counts with the centered log ratio transform, and then performing a principal component analysis. The variance explained by each genus is overlaid on the same principal component analysis plot. Note that the variance explained by the first and the second component is 13.6% and 8.4% respectively, indicating that there is not a clear unidirectional separation between groups. Samples from healthy controls selected for the metagenomic study are colored black while samples from patients with extreme NASH are colored dark red. The remaining healthy controls are colored gray, and the remaining NASH samples are colored bright red. Samples from patients with simple steatosis (SS) are colored pink.

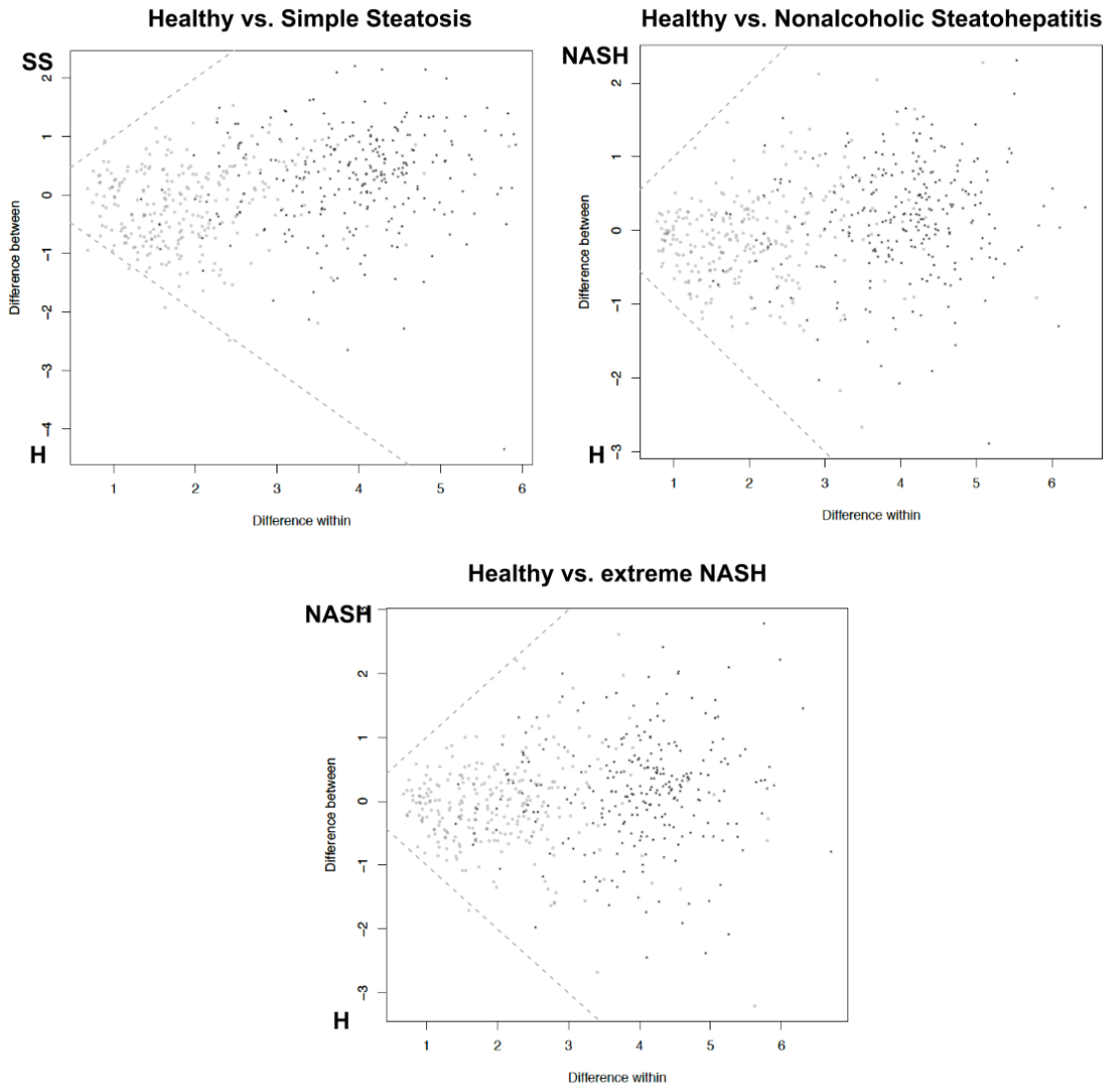


Figure 3.5: **Effect plot showing difference within vs. difference between groups.** Each point represents one OTU, and the dispersion of that OTU within groups is plotted against the difference in abundance between groups. None of the OTUs are more different between groups than within groups. The healthy 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. The dashed lines represent where the difference in abundance between groups and within groups are equal.

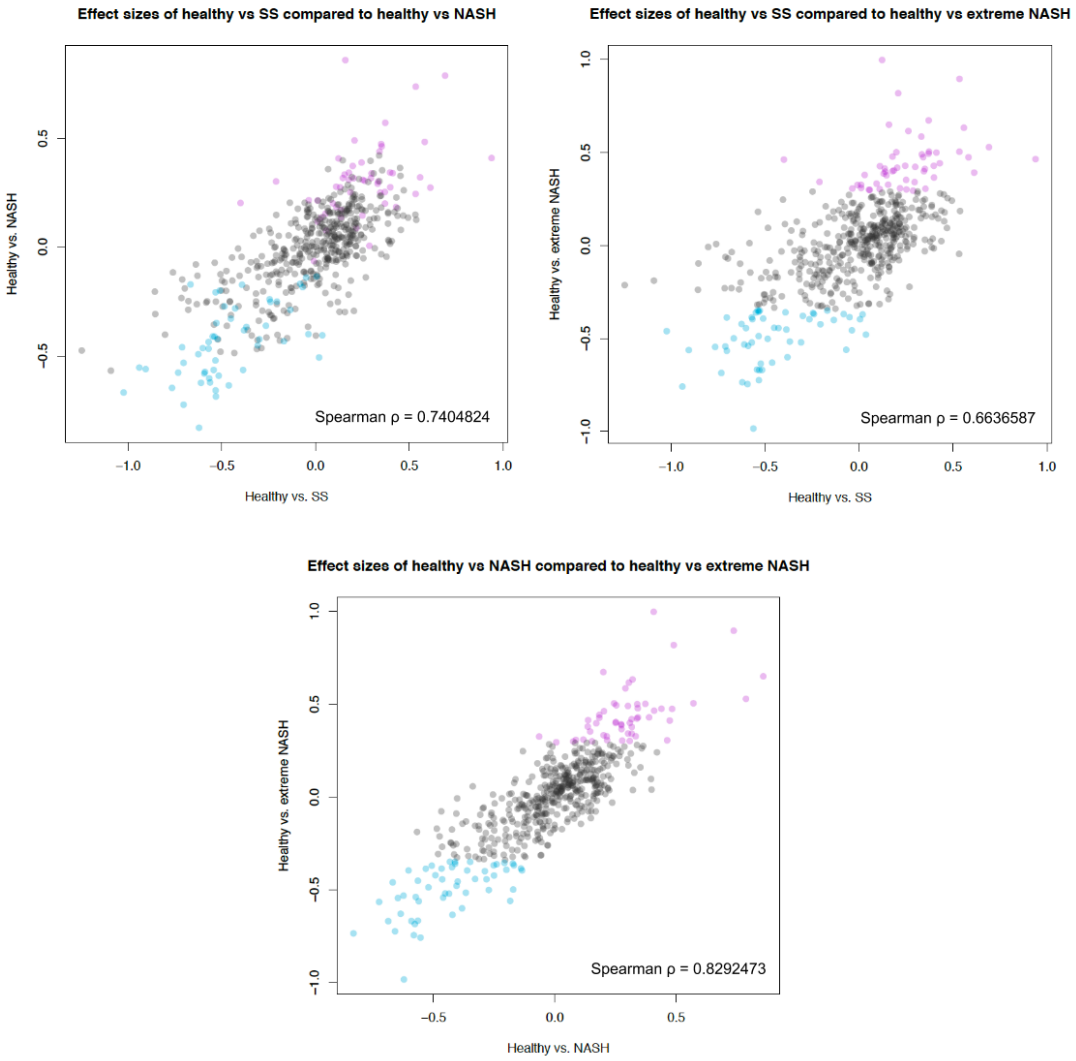


Figure 3.6: **Correlation in effect sizes of different group experiments.** Each point represents one OTU, and the effect size of that OTU in one comparison (for example, comparing the gut microbiome of healthy patients with patients who have simple steatosis) is plotted against the effect size of that 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 metegenomic study. The top decile of OTUs relatively increased in NASH for the metagenomic experiment are colored pink, and the top decile of OTUs relatively increased healthy for the metagenomic experiment are colored blue. The median difference in the effect sizes of the 'Healthy vs. NASH' - 'Healthy vs. SS' is 0.12370 for the pink points, and 0.02626 for the blue points. The median difference in the effect sizes of the 'Healthy vs. extreme NASH' - 'Healthy vs. SS' is 0.4014 for the pink points, and -0.3513 for the blue points. The median difference in the effect sizes of the 'Healthy vs. extreme NASH' - 'Healthy vs. NASH' is 0.3799 for the pink points, and -0.3742 for the blue points.

Here we have looked at three different comparisons: healthy compared to simple steatosis (SS), healthy compared to non alcoholic steatohepatosis (NASH), and healthy compared to extreme NASH. The healthy controls have been selected so that different healthy controls are used in each comparison. The ALDEx analysis in Figure 3.5 showed that there are no OTUs significantly differentially abundant between conditions.

Effect size is a standardized difference between two groups, calculated by dividing difference by dispersion. If there were truly no effect, then there would be no correlation between the effect sizes for each OTU in the different comparisons. However, in Figure 3.6, we show that the effect sizes for each OTU in each comparison are correlated. The effect sizes are higher in the Healthy vs. extreme NASH compared to the Healthy vs. SS or Healthy vs. NASH comparison at the extreme deciles.

OTU family	OTU genus	SILVA bootstrap value	H Vs. SS effect sizes	H Vs. NASH effect sizes	H vs. extreme NASH effect sizes
Acidaminococcaceae	Phascolarctobacterium	100	0.122	0.407	0.998
Lactobacillaceae	Lactobacillus	97	0.534	0.736	0.896
Prevotellaceae	Paraprevotella	100	0.208	0.489	0.819
Lachnospiraceae	Incertae Sedis	98	0.37	0.2	0.673
Lachnospiraceae	Marvinbryantia	77	0.159	0.858	0.65
Lachnospiraceae	Incertae Sedis	73	0.557	0.32	0.634
Bifidobacteriaceae	Bifidobacterium	100	0.262	0.304	0.616
Ruminococcaceae	Incertae Sedis	72	0.331	0.291	0.586
Prevotellaceae	Paraprevotella	100	0.691	0.787	0.529
Lachnospiraceae	unclassified	100	0.371	0.571	0.505
unclassified	unclassified	72	0.533	0.244	0.505
Ruminococcaceae	Butyricicoccus	71	0.198	0.372	0.502
Lachnospiraceae	Incertae Sedis	91	0.411	0.339	0.5
Ruminococcaceae	Ruminococcus	93	0.369	0.253	0.494
Coriobacteriaceae	unclassified	97	0.334	0.301	0.491
Lachnospiraceae	unclassified	98	0.178	0.341	0.478
Lactobacillaceae	Lactobacillus	98	0.341	0.439	0.476
Ruminococcaceae	Subdoligranulum	87	0.582	0.483	0.475
Ruminococcaceae	Incertae Sedis	98	0.939	0.409	0.465
Ruminococcaceae	unclassified	100	−0.4	0.202	0.462
Coriobacteriaceae	Olsenella	91	0.429	0.183	0.443
Ruminococcaceae	Subdoligranulum	98	0.245	0.388	0.429
Lachnospiraceae	unclassified	100	0.397	0.342	0.429
Prevotellaceae	Prevotella	99	0.111	0.183	0.427
Lachnospiraceae	Anaerostipes	100	0.298	0.338	0.423
Prevotellaceae	unclassified	70	0.203	0.316	0.419
unclassified	unclassified	92	0.136	0.137	0.414
Ruminococcaceae	Incertae Sedis	99	0.35	0.473	0.412
Ruminococcaceae	Faecalibacterium	100	0.185	0.251	0.404
Alcaligenaceae	Sutterella	100	0.177	0.309	0.4
unclassified	unclassified	73	0.345	0.25	0.4
Rikenellaceae	Alistipes	100	0.139	0.17	0.397
Lachnospiraceae	Roseburia	98	0.612	0.273	0.392
Prevotellaceae	unclassified	75	0.131	0.274	0.387
Coriobacteriaceae	Enterorhabdus	72	0.029	0.135	0.38
Ruminococcaceae	Ruminococcus	100	0.146	0.317	0.378
unclassified	unclassified	98	0.398	0.275	0.366
Veillonellaceae	Dialister	100	0.25	0.145	0.353
Lachnospiraceae	unclassified	100	−0.211	0.301	0.342
Family XIII	Incertae Sedis	100	0.295	0.317	0.341
Bacteroidaceae	Bacteroides	100	0.093	0.201	0.333
Lachnospiraceae	unclassified	100	0.155	0.333	0.327
Lachnospiraceae	unclassified	100	0.01	0.214	0.327
Desulfovibrionaceae	Desulfovibrio	100	−0.009	−0.065	0.326
Lachnospiraceae	unclassified	100	0.011	0.117	0.309
Lachnospiraceae	Blautia	96	0.218	0.087	0.308
Lachnospiraceae	Blautia	97	−0.036	0.215	0.306
Ruminococcaceae	unclassified	100	0.353	0.462	0.305
Christensenellaceae	unclassified	99	0.11	0.277	0.303
Alcaligenaceae	Parasutterella	100	0.252	0.308	0.302
Lachnospiraceae	Incertae Sedis	100	0.052	0.153	0.3
Ruminococcaceae	Subdoligranulum	92	0.056	0.076	0.299
Acidaminococcaceae	Acidaminococcus	100	0.287	0.006	0.295

Table 3.3: **Top decile of OTUs relatively increased in NASH based on effect size from healthy vs. NASH comparison.** This table lists the OTUs and their effect sizes in all the comparisons. The OTUs were picked by open reference, by clustering and comparison with the SILVA database [101]. Positive effect sizes indicate that the feature was found to be relatively increased in NASH while negative effect sizes indicate that the feature was found to be relatively increased in healthy. OTUs were annotated with SILVA, and a confidence percentage is reported based on the provided bootstrapping algorithm.

OTU family	OTU genus	SILVA bootstrap value	H Vs. SS effect sizes	H Vs. NASH effect sizes	H vs. extreme NASH effect sizes
Ruminococcaceae	Incertae Sedis	100	−0.539	−0.411	−0.349
Verrucomicrobiaceae	Akkermansia	100	−0.169	−0.433	−0.35
Porphyromonadaceae	Parabacteroides	100	−0.529	−0.349	−0.35
Rikenellaceae	Alistipes	100	−0.534	−0.208	−0.356
Lachnospiraceae	Incertae Sedis	73	−0.392	−0.173	−0.36
Lachnospiraceae	unclassified	100	−0.549	−0.411	−0.361
Streptococcaceae	Streptococcus	100	−0.245	−0.24	−0.363
Lachnospiraceae	unclassified	100	−0.082	−0.169	−0.369
Lachnospiraceae	Dorea	100	−0.241	−0.252	−0.369
Lachnospiraceae	Roseburia	83	0.019	−0.507	−0.371
Lachnospiraceae	Incertae Sedis	82	−0.3	−0.424	−0.379
Christensenellaceae	unclassified	98	−0.051	−0.14	−0.386
Christensenellaceae	unclassified	100	−0.571	−0.467	−0.387
Lachnospiraceae	Blautia	93	−0.704	−0.532	−0.387
Ruminococcaceae	unclassified	100	−0.511	−0.2	−0.393
Lachnospiraceae	Roseburia	91	−0.568	−0.603	−0.397
Porphyromonadaceae	Odoribacter	100	0.005	−0.135	−0.397
Lachnospiraceae	Incertae Sedis	85	−0.265	−0.362	−0.397
Lachnospiraceae	unclassified	98	−0.135	−0.289	−0.401
Porphyromonadaceae	Odoribacter	100	−0.625	−0.492	−0.422
Erysipelotrichaceae	Turicibacter	100	−0.206	−0.251	−0.424
Christensenellaceae	unclassified	100	−0.452	−0.329	−0.443
Ruminococcaceae	Ruminococcus	100	−0.429	−0.282	−0.444
Christensenellaceae	unclassified	99	−0.602	−0.464	−0.445
Lachnospiraceae	unclassified	100	−0.387	−0.564	−0.452
Bacteroidaceae	Bacteroides	100	−0.038	−0.4	−0.456
Ruminococcaceae	Incertae Sedis	84	−1.024	−0.668	−0.461
Veillonellaceae	Dialister	100	0.036	−0.405	−0.479
Bacteroidaceae	Bacteroides	100	−0.534	−0.521	−0.488
Prevotellaceae	Alloprevotella	100	−0.667	−0.172	−0.5
Bacteroidaceae	Bacteroides	100	−0.487	−0.272	−0.502
Rikenellaceae	Alistipes	100	−0.369	−0.367	−0.517
Coriobacteriaceae	Adlercreutzia	100	−0.309	−0.452	−0.521
Ruminococcaceae	unclassified	100	−0.571	−0.436	−0.522
Family XIII	Anaerovorax	91	−0.611	−0.624	−0.533
Ruminococcaceae	unclassified	76	−0.59	−0.573	−0.54
Lachnospiraceae	Pseudobutyrivibrio	98	−0.712	−0.46	−0.543
Bacteroidaceae	Bacteroides	100	−0.766	−0.647	−0.546
Bacteroidaceae	Bacteroides	100	−0.069	−0.184	−0.561
Lachnospiraceae	Blautia	85	−0.906	−0.561	−0.562
Ruminococcaceae	Faecalibacterium	100	−0.704	−0.724	−0.566
unclassified	unclassified	85	−0.381	−0.382	−0.601
Ruminococcaceae	Incertae Sedis	100	−0.463	−0.635	−0.63
Ruminococcaceae	Subdoligranulum	99	−0.523	−0.422	−0.636
Ruminococcaceae	Incertae Sedis	97	−0.544	−0.565	−0.668
Ruminococcaceae	Ruminococcus	100	−0.516	−0.591	−0.669
Lachnospiraceae	Incertae Sedis	85	−0.532	−0.686	−0.67
Lachnospiraceae	Coprococcus	91	−0.733	−0.577	−0.686
Ruminococcaceae	unclassified	74	−0.533	−0.658	−0.725
Erysipelotrichaceae	unclassified	100	−0.621	−0.83	−0.736
Ruminococcaceae	Subdoligranulum	85	−0.594	−0.581	−0.746
Lachnospiraceae	Dorea	100	−0.94	−0.554	−0.759
Family XIII	Incertae Sedis	91	−0.563	−0.622	−0.985

Table 3.4: **Top decile of OTUs relatively increased in healthy based on effect size from healthy vs. NASH comparison.** This table lists the OTUs and their effect sizes in all the comparisons. The OTUs were picked by open reference, by clustering and comparison with the SILVA database [101]. Positive effect sizes indicate that the feature was found to be relatively increased in NASH while negative effect sizes indicate that the feature was found to be relatively increased in healthy. OTUs were annotated with SILVA, and a confidence percentage is reported based on the provided bootstrapping algorithm.

3.3.3 Metagenomic experiment

Functional analysis

The reads were annotated into SEED subsystems [90]. The most descriptive and specific SEED subsystem is subsystem 4. Each read is categorized into at most one subsystem 4. Subsystems 1, 2, and 3 are broader, and a read can be categorized into multiple higher categories. We use ALDEx to test for differential expression at the subsystem 4 categorization level (Fig. 3.7). Because the reads do not have unique categorizations at the higher levels, we use stripcharts to examine these (Fig. 3.8).

ALDEx analysis results show that no functional annotations are differentially abundant between the healthy and the NASH conditions (Fig. 3.7). In the principal components analysis, the healthy samples appear to be clustered in the middle, with sample HLD_23 as an outlier. The extreme NASH samples are more spread out, and the functional annotations that contribute the most to their spread are written in blue in Figure. 3.9. We were particularly interested in the subset of functional annotations related to carbohydrates and lipids, as they are relevant to obesity and potentially NASH. When the carbohydrate or lipid subset is analyzed separately (Figures 3.10 and 3.11), the samples are still arranged in the same way as the full data set in the principal components analysis. What differentiates the samples in the full set is not lost in the carbohydrate or the lipid subset, even though the lipid subset is much smaller and does not include functional annotations with an effect size greater than 0.68.

The Boursier et al. [8] paper imputed the metagenomic analysis, and reported that KEGG pathways for carbohydrate, lipid, and amino acid metabolism were significantly differential between groups. We did a principal components analysis on the Amino Acids and Derivatives subsystem 1 category (Figure 3.12). This produced a spread of samples less similar to the principal components analysis with all the functional annotations (Fig. 3.9), compared to the lipid subset (Fig. 3.11), despite the amino acid subset containing more annotations than the lipid subset (449 compared to 134 annotations).

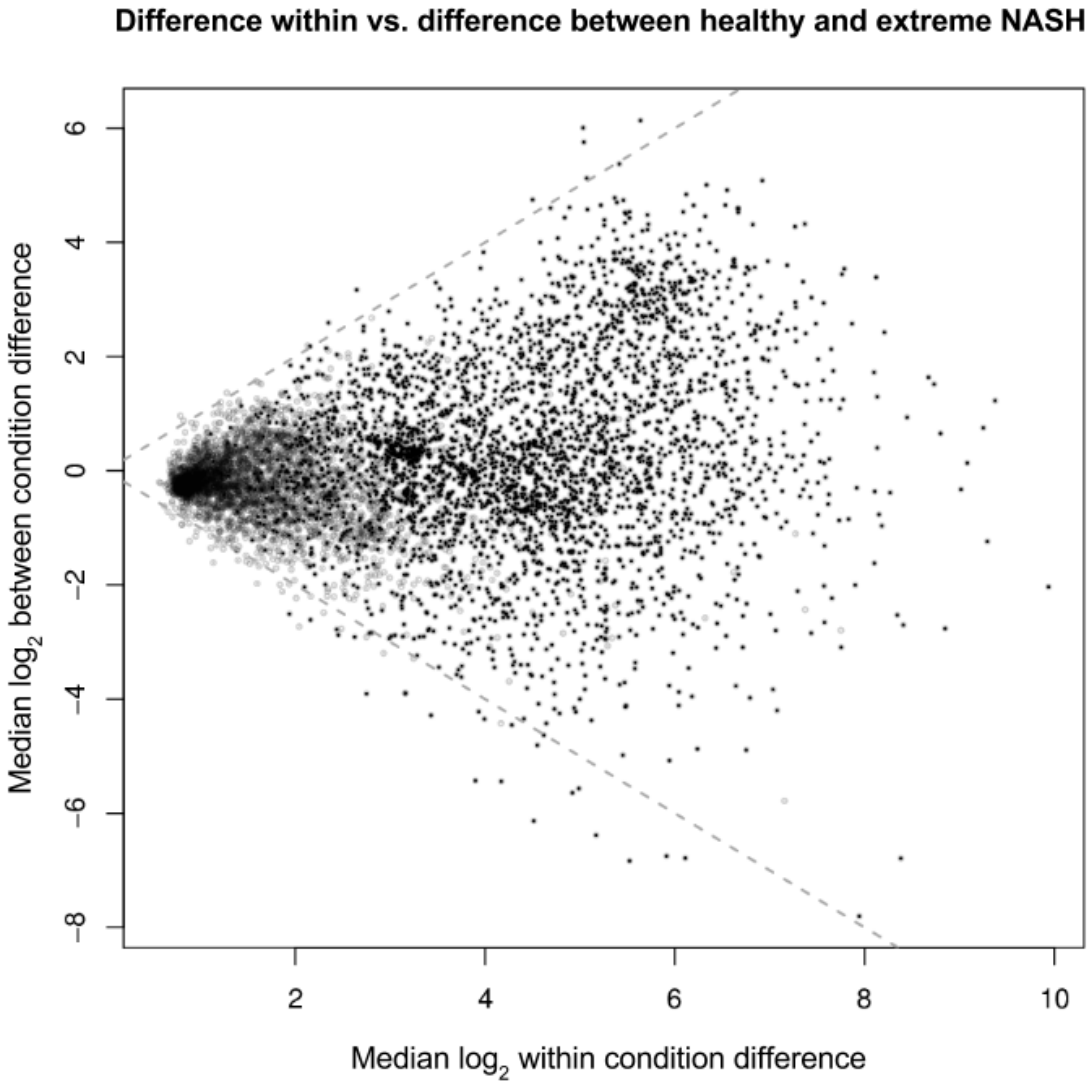


Figure 3.7: **Difference within vs. difference between healthy and extreme NASH for metagenomic data.** Each point on this plot represents a functional annotation at the SEED subsystem 4 level. None of the subsystem 4 functional annotations have a significantly greater abundance between groups than within groups.

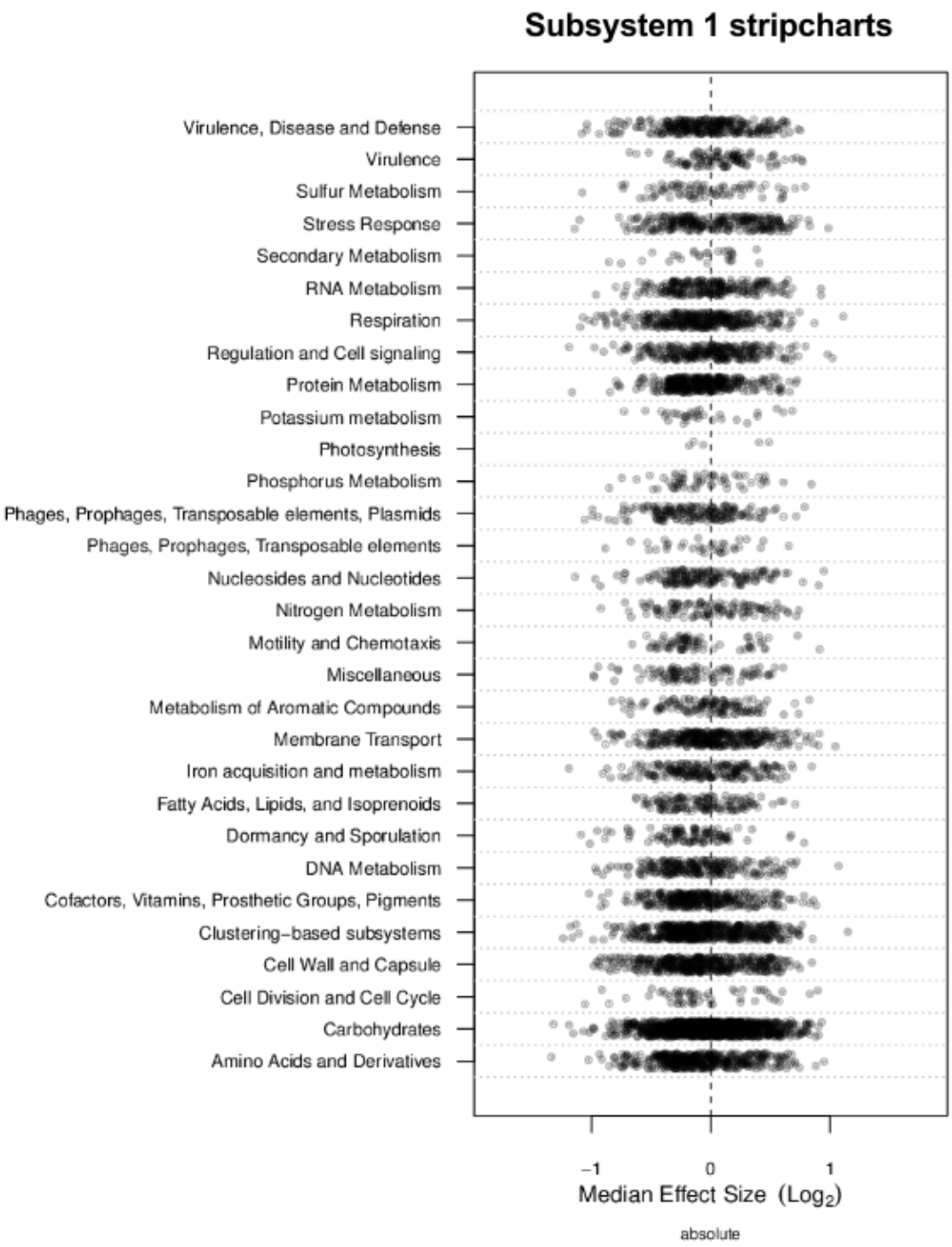


Figure 3.8: **Difference within vs. difference between healthy and extreme NASH for metagenomic data.** Each point on this plot represents a functional annotation at the SEED subsystem 4 level. The different rows of the plots show the subsystem 4 annotations within each subsystem 1 group. Points on the positive side were more abundant in the extreme NASH condition while points on the negative side were more abundant in the healthy condition. None of the points were significantly differentially abundant between groups.

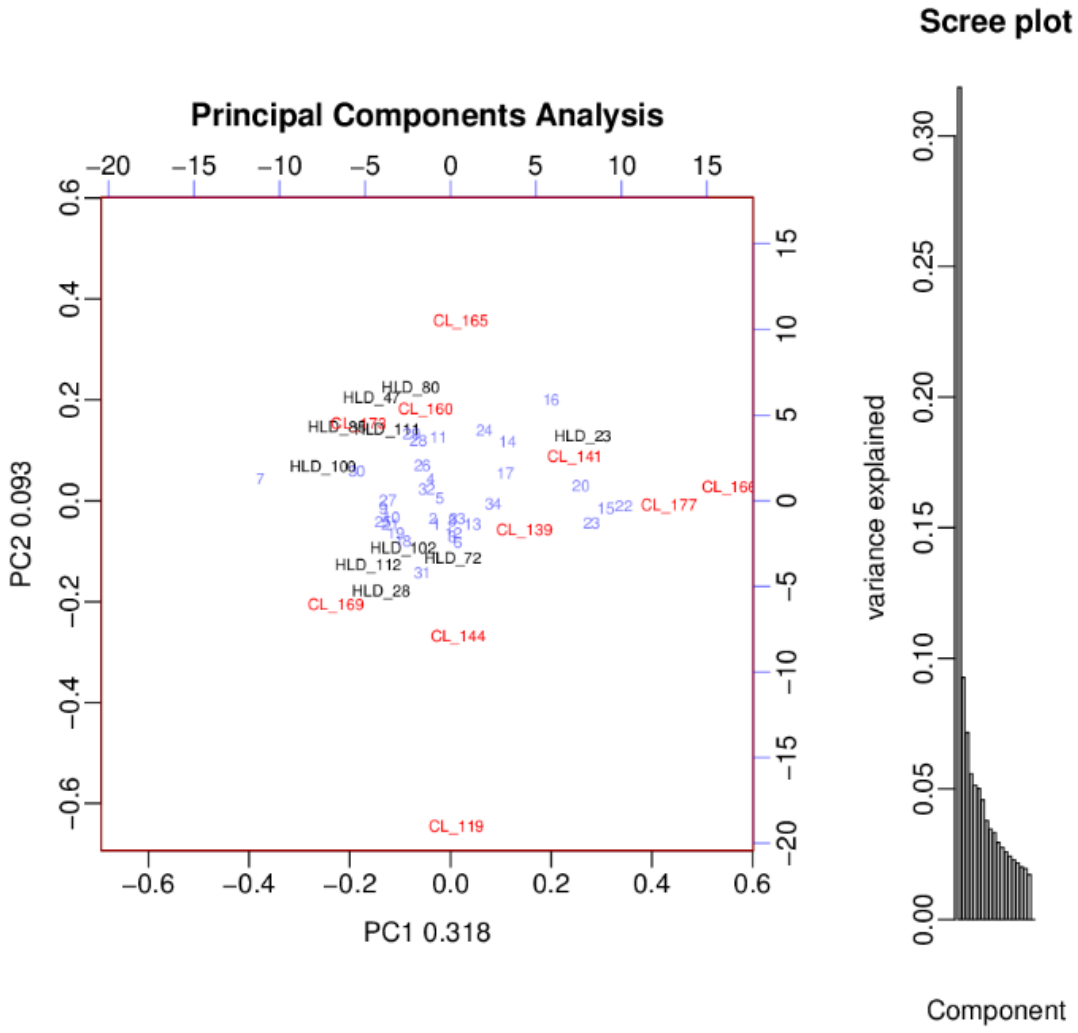


Figure 3.9: **Principal components analysis of metagenomic data.** In this biplot, the healthy samples are shown in black, while the extreme NASH samples are shown in red. Samples CL_119, CL_139, CL_160, and CL_165 have a steatosis grading of 3. Samples CL_141, CL_144, CL_169, CL_173, and Cl_177 have a steatosis grading of 2. Sample CL_166 has a steatosis grading of 1. The steatosis grading doesn't appear to separate the samples. The location of the subsystem 4 labels (in blue) show which direction they pull the samples in, in terms of variance. The biplot shows only the thirty four subsystem 4 categories with an effect size greater than 1, out of 7026 categories in total. The key for the subsystem 4 labels is Table 3.5.

Function ID	Function name
1	DNA topoisomerase III (EC 5.99.1.2)
2	3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100)
3	2-hydroxy-3-oxopropionate reductase (EC 1.1.1.60)
4	Imidazolonepropionase (EC 3.5.2.7)
5	Flagellar basal-body rod modification protein FlgD
6	HflC protein
7	Diaminobutyrate-pyruvate aminotransferase (EC 2.6.1.46)
8	Metal-dependent hydrolases of the beta-lactamase superfamily II
9	pyoverdine-specific efflux macA-like protein
10	Homoserine O-acetyltransferase (EC 2.3.1.31)
11	Threonine kinase in B12 biosynthesis
12	Molybdopterin biosynthesis protein MoeB
13	D-allose kinase (EC 2.7.1.55)
14	Gluconate permease, Bsu4004 homolog
15	2-methylisocitrate dehydratase (EC 4.2.1.99)
16	Uncharacterized protein ImpH/VasB
17	Phosphate starvation-inducible ATPase PhoH with RNA binding motif
18	Spore germination protein GerLC
19	Phosphohistidine phosphatase SixA
20	Glutathione-regulated potassium-efflux system ancillary protein KefG
21	Histidyl-tRNA synthetase, archaeal-type paralog (EC 6.1.1.21)
22	3-hydroxydecanoyl-[acyl-carrier-protein] dehydratase (EC 4.2.1.60)
23	FKBP-type peptidyl-prolyl cis-trans isomerase slpA (EC 5.2.1.8)
24	3-oxoacyl-[ACP] synthase
25	IncW plasmid conjugative relaxase protein TrwC (TraI homolog)
26	CO dehydrogenase iron-sulfur protein CooF (EC 1.2.99.2)
27	Transcriptional regulator of AraC family, enterobactin-dependent, predicted
28	Antitoxin YgiT
29	acyl-acyl carrier protein synthetase (EC 6.2.1.20)
30	Outer membrane protein GumB, involved in the export of xanthan
31	C3 family ADP-ribosyltransferase (EC 2.4.2.-)
32	Hypothetical SAV0786 homolog in superantigen-encoding pathogenicity islands SaPI
33	Predicted cellobiose ABC transport system, sugar-binding protein
34	Lipoprotein Bor

Table 3.5: **Subsystem 4 label key for Figure 3.9.** This table lists the subsystem 4 labels and corresponding names.

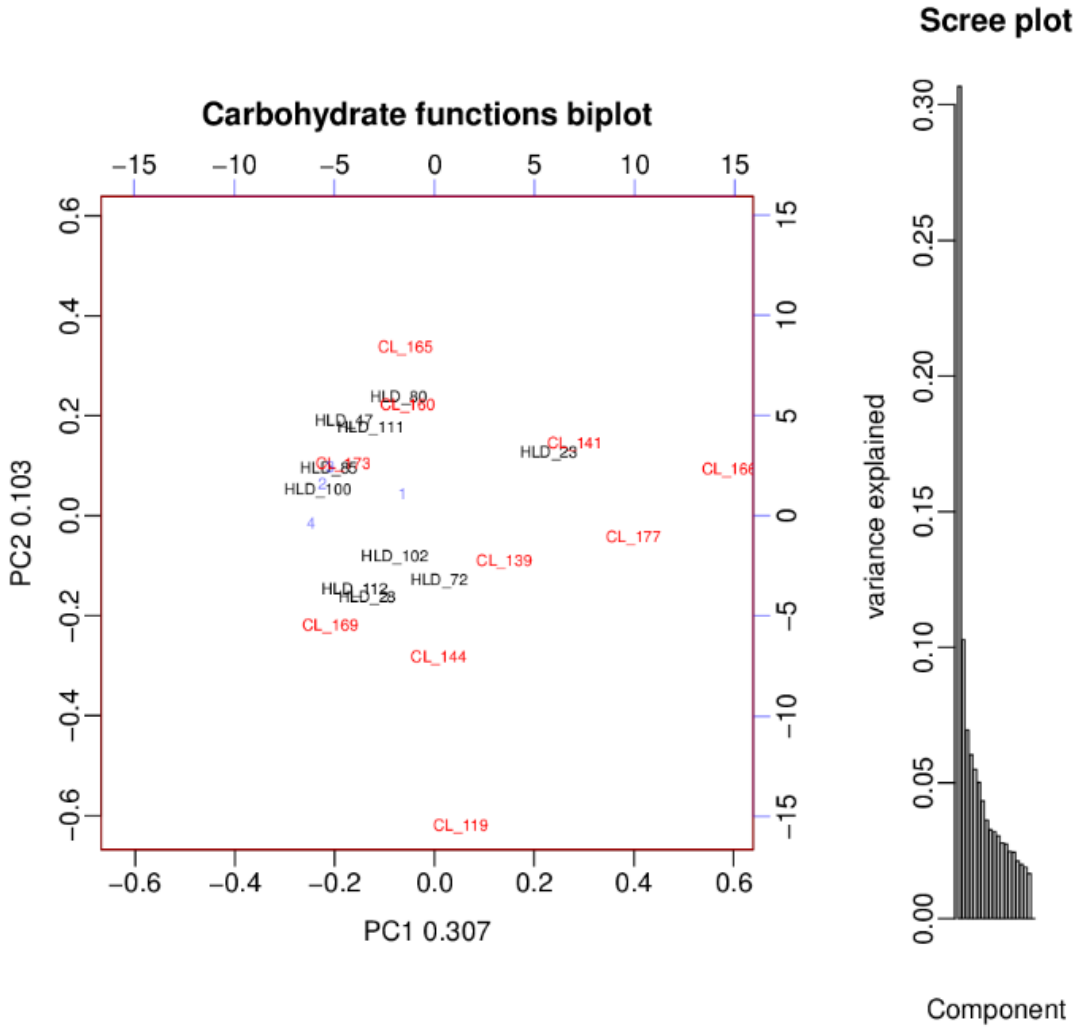


Figure 3.10: **Principal components analysis of carbohydrate metagenomic data.** This biplot shows the sample separation produced when only the 1164 subsystem 4 categories under the 'Carbohydrates' subsystem 1 categorization are used. In this biplot, the healthy samples are shown in black, while the extreme NASH samples are shown in red. The location of the subsystem 4 labels (in blue) show which direction they pull the samples in, in terms of variance. The biplot shows only the thirty four subsystem 4 categories with an effect size greater than 1. Below is the key for the subsystem 4 labels.

Function ID	Function name
1	Maltodextrin phosphorylase (EC 2.4.1.1)
2	Dihydrolipoamide acyltransferase component of branched-chain alpha-keto acid dehydrogenase complex (EC 2.3.1.168)
3	Predicted L-arabinose isomerase (EC 5.3.1.4)
4	Predicted galactoside ABC transporter, permease protein 2

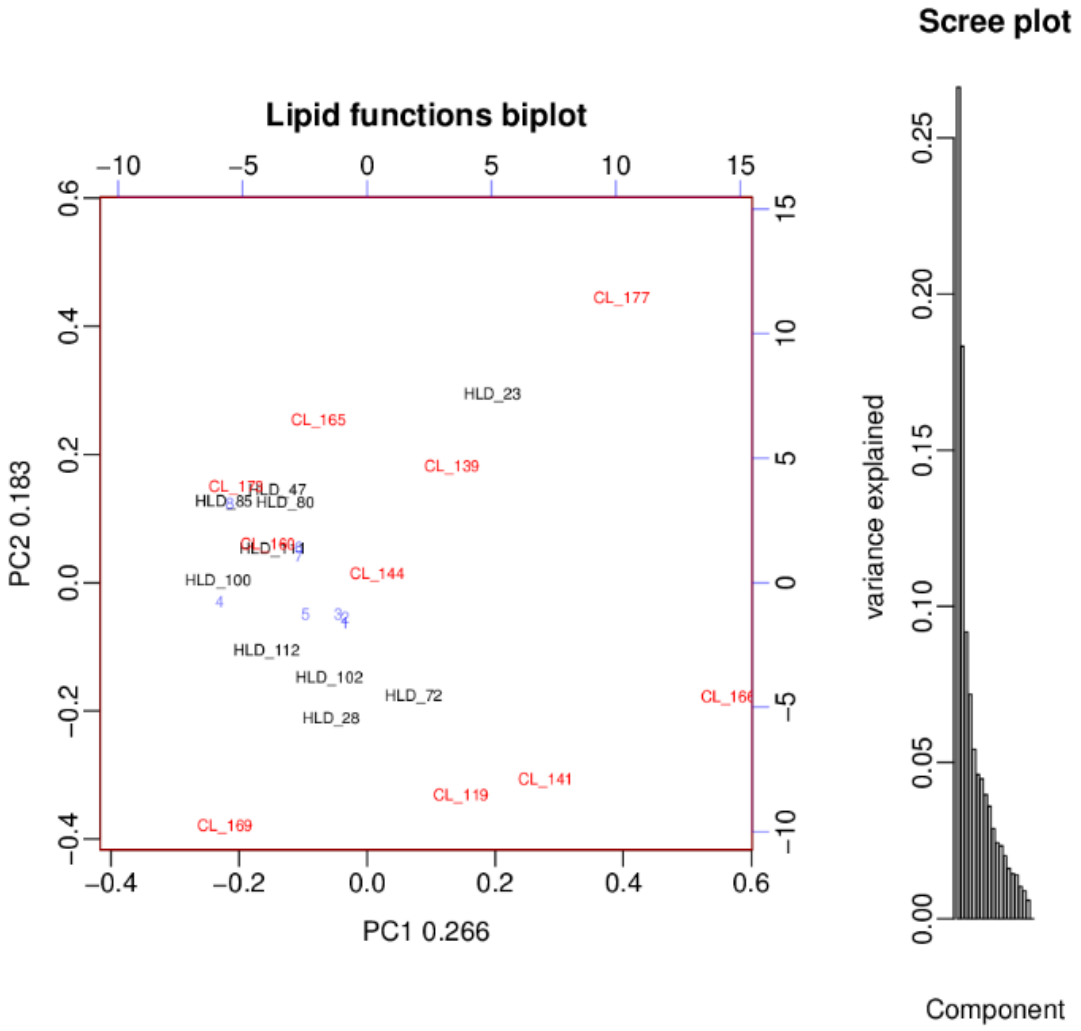


Figure 3.11: **Principal components analysis of lipid metagenomic data.** This biplot shows the sample separation produced when only the 134 subsystem 4 categories under the 'Fatty Acids, Lipids, and Isoprenoids' subsystem 1 categorization are used. In this biplot, the healthy samples are shown in black, while the extreme NASH samples are shown in red. The location of the subsystem 4 labels (in blue) show which direction they pull the samples in, in terms of variance. The biplot shows only the thirty four subsystem 4 categories with an effect size greater than 0.5. Below is the key for the subsystem 4 labels.

Function ID	Function name
1	3-oxoacyl-[acyl-carrier-protein] synthase, KASII (EC 2.3.1.41)
2	3-oxoacyl-[acyl-carrier-protein] synthase, KASIII (EC 2.3.1.41)
3	Acetyl-coenzyme A carboxyl transferase alpha chain (EC 6.4.1.2)
4	3-ketoacyl-CoA thiolase (EC 2.3.1.16)
5	Squalene–hopene cyclase (EC 5.4.99.17)
6	Hydroxyneurosporene methyltransferase (EC 2.1.1.-)
7	FIG143263: Glycosyl transferase
8	Phytoene dehydrogenase and related proteins

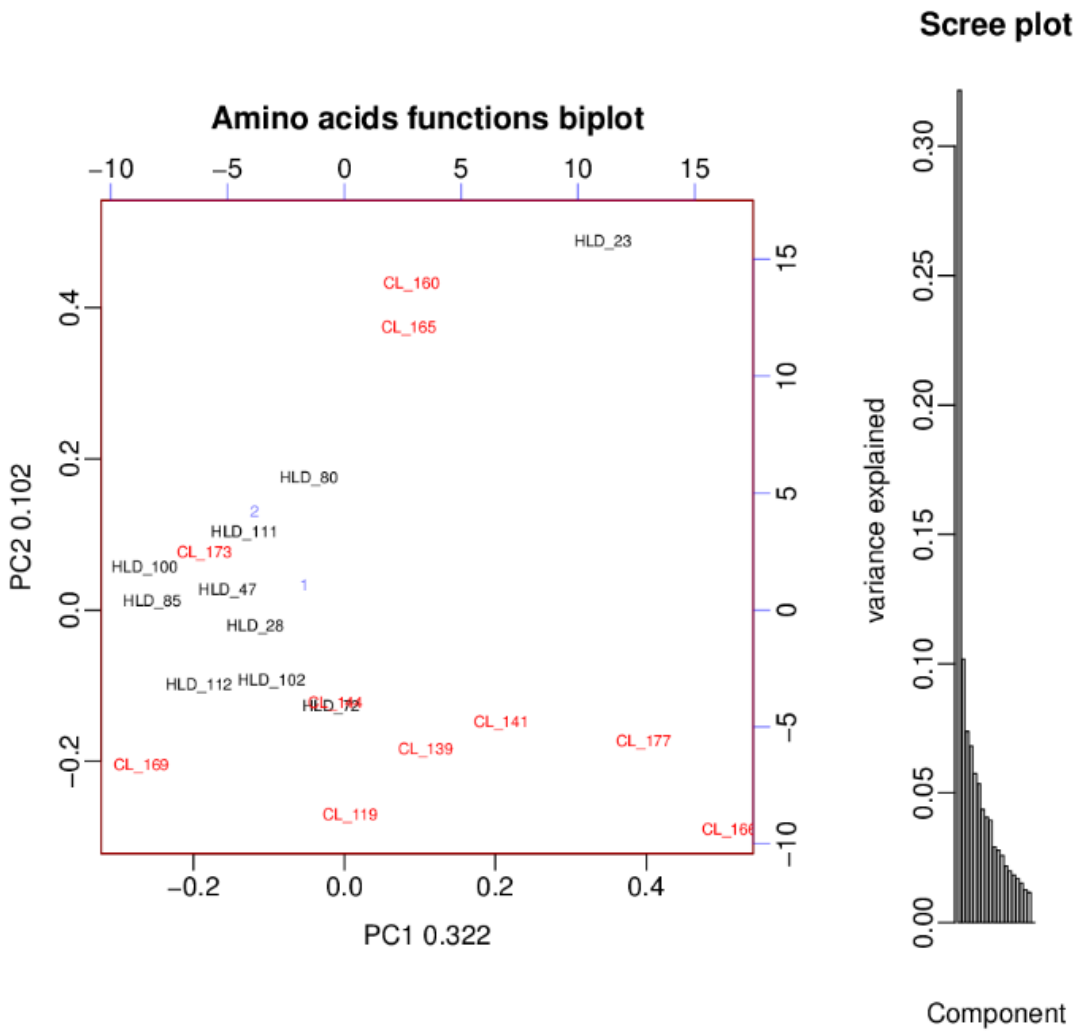


Figure 3.12: **Principal components analysis of lipid metagenomic data.** This biplot shows the sample separation produced when only the 449 subsystem 4 categories under the 'Amino Acids and Derivatives' subsystem 1 categorization are used. In this biplot, the healthy samples are shown in black, while the extreme NASH samples are shown in red. The location of the subsystem 4 labels (in blue) show which direction they pull the samples in, in terms of variance. The biplot shows only the thirty four subsystem 4 categories with an effect size greater than 1. Below is the key for the subsystem 4 labels.

Function ID	Function name
1	Phosphoadenylyl-sulfate reductase [thioredoxin] (EC 1.8.4.8)
2	2,3-diketo-5-methylthiopentyl-1-phosphate enolase

MetaPhlAn

We ran the sequences from the metagenomic sequencing experiment through MetaPhlAn. MetaPhlAn generates taxonomy profiles by inferring what the results would be with 16S rRNA gene tag sequencing.

The operational taxonomic units in the MetaPhlAn and 16S rRNA gene analysis were derived from different databases, and can not be compared directly. Note that OTUs can reside in between genera, such that the genus classification is not perfectly concordant between the two comparisons. The top four relatively abundant genera from the MetaPhlAn analysis were *Ruminococcus*, *Eubacterium*, *Coprococcus*, and *Bacteroides*. Only *Bacteroides* is also on the top four relatively abundant genera from the 16S rRNA gene tag sequencing experiment.

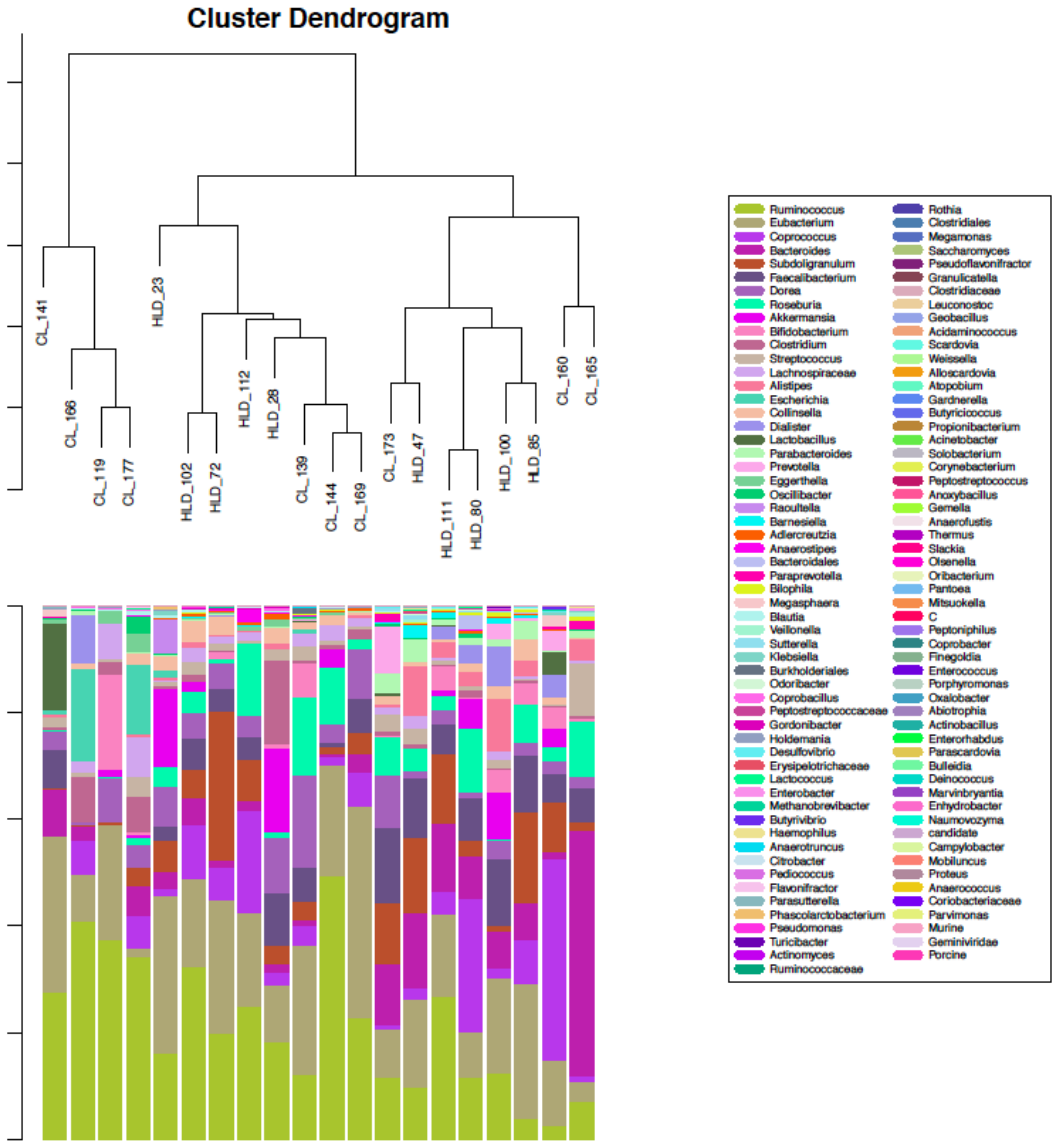


Figure 3.13: **Taxa barplot dendogram derived from MetaPhlAn.** The metagenomic reads were input into MetaPhlAn to generate a count table. The taxa in the count table were filtered such that only taxa with at least 1% abundance in any sample was kept. In this barplot, each bar represents one sample, and each color represents one genus, with the size of the colored segments corresponding with the relative abundance of the genus.

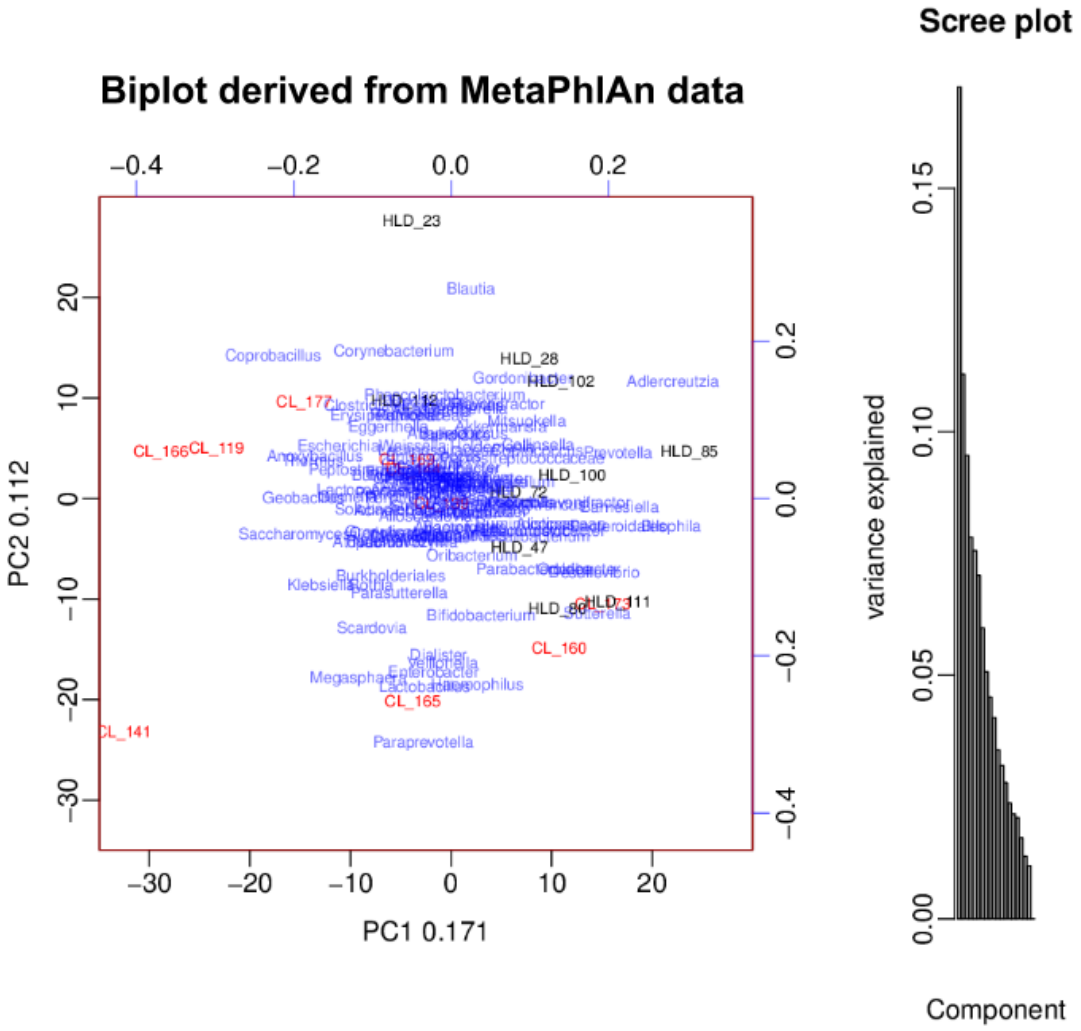


Figure 3.14: **Biplot derived from MetaPhlAn.** Compositional data analysis is done by transforming the counts with the centered log ratio transform, and then performing a principal component analysis. The variance explained by each genera is overlayed on the same principal component analysis plot. This biplot was generated from the count table inferred by MetaPhlAn. Note that the variance explained by the first and the second component is 17% and 11% respectively, 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. The two groups appear to have a nice split, but this is contradicted by the experimental 16S rRNA gene sequencing results in Figure 3.15

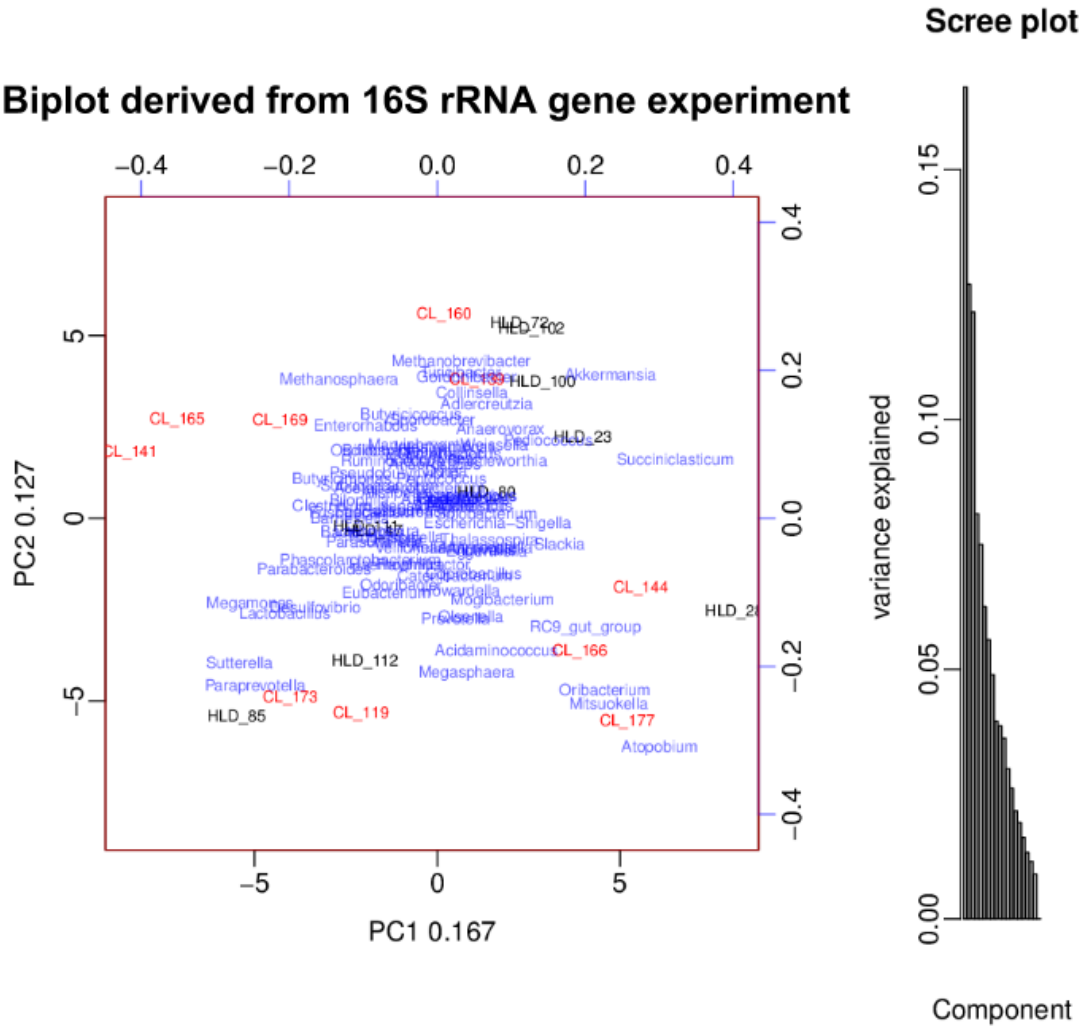


Figure 3.15: **Biplot derived from 16S rRNA gene experiment.** Compositional data analysis is done by transforming the counts with the centered log ratio transform, and then performing a principal component analysis. The variance explained by each genera is overlaid on the same principal component analysis plot. This biplot was generated from the count table generated by the 16S rRNA gene sequencing experiment. Note that the variance explained by the first and the second component is 16.7% and 12.7% respectively, 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. This plot only shows the 20 samples selected for metagenomic sequencing.

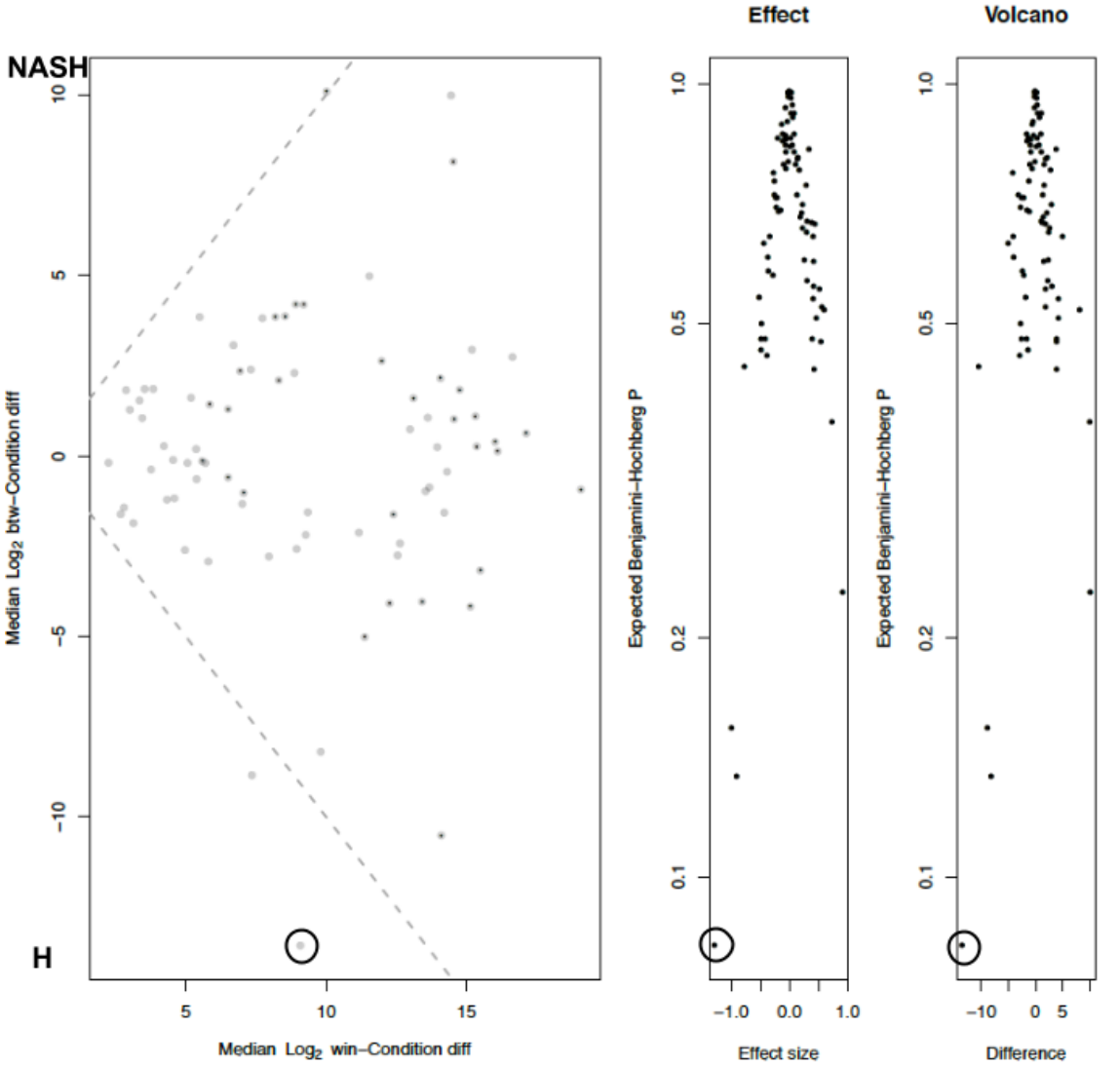


Figure 3.16: **Difference within groups vs. difference between groups per taxa, derived from MetaPhlAn.** This plot was generated from the count table inferred by MetaPhlAn, with taxa filtered such that only taxa with at least 1% abundance in any sample was kept. No taxa are more differential between groups than within groups. A positive difference between indicates that the taxa was relatively increased in NASH while a negative difference between indicates that the taxa was relatively increased in healthy. This analysis was done at the inferred species level (using operational taxonomic units) rather than at the genus level. The circled taxa is *Alistipes shahii*, and has an expected Benjamini-Hochberg p-value of 0.0776.

3.4 Discussion

Given the inconsistency in the five papers that have been published about NAFLD and the gut microbiome, we have performed our analysis in a rigorous manner in an effort to find OTUs with true effects. We found that there was no significant difference between groups by sample clustering (Fig. 3.2) or at the level of the individual OTUs (Fig. 3.5). The principal component analysis done with MetaPhlAn appeared to show the samples separating by group (Fig. 3.14). One taxa, *Alistipes shahii*, was nearly significantly increased in the healthy condition in the MetaPhlAn analysis, which infers a taxonomic profile from metagenomic sequencing. However, when we sequenced the true taxonomic profile through 16S rRNA gene sequencing, no taxa were as differential (Fig. **??**), and the groups were not separated (Fig. 3.15). Members of the *Alistipes* genus appear in the top decile of taxa differentially abundant in both NASH (Fig. 3.3) and healthy (Fig. 3.4) conditions.

We conclude that 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. The literature shows that differences in the gut microbiome are often driven by diet [19]. Additionally, the nature of microbiome data is that there are very many more variables (in the form of OTUs or annotated gene functions) than samples, and the power of the study is inversely proportional to the number of variables.

From Fig. 3.6, 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 taxonomic profile 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.

In our metagenomic analysis, we performed a principal component analysis and found that our healthy samples clustered together into two groups (with the exception of one outlier). The extreme NASH samples were much more spread out. The samples were not arranged this way in the principal component analysis of the 16S rRNA gene experiment. This could be indicative of a core healthy microbiome profile that is more distinguishable by metabolic potential (gene content profile), and potentially by transcriptomic profile, than the taxonomic profile. Rather than deviating from the healthy samples in a specific direction, the microbiome of extreme NASH patients could deviate from the healthy microbiome in a number of directions and yet produce the same clinical outcomes. One other example of this is bacterial vaginosis, which may differ taxonomically from the typical lactobacillus-dominated microbiome in many different ways, but has a common transcriptomic profile that lead to the same clinical symptoms [70].

Boursier et al. [8] imputed a metagenomic analysis with PICRUSt [60], and found significantly differentially abundant functional annotations in KEGG pathways for carbohydrate, lipid, and amino acid metabolism. We performed the metagenomic experiment with principal component analysis on the full set of functional annotations, and subsets of annotations categorized under carbohydrate, lipid, and amino acid. We found that the carbohydrate subset and the lipid subset spread out the samples in a pattern closer to that of the full set of functional annotations, compared to the amino acid subset. This makes sense for the carbohydrate subset

which contains over one seventh of all the functional annotations. However, the amino acid subset actually contains several times more annotations (449) compared to the lipid subset (134). Still, the healthy samples cluster together (with the exception of one outlier), and the extreme NASH samples are more spread out.

The next step in exploring the relationship between the gut microbiome and non alcoholic fatty liver disease progression is to perform studies with a more homogenous population that have more extreme phenotypic differences, and to include transcriptomic as well as metagenomic experiments. Additionally, we implore other research groups to employ multiple test corrections in their statistical analysis so that we are not chasing false positives, and our research will be replicable.

Supporting Information

S1 Fig.

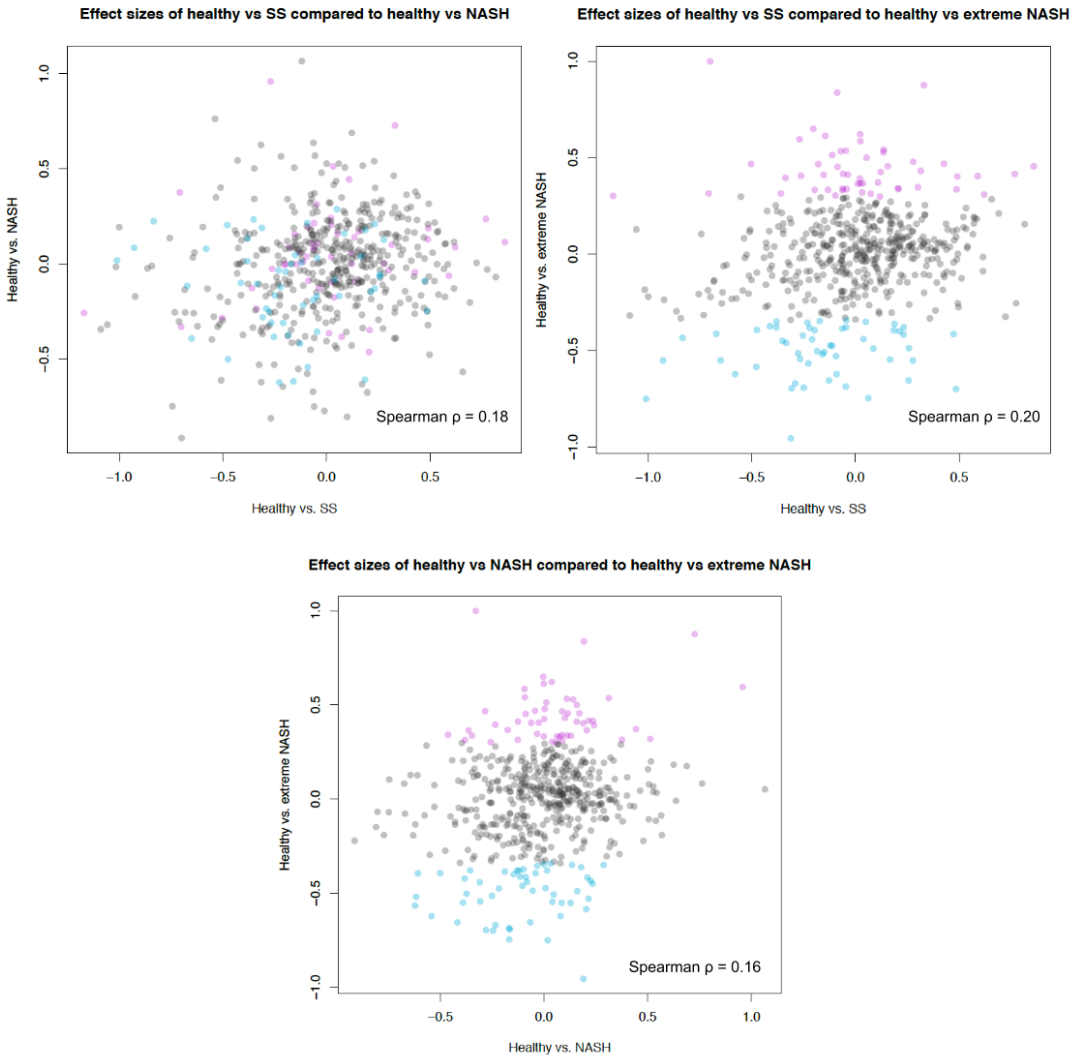


Figure 3.17: **NAFLD comparison effect sizes, with non overlapping healthy samples.** In these plots, the healthy and extreme NASH comparison is the same as in Figure 3.6. To show what the correlations look like without the spurious correlation effects of overlapping samples, the healthy samples not used in the healthy vs. extreme NASH comparison have been partitioned between the healthy vs. SS and the healthy vs. NASH comparison. Additionally, only the NASH samples not used in the extreme NASH comparison are used in the NASH comparison. This way none of the comparisons have overlapping samples.

Chapter 4

Discussion

4.1 Lack of reproducibility

It is clear that more robust analysis methods are necessary in this field. A lot of microbiome research is irreproducible due to improper use of statistics and high variability in experimental design. This is problematic as groups are using this research to patent and produce medical interventions such as probiotics or fecal transplants.

The chapter on expanding the UnIFrac toolbox highlighted examples of misuse of unweighted UniFrac in papers published in top journals. One claimed to find differences in the gut microbiome of mice modelling autism spectrum disorder, compared to healthy controls. The taxa purported to be significantly differentially abundant (*Bacteroides fragilis*), was promptly patented. The other claimed to find differences in the gut microbiome of humanized mice fed a more traditional fibrous diet compared to mice compared to mice fed a diet similar in composition to the Western diet. These studies had small sample sizes (n = 20 and 10 respectively), used unweighted UniFrac (which we have shown to be unreliable), and had a low amount of variance explained by the principal components axes (14% on PC1).

The chapter about nonalcoholic fatty liver disease (NAFLD) showcased five studies which all claimed to have found a difference in the gut microbiome of patients with nonalcoholic fatty liver disease compared to healthy controls, but with almost non-overlapping results (Fig. 3.1). Some of the variation can be explained by differences in sequencing platform (Roche 454 vs. Illumina MiSeq). More variation can be explained by the variable region of the 16S rRNA gene chosen for sequencing - one study used V1-2, one study used V3, one study used V4, and the other two studies did not report which variable region was used. Three out of five studies used healthy controls with a lower BMI than the NAFLD group, such that differences due to level of obesity could not be distinguished from differences due to NAFLD. Lastly, only one of the studies performed a multiple test correction, so most of the results could not be distinguished from false positives.

Recently the social sciences, particularly psychology, has come under fire for producing irreproducible results, to the point where some claim that most findings are actually false [51]. The biomedical sciences suffer similar issues, prompting Nature to publish a collection of statistics for biologists (http://www.nature.com/collections/qghhqm/). Collectively these papers argue that statistical standards should be set and met, in order to encourage high

quality scientific work.

4.2 Recommendations

Throughout this thesis we have made a case for compositional data analysis. Currently the analytical tools with the most widespread use in the field are the unweighted and weighted UniFrac distance, as well as the Bray Curtis dissimilarity for principal coordinates analysis. Other software such as metagenomeSeq [94], DESeq2 [65], and Metastats [93] are used for differential expression analysis. These are commonly accessed through pipelines such as QIIME and mothur. Many of these have roots in ecology, for example, diversity and species richness. Certain types of diversity measurements, especially those similar to species richness which rely on accurately counting rare taxa, do not make sense for complex biological samples where diversity can be increased by performing deeper sequencing to uncover more bacterial taxa [46]. Shannon and Simpson diversity are the most stable in microbiome experiments performed with next generation sequencing [46].

While compositional data analysis may not be at a stage where it is ready to set as the standard analytical tool, we believe that this model is much closer to the correct answer than the standard toolkit used by microbiome researchers. Recommended compositional data resources include the book *Analyzing compositional data with R* [129], the 16S rRNA gene sequencing compositional analysis workshop (hosted online on GitHub) and all the other resources hosted by the CoDa organization (`http://www.compositionaldata.com/`). Recommended software and tools for microbial network correlations include SPARCC [35], SpiecEasi [59] or the phi metric [66]. Recommended software and tools for differential expression analysis include the analysis of composition of microbiomes (ANCOM) [71] and ALDEx2 for differential expression analysis [32].

There is also a dependance on the p-value for statistical analysis, which may not make sense in microbiome research where the number of variables being compared is far greater than the number of samples. Generally in statistical analysis, it has been found that using p-value based approaches with a 0.05 cut off corresponds to a Bayes factor of 3 to 5, which means that the odds only favor the hypothesis between 3 to 1 and 5 to 1. An estimated 17-25% of such reported results are expected to be wrong, even without p-hacking [53]. The discrepancy here is that p-values represent the probability of observing a test statistic as extreme or more extreme than what was observed in the null hypothesis, while Bayes factors examine the probability of observing the test statistic by the null hypothesis compared to the experimental hypothesis. We recommend other approaches such as looking at patterns in effect size as with the NAFLD study, where we found that the effect size of the OTUs relatively increased in one condition tended to increase with the severity of the disease.

Additionally, the use of pipelines make it easy for researchers to attempt to analyze their data without looking at the data raw. We recommend visualizing the data in bar graphs (as in Fig. 3.2), principal components (as in Fig. **??**), as well as looking at the raw counts throughout the analysis process. This way the research can identify outliers that may not be obvious by conventional analytical techniques (Fig. 2.7), correct data formatting errors, and ensure that filters and other data transformations are not removing all of the useful information.

4.3 Summary

In this work we have done some methods development and applied it to a study on the gut microbiome of patients with nonalcoholic fatty liver disease (NAFLD). Specifically, we investigated alternate weightings for the UniFrac distance metric (information and ratio UniFrac), allowing the visualization of outliers in certain cases (Fig. 2.7), as well as the spread of similar but non-identical data (Fig. 2.8). In the NAFLD study, ratio UniFrac produces a principal component analysis with 34.8% of the variance explained in the first component, compared to 24.4% for weighted UniFrac and 14.4% in unweighted UniFrac.

We have also found that many studies in the field are not performed in a statistically sound way, publishing results that cannot be reproduced. Resources, software, and tool recommendations are made in the previous section to prevent this.

The field of microbiome research is in need of standards, such as those set for clinical genomics. When clinical genomics was a budding field, many genome wide association studies were published claiming to have found single nucleotide polymorphisms (SNPs) corresponding to genetic conditions. Discordant results between similar studies prompted the development of standards to ensure statistical validity in analysis and reproducible results. The quality of this information was paramount as study results moved from research labs to clinics for patient genetic counselling. Factors contributing to irreproducibility included batch effects from sample processing [63], ancestry differences [98], and variations in genotype calling methods [84], and recommendations were made to avoid pooling the sequences together [77], and for using sample sizes in the thousands [9], stratification detection [98], and technical replicates [48], and experimental validation [77]. Patient genomes must be sequenced at 30 times coverage or higher to validate the presence of SNPs [103]

Interestingly, the field of microbiome research seems to have standardized too early. Efforts such as the Human Microbiome Project, set a precedent for the types of analyses performed, as well as the tools and techniques researchers use. Some of these have foundations in ecology and are not necessarily applicable to microbiome research, and only a limited number of alternatives have been discussed in the literature. Pipelines such as QIIME [12] and mothur [109] make it comparatively difficult for researchers to explore other analysis options, both in terms of analyzing the data, but also in terms of getting alternative options published, due to bias from peer reviews for the standard techniques.

The field of microbiome research has shown lots of promise, yeilding findings such as an obesity-associated increased capacity for energy harvest [124], and leading to clinical interventions for diseases such as *C. diff* [96]. With more time and more research, tools and techniques will be developed to perform robust microbiome research, potentially leading to methods to modulate the microbiome and increase quality of life through preventative and restorative medical interventions.

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] Edward W Beals. “Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data”. In: *Advances in Ecological Research* 14.1 (1984), p. 55.

[6] David R Bentley et al. “Accurate whole human genome sequencing using reversible terminator chemistry”. In: *nature* 456.7218 (2008), pp. 53–59.

[7] Alain Berson et al. “Steatohepatitis-inducing drugs cause mitochondrial dysfunction and lipid peroxidation in rat hepatocytes”. In: *Gastroenterology* 114.4 (1998), pp. 764–774.

[8] 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).

[9] Paul R Burton et al. “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls”. In: *Nature* 447.7145 (2007), pp. 661–678.

[10] Benjamin J Callahan et al. “DADA2: High resolution sample inference from amplicon data”. In: *bioRxiv* (2015), p. 024034.

[11] J Gregory Caporaso et al. “Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample”. In: *Proceedings of the National Academy of Sciences* 108.Supplement 1 (2011), pp. 4516–4522.

[12] J Gregory Caporaso et al. “QIIME allows analysis of high-throughput community sequencing data”. In: *Nature methods* 7.5 (2010), pp. 335–336.

[13] 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.

[14] 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.

[15] Albi Celaj et al. “Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation”. In: *Microbiome* 2.1 (2014), p. 1.

[16] Jun Chen et al. “Associating microbiome composition with environmental covariates using generalized UniFrac distances”. In: *Bioinformatics* 28.16 (2012), pp. 2106–2113.

[17] Francesca D Ciccarelli et al. “Toward automatic reconstruction of a highly resolved tree of life”. In: *science* 311.5765 (2006), pp. 1283–1287.

[18] 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.

[19] Lawrence A David et al. “Diet rapidly and reproducibly alters the human gut microbiome”. In: *Nature* 505.7484 (2014), pp. 559–563.

[20] Arthur L Delcher et al. “Identifying bacterial genes and endosymbiont DNA with Glimmer”. In: *Bioinformatics* 23.6 (2007), pp. 673–679.

[21] 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.

[22] 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.

[23] 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.

[24] Gilbert GG Donders et al. “Definition of a type of abnormal vaginal flora that is distinct from bacterial vaginosis: aerobic vaginitis”. In: *BJOG: An International Journal of Obstetrics & Gynaecology* 109.1 (2002), pp. 34–43.

[25] Robert C Edgar. “MUSCLE: multiple sequence alignment with high accuracy and high throughput”. In: *Nucleic acids research* 32.5 (2004), pp. 1792–1797.

[26] Robert C Edgar. “Search and clustering orders of magnitude faster than BLAST”. In: *Bioinformatics* 26.19 (2010), pp. 2460–2461.

[27] Robert C Edgar. “UPARSE: highly accurate OTU sequences from microbial amplicon reads”. In: *Nature methods* 10.10 (2013), pp. 996–998.

[28] Robert C Edgar et al. “UCHIME improves sensitivity and speed of chimera detection”. In: *Bioinformatics* 27.16 (2011), pp. 2194–2200.

[29] JJ Egozcue and V Pawlowsky-Glahn. “Evidence information in bayesian updating”. In: *Proceedings of the 4th International Workshop on Compositional Data Analysis*. 2011.

[30] 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.

[31] Andrew D Fernandes et al. “ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq”. In: *PLoS One* 8.7 (2013), e67019.

[32] 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.

[33] 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.

[34] 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.

[35] Jonathan Friedman and Eric J Alm. “Inferring correlation networks from genomic survey data”. In: *PLoS Comput Biol* 8.9 (2012), e1002687.

[36] Jack A Gilbert, Janet K Jansson, and Rob Knight. “The Earth Microbiome project: successes and aspirations”. In: *BMC biology* 12.1 (2014), p. 69.

[37] Steven R Gill et al. “Metagenomic analysis of the human distal gut microbiome”. In: *science* 312.5778 (2006), pp. 1355–1359.

[38] Gregory B Gloor and Gregor Reid. “Compositional analysis: a valid approach to analyze microbiome high throughput sequencing data”. In: *Canadian Journal of Microbiology* ja (2016).

[39] Gregory B Gloor et al. “It’s all relative: analyzing microbiome data as compositions”. In: *Annals of Epidemiology* (2016).

[40] Gregory B Gloor et al. “Microbiome profiling by illumina sequencing of combinatorial sequence-tagged PCR products”. In: *PloS one* 5.10 (2010), e15406.

[41] Anastassia Gorvitovskaia, Susan P Holmes, and Susan M Huse. “Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle”. In: *Microbiome* 4.1 (2016), p. 1.

[42] 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.

[43] 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.

[44] Francisco Guarner and Juan-R Malagelada. “Gut flora in health and disease”. In: *The Lancet* 361.9356 (2003), pp. 512–519.

[45] 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.

[46] Bart Haegeman et al. “Robust estimation of microbial diversity in theory and in practice”. In: *The ISME journal* 7.6 (2013), pp. 1092–1101.

[47] Lewis G Halsey et al. “The fickle P value generates irreproducible results”. In: *nature methods* 12.3 (2015), pp. 179–185.

[48] Huixiao Hong et al. “Technical reproducibility of genotyping SNP arrays used in genome-wide association studies”. In: *PLoS One* 7.9 (2012), e44483.

[49] Elaine Y Hsiao et al. “Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders”. In: *Cell* 155.7 (2013), pp. 1451–1463.

[50] Ruben Hummelen et al. “Deep sequencing of the vaginal microbiota of women with HIV”. In: *PloS one* 5.8 (2010), e12078.

[51] John PA Ioannidis. “Why most published research findings are false”. In: *PLoS Med* 2.8 (2005), e124.

[52] 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).

[53] Valen E Johnson. “Revised standards for statistical evidence”. In: *Proceedings of the National Academy of Sciences* 110.48 (2013), pp. 19313–19317.

[54] Takahiro Kanagawa. “Bias and artifacts in multitemplate polymerase chain reactions (PCR)”. In: *Journal of bioscience and bioengineering* 96.4 (2003), pp. 317–323.

[55] Minoru Kanehisa and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[56] R Koch. “Uber bakteriologische Forschung Verhandlung des X Internationalen Medichinischen Congresses, Berlin, 1890, 1, 35. August Hirschwald, Berlin”. In: *German.) Xth International Congress of Medicine, Berlin*. 1891.

[57] 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.

[58] HA Krebs and JR Perkins. “The physiological role of liver alcohol dehydrogenase”. In: *Biochemical Journal* 118.4 (1970), pp. 635–644.

[59] Zachary D Kurtz et al. “Sparse and compositionally robust inference of microbial ecological networks”. In: *PLoS Comput Biol* 11.5 (2015), e1004226.

[60] 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.

[61] Ben Langmead and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature methods* 9.4 (2012), pp. 357–359.

[62] 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.

[63] Jeffrey T Leek et al. “Tackling the widespread and critical impact of batch effects in high-throughput data”. In: *Nature Reviews Genetics* 11.10 (2010), pp. 733–739.

[64] 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.

[65] 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.

[66] David Lovell et al. “Proportionality: a valid alternative to correlation for relative data”. In: *PLoS Comput Biol* 11.3 (2015), e1004075.

[67] Catherine A Lozupone et al. “Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities”. In: *Applied and environmental microbiology* 73.5 (2007), pp. 1576–1585.

[68] 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.

[69] Catherine Lozupone et al. “UniFrac: an effective distance metric for microbial community comparison”. In: *The ISME journal* 5.2 (2011), p. 169.

[70] 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.

[71] 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).

[72] Elaine R Mardis. “Next-generation DNA sequencing methods”. In: *Annu. Rev. Genomics Hum. Genet.* 9 (2008), pp. 387–402.

[73] Marcel Margulies et al. “Genome sequencing in microfabricated high-density picolitre reactors”. In: *Nature* 437.7057 (2005), pp. 376–380.

[74] 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.

[75] 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.

[76] Andre P Masella et al. “PANDAseq: paired-end assembler for illumina sequences”. In: *BMC bioinformatics* 13.1 (2012), p. 31.

[77] Mark I McCarthy et al. “Genome-wide association studies for complex traits: consensus, uncertainty and challenges”. In: *Nature reviews genetics* 9.5 (2008), pp. 356–369.

[78] Philip McKenna et al. “The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis”. In: *PLoS Pathog* 4.2 (2008), e20.

[79] Kevin McKernan et al. *Reagents, methods, and libraries for bead-based sequencing*. US Patent App. 12/628,209. Nov. 2009.

[80] Paul J McMurdie and Susan Holmes. “Waste not, want not: why rarefying microbiome data is inadmissible”. In: *PLoS Comput Biol* 10.4 (2014), e1003531.

[81] NI McNeil. “The contribution of the large intestine to energy supplies in man.” In: *The American journal of clinical nutrition* 39.2 (1984), pp. 338–342.

[82] Nathan P McNulty et al. “The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins”. In: *Science translational medicine* 3.106 (2011), 106ra106–106ra106.

[83] Folker Meyer et al. “The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes”. In: *BMC bioinformatics* 9.1 (2008), p. 386.

[84] K Miclaus et al. “Variability in GWAS analysis: the impact of genotype calling algorithm inconsistencies”. In: *The pharmacogenomics journal* 10.4 (2010), pp. 324–335.

[85] 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.

[86] KM Neufeld et al. “Reduced anxiety-like behavior and central neurochemical change in germ-free mice”. In: *Neurogastroenterology & Motility* 23.3 (2011), 255–e119.

[87] Els van Nood et al. “Duodenal infusion of donor feces for recurrent Clostridium difficile”. In: *New England Journal of Medicine* 368.5 (2013), pp. 407–415.

[88] Jari Oksanen et al. “The vegan package”. In: *Community ecology package* (2007), pp. 631–637.

[89] 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.

[90] 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.

[91] Lior Pachter. “Models for transcript quantification from RNA-Seq”. In: *arXiv preprint arXiv:1104.3889* (2011).

[92] Javier Palarea-Albaladejo and Josep Antoni Martın-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.

[93] Joseph N Paulson, Mihai Pop, and Hector Corrada Bravo. “Metastats: an improved statistical method for analysis of metagenomic data”. In: *Genome biology* 12 (2011), pp. 1–27.

[94] Joseph Nathaniel Paulson. “metagenomeSeq: Statistical analysis for sparse high-throughput sequencing”. In: *Bioconductor package* 1 (2014).

[95] 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.

[96] 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.

[97] 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.

[98] Alkes L Price et al. “Principal components analysis corrects for stratification in genome-wide association studies”. In: *Nature genetics* 38.8 (2006), pp. 904–909.

[99] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. “FastTree 2–approximately maximum-likelihood trees for large alignments”. In: *PloS one* 5.3 (2010), e9490.

[100] Albert Propst et al. “Prognosis and life expectancy in chronic liver disease”. In: *Digestive diseases and sciences* 40.8 (1995), pp. 1805–1815.

[101] 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.

[102] 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.

[103] Heidi L Rehm et al. “ACMG clinical laboratory standards for next-generation sequencing”. In: *Genetics in Medicine* 15.9 (2013), pp. 733–747.

[104] Jai Ram Rideout et al. “Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences”. In: *PeerJ* 2 (2014), e545.

[105] Christian S Riesenfeld, Patrick D Schloss, and Jo Handelsman. “Metagenomics: genomic analysis of microbial communities”. In: *Annu. Rev. Genet.* 38 (2004), pp. 525–552.

[106] 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.

[107] 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.

[108] Susannah J Salter et al. “Reagent and laboratory contamination can critically impact sequence-based microbiome analyses”. In: *BMC biology* 12.1 (2014), p. 87.

[109] 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.

[110] Nicola Segata et al. “Metagenomic biomarker discovery and explanation”. In: *Genome Biol* 12.6 (2011), R60.

[111] Nicola Segata et al. “Metagenomic microbial community profiling using unique clade-specific marker genes”. In: *Nature methods* 9.8 (2012), pp. 811–814.

[112] 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.

[113] Claude Elwood Shannon. “A mathematical theory of communication”. In: *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1 (2001), pp. 3–55.

[114] Jay Shendure and Hanlee Ji. “Next-generation DNA sequencing”. In: *Nature biotechnology* 26.10 (2008), pp. 1135–1145.

[115] Jay Shendure et al. “Accurate multiplex polony sequencing of an evolved bacterial genome”. In: *Science* 309.5741 (2005), pp. 1728–1732.

[116] Daniel Simberloff. “Use of rarefaction and related methods in ecology”. In: *Biological data in water pollution assessment: quantitative and statistical analyses*. ASTM International, 1978.

[117] Michelle I Smith et al. “Gut microbiomes of Malawian twin pairs discordant for kwashiorkor”. In: *Science* 339.6119 (2013), pp. 548–554.

[118] Se Jin Song et al. “Cohabiting family members share microbiota with one another and with their dogs”. In: *Elife* 2 (2013), e00458.

[119] Erica D Sonnenburg et al. “Diet-induced extinctions in the gut microbiota compound over generations”. In: *Nature* 529.7585 (2016), pp. 212–215.

[120] 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.

[121] 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).

[122] Robert Tibshirani and Guenther Walther. “Cluster validation by prediction strength”. In: *Journal of Computational and Graphical Statistics* 14.3 (2005), pp. 511–528.

[123] 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.

[124] Peter J Turnbaugh et al. “An obesity-associated gut microbiome with increased capacity for energy harvest”. In: *nature* 444.7122 (2006), pp. 1027–131.

[125] 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.

[126] 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.

[127] 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.

[128] 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.

[129] K Gerald Van den Boogaart and Raimon Tolosana-Delgado. *Analyzing compositional data with R*. Springer, 2013.

[130] 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.

[131] Ruth Grace Wong. *UniFrac workshop*. http://dx.doi.org/10.5281/zenodo.50248. 2016. DOI: 10.5281/zenodo.50248.

[132] Ruth Grace Wong and Jia Rong Wu. *Scripts for generating paper figures*. http://dx.doi.org/10.5281/zenodo.50629. 2016. DOI: 10.5281/zenodo.50629.

[133] 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.

[134] 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.

[135] Tanya Yatsunenko et al. “Human gut microbiome viewed across age and geography”. In: *Nature* 486.7402 (2012), pp. 222–227.

[136] 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.

[137] Marcela Zozaya-Hinchliffe et al. “Quantitative PCR assessments of bacterial species in women with and without bacterial vaginosis”. In: *Journal of clinical microbiology* 48.5 (2010), pp. 1812–1819.

Appendix A

Workflows

A.1 Non-alcoholic fatty liver disease metagenomic workflow

In this workflow, two annotation strategies were used. The first strategy is to create an annotated reference library to map sequenced reads to, and the second strategy is to annotate sequences assembled de novo from the reads, and map sequenced reads to these. The experiment could have been performed with only the second strategy and yielded a similar number of annotated reads. However, the second strategy is much more computationally intensive, so depending on the computational resources available, it may be worth using both strategies in conjunction, as described below.

A.1.1 Filter OTUs

In this experiment, the sequencing depth is expected to have the power to detect a 2 fold change up or down in bacteria that are 0.2% abundant in a sample. The OTUs were filtered to remove any with an abundance lower than 0.2% in all samples, and the OTU seed sequences were retrieved.

A.1.2 Reference library annotation strategy

Get reference library genomes
The list of genomes used in the reference library was created using two sources: the Human Microbiome Project gut reference genomes (http://hmpdacc.org/HMRGD/healthy/), and the NCBI complete and draft bacterial genomes (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE$_TYPE$ = *BlastSearch*BLAST$_S$PEC = *MicrobialGenomes*).

Human Microbiome Project
The Human Microbiome Project gut reference genomes (http://hmpdacc.org/HMRGD/healthy/) were all added to the reference library genome list for the metagenomic study.

NCBI complete and draft bacterial genomes
The draft and complete bacterial genomes can be queried here: http://blast.ncbi.nlm.

nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=MicrobialGenomes. During this process, we ran into a bug using the NCBI webtool and had to search once through the wgs database, and once with Complete Genomes to get both the draft and the complete genomes.

The BLAST output can be downloaded. In this case we were only interested in the genomes that matched with 98% identity or greater. For these genomes we extracted the GI number, and performed web scraping in Python to visit http://www.ncbi.nlm.nih.gov/nuccore/GInumber\mskip\medmuskip and programatically retrieve the taxon ID. The taxon ID is found in ftp://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/assembly_summary_genbank.txt and the corresponding FTP link is used to download the genome. For each species found by this method, the genomes for 10 random strains are downloaded (or all of the strains if there are less than 10), to increase the coverage of the library.

Get reference library coding sequences
Some of the genomes have a .gff file which includes the locations of the coding sequences already. For the rest, we used Glimmer [20] to predict open reading frames (ORFs).

Annotate reference library coding sequences
Annotation was performed by querying the SEED database [90] using command line BLAST (http://www.ncbi.nlm.nih.gov/books/NBK279690/). This is the most computationally intensive part of the process and can take a number of days, depending on your computing platform. The specific SEED database we used was downloaded June 2013, and had the fig.pegs from the 2010 SEED database which are missing from the 2013 database manually added in. For each ORF, we retained the top 10 hits in the SEED database with an e-value cutoff of 10^{-3}.

A.1.3 De novo assembly annotation strategy

Assembly
The reads were assembled per sample using Trinity [45].

Subtraction of reference library
To prevent double counting, the assembled sequences were queried to the reference library. Any assembled sequences that matched a reference sequence with at least 90% identity was removed.

Annotate de novo assembly
Annotation was performed by querying the SEED database [90] using command line BLAST (http://www.ncbi.nlm.nih.gov/books/NBK279690/). This is the most computationally intensive part of the process and can take a number of days, depending on your computing platform. The specific SEED database we used was downloaded June 2013, and had the fig.pegs from the 2010 SEED database which are missing from the 2013 database manually added in. Because we did not isolate the coding sequences beforehand, we devised a recursive BLAST strategy. For each round, we retained the top 10 hits in the SEED database per sequence with an e-value cutoff of 10^{-3}. The portions of sequence greater than 500 nucleotides long that did not match the SEED hits were run through BLAST again.

A.1.4 Map sequenced reads to reference library

The reference library and the de novo assembly were amalgamated, and used to create a Bowtie2 index. The sequenced reads were mapped to this using Bowtie2 [61].

Custom scripts were used to convert the mapping output to a table of counts per annotation per sample, which can then be analyzed with differential expression tools such as ALDEx2 [32].

All of the custom scripts used to perform the above for the metagenomic non-alcoholic fatty liver disease experiment can be found on GitHub. Specifically, scripts for building the reference library can be accessed at https://github.com/ruthgrace/make_functional_mapping_library, scripts for annotating the reference library and mapping the sequenced reads can be accessed at https://github.com/ruthgrace/mapping_library_annotated_counts, and scripts for the de novo assembly annotation strategy can be found at https://github.com/ruthgrace/exploring_nafld_assembly.

Appendix B

NAFLD study data collection

The entire contents of Appendix B were written by Hannah Da Silva from Allard research group in Toronto, and provided through personal correspondence.

B.0.1 Study Participants

This cross-sectional study includes 39 NAFLD patients (15 SS, 24 NASH) from the University Health Network (UHN) outpatient liver clinics and 30 healthy living liver donors as healthy controls (HC) from the UHN Liver Transplant Clinic. Patient recruitment occurred from July 2010 to May 2014. The University Health Network and University of Toronto Research Ethics Boards approved this research. All subjects provided their informed written consent.

Patients suspected of having NAFLD due to persistently elevated liver enzymes after ruling out other causes were recruited at their hepatologist appointment prior to liver biopsy. Following a detailed explanation of the study and consent form, and answering of any questions, informed consent was obtained. Eligibility to participate in the study was then considered. For NAFLD patients the inclusion criteria were: a biopsy proven diagnosis of NAFLD (determined during the study), age greater than or equal to18 years, alcohol consumption of less than 20g per day. NAFLD patients were excluded from the study if they had a diagnosis of any other liver disease or HIV infection, if liver transplant was expected to be required within one year, if they had significant liver complications (e.g. variceal bleeding, jaundice, etc.) or any contraindications for liver biopsy, if they were pregnant or lactating, if they had any gastrointestinal disease, or if they were taking any medications known to cause steatohepititis, insulin, NSAIDS, antibiotics, prebiotics, probiotics, or experimental drugs within the last three months. Healthy living liver donors were approached at their first screening appointment for liver donation. Upon completion of informed consent healthy donors were screened for research eligibility. Inclusion and exclusion criteria were the same with the additional inclusion criteria that they must be eligible for live liver donation.

B.0.2 Study Visits

Each participant attended three study visits as outlined in Figure 1. For NAFLD patients, after informed consent was obtained, they were provided with detailed instructions for the completion of a 7-day food record, 7-day activity record, and environmental questionnaire (see below for

further details). Patients were also provided with instructions for the collection and transport of their stool sample, which they were requested to return on the day of their scheduled liver biopsy. Usual clinic blood work was also collected at this initial visit for general markers of liver health. On their second visit fasting study-specific blood work was collected and anthropometric measurements were completed. The third visit was the day of their liver biopsy when patients also returned their stool sample and food and activity logs. Healthy living liver donors followed a similar schedule, however, stool samples were typically collected on visit two and the liver biopsy was taken intraoperatively. For further details on all study measurements and processing see below.

B.0.3 Clinical Data, Environmental Questionnaire, and Anthropometric Measurements

Clinical data was collected on study visit one. Study participant's smoking and alcohol consumption history and medication and supplement use were reviewed, including medications taken in the last three months. Study participants were also asked to answer a number of questions regarding their personal and family history of disease. Age, ethnicity, and menstrual history were also recorded.

An environmental questionnaire was completed and returned with the stool sample. This questionnaire collected information that may affect an individual's IM composition including country of origin, method of birth (vaginal versus caesarian section), whether they were breastfed as an infant, what kind of pets they have at home, and others.

Anthropometric measurements including height (ht), weight (wt), waist circumference (WC), hipcircumference, and weight-to-hip ratio (WHR) were measured by a trained research professional. Weight was measured using a calibrated hospital-grade chair scale; height was measured using a standometer. Waist circumference was measured at the umbilicus level and hip circumference was measured at the widest point over the buttock. All measurements were taken in triplicate and the average value was used.

B.0.4 Nutrition and Activity Assessment

Each participant was given a food record and activity log to complete in the weekprior to returning their stool sample. Detailed instructions were provided for the completion of both of these tools. The food log included all food and beverages consumed each 24 hours for seven days. In cases where time was insufficient a three-day food record including one weekend day was completed. Participants used the 2D Food Portion Visual Chart (Nutrition Consulting Enterprises, Framingham, MA) to estimate portion sizes. This is a validated tool which has been used in our previous studies [59, 187, 188]. Food records were reviewed by an experienced registered dietitian and were analyzed using Food Processor Diet and Nutrition Analysis Software (Version 7, ESHA Research, Salem, OR).

Physical activity logs were recorded for 7 days concurrent with the food records. Participants were asked to record any activity, including household chores, the duration of the activity, and the intensity level. Detailed instructions were provided including examples for each intensity level (mild, moderate, strenuous, and very strenuous). This information was used to calculate daily

physical activity units: 1 unit = 30 minutes mild, 20 minutes moderate, 10 minutes strenuous, or 5 minutes very strenuous activity. This is a validated method for measuring physical activity level [189]. Basal metabolic rate (BMR) was calculated using the Harris-Benedict equation: BMR for men = 66.5 + [13.75 × wt(kg)] + [5.003 × ht(cm)] – [6.755 × age(y)], BMR for women = 655.1 [9.563 × wt(kg)] + [1.850 × ht(cm)] – [4.676 × age(y)]. Estimated energy expenditure (EER) was calculated using Health Canada Guidelines: EER for men = 662 – [9.53 x age(y)] + PA x [15.91 x wt(kg])] + [539.6 x ht(m)], and EER for women = 354 – [6.91 x age(y)] + PA x [9.36 x wt(kg)] + (726 x ht(m)] where PA is the physical activity coefficients.

B.0.5 Biochemistry

Routine and study specific blood work was drawn after a 12 hour overnight fast on study visittwo. Liver markers were drawn as a routine clinical measure and included: aspartate transaminase (AST), alanine transaminase (ALT), alkaline phosphatase (ALP), and bilirubin. Measures of glucose metabolism included plasma glucose, insulin, hemoglobin A1c (HbA1c), and HOMA-IR which was calculated as fasting glucose (mmol/L) × fasting insulin (mU/L)/22.5 [191]. A lipid profile was also conducted, including total cholesterol, low density lipoprotein (LDL), high density lipoprotein (HDL), and triglycerides (TG). These analyses were conducted by the UHN Laboratory Medicine Program. Liver enzymes and lipid profile were measured using the Architect c8000 system (Abbott Laboratories). LDL was calculated from total cholesterol – HDL. Fasting plasma glucose and plasma insulin were measured by the enzymatic hexokinase method and radioimmunoassay, respectively.

B.0.6 Serum Metabolites

Serum metabolites, including choline, ethanol, and TMA, were measured to evaluate potential implications of bacterial metabolism at a systemic level. For the full list of the 41 metabolites measured see Table 4. Serum was drawn in a fasting state using a gel serum separation vacutainer and was immediately placed in an insulated container with cooling elements. Blood was separated by centrifuge at 4°C at 2800 x g for 20 minutes. Serum was then aliquoted and stored at -80°C until all study samples were collected. Serum was then shipped to the Metabolomic Innovation Centre (Edmonton, AB) were metabolites were analyzed using nuclear magnetic resonance spectrometry (NMR), a method which this centre has perfected [192]. The following methods were used by the centre and are stated in their own words [192]: All serum samples were deproteinized using ultrafiltration. Prior to filtration, two 0.5 mL, 3 KDa cut-off centrifugal filter units (Millipore Microcon YM-3) were rinsed four times each with 0.5 mL of water, then centrifuged at 11 000 rpm for 1 hour, to remove residual glycerol bound to the filter membranes. Two 150 L aliquots of each serum sample were then transferred into the two centrifuge filter devices. The samples were then spun at a rate of 11 000 rpm for 140 minutes, to remove macromolecules (primarily proteins and lipoproteins) from the sample. The subsequent filtrates were then checked visually for a red tint, which indicates that the membrane was compromised. For those “membrane compromised” samples, we repeated the filtration process with a different filter and inspected the filtrate again. We then pooled the filtrates that passed the inspections and recorded the volume. If the total volume of the sample was under 300 L, we added an appropriate amount from a 50 mM NaH2PO4 buffer (pH 7) to the sample until

the total volume was 300 L. Subsequently, 35 L of D2O and 15 L of a standard buffer solution [11.667 mM DSS (disodium-2,2-dimethyl-2-silapentane-5-sulphonate), 730 mM imidazole, and 0.47% NaN3 in H2O] was added to the sample. The serum sample (350 L) was then transferred to a standard Shigemi microcell NMR tube for subsequent spectral analysis.

All 1H-NMR spectra were collected on a 500 MHz Inova (Varian Inc., Palo Alto, CA) spectrometer equipped with either a 5 mm HCN Z-gradient pulsed-field gradient (PFG) room-temperature probe or a Z-gradient PFG Varian cold-probe. 1H-NMR spectra were acquired at 25°C using the first transient of the tnnoesy-presaturation pulse sequence, which was chosen for its high degree of quantitative accuracy [193]. Spectra were collected with 128 transients and 8 steady-state scans using a 4 second acquisition time and a 1 second recycle delay.

All FIDs were zero-filled to 64k data points and subjected to line broadening of 0.5 Hz. The singlet produced by a known quantity the DSS methyl groups was used as an internal standard for chemical shift referencing (set to 0 ppm) and for quantification. All 1H-NMR spectra were processed and analyzed using the Chenomx NMR Suite Professional software package version 6.0 (Chenomx Inc., Edmonton, AB), as previously described [194]. Each spectrum was processed and analyzed by at least two experienced NMR spectroscopists to minimize compound mis-identification and mis-quantification.

Serum trimethylamine N-oxide (TMAO) was not detectable using NMR therefore TMAO was measuring using TMAO using a targeted quantitative metabolomics approach by Liquid Chromatography Mass Spectrometry (LCMS). Isotopically-labeled internal standards were added to the serum to facilitate metabolite quantification. Sample extraction was performed on a 96 well plate with a 0.2 μm solvent filter. 10 μL of serum was spiked with the internal standard (TMAO D9) and then 150 μL of methanol with 10 mM ammonium acetate was added for extraction. The plate was shaken for 10 min and centrifuged at 500 rpm for 5 minutes at 4 ºC. Each sample was diluted with 150 L of water. Seven calibrant solutions with known concentrations went through the same extraction steps. LCMS analysis was performed on AB SCIEX 4000 QTrap mass spectrometer with Agilent 1100 HPLC. 10 μL of the extracted samples were injected onto the Kinetex C18 (2.6 μm, 3.0x100mm, 100A) Column with guard column. Isobaric elution was performed with 90% A (10 mM ammonium formate in water, PH3) and 10% B (10 mM ammonium formate in 90:10 Acetonitrile:water, PH3). Total LC method run time was 3 min with flowrate of 500 μl/min. A seven-point calibration curve was generated to quantify the concentration of TMAO in samples.

B.0.7 Liver Histology

Liver biopsies were taken percutaneously (needle biopsy) for NAFLD patients and intraoperatively (wedge biopsy) for HC and preserved immediately in formalin. Liver biopsies were assessed by the same pathologist using standard stains for the diagnosis of NAFLD, morphologic evaluation, and to rule out any iron overload. The evaluation of NAFLD related measures of steatosis, inflammation, and fibrosis were conducted using the validated and reproducible Brunt system. Disease severity was also evaluated using the NAFLD Activity Score (NAS) which accounts for degree of steatosis, lobular inflammation, and hepatocellular ballooning for a final score of 0-8.

B.0.8 Stool Sample Collection and Analysis

On study visit one participants received a stool collection kit, including a plastic collection/storage container with a tightly closing lid, an insulated bag, and cooling elements. Within 24 hours of their next appointment they collected one stool sample, which was frozen immediately after defecation in the patient's home freezer (-20°C). Participants brought the frozen sample in the insulated bag with cooling elements to their appointment at the hospital, where it will stored at -80C until homogenization.

B.0.9 Stool Homogenization

Stool samples were homogenized prior to DNA extraction for IM sequencing and metabolite-measurements. The entire sample was first transferred into a sterile masticator bag. The sample was allowed to thaw until a smooth consistency was reached, typically 2-3 hours depending on sample size. Once thawed excess air was released from the bag and the sample was homogenized for two minutes using a masticator blender (IUL, S.A., Barcelona, Spain). This was followed by one minute of hand mastication of any areas that were missed. The corner of the masticator bag was then cut and the sample was aliquoted: 1-2 g samples were stored for metabolite analysis and 0.1-0.2 g aliquots were stored for DNA extraction. Samples were immediately placed on dry ice and then transferred to -80°C for storage until analysis. Weight was recorded for each aliquot and pH was measured for each sample.

B.0.10 DNA Extraction

DNA was extracted using the E.Z.N.A. Stool DNA Kit (Omega Bio-Tek, Norcross, GA) and amodified manufacturer's protocol. Briefly, 200 mg of glass beads and 600 μL of SLB buffer were added to the sample and vortexed at maximum speed for 15 minutes. 20 μL of lysozyme (20mg/ μL) was added and flicked to mix then incubated at 37°C for 30 minutes. 60 μL DS Buffer and 20 μL Proteinase K were added and vortexed to mix then incubated at 70°C, 300 rpm, for 13 min, vortexing at T=6.5 minutes and T=13 minutes. The incubation temperature was then increased to 95°C for an additional 5 minutes. 200 μL SP2 Buffer was added and mixed by vortex for 30 seconds then put on ice for 5 minutes. The mixture was then centrifuged at 21 000 rcf for 7 minutes and the supernatant was transferred to a new 1.5 mL centrifuge tube while the old tube was discarded. 200 μL HTR Reagent was added and vortexed at maximum speed for 10 seconds, then incubated at room temperature for two minutes and centrifuged for an additional two minutes. The supernatant was again transferred to a new 1.5 mL centrifuge tube and the addition of HTR Reagent and following steps were repeated once. After transfer to the last 1.5 mL tube 250 μL of BL buffer and absolute ethanol were added and vortexed for 10 seconds.

The DNA column was placed into a collection tube and 100 μL of 3M NAOH was added. This was incubated for four minutes and centrifuged for one minute. 100 μL of distilled water was added to the DNA column and centrifuged for one minute. 800 μL of sample was transferred into the DNA column and centrifuged for one minute. The contents of collection tube was discarded and the remainder of the sample was transferred into the DNA column and again centrifuged for one minutes. The flow-through and collection tube were discarded. The column

was then placed into a new 2 mL collection tube and 500 μL of VHB Buffer was added to the column. This was centrifuged for 30 seconds and the flow-through was discarded. 700 μL of DNA Wash Buffer was then added to the DNA column and centrifuged for one minute and the flow-through and tube was discarded. This washing stage was repeated once. The column was then transferred to a new 2 mL collection tube, centrifuged for one minute and the flow-through was discarded. The tube was then centrifuged again for two minute, this time with the cap open, to the dry the column. The column was finally transferred to a new 1.5 mL tube, 100 uL of distilled water was added, this was incubated for five minutes, centrifuged for two minutes, the column was removed and then the final DNA sample was analyzed for purity and concentration using the Nanodrop 1000 Spectrophotometer (ThermoScientific, Rockford, IL). DNA samples were stored at -80°C.

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.